Extensions and Applications of the Tail Pairwise Dependence Matrix

Matthew Pawley

October 12, 2024

Table of contents

List of Figures

List of Tables

Preface

Draft thesis of Matthew Pawley, created on October 12, 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, with distribution function

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

The parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are called the location, scale, and shape, respectively. The shape dictates which sub-class the asymptotic distribution belongs to: $\xi > 0$ corresponds to the heavy-tailed Fréchet class; $\xi = 0$ (interpreted as $\xi \to 0$) corresponds to the exponential-tailed Gumbel class; $\xi < 0$ corresponds to 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 $x_1, \dots, x_n$ denote independent observations of $X_1, \dots, X_n$. 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 $x_1, \dots, x_n$, 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 will accentuate particular features of the extreme values. For example, heavy-tailed Fréchet margins will highlight the most extreme values, while 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

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

where

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

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

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

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 Jonathan A 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 (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

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

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 C. 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

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

   where $\xrightarrow{\mathrm{v}}$ 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

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

   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:

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

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 (S. 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 ??.

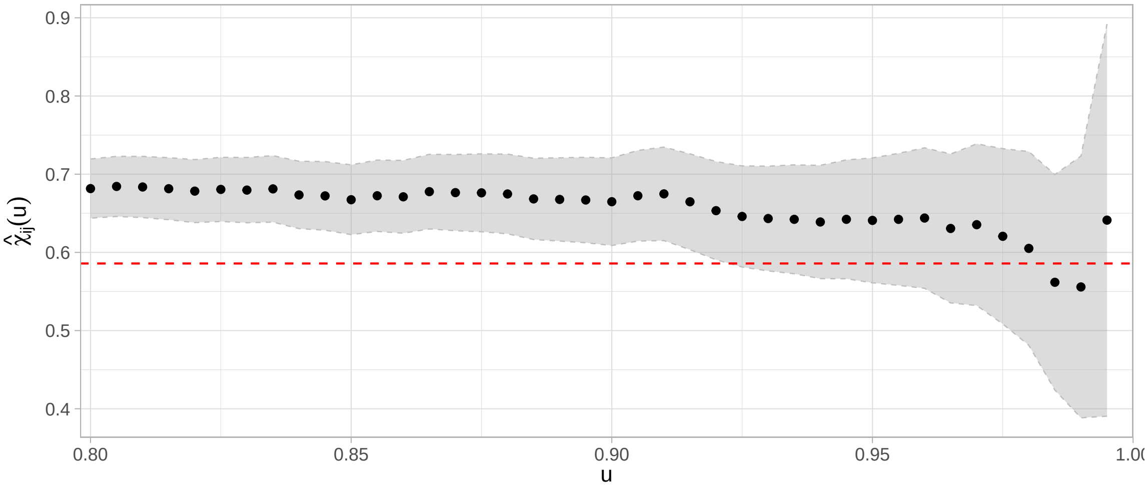

Figure 2.1: Symmetric logistic with $\alpha = 0.5$. True $\chi = 2 - 2^{\alpha} \approx 0.586$. Estimates based on data $n = 5000$. 95% Wald CI based on `mev::taildep` function.

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

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

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. Stéphan 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*

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

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*

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

Then

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

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

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

This definition was generalised to permit any $\alpha \geq 1$ by Kiriliouk and C. 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 C. 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 $i = 1, \ldots, d$, 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 $c > 0$,

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

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 $i = 1, \ldots, d$, $\sigma_{ii} = 1$.*
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

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

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

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

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 $\alpha \in (0, 1)$. Then

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

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

The limiting cases are $\lim_{\alpha\to 0}\sigma_{ij} = 1$ and $\lim_{\alpha\to 1}\sigma_{ij} = 0$.

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

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

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

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.

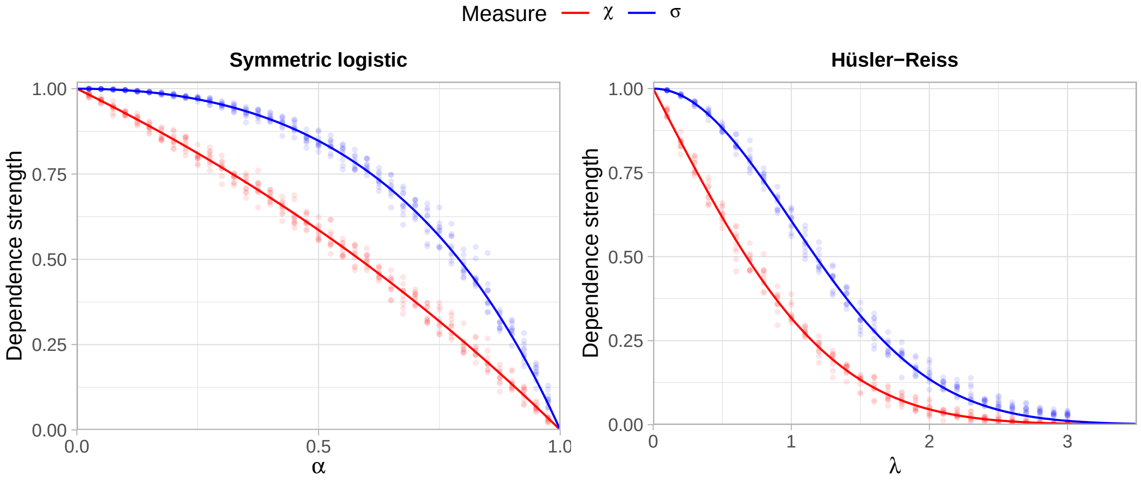

Figure 2.2: Blah

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 C. 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 $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. 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 C. 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 C. 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 $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$,

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

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

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

where

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

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 $\operatorname{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$, we have

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

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,

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

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 $i,j=1,\ldots,d$, we have $\sigma_{ij} = \sum_{l=1}^{q} a_{il}^{\alpha/2}a_{jl}^{\alpha/2}$ and

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

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

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

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 subpanels. 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,

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

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.*

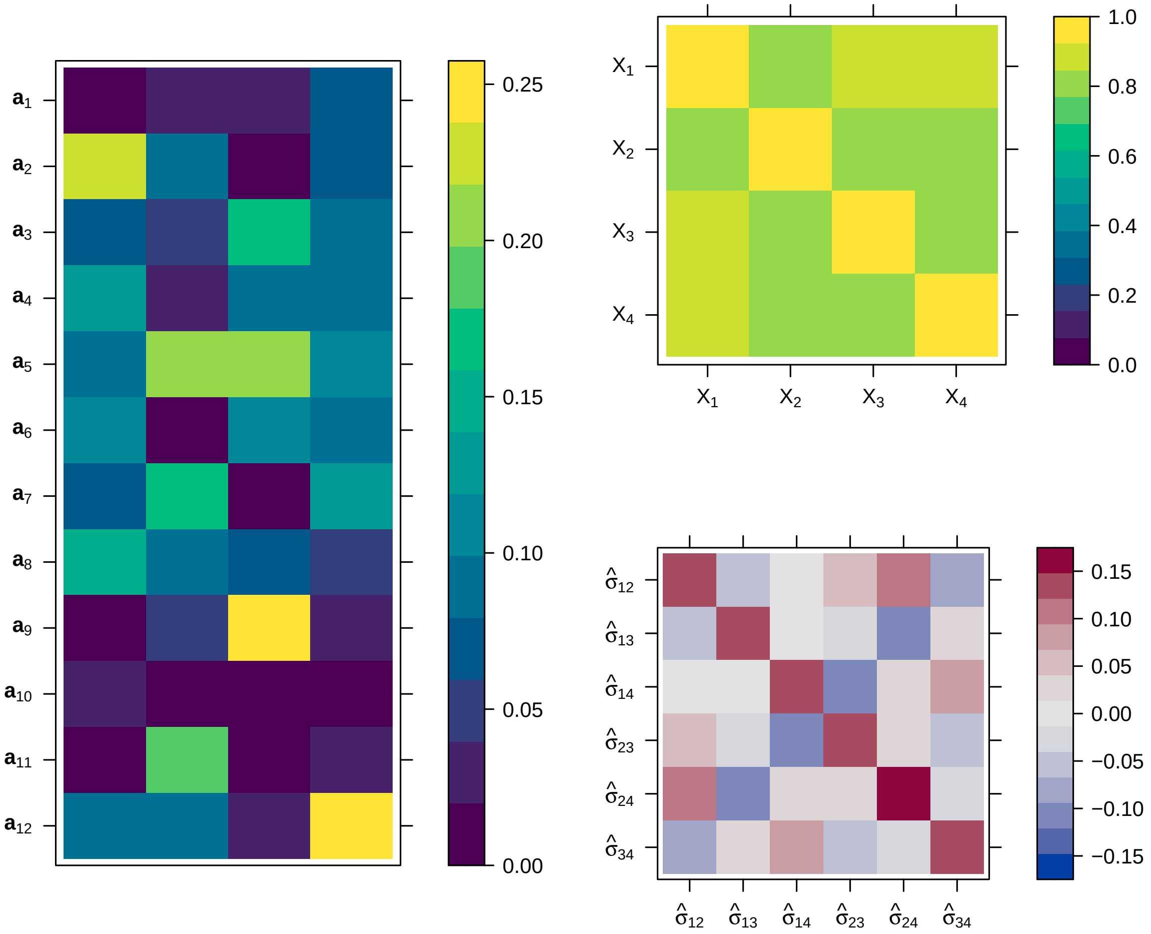

Figure 2.3: Blah

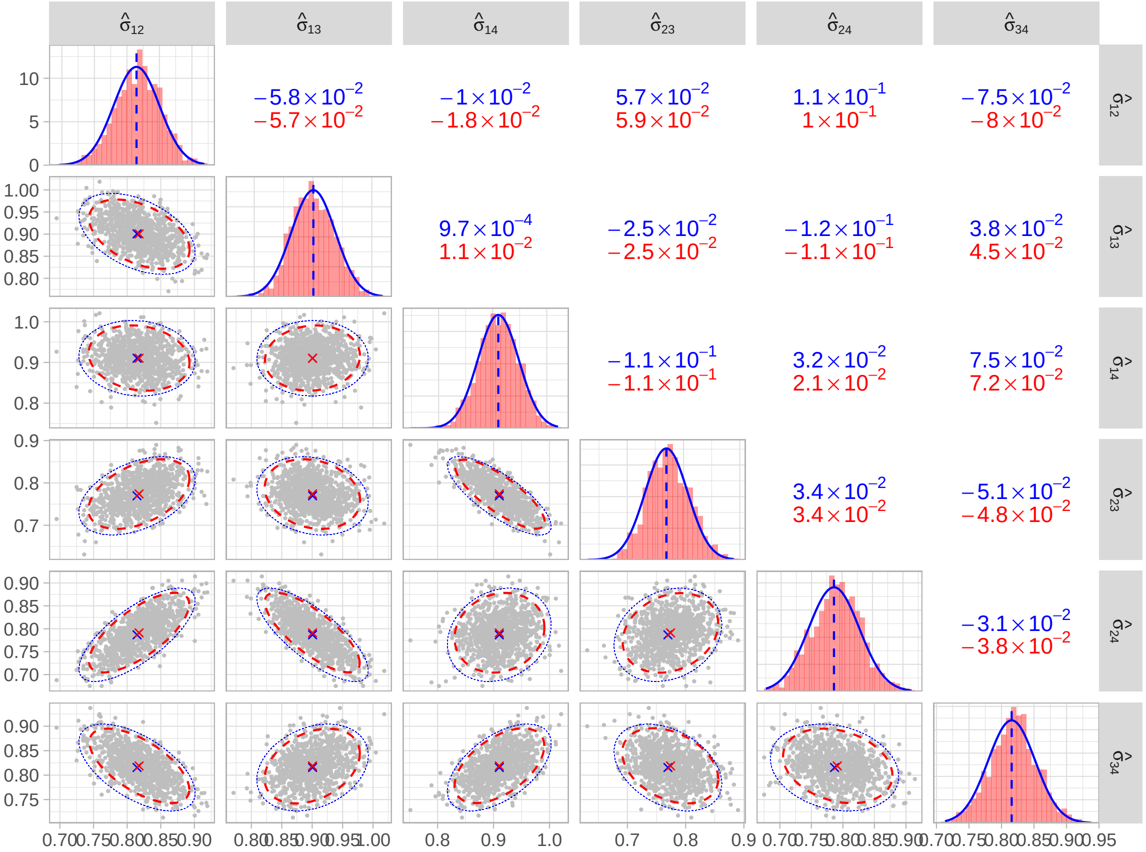

Figure 2.4: Blah

2.4 Existing applications and extensions of the TPDM

Copy this section over.

2.5 Bias in the empirical TPDM in weak-dependence scenarios

Copy this section over.

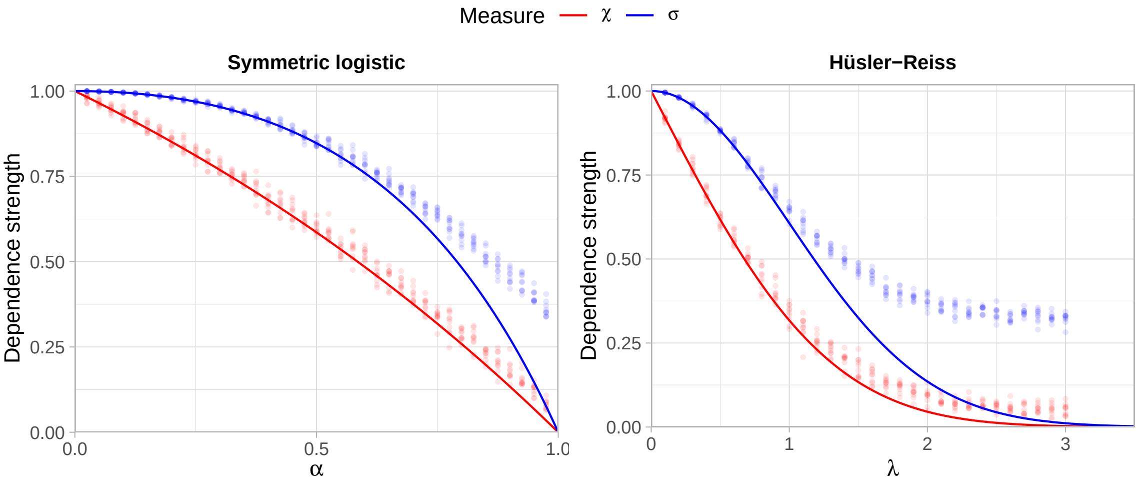

Figure 2.5: Blah

3 Testing for time-varying extremal dependence

3.1 Motivation

Inference for multivariate extremes typically assumes that the observed data are independent and identically distributed. The latter property requires that the extremal dependence structure is fixed throughout the observation period, but in some contexts the validity of this assumption may be doubtful. For example, there is evidence that anthropogenic climate change is driving changes in the spatial structure of climate extremes (S. Zhou et al. 2023), while regulatory changes can cause structural changes in the joint tail behaviour of financial asset prices (Poon et al. 2003). Testing for or modelling changing extremal dependence represents a challenging statistical problem: the underlying signal will is likely fairly weak (e.g. the underlying physical processes driving climate change only manifest over long observation periods) and inference is hampered by the inherent scarcity of relevant data.

Most research in the realm of non-stationary extremes relates to changes in the marginal distributions (CITE), while the problem of time-varying dependence has received comparatively little attention. The regression model of Castro-Camilo et al. (2018) or the semi-parametric spectral density ratio model of can incorporate the effects of covariates, including time, are capable of modelling non-stationary dependence structures in low-dimensional settings using parametric models. The non-parametric procedure of Drees (2023) can detect (i.e. test for) dependence changes, but is similarly restricted to a small number of dimensions. Our contribution is extending/adapting their method to facilitate testing in moderate- to high-dimensional settings. To achieve this, we leverage the tail pairwise dependence matrix (Cooley and Thibaud 2019; Larsson and Resnick 2012), a covariance-like matrix for extremes that has garnered widespread use for analysing the joint tail distribution of high-dimensional random vectors (Fomichov and Ivanovs 2023; Kiriliouk and C. Zhou 2022; Rohrbeck and Cooley 2023).

Notation. All random elements are defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

3.1.1 Framework and hypothesis

Suppose $\{\boldsymbol{X}(t) = (X_1(t), \dots, X_d(t)) : t \in [0,1]\}$ is an $\mathbb{R}_+^d$-valued, continuous time stochastic process with no serial dependence. Let $\|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^d$. For $t \in [0,1]$, assume that the random vector $\boldsymbol{X}(t)$ is multivariate regularly varying (MRV) with index of regular variation $\alpha(t) = 2$ and angular measure $H(\cdot\,;t)$ on $\mathbb{S}_+^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_2 = 1\}$. Denoting by $R(t) := \|\boldsymbol{X}(t)\|_2$ and $\boldsymbol{\Theta}(t) := \boldsymbol{X}(t)/\|\boldsymbol{X}(t)\|_2$ the radial and angular components of $\boldsymbol{X}(t)$, respectively, the MRV property states that for all $z > 0$ and Borel sets $\mathcal{B} \subset \mathbb{S}_+^{d-1}$,

$$\lim_{u\to\infty} \frac{\mathbb{P}(R(t) > zu, \boldsymbol{\Theta}(t) \in \mathcal{B})}{\mathbb{P}(R(t) > u)} = z^{-\alpha(t)} H(\mathcal{B}; t). \tag{3.1}$$

We assume $\boldsymbol{X}(t)$ is on Fréchet margins with shape parameter $\alpha(t) = 2$, perhaps after a suitable marginal transformation. With this scaling, the angular measure of $\boldsymbol{X}(t)$ satisfies $H(\mathbb{S}_+^{d-1}; t) = d$ for all $t \in [0,1]$. In the MRV paradigm, extremal dependence is fully characterised by the angular measure. Our working null and alternative hypotheses can be stated formally as

$$\mathrm{H}_0 \ : \ \forall t \in [0,1],\, H(\cdot\,;t) = H(\cdot\,;1), \tag{3.2}$$

$$\mathrm{H}_1 \ : \ \exists t,\, H(\cdot\,;t) \neq H(\cdot\,;1). \tag{3.3}$$

Our goal is to devise a statistical procedure for testing these hypotheses given a discretised sample path of $\{\boldsymbol{X}(t) : t \in [0,1]\}$.

3.1.2 Background and outlook

Drees (2023) tests (3.2) against (3.3) via a large family $\mathcal{A}$ of subsets of $\mathbb{S}_+^{d-1}$ and suitably rescaled versions of stochastic processes

$$\left\{\int_0^t \hat{H}(A;s)\,\mathrm{d}s - t\int_0^1 \hat{H}(A;s)\,\mathrm{d}s : t \in [0,1]\right\}, \qquad (A \in \mathcal{A}). \tag{3.4}$$

Here $\hat{H}(A;s)$ denotes a non-parametric estimate of the angular measure $H(A;s)$ at time $s \in [0,1]$ – see (3.6) for a formal definition. The null is rejected if any paths in (3.4) deviate from what would typically occur under the null. If $\mathcal{A}$ is sufficiently rich, then even very subtle dependence changes may be revealed, in principle. However, as the dimension d increases the family of sets grows rapidly, typically $|\mathcal{A}| = \mathcal{O}(2^d)$. Consequently, the underlying computations become prohibitively intensive and the convergence $H(\hat{A};t) \to H(A;t)$ of the non-parametric estimators is too slow. Thus, their method is primarily intended for the bivariate setting and is restricted to $d \leq 5$ in practice. Fundamentally, this limitation stems from the curse of dimensionality inherent to estimation of the angular measure. This impediment is exacerbated by the fact that inference must be performed *locally*, i.e. using only (extreme) observations lying within some small temporal neighbourhood.

Our approach mitigates this issue by concentrating on bivariate summaries of tail dependence instead of the full dependence structure. The $\mathcal{O}(d^2)$ coefficients of the TPDM encode second-order information about the local angular measure and can be more reliably estimated in high dimensions. The downside is that the TPDM contains incomplete information about the angular measure. This means our test is powerless in certain circumstances; a class of examples is provided in Section 3.3.

3.1.3 The local (integrated) TPDM

Non-stationary dependence as in (3.1) necessitates a time-dependent version of the TPDM. This is naturally defined via an integral with respect to the local angular measure.

Definition 3.1. For $t \in [0,1]$, the local TPDM is the $d \times d$ matrix given by

$$\sigma_{ij}(t) = \int_{\mathbb{S}_+^{d-1}} \theta_i\theta_j\,\mathrm{d}H(\boldsymbol{\theta};t), \qquad \Sigma(t) = (\sigma_{ij}(t))_{i,j=1,\ldots,d}. \tag{3.5}$$

The local TPDM summarises the tail dependence strength between pairs of components of $\boldsymbol{X}(t)$. Since $H(\cdot\,;t)$ is a valid angular measure, the local TPDM satisfies all the usual mathematical properties of a TPDM (Cooley and Thibaud (2019)).

While our principle objective is to detect changes in the local TPDM, it is common to devise statistical tests based on integrated versions of the quantity of interest. This strategy is employed by Drees (2023) – consider (3.4). This motivates the introduction of an integrated TDPM.

Definition 3.2. For $t \in [0,1]$, the integrated TPDM is the $d \times d$ matrix given by

$$\psi_{ij}(t) = \int_0^t \sigma_{ij}(s)\,\mathrm{d}s, \qquad \Psi(t) = (\psi_{ij}(t))_{i,j=1,\ldots,d}.$$

The integrated TPDM is symmetric, positive semi-definite, and possesses the property that $\psi_{ij}(t) = 0$ if and only if $X_i(s)$ and $X_j(s)$ are asymptotically independent for all $s \leq t$. Beyond this, it has no obvious interpretation. With standardised margins, we have that $\sigma_{ii}(t) = 1$ and hence $\psi_{ii}(t) = t$ for all $i = 1,\ldots,d$ and $t \in [0,1]$. The integrated TPDM can be equivalently defined via the so-called integrated angular measure of Drees (2023), since

$$\psi_{ij}(t) = \int_0^t \int_{\mathbb{S}_+^{d-1}} \theta_i\theta_j \,\mathrm{d}H(\boldsymbol{\theta};s)\,\mathrm{d}s = \int_{\mathbb{S}_+^{d-1}} \theta_i\theta_j \int_0^t \mathrm{d}H(\boldsymbol{\theta};s)\,\mathrm{d}s.$$

Due to this connection, many of the theoretical results contained in Drees (2023) transfer immediately to our methodology.

The matrices $\Sigma(t)$ and $\Psi(t)$ are symmetric with known diagonal entries, so nothing is lost by focussing exclusively on their strictly upper triangular elements. We introduce the following notation for referring to these entries. If $M = (m_{ij})$ denote an arbitrary $d \times d$ (random) matrix, the (random) vector obtained by row-wise vectorisation of its upper triangular elements shall be denoted by

$$\mathrm{vecu}(M) := (m_{12}, m_{13}, \ldots, m_{1d}, m_{23}, \ldots, m_{2d}, \ldots, m_{d-1,d}).$$

The components of $\boldsymbol{m} := \mathrm{vecu}(M)$ are indexed according to the sub-indices of M, e.g. the first element of is m_{12} rather than m_1. The upper-vectorised local TPDM and integrated

TPDM are denoted by

$$\begin{aligned}\boldsymbol{\sigma}(t) &:= \operatorname{vecu}(\Sigma(t)) = (\sigma_{12}(t), \sigma_{13}(t), \dots, \sigma_{1d}(t), \sigma_{23}(t), \dots, \sigma_{2d}(t), \dots, \sigma_{d-1,d}(t)), \\ \boldsymbol{\psi}(t) &:= \operatorname{vecu}(\Psi(t)) = (\psi_{12}(t), \psi_{13}(t), \dots, \psi_{1d}(t), \psi_{23}(t), \dots, \psi_{2d}(t), \dots, \psi_{d-1,d}(t)).\end{aligned}$$

The dimension of these vectors is

$$\mathcal{D} := |\{(i,j) : 1 \leq i < j \leq d\}| = \binom{d}{2} = \frac{1}{2}d(d-1).$$

The following section concerns the estimation of these quantities.

3.1.4 Inference

Suppose we observe a sample path of $\{\boldsymbol{X}(t) : t \in [0,1]\}$ along n discrete time-points according to an equidistant sampling scheme, yielding a collection of independent random vectors $\{\boldsymbol{X}(i/n) : i = 1, \dots, n\}$. Our methodology could accommodate more general sampling schemes, but this one is the simplest and most commonly encountered. The general principle underlying the following is that extremal dependence at time $t \in [0,1]$ may be inferred from the k most extreme observations lying within in a h-neighbourhood of t. The hyperparameters $h > 0$ and $k \geq 1$ are called the *bandwidth* and *level*, respectively. Specifically, we define

$$\mathcal{I}(t) := \{i \in \{1, \dots, n\} : i/n \in (t-h, t+h]\},$$

and among the observations $\{\boldsymbol{X}(i/n) : i \in \mathcal{I}(t)\}$, only those whose norm exceeds a specified radial threshold will enter into our estimators. The threshold $\hat{u}(t)$ is set as the $k+1$ largest order statistic among $\{R(i/n) : i \in \mathcal{I}(t)\}$; by construction, there will be exactly k radial threshold exceedances. Selecting the level and bandwidth involves managing trade-offs between retaining an adequate number of samples (by increasing h and k) while ensuring that estimation remains time-localised (reducing h) and free of bias due to observations from the distributional bulk (reducing k).

Our estimator for the local TPDM is founded on the empirical local angular measure defined in Drees (2023). For any $t \in [0,1]$, this random measure is given by

$$\hat{H}(\cdot\,;t) := \frac{d}{k}\sum_{i\in\mathcal{I}(t)} \mathbf{1}\{R(i/n) > \hat{u}(t), \boldsymbol{\Theta}(i/n) \in \cdot\}. \tag{3.6}$$

Substituting (3.6) into (3.5) results in the following definition.

Definition 3.3. For $t \in [0,1]$, the empirical local TPDM is the $d \times d$ matrix given by

$$\hat{\sigma}_{ij}(t) := \int_{\mathbb{S}_+^{d-1}} \theta_i\theta_j \,\mathrm{d}\hat{H}(\boldsymbol{\theta};t) = \frac{d}{k}\sum_{l\in\mathcal{I}(t)} \Theta_i(l/n)\Theta_j(l/n)\mathbf{1}\{R(l/n) > \hat{u}(t)\}, \qquad \hat{\Sigma}(t) = (\hat{\sigma}_{ij}(t)). \tag{3.7}$$

We recognise (3.7) as simply a time-localised version of the familiar empirical TPDM (Equation 5 in Cooley and Thibaud (2019)). Thus it retains all the usual properties of an empirical TPDM.

Estimating the integrated TPDM is slightly more complicated, because $\Psi(t)$ depends on the full history of $\Sigma(s)$ over the continuous interval $s \in [0,t]$. This is achieved by the same block-based construction used by Drees (2023). First, we partition the full observation period $[0,1]$ into blocks of width $2h$; each block contains $b := 2nh$ observations. (Henceforth, assume that the number of blocks $t/(2h) = n/b$ is an integer.) Next, we estimate the local TPDM at at the block centres $t \in \{h, 3h, \dots, (2n/b-1)h\}$ using (3.7) with the bandwidth in $\mathcal{I}(t)$ set equal to half the block width (that is, h). Then, we interpolate the local TPDM estimates according to an assumption of constant dependence within each block, so that for any index pair $1 \leq i < j \leq d$, $\hat{\sigma}_{ij}(s)$ constitutes a piecewise constant function of s on $[0,1]$. The entries of the empirical integrated TPDM are given by the corresponding time-integrals of these functions.

Definition 3.4. For $t \in [0,1]$, the empirical integrated TPDM is the $d \times d$ matrix given by

$$\hat{\psi}_{ij}(t) := \int_0^t \hat{\sigma}_{ij}((2\lceil s/(2h)\rceil - 1)h)\,\mathrm{d}s, \qquad \hat{\Psi}(t) = (\hat{\psi}_{ij}(t)). \tag{3.8}$$

This can be equivalently and more conveniently expressed as

$$\hat{\Psi}(t) := 2h \sum_{l=1}^{L(t)} \hat{\Sigma}(s_l) + (t - 2hL(t))\hat{\Sigma}(s_{L(t)+1}), \qquad L(t) := \lfloor t/(2h) \rfloor, \qquad s_l := (2l-1)h. \tag{3.9}$$

The first term in (3.9) corresponds to the $L(t)$ whole blocks in $[0, t]$, each of which receive a full weighting of $2h$. The second term receives a reduces weight since it relates to the partial block containing t.

While the formulation (3.8) looks less cumbersome, (3.9) will prove more convenient, both computationally and mathematically. In particular, it reveals that the $\hat{\psi}_{ij}(t)$ is a weighted sum of the independent random variables $\sigma_{ij}(s_1), \ldots, \sigma_{ij}(s_{L(t)+1}$. Independence is due to the blocks being non-overlapping and is crucial in the elicitation of the asymptotic results to follow.

3.1.5 Asymptotic theory

We now formulate the asymptotic properties of the estimators $\hat{\boldsymbol{\sigma}}(t) := \mathrm{vecu}(\hat{\Sigma}(t))$ and $\hat{\boldsymbol{\psi}}(t) := \mathrm{vecu}(\hat{\Psi}(t))$. Henceforth, the bandwidth and level are sequences satisfying $h \to 0$, $nh \to \infty$, $k \to \infty$, and $k/(nh) \to 0$ as $n \to \infty$.

The first result concerns asymptotic normality of the empirical local TPDM. As remarked earlier, $\hat{\Sigma}(t)$ is simply an empirical TPDM based on the data subset $\{\boldsymbol{X}(i/n) : i/n \in (t-h, t+h]\}$, which comprises $2nh$ observations. Asymptotically, by assumption, the size of this restricted sample $2nh \to \infty$, the number of observations entering the estimator $k \to \infty$, and the proportion of observations entering the estimator $k/(2nh) \to 0$. Thus all the conditions required for asymptotic normality of the empirical TPDM hold (see Larsson and Resnick (2012) and Section 6.1 in Lee and Cooley (2023)).

Proposition 3.1. *For any* $t \in [0, 1]$,

$$k^{1/2}(\hat{\boldsymbol{\sigma}}(t) - \boldsymbol{\sigma}(t)) \to N(\mathbf{0}, V(t)) \tag{3.10}$$

as $n \to \infty$. The $\mathcal{D} \times \mathcal{D}$ asymptotic covariance matrix is given by

$$V(t) := \mathrm{Cov}(\mathrm{vecu}(d\tilde{\boldsymbol{\Theta}}(t)\tilde{\boldsymbol{\Theta}}(t)^T)), \qquad \tilde{\boldsymbol{\Theta}}(t) \sim d^{-1}H(\cdot\,;t). \tag{3.11}$$

The diagonal entries $V_{ij,ij}(t)$ of $V(t)$ relate to the asymptotic variance of the estimators $\hat{\sigma}_{ij}(t)$. The off-diagonal entries $V_{ij,lm}(t)$ relate to the asymptotic covariance between $\hat{\sigma}_{ij}(t)$ and $\hat{\sigma}_{lm}(t)$. Ordinarily $V(t)$ is unknown but will be present as a nuisance parameter in our test statistics. For now we assume that $V(t)$ is known; later it will be replaced by a plug-in estimator.

Considering (3.9) and Proposition 3.1, the components of $\boldsymbol{\psi}(t)$ are weighted sums of independent, asymptotically normal random variables. By a functional central limit theorem type argument it follows that, with suitable rescaling, the stochastic process $\{\hat{\psi}_{ij}(t) : t \in [0,1]\}$ converges in distribution to a Gaussian processes.

Proposition 3.2. *The $\mathcal{D}$-dimensional, continuous-time stochastic process*

$$\left\{ \left(\frac{k}{h}\right)^{1/2} \left(\hat{\boldsymbol{\psi}}(t) - \boldsymbol{\psi}(t)\right) : t \in [0,1] \right\}, \tag{3.12}$$

converges to the $\mathcal{D}$-dimensional centred Gaussian process $\{\boldsymbol{Y}(t) : t \in [0,1]\}$ with covariance function

$$\mathrm{Cov}(Y_{ij}(s), Y_{lm}(t)) = 2\int_0^{\min(s,t)} V_{ij,lm}(\tau)\,\mathrm{d}\tau. \tag{3.13}$$

Proof. Write proof here.

□

The drift and diffusion coefficients associated with each univariate process $\{Y_{ij}(t) : t \in [0,1]\}$ are controlled by underlying integrated TPDM $\{\psi_{ij}(t) : t \in [0,1]\}$ and asymptotic variance process $\{V_{ij,ij}(t) : t \in [0,1]\}$, respectively. Meanwhile, the cross-correlation between $\{Y_{ij}(t) : t \in [0,1]\}$ and $\{Y_{lm}(t) : t \in [0,1]\}$ is determined by the asymptotic covariance $\{V_{ij,lm}(t) : t \in [0,1]\}$.

Under the null hypothesis (3.2), the asymptotic variance-covariance matrix $V = V(t)$ is independent of time and the covariance function (3.13) simplifies to $\mathrm{Cov}(\boldsymbol{Y}(s), \boldsymbol{Y}(t)) =$

$2V\min(s,t)$. Thus, upon pre-multiplying (3.12) by $(2V)^{-1/2}$, the distribution of the limiting process equals that of a standard $\mathcal{D}$-dimensional standard Brownian motion.

3.1.6 Hypothesis testing

In view of Proposition 3.2 and the ensuant discussion, we define a $\mathcal{D}$-dimensional test process $\{\hat{\boldsymbol{Z}}(t) : t \in [0,1]\}$ as

$$\hat{\boldsymbol{Z}}(t) := \left(\frac{k}{2h}\right)^{1/2} V(t)^{-1/2}(\hat{\boldsymbol{\psi}}(t) - t\hat{\boldsymbol{\psi}}(1)). \tag{3.14}$$

The nuisance parameter $V(t)$ standardises the processes $\hat{Z}_{ij}(t)$ and removes cross-correlation between them. Its inclusion is vital for ensuring a convenient asymptotic null distribution for our test statistics and thus allowing critical values to be readily available without recourse to simulation. Generally, $V(t)$ may be assumed to be invertible, since the off-diagonal TPDM entries are not constrained to equal any particular value. This would not be the case had the vectorised quantities $\boldsymbol{\sigma}(t), \boldsymbol{\psi}(t)$ included components pairs $i = j$: the diagonal entries of the TPDM satisfy $\operatorname{trace}(\Sigma(t)) = \sigma_{11}(t) + \ldots + \sigma_{dd}(t) = d$ for all $t \in [0,1]$, forming a linear combination of components with zero variance.

From the test process we define Kolmogorov-Smirnov (KS) and Cramér-von-Mises (CM) type test statistics by

$$T^{(KS)} := \sup_{t\in[0,1]} \left\|\hat{\boldsymbol{Z}}(t)\right\|_{\infty}, \tag{3.15}$$

$$T^{(CM)} := \sup_{1\leq i<j\leq d} \left\|\hat{Z}_{ij}(t)\right\|^2_{L^2[0,1]}, \tag{3.16}$$

where $\|\boldsymbol{x}\|_\infty := \max\{|x_i| : i = 1, \ldots, \mathcal{D}\}$ denotes the sup-norm in $\mathbb{R}^{\mathcal{D}}$ and $\|Y(t)\|^2_{L^2[0,1]} := \int_0^1 |Y(t)|^2 \,\mathrm{d}t$ denotes the L^2-norm of a stochastic process on $[0,1]$. Their asymptotic null distributions are given below.

Proposition 3.3. *Under the null hypothesis* (3.2)*,*

$$T^{(KS)} \to \sup_{t\in[0,1]} \|\boldsymbol{B}(t)\|_\infty \overset{d}{=} \sup_{1\leq i<j\leq d} K_{ij}, \qquad T^{(CM)} \to \sup_{1\leq i<j\leq d} \|B_{ij}(t)\|^2_{L^2[0,1]}, \tag{3.17}$$

where $\boldsymbol{B}(t) = (B_{ij}(t) : 1 \leq i < j \leq d)$ denotes a standard $\mathcal{D}$-dimensional Brownian bridge and $\{K_{ij} : 1 \leq i < j \leq d\}$ is a collection of $\mathcal{D}$ independent Kolmogorov random variables.

Proof. Under the null hypothesis, $\boldsymbol{\psi}(t) = t \cdot \boldsymbol{\psi}(1)$ and therefore

$$\begin{aligned}\hat{\boldsymbol{Z}}(t) &= (2V)^{-1/2}\left(\frac{k}{h}\right)^{1/2}\left(\hat{\boldsymbol{\psi}}(t) - \boldsymbol{\psi}(t) - t(\hat{\boldsymbol{\psi}}(1) - \boldsymbol{\psi}(1))\right) \\ &\to (2V)^{-1/2}(\boldsymbol{Y}(t) - t\boldsymbol{Y}(1)) \\ &\overset{d}{=} \boldsymbol{W}(t) - t\boldsymbol{W}(1) \\ &\overset{d}{=} \boldsymbol{B}(t),\end{aligned}$$

where $\boldsymbol{W}(t) = (W_{ij}(t) : 1 \leq i < j \leq d)$ denotes a standard $\mathcal{D}$-dimensional Brownian motion. The independent random variables $K_{ij} := \sup_{t\in[0,1]} |B_{ij}(t)|$, for $1 \leq i < j \leq d$, are Kolmogorov distributed by definition.

□

Denoting the Kolmogorov distribution function by F_K,

$$\mathbf{1}\{T^{(KS)} > c_\alpha\}, \qquad c_\alpha = F_K^{-1}((1-\alpha)^{1/\mathcal{D}}) \tag{3.18}$$

constitutes an asymptotic level α test. The critical value c_α represents the value for which the probability that a set of $\mathcal{D}$ independent one-dimensional Brownian bridges all remain in the region $(-c_\alpha, c_\alpha)$ equals $1-\alpha$. A CM-type test is constructed analogously, except the distribution of the L^2-norm of a Brownian bridge is unknown, so the critical values must be obtained via simulation. To this end, we generate 50,000 Brownian bridge sample paths on a fine mesh, compute the appropriate L^2 norms via numerical integration, and obtain critical values by estimating the various quantiles of interest. Critical values for selected dimensions and significance levels are listed in Table 3.1.

It remains to explain how we deal with the nuisance parameter(s) $\{V(t) : t \in [0,1]\}$. Our approach is simply to estimate it from the data. There are various ways this could be done, but we find the following works well in practice. Under the null, the (single) nuisance parameter $V = V(t)$ represents the covariance matrix of $\text{vecu}(d\boldsymbol{\Theta}\boldsymbol{\Theta}^T)$, where the redundant time-dependence in $\boldsymbol{\Theta} = \boldsymbol{\Theta}(t)$ is suppressed. Our estimator for V will be the

Table 3.1: Asymptotic critical values for selected dimensions and significance levels.

(a) Critical values for selected dimensions and significance levels.

		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
d	$\mathcal{D}$	CM	KS	CM	KS	CM	KS
2	1	0.743	1.628	0.460	1.358	0.346	1.224
3	3	0.953	1.788	0.648	1.544	0.524	1.425
4	6	1.086	1.882	0.775	1.652	0.643	1.540
5	10	1.173	1.949	0.874	1.727	0.733	1.620
10	45	1.479	2.133	1.152	1.933	1.024	1.837
15	105	1.623	2.230	1.310	2.039	1.174	1.949
20	190	1.724	2.296	1.433	2.111	1.287	2.024
25	300	1.824	2.345	1.532	2.164	1.370	2.079

associated sample covariance matrix estimated from the entire set of $k_{\text{total}} := kn/b$ radial threshold exceedances taken from all blocks. That is

$$\hat{V} := \frac{1}{k_{\text{total}}} \sum_{l=1}^{n} W_l W_l^T \mathbf{1}\{R(l/n) > \hat{u}((2\lceil l/b \rceil - 1)h)\}$$

$$W_l := \text{vecu}(d\boldsymbol{\Theta}(l/n)\boldsymbol{\Theta}(l/n)^T) - \hat{\boldsymbol{\sigma}}((2\lceil l/b \rceil - 1)h).$$

Provided the *rank condition* $k_{\text{total}} > \mathcal{D}$ is satisfied, the estimator $\hat{V}$ is full-rank and therefore invertible. For a fixed sample size and set of tuning parameters, the rank condition imposes an upper limit on the dimension, roughly $d < \sqrt{2k_{\text{total}}}$. It seems natural that such a restriction should exist: reliable inference in high-dimensional settings requires commensurate data. For fixed n, we may reduce b and/or increase k in order to enlarge the effective sample size, but these parameters are subject to their own particular trade-offs that will influence the performance of the test. Alternatively, one could substitute V^{-1} with the pseudoinverse to circumvent invertibility concerns. This avenue is not explored and in any case it doesn't seem sensible to proceed with the test in circumstances where violation of the rank condition indicates there is insufficient data for the task at hand.

3.2 Simulation experiments

In this section, we present a series of numerical experiments demonstrating our method's performance and, where applicable, draw conclusions regarding its relative merits compared to Drees (2023).

3.2.1 Data generating processes

Suppose $\boldsymbol{X}(t)$ has dimension d and its extremal dependence structure is parametrised by $\vartheta(t) \in \Omega$, where Ω is a convex parameter space. Let $\vartheta_0, \vartheta_1 \in \Omega$ denote arbitrary parameters. We consider three scenarios for how the dependence of $\boldsymbol{X}(t)$ varies over time:

1. **Constant:** the parameter is fixed, i.e. $\vartheta(t) = \vartheta_0$.
2. **Jump:** the parameter changes (instantaneously) from ϑ_0 to ϑ_1 at a change point $\tau \in (0,1)$, i.e. $\vartheta(t) = \vartheta_0 \mathbf{1}\{t < \tau\} + \vartheta_1 \mathbf{1}\{t \geq \tau\}$. In all experiments we set $\tau = 0.5$.
3. **Linear:** the parameter evolves linearly from ϑ_0 to ϑ_1, i.e. $\vartheta(t) = \vartheta_0 + t(\vartheta_1 - \vartheta_0)$. Convexity of Ω guarantees that $\vartheta(t) \in \Omega$ for all $t \in [0,1]$.

The parametric models we consider are as follows:

1. **Symmetric logistic (SL):** the dependence structure is characterised via the extreme value copula given by

$$C(u_1, \ldots, u_d) = \exp\left(- \left[\sum_{j=1}^{d} (-\log u_j)^{\vartheta(t)} \right]^{1/\vartheta(t)} \right).$$

   The parameter space is $\Omega = [1, \infty)$, with asymptotic independence when $\vartheta(t) = 1$ and complete asymptotic dependence as $\vartheta(t) \to \infty$.

2. **Hüsler-Reiss (HR):** the dependence structure is characterised by the variogram $\Gamma(t) = \vartheta(t)\Gamma_0$, where $\Gamma_0 \in \mathbb{R}^{d \times d}$ is a conditionally negative definite matrix and $\vartheta(t) \in \Omega = (0, \infty)$. Under this model, the extremal dependence coefficient between X_i and X_j at time $t \in [0,1]$ is $\chi_{ij}(t) = 2\bar{\Phi}(\sqrt{\Gamma_{ij}(t)}/2)$, where $\bar{\Phi}$ is the survival function of the standard normal distribution. Asymptotic independence between X_i and X_j occurs as $\Gamma_{ij}(t) \to \infty$ and complete asymptotic dependence occurs if $\Gamma_{ij}(t) = 0$. The multiplicative scalar $\vartheta(t)$ has the effect of increasing ($0 < \vartheta(t) < 1$) or decreasing ($\vartheta(t) > 1$) the strength of all pairwise dependencies (relative to Γ_0). While not strictly necessary, we take $\vartheta_0 = 1$ so that Γ_0 parametrises the dependence at time $t = 0$. For fixed d, the elements of the initial variogram Γ_0 are generated randomly using (elements of) the procedure outlined in Appendix B1 in Fomichov and Ivanovs

(2023). Specifically, we set $\Gamma_{0,ij} = \frac{3}{d}\|\boldsymbol{h}_i - \boldsymbol{h}_j\|_2^2$, where $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d$ are independent d-dimensional random vectors whose components are independent, identically distributed Pareto random variables with shape parameter equal to 2.5. The scaling factor $3/d$ ensures a suitable distribution for the extremal dependence coefficients.

Data are generated via the `rmev` function in the `mev` package. Our nomenclature for referring to the six qualitatively different models is as follows: HR-jump refers to the Hüsler-Reiss model with a jump change in dependence, SL-linear refers to the symmetric logistic model with linearly evolving dependence and so on.

For bivariate experiments Drees (2023) is included as a comparator. Results pertaining to their test are based on $\mathcal{A} = \{A_y : y = 0.01, 0.02, \ldots, 0.99\}$, where $A_y := \{\boldsymbol{\theta} \in \mathbb{S}_+^1 : \theta_1 \leq y\} \subset \mathbb{S}_+^1$.

3.2.2 Results: asymptotic (large sample) performance

In an idealised setting with infinite data, the asymptotic theory in the previous section holds exactly. In practice we are naturally limited to finite samples, but we can validate our theoretical results empirically by taking n sufficiently large and choosing k and b are appropriately. Specifically, the test's p-values under (3.17) should be uniformly distributed under the null and the test's empirical power under fixed alternatives should converge to 100%.

First, we examine the asymptotic empirical distribution of the test statistics under the null. We generate 350 samples of size $n = 10^6$ from the SL-constant ($\vartheta_0 = 2$) and HR-constant ($\vartheta_0 = 1$) models in dimensions $d \in \{2, 5\}$. The bandwidth is $h = 10^{-3}$ and the level is $k = 50$. This yields 500 blocks of size $b = 2,000$, a sampling fraction $k/b = 2.5\%$, and an overall effective sample size of $k_{\text{total}} = 25,000$. Figure 3.1 depicts the empirical quantile functions of the p-values (upper plots) and test statistics (lower plots) against their theoretical counterparts. For the KS-type test, the theoretical quantiles in the QQ plots are computed using the exact Kolmogorov quantile function; for the CM-type test they are estimated from the set of simulated Brownian bridges discussed earlier. For both dimensions and models, the empirical p-values are approximately uniformly distributed. This indicates that for all nominal sizes the corresponding tests will approximately maintain the desired level. Analogous plots for Drees (2023) method can be found in Figure 7 within their Supplementary Material.

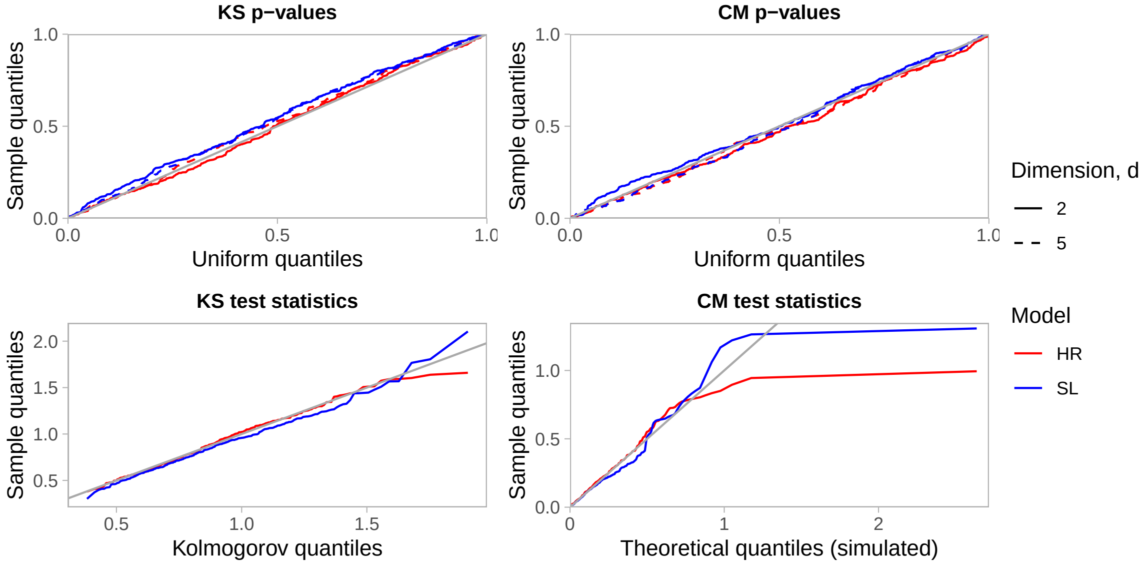

Figure 3.1: Large sample QQ plots for the p-values (top) and test statistics (bottom) associated with the KS test (left) and CM test (right). Based on 350 simulations from the SL- and HR-constant models with $n = 10^6$, $b = 2000$ and $k = 50$.

Next, we check that our procedure can leverage abundant information to detect dependence changes with high probability (i.e. is consistent under certain fixed alternatives). The experimental procedure is unchanged, except that data are now generated from SL-jump ($\vartheta_0 = 2$, $\vartheta_1 = 2.5$) and HR-jump ($\vartheta_0 = 1$, $\vartheta_1 = 1.5$) models. These values are chosen to bring about relatively subtle shifts in the dependence structure. Nevertheless our method consistently and overwhelmingly identifies non-stationary dependence. For the SL-jump data all p-values equal zero, up to numerical precision. For HR-jump data, the 90% empirical quantile of the p-values is 2×10^{-6}.

3.2.3 Results: finite sample performance

In finite sample settings, empirical size of an asymptotic test will generally differ from the nominal size; we only guarantee that the correct level is attained asymptotically. The hope is that convergence occurs with sufficient rapidity that this difference is acceptably small. We conduct repeated simulations from the SL-constant ($\vartheta_0 = 2$) and HR-constant ($\vartheta_0 = 1$) models in dimensions $d \in \{2, 5, 10, 25\}$ with sample sizes $n \in \{2500, 5000, 10000\}$.

For each data set, we apply hypothesis tests with nominal level 5%, based on various combinations of hyperparameters b and k. Specifically, the number of blocks is $n/b \in \{25, 50\}$ and the proportion of extreme observations within each block is $k/b \in \{0.05, 0.10, 0.15\}$. Table 3.2 reports the empirical Type I error rates of these tests. Blank cells indicate that the corresponding tuning parameters violate the rank condition or the condition $k \leq d$. Large sample results are included in the tables for completeness – recall that these are only available in dimensions $d \in \{2, 5\}$.

The size of our test exceeds the nominal level by at most 1.4% and 3.3% for the SL and HR models, respectively. Moreover, and arguably more pertinently, under any scenario (i.e. model, sample size and dimension) there exists hyperparameters for which this difference is at most 0.6%. This suggests that where there is a large discrepancy between the nominal and empirical error rates, suboptimal hyperparameter selection may be the proximate cause. Having said that, the general stability in the empirical error rates demonstrates a certain degree of robustness to hyperparameter choices. The KS-based test is universally more conservative than the CM-based test, particularly for larger block lengths. The same pattern is observed in Drees (2023); it stems from the fact that a coarsely discretised path may only attain its supremum at a small number of points, whereas the corresponding critical values arise from suprema of continuous processes.

Next we examine the empirical power under alternatives. Figure 3.2 shows the power across a range of scenarios where data generating process undergoes a jump/linear dependence change of varying magnitude. All values are based on 1000 bivariate datasets of size $n = 2500$; the six panels within each sub-plot illustrate the power for various hyperparameter choices. The nominal size of the tests (5%) is indicated by the grey dashed line. Of course, when the null hypothesis is true ($\vartheta_1 = 2$ for SL, $\vartheta_1 = 1$ for HR), the power reverts to approximately this level.

The power doesn't appear to be overly sensitive to the choice of b and k, but is generally underpowered (relative to other choices) when $n/b = 25$, $k/b = 0.05$. This is because the test only has a small number of noisy TPDM estimates at its disposal. Conversely, the power tends to be marginally greatest when $n/b = 50$ and $k/b = 0.15$. However, with such a large effective sample size ($k_{\text{total}} = \lfloor 0.15 \times 2500/50 \rfloor \times 50 = 350$) we might suspect that observations from the bulk are biasing the results. In this instance the bias appears to

Table 3.2: Empirical Type I error rates (%) across repeated simulations. The number of simulations is $N = 1000$ if $n \leq 10^4$ and $d \leq 5$, or $N = 300$ otherwise. All tests have nominal size 5%.

(a) SL-constant.

			$d = 2$				$d = 5$		$d = 10$		$d = 25$	
			Drees		Pawley		Pawley		Pawley		Pawley	
n	n/b	k/b	CM	KS	CM	KS	CM	KS	CM	KS	CM	KS
2,500	25	0.050	3.2	2.4	5.5	2.9						
		0.100	3.9	2.2	5.0	2.7	4.5	1.2				
		0.150	3.6	2.1	5.6	3.7	6.1	3.0	2.9	0.6		
	50	0.100	3.6	2.6	6.4	3.5						
		0.150	2.9	2.1	5.8	3.9	3.8	2.2				
5,000	25	0.050	3.7	1.0	4.7	2.9	4.5	1.4				
		0.100	3.5	2.2	4.9	2.5	4.5	1.5	3.1	1.7		
		0.150	3.5	2.0	5.4	2.7	5.1	2.3	3.7	0.9	4.0	0.3
	50	0.050	4.1	3.1	5.5	3.7						
		0.100	3.7	3.2	5.5	3.5	4.2	2.6				
		0.150	3.9	2.5	5.8	3.4	5.9	2.9	4.0	0.9		
10,000	25	0.050	4.6	3.0	4.5	2.1	4.5	2.0	4.0	0.9		
		0.100	4.5	2.5	4.4	2.5	4.3	2.3	3.4	1.1	1.7	0.6
		0.150	3.6	2.1	4.5	2.2	4.0	1.9	5.7	2.3	3.4	0.6
	50	0.050	3.6	2.5	4.1	2.7	5.5	3.2				
		0.100	4.5	2.9	5.9	2.8	5.1	2.4	5.4	1.7		
		0.150	3.9	2.4	5.1	2.8	6.0	3.0	3.1	0.9	5.4	2.6
1,000,000	500	0.025	2.9	3.1	4.3	3.4	5.1	4.6				

(b) HR-constant.

			$d = 2$				$d = 5$		$d = 10$		$d = 25$	
			Drees		Pawley		Pawley		Pawley		Pawley	
n	n/b	k/b	CM	KS	CM	KS	CM	KS	CM	KS	CM	KS
2,500	25	0.050	3.7	2.0	5.5	3.6						
		0.100	4.7	2.9	6.9	4.2	4.8	1.5				
		0.150	3.8	2.4	6.5	3.8	4.9	1.9	4.6	1.7		
	50	0.100	3.4	2.6	6.6	4.5						
		0.150	4.6	2.9	7.4	5.4	4.6	2.5				
5,000	25	0.050	3.6	1.8	5.6	3.8	4.5	2.9				
		0.100	4.0	2.5	8.3	4.9	5.0	2.4	1.7	0.0		
		0.150	4.4	2.6	6.4	3.3	4.5	2.1	4.3	2.6	1.4	0.3
	50	0.050	4.1	2.7	7.5	5.3						
		0.100	5.3	3.8	8.2	5.3	4.2	2.2				
		0.150	4.7	3.8	6.1	4.8	5.2	2.5	2.6	1.4		
10,000	25	0.050	4.4	2.6	6.6	4.2	4.5	1.8	4.0	0.6		
		0.100	5.8	2.8	5.6	3.0	5.2	2.5	3.7	2.0	0.9	0.0
		0.150	5.1	3.4	6.9	3.6	5.3	1.9	6.3	2.6	2.0	1.1
	50	0.050	4.4	3.2	6.9	5.1	5.1	2.4				
		0.100	3.8	2.8	5.8	3.5	5.7	3.0	4.9	2.6		
		0.150	4.6	2.9	5.4	3.5	4.9	3.5	6.0	4.6	2.0	0.6
1,000,000	500	0.025	6.0	4.9	5.4	4.9	6.6	4.6				

enhance the power, but it need not, since changes in dependence in the bulk and in the tail are generally two separate matters.

When dependence changes abruptly (SL-jump and HR-jump), the CM- and KS-based tests perform equally well. Upon further investigation, we find that for these models that paths $\{\hat{Z}_{ij}(t) : t \in [0,1]$ are roughly $\wedge$-shaped curves attaining their suprema at $t = 0.5$ i.e. when the changepoint occurs. Very loosely speaking, the CM-type test statistic corresponds to the largest area under these (squared) curves, while the KS-type test statistic corresponds to the largest supremum. By picturing $\{\hat{Z}_{ij}(t) : t \in [0,1]$ as a triangle of width one and height $Z_{ij}(0.5)$, it becomes apparent that both test statistics are simply functions of $Z_{ij}(0.5)$ and thus contain equivalent information. For the linear dependence changes, the CM-based test is superior.

Empirically, our test is more highly powered than Drees (2023). It achieves near full power for the SL-jump change; in the more challenging case of the HR-linear model, for which Drees' test is virtually powerless, our CM-type test discerns a signal more often that not. Initially, this might seem rather counterintuitive, since Drees (2023) leverages the full angular measure, whereas we rely solely on summary information. However, one can think of our method as imposing some additional structure or information, namely that dependence is captured via the TPDM. When this assumption is fulfilled, a method that incorporates it will generally be superior to a fully non-parametric method that doesn't. In the case of the HR model this assumption does hold exactly, since dependence is fully characterised by the variogram $\Gamma(t)$, which is in one-to-one correspondence with the set of TPDMs. *(If more detail is needed to substantiate this claim, use Section 2.3 in https://arxiv.org/pdf/1207.6886 and Section 3 in Supp. Material of Cooley.)*

The Q-Q plots in Figure 3.3 show how the power improves as more data is acquired. For a given ϑ_1, as n increases, the curves shift further below the main diagonal. When the curves lies below the diagonal, this indicates that the rejection rate exceeds the nominal level and the test has power.

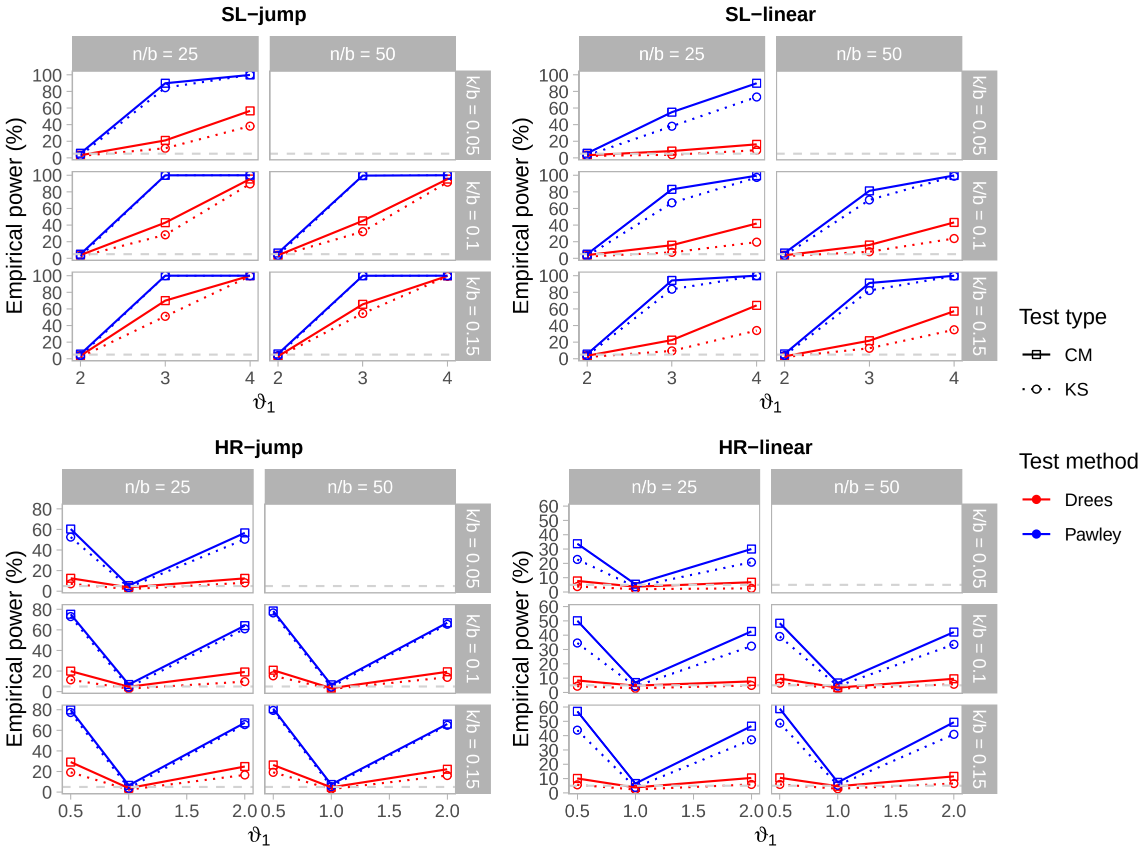

Figure 3.2: Empirical power against the dependence parameter ϑ_1. Based on 1000 simulations with $n = 2,500$ and $d = 2$.

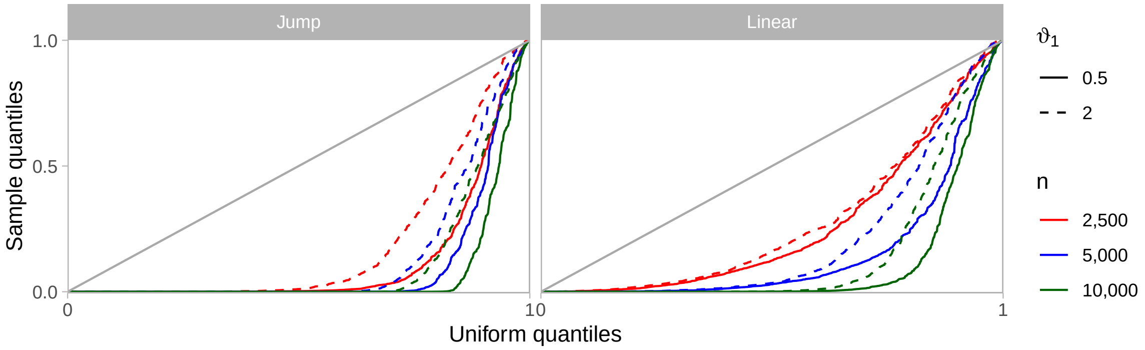

Figure 3.3: QQ-plots for the KS p-values with varying sample size. Based on 1000 simulations from the HR-jump (left) and HR-linear (right) models with $\vartheta_1 \in \{0.5, 2\}$, $n/b = 25$ and $k/b = 0.1$.

Discussion of computation time here, referring to Figure 3.4.

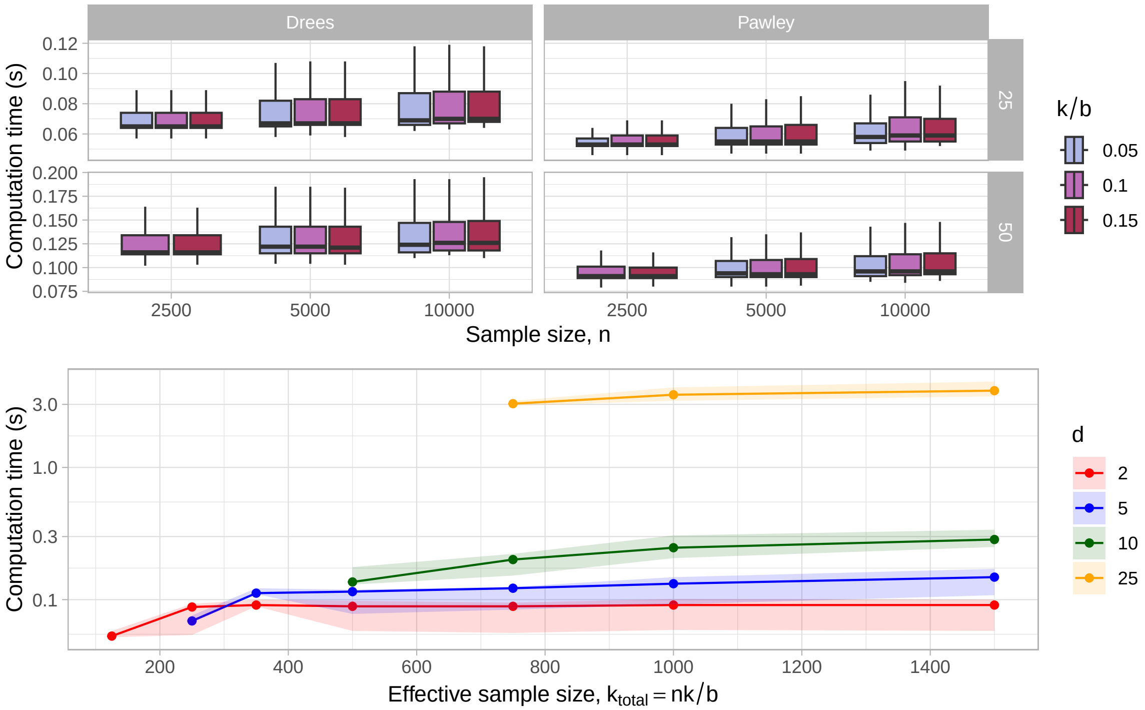

Figure 3.4: Average computation time across numerical experiments.

3.3 No free lunch: constant TPDM with changing dependence

Our proposed extension to Drees (2023) affords many advantages, most notably the ability to conduct tests in high dimensions. The price paid is that we forgo the ability to detect TPDM-invariant dependence changes. (The existence of such changes is consequence of the many-to-one correspondence between angular measures and TPDMs.) For this class of alternatives our test will be inherently predisposed to commit Type II errors. In this section, we illustrate this flaw by constructing a sub-class of examples based on a time-dependent version of the max-linear model.

Suppose $\{\boldsymbol{X}(t) : t \in [0,1]\}$ is a d-dimensional stochastic process defined by

$$\boldsymbol{X}(t) = A(t) \times_{\max} \boldsymbol{Z}(t), \qquad A(t) = A_0 \mathbf{1}\{t < 0.5\} + A_1 \mathbf{1}\{t \geq 0.5\}. \tag{3.19}$$

The stochastic innovations process $\{\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_q(t)) : t \in [0,1]\}$ is a collection of independent random vectors; for any $t \in [0,1]$ the $q \geq 1$ components of $\boldsymbol{Z}(t)$ are independently Fréchet distributed with shape parameter equal to 2. The dependence structure of $\boldsymbol{X}(t)$ is characterised by the parameter matrix $A(t) = (a_{ij}(t)) \in \mathbb{R}_+^{d\times q}$. Under the model (3.19), the dependence parameter undergoes a jump-change from $A_0 \in \mathbb{R}_+^{d\times q}$ to $A_1 \in \mathbb{R}_+^{d\times q}$ at time $t = 0.5$. More flexible models can easily be conceived, whereby $A(t)$ evolves smoothly, perhaps even with a varying number of factors $q = q(t)$, but the simple model above will suffice for our aims. The local angular measure associated with (3.19) can be expressed in terms of the columns $\boldsymbol{a}_1(t), \dots, \boldsymbol{a}_q(t) \in \mathbb{R}_+^d$ of $A(t)$ as

$$H(\cdot\,;t) = \sum_{j=1}^{q} \|\boldsymbol{a}_j(t)\|_2^2 \delta_{\boldsymbol{a}_j(t)/\|\boldsymbol{a}_j(t)\|_2}(\cdot).$$

The local TPDM is given by $\Sigma(t) = A(t)A(t)^T$ and the diagonal and off-diagonal entries of its asymptotic covariance $V(t)$ matrix are given by **krali**

$$k\mathrm{Cov}(\hat{\sigma}_{ij}(t), \hat{\sigma}_{lm}(t)) \to \begin{cases} d\sum_{s=1}^{q} \frac{a_{is}(t)^2 a_{js}(t)^2}{\|\boldsymbol{a}_s(t)\|_2^2} - \sigma_{ij}(t)^2, & i = l, j = m, \\ d\sum_{s=1}^{q} \frac{2a_{is}(t)a_{js}(t)a_{ls}(t)a_{ms}(t)}{\|\boldsymbol{a}_s(t)\|_2^2} - 2\sigma_{ij}(t)\sigma_{lm}(t), & \text{otherwise.} \end{cases} \tag{3.20}$$

If A_0 and A_1 are distinct (up to permutations of their columns) yet carefully chosen so that $A_0A_0^T = A_1A_1^T$ and (3.20) yields identical asymptotic covariances, then the alternative hypothesis (3.3) is true but the convergences (3.17) still hold. Finding non-trivial (i.e. $q > 2$) pairs A_0, A_1 by hand would be extremely laborious, if not impossible, so we resort to a computational approach. We generate $N \gg 1$ candidate $d \times q$ matrices with uniformly distributed entries; the rows of each matrix are subsequently normalised to ensure the resulting TPDM is properly scaled. Then we search for pairs of matrices satisfying (within some small tolerance) the required conditions. Using this procedure with $d = 2$, $q = 20$, and $N = 50,000$, we find a suitable matrix pair for which $\sigma_{12}(t) = 0.1000$ and $k\mathrm{Var}(\hat{\sigma}_{12}(t)) \to 0.060$, to three decimal places. We generate 1000 datasets, each with $n = 10,000$ samples, from the model (3.19) using the`SpatialExtremes` package. For each dataset, we apply our test and that of Drees (2023) with $b = 400$ and $k = 40$.

The diagnostic plots in Figure 3.5 illustrate the computations underlying our testing procedure when applied to one of these datasets. The top-left panel depicts the empirical local TPDM over time. (Since $d = 2$, there is only one component pair to consider.) The range of values of $\hat{\sigma}_{12}(t)$ is consistent with (3.10), which implies that

$$\left(0.1000 - \Phi^{-1}(0.975)\sqrt{\frac{0.060}{40}}, 0.1000 + \Phi^{-1}(0.975)\sqrt{\frac{0.060}{40}}\right) \approx (0.724, 0.876)$$

represents a 95% asymptotic confidence interval for $\sigma_{12}(t)$. Indeed, the empirical coverage of the interval, based on the $1000 \times n/b = 25000$ estimates of $\sigma_{12}(t)$ from across the full set of simulations, equals 93.56%. There is no temporal trend in the blocks' TPDMs, so the integrated TPDM (top-right panel) is a straight line and the test process $Z_{12}(t)$ (bottom-left) resembles a typical Brownian bridge sample path. The bottom-right panel depicts $\int_0^t |\hat{Z}_{12}(s)|\,\mathrm{d}s$ (CM, upper sub-panel) and $\sup_{0\leq s\leq t} |\hat{Z}_{12}(s)|$ (KS, lower sub-panel) as functions of t. The maximal values of these processes do not exceed the associated critical values at the 5% level, marked by the dashed lines. We conclude there is insufficient evidence to conclude that dependence is changing and commit a Type II error. The empirical Type II error rates across all 1000 replications of the experiment are 94.5% (CM) and 96.5% (KS). As expected, the empirical power of the test is approximately the desired Type I error rate.

Figure 3.6 shows the analogous plots corresponding to Drees' method applied to the same data. The left-hand plot depicts the empirical integrated angular measure as a function of t. Each curve corresponds to a particular set $A_y \in \mathcal{A}$, with darker colours indicating larger values of y. Close inspection of these curves reveals a slight kink at $t = 0.5$. The dependence change is more apparent in the middle panel, which depicts the corresponding $|\mathcal{A}|$-dimensional test process. This process is analogous to (3.14), but its interpretation is less straightforward because the curves are cross-correlated. Computing the relevant time-integrals of these processes yields the curves in the right-hand plot. For both the CM- and KS-based tests, there exists a curve that enters the rejection region demarcated by the dashed lines, so according to either test we would (correctly) reject the null hypothesis at the 5% level. Upon repeating this 1000 times, the test's empirical power is found to be 100% (CM) and 99.8% (KS).

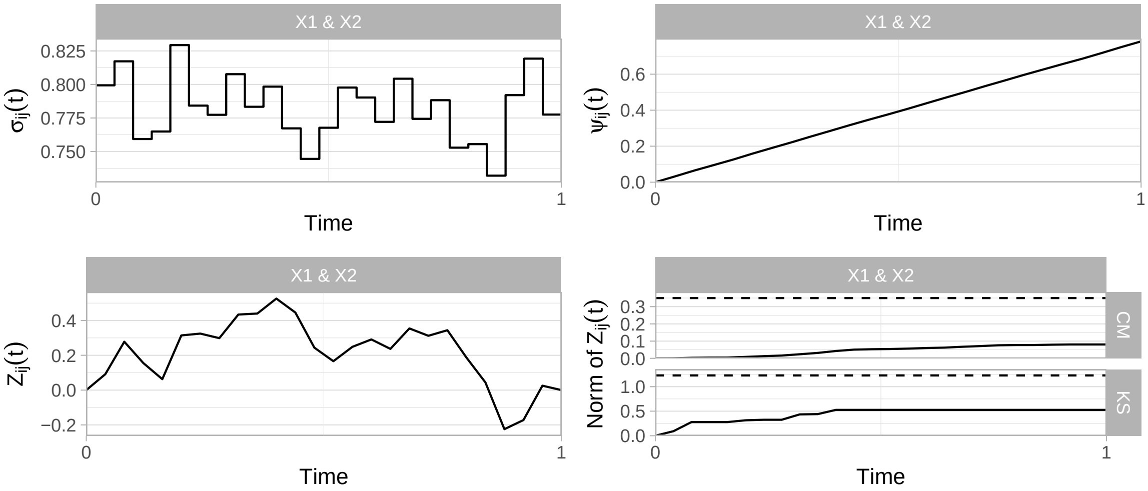

Figure 3.5: Diagnostic plots for our test for data from (3.19) with $n = 10,000$, $b = 400$, $k = 40$.

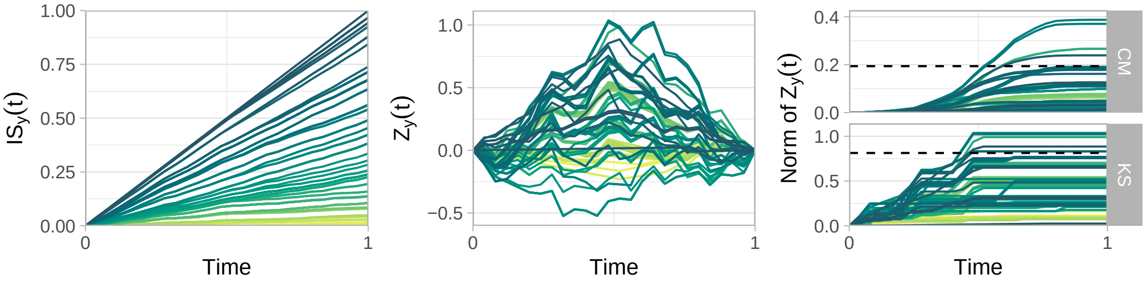

Figure 3.6: Diagnostic plots for Drees' test for data from (3.19) with $n = 10,000$, $b = 400$, $k = 40$. Each curve represents a set A_y with darker colours indicating larger values of y.

3.4 Application: extreme Red Sea surface temperatures

We now apply our methodology to test for changing dependence in extreme Red Sea surface temperature anomalies. The dataset has been widely studied in the extremes community, primarily because it was the focus of the EVA 2019 Data Challenge but also because extreme temperatures are related to ecological issues such as coral bleaching. Further details about the data collection and pre-processing can be found in Huser (2020).

Previous investigations by Simpson and Wadsworth (2020) and Huser (2020) conclude that surface temperature extremes exhibit differing behaviour in the north and south, so it is advisable to treat these areas separately. We divide the spatial domain into northerly and southerly sub-regions, each comprising 70 sites whose are shown in Figure 3.7.

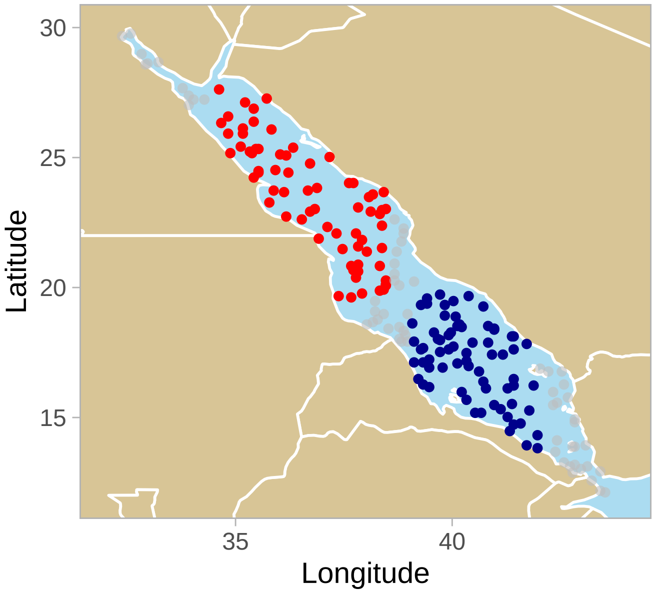

Figure 3.7: Locations of the 70 sites in each of the two sub-regions in the Red Sea.

At any particular location, daily maxima are known to occur in (temporal) clusters, meaning high temperatures may persist across several days (Simpson and Wadsworth, 2020). We address this by working with weekly maxima, so that observations are approximately independent over time. This yields $n = 1605$ samples spanning approximately 31 years. Let $X_i^{(\text{north})}(t)$ and $X_i^{(\text{south})}(t)$ denote the surface temperature anomaly (on stationary Fréchet margins) at site $i \in \{1, \dots, 70\}$ and time $t \in [0, 1]$ in the north and south sub-regions, respectively.

Our goal is to determine whether either of

$$\boldsymbol{X}^{(\text{north})}(t) = \{X_i^{(\text{north})}(t) : i = 1, \dots, 70\}, \qquad \boldsymbol{X}^{(\text{south})}(t) = \{X_i^{(\text{south})}(t) : i = 1, \dots, 70\}$$

exhibit evidence for stationary or changing extremal dependence. To this end, we will apply our test using $b = 107$ and $k = 20$, yielding 15 blocks and an effective sample size of $k_{\text{total}} = 15 \times 20 = 300$. The rank condition (and other considerations) restricts us to testing up to 17 sites at a time; it is not possible/advisable to test for changing dependence in each region using all 70 sites. Our strategy will be to repeatedly sample $2 \leq d \leq 17$ sites from each region. We apply this procedure $N = 1000$ times for $d \in \{5, 10, 15\}$, perform our test, and collate the resulting p-values. Their distributions are shown in Figure 3.8.

Rough summary of conclusions: North shows evidence of changing dependence, South not so much; results for $d = 15$ are unreliable as convergence unlikely (based on earlier tables etc.); CM has higher rejection rate than KS (aligns with sim studies that shows CM has greater power than KS when dependence change is gradual, as is likely the case here). For $d = 5$, the p-values are strongly skewed towards zero, resulting in a rejection rate of approximately 60% (for both KS and CM).

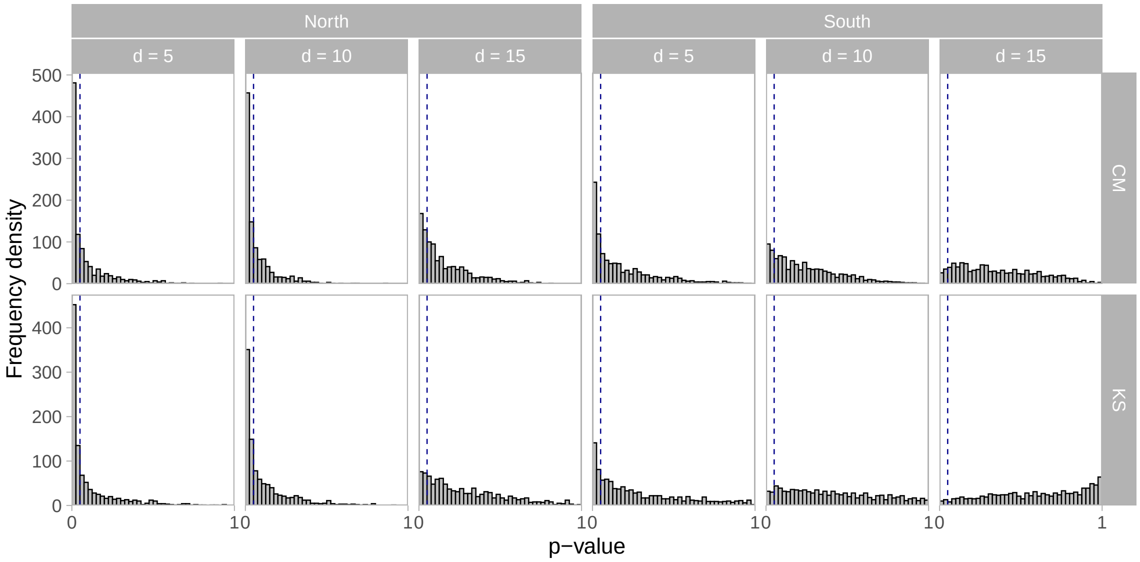

Figure 3.8: Blah.

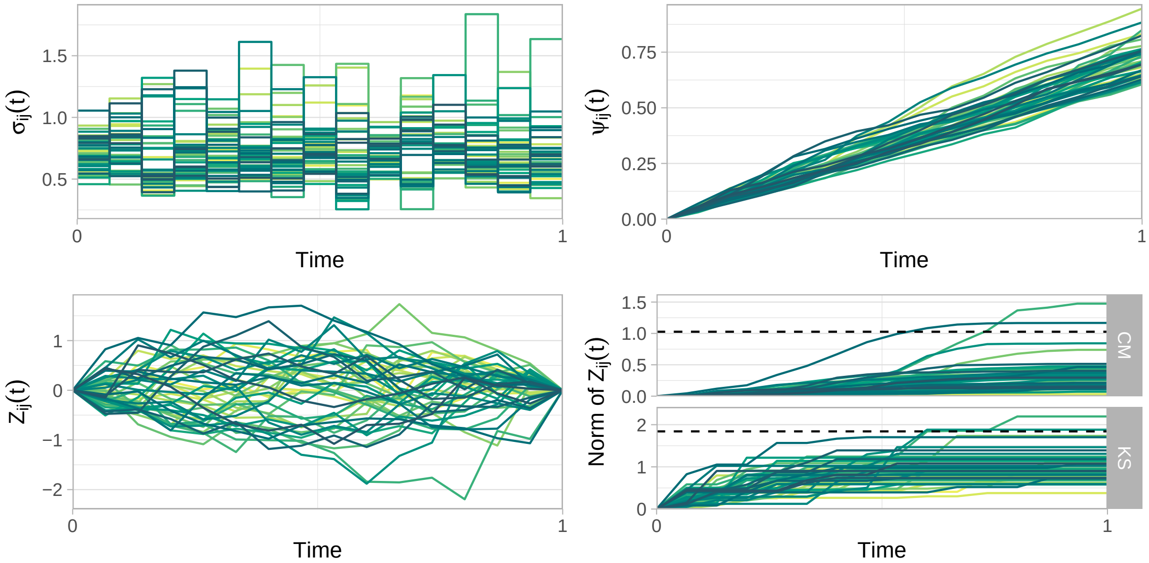

Figure 3.9: Diagnostic plots for our test, based $b = 107$ and $k = 20$, applied to data from $d = 10$ randomly selected northerly sites in the Red Sea. Each curve corresponds one of the $\mathcal{D} = 45$ component pairs.

3.5 Extensions and modifications

3.5.1 Alternative dependence measures

Our method considers the time-evolution of the dependence between X_i and X_j according to the measure

$$\sigma_{ij}(t) = \lim_{r\to\infty} \mathbb{E}[f(\boldsymbol{\Theta}(t)) \mid R(t) > r], \tag{3.21}$$

where $f : \mathbb{S}_+^{d-1} \to \mathbb{R}_+$ is defined by $f(\boldsymbol{\Theta}) = d\theta_i\theta_j$. However, the EDM/TPDM is just one measure of extremal dependence among a large class. Alternative measures can be generated by replacing f in (3.21) with other functions $g : \mathbb{S}_+^{d-1} \to \mathbb{R}_+$ (Larsson and Resnick 2012). Provided g satisfies the conditions of Theorem 4 in Klüppelberg and Krali (2021), the theory underpinning our testing methodology holds. That these alternative measures lack the nice properties of the TPDM, such as positive definiteness, is not particularly relevant for the task-at-hand. The circumstances under which a particular measure is inferior/superior (in terms of power, say) to others is governed by the nature of the associated function g. A practitioner working in a particular setting, where the dependence structures and dependence changes tend to be of a certain nature, might wish to tailor the dependence measure to suit their purposes. This could be achieved by running a series of numerical experiments, designed to mimic the scenarios they typically encounter, and choosing g optimally among some (parametric) subfamily according to some performance metric.

I could illustrate this process with a simple toy example, e.g. take $g(\boldsymbol{\theta};\gamma) = d\theta_i^\gamma\theta_j^\gamma$ and find $\gamma \in (0,4]$ that achieves maximal empirical power on a particular model.

3.5.2 Changepoint detection

Our primary objective was to devise a test to ascertain whether or not an assumption of constant tail dependence is reasonable. In certain applications (e.g. finance), it may be more interesting to ask *when*, not if, dependence has changed. This is the realm of changepoint detection. Suppose the angular measure of $\{\boldsymbol{X}(t) : t \in [0,1]\}$ is given by

$H(t) = H_0\mathbf{1}\{t \leq \tau\} + H_1\mathbf{1}\{t > \tau\}$ for some $\tau \in (0,1)$. Then

$$\hat{\tau} = \ldots$$

is a CUSUM-type estimator of τ. *Discuss (and illustrate) how this estimator is biased towards the centre of the time interval, and the modifications that would be needed to remedy this.*

3.5.3 Robustness

Test robustness to serial dependence. (e.g. simulation from AR process)

3.6 Other things could do

Do example where dependence only changes in a subset of components. How does power vary against proportion of pairs that undergo change?

4 Compositional perspectives on extremes

The primary object of interest in multivariate extremes is the angular measure, which represents the limiting conditional distribution of $\boldsymbol{\Theta} \mid R > t$, where $\boldsymbol{\Theta} := \boldsymbol{X}/\|\boldsymbol{X}\|_\alpha$ and $R := \|\boldsymbol{X}\|_\alpha$, as $t \to \infty$. Taking $\alpha = 1$ and $\boldsymbol{X} \in \mathrm{RV}_+^d(1)$ on unit Fréchet margins, the following statements are true:

1. $\boldsymbol{\Theta}$ lies in the $(d-1)$-dimensional simplex $\mathbb{S}_+^{d-1} = \{\boldsymbol{\theta} \in \mathbb{R}^d : \theta_j \geq 0, \sum_{j=1}^d \theta_j = 1\}$ in $\mathbb{R}^d$.
2. In the limit as $t \to \infty$, the angle of threshold exceedances $\boldsymbol{\Theta} \mid R > t$ is independent of R.
3. The (normalised) angular measure satisfies $\int_{\mathbb{S}_+^{d-1}} \theta_j \, \mathrm{d}H(\boldsymbol{\theta}) = 1/d$ for all $j = 1, \ldots, d$.

Property (i) states that $\boldsymbol{\Theta}$ is a d-part *random composition* with non-negative components summing to unity. In his seminal paper, Aitchison (1982) contended that analysing such data using standard methodology designed for unconstrained vectors can lead to misleading inferences. Compositional data analysis (CoDA) subsequently emerged as a discipline for developing statistical theory and techniques tailored to account for the geometry of the simplex. Statement (ii) expresses that, in the limit, MRV random vectors satisfy a fundamental principle of CoDA called *scale invariance*: the distribution of the composition (angular component) of is independent of the absolute size (radial component). Finally, the moment constraint (iii) means that the centre of mass of any valid angular measure (with respect to $\|\cdot\|_1$) lies at the barycentre of the simplex.

There is a clear connection between compositional data analysis and multivariate extreme value statistics. We are not the first to notice this link. S. G. Coles and Jonathan A. Tawn (1991) remark that parametric CoDA models would be useful for modelling extremal dependence, were it not for the fact that they typically violate the moment constraint. In a paper outlining potential future avenues for research in extremes, **longin2016** suggests applying "PCA for compositional data ...to the pseudo-angles" to "disentangle dependence into components ...of practical interest". Finally, Serrano (2022) leverage CoDA techniques (e.g. log-ratio transformations and compositional splines) to construct bivariate extreme value copulas.

This chapter aims to explore the link between these two statistical disciplines more thoroughly. In particular, we apply a CoDA lens to two statistical learning problems within multivariate extremes: tail dimension reduction via principal components analysis (Stephan Clémençon et al. 2024; Cooley and Thibaud 2019; Drees and Sabourin 2021) and binary classification in extreme regions (Jalalzai et al. 2018). We demonstrate that off-the-shelf CoDA methods are readily applicable to these problems and compare their performance against existing state-of-the-art methods from the extremes literature.

4.1 Compositional data analysis

Aitchison (1982) argues that standard multivariate data analysis techniques are inappropriate for modelling compositional data because they are designed for unconstrained data. Neglecting the compositional constraint causes an array of difficulties: spurious correlations (Pearson 1897; Aitchison 1982), a failure to capture the marked curvature characteristic of compositional data sets (Aitchison 1983), or contradictory conclusions between analyses depending on which components are included in the composition (Aitchison 1986). These issues are addressed by devising a statistical framework tailored to the algebraic-geometric structure of the underlying sample space: the unit simplex. This section reviews the basic concepts and principles of CoDA.

4.1.1 Compositions

Compositional analysis concerns vectors with strictly positive components for which all relevant information is *relative*, i.e. conveyed by ratios between components.

Definition 4.1. Vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^d := (0,\infty)^d$ are *compositionally equivalent*, denoted $\boldsymbol{x} \sim \boldsymbol{y}$, if there exists $c > 0$ such that $\boldsymbol{y} = c\boldsymbol{x}$.

The equivalence relation $\sim$ defines equivalence classes on $\mathbb{R}_+^d$.

Definition 4.2. For $\boldsymbol{x} \in \mathbb{R}_+^d$, the compositional class $[\boldsymbol{x}] := \{c\boldsymbol{x} : c > 0\}$ is represented on the d-part unitary simplex by its *closed composition* given by

$$\mathcal{C}\boldsymbol{x} := \frac{(x_1, \dots, x_d)}{\sum_{i=1}^d x_i}.$$

4.1.2 Aitchison geometry

Adhering to the principles propounded by Aitchison (1986) necessitates introducing alternative notions of mean/variance, distance, projections, etc. that make sense for compositions. Formally, this involves constructing a Hilbert space structure on $\mathbb{S}_+^{d-1}$ (Aitchison 1986; Pawlowsky-Glahn and Egozcue 2001). This is achieved by defining a vector space structure on $\mathbb{S}_+^{d-1}$ and endowing it with a suitable inner product, norm and distance.

Definition 4.3. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}_+^{d-1}$ be closed compositions and $\alpha \in \mathbb{R}$ a scalar. The perturbation and power operations are defined by

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \dots, x_d y_d), \qquad \alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \dots, x_d^\alpha).$$

It is straightforward to show that $(\mathbb{S}_+^{d-1}, \odot, \oplus)$ is a real vector space. The additive identity and inverse elements are $\boldsymbol{e}_a := \mathcal{C}(1, \dots, 1)$ and $\ominus\boldsymbol{x} := \mathcal{C}(x_1^{-1}, \dots, x_d^{-1})$, respectively.

Definition 4.4. The *centred log-ratio (CLR) transformation* is

$$\operatorname{clr} : \mathbb{R}^d \to \mathbb{R}^d, \qquad \boldsymbol{x} \mapsto \log\left(\frac{\boldsymbol{x}}{\bar{g}(\boldsymbol{x})}\right),$$

where $\bar{g}(\boldsymbol{x}) := (\prod_{i=1}^d x_i)^{1/d}$ denotes the geometric mean of the components of $\boldsymbol{x}$.

CLR-transformed vectors lie in the hyperplane $\mathcal{T}^{d-1} := \{\boldsymbol{y} \in \mathbb{R}^d : y_1 + \ldots + y_d = 0\} \subset \mathbb{R}^d$. The transformation can be inverted to recover the original (closed) composition via

$$\operatorname{clr}^{-1} : \mathcal{T}^{d-1} \to \mathbb{S}_+^{d-1}, \qquad \boldsymbol{v} \mapsto \mathcal{C}\exp(\boldsymbol{v}).$$

Definition 4.5. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}_+^{d-1}$ be closed compositions. Let $\langle \cdot, \cdot \rangle_e$ denote the Euclidean inner product in $\mathbb{R}^d$. The *Aitchison inner product* in $\mathbb{S}_+^{d-1}$ is

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a := \langle \mathrm{clr}(\boldsymbol{x}), \mathrm{clr}(\boldsymbol{y}) \rangle_e = \sum_{i=1}^{d} \log\left(\frac{x_i}{\bar{g}(\boldsymbol{x})}\right) \log\left(\frac{y_i}{\bar{g}(\boldsymbol{y})}\right).$$

The *Aitchison norm* and *Aitchison distance* are the metric elements induced by $\langle \cdot, \cdot \rangle_a$, that is

$$\|\boldsymbol{x}\|_a := \langle \boldsymbol{x}, \boldsymbol{x} \rangle_a^{1/2}, \qquad d_a(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a. \tag{4.1}$$

The CLR-transformation is an isometry between $\mathcal{T}^{d-1}$ equipped with Euclidean geometry and $\mathbb{S}_+^{d-1}$ equipped with Aitchison geometry.

Definition 4.6. The centre and total variance of a random composition $\boldsymbol{X}$ are given by

$$\mathrm{cen}_a(\boldsymbol{X}) := \operatorname*{arg\,min}_{y \in \mathbb{S}_+^{d-1}} \mathbb{E}[d_a^2(\boldsymbol{X}, \boldsymbol{y})] = \mathcal{C}(\exp(\mathbb{E}[\log(\boldsymbol{X})]),$$

$$\mathrm{totVar}_a(\boldsymbol{X}) := \mathbb{E}[d_a^2(\boldsymbol{X}, \mathrm{cen}_a(\boldsymbol{X}))] = \sum_{j=1}^{d} \mathrm{Var}([\mathrm{clr}(\boldsymbol{X})]_j).$$

4.1.3 Compositional principal components analysis

Compositional principal component analysis (CoDA-PCA) aims at finding low-dimensional descriptions of compositional data which retain most of the variability in the original data (Aitchison 1983). Classical PCA based on Euclidean geometry is ill-suited to this task for two main reasons. First, PCA is typically used as an exploratory tool for understanding the correlation structure among a set of variables, but the compositional constraint places restrictions on this structure and spurious correlations arise when these are not properly accounted for (Aitchison 1982). Second, the Hilbert space in which PCA is conducted should conform to the underlying sample space to facilitate a consistent and interpretable analysis. For example, compositional data often exhibit curvature that cannot be captured by Euclidean straight lines, and Euclidean projections of the data may lie outside of the simplex (Aitchison 1983).

Let $\boldsymbol{X} = (X_1, \dots, X_d)$ denote a d-part centred random composition with finite second order moments, that is $\mathrm{cen}_a(\boldsymbol{X}) = \boldsymbol{e}_a$ and $\mathbb{E}[\|\boldsymbol{X}\|_a^2] < \infty$. Let $\Gamma = \mathrm{Cov}(\mathrm{clr}(\boldsymbol{X}))$ denote the CLR-covariance matrix and $\Gamma = U\Lambda U^T$ its eigendecomposition, where Λ is a diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{d-1} \geq \lambda_d = 0$ and U is an orthonormal $d \times d$ matrix whose columns are the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d-1} \in \mathcal{T}^{d-1}$ and $\boldsymbol{u}_d \propto \mathbf{1}_d$. CoDA-PCA consists in retaining the leading $p \leq d-1$ (back-transformed) eigenvectors to account for a desired proportion of the total variability of $\boldsymbol{X}$ (Wang et al. 2015). For any linear subspace $V \subset \mathbb{S}_+^{d-1}$, let Π_V denote the orthogonal projection onto V. Then,

$$\mathcal{V}_p := \operatorname{span}_a(\operatorname{clr}^{-1}(\boldsymbol{u}_1), \ldots, \operatorname{clr}^{-1}(\boldsymbol{u}_p)) := \left\{ \bigoplus_{j=1}^{p} (\alpha_j \odot \operatorname{clr}^{-1}(\boldsymbol{u}_j)) : \alpha_1, \ldots, \alpha_p \in \mathbb{R} \right\}$$

minimises the expected squared (Aitchison) reconstruction error among all p-dimensional linear simplicial subspaces. The low-dimensional approximation $\Pi_{\mathcal{V}_p}\boldsymbol{X}$ accounts for a proportion

$$\frac{\operatorname{totVar}(\Pi_{\mathcal{V}_p}\boldsymbol{X})}{\operatorname{totVar}(\boldsymbol{X})} = \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{d-1} \lambda_j}$$

of the total variance. The compositional line $\{\tau \odot \operatorname{clr}^{-1}(\boldsymbol{v}_j) : \tau \in \mathbb{R}\} \subset \mathbb{S}_+^{d-1}$ represents the trend described by the jth principal component.

4.1.4 Compositional classification based on the α-metric

The k-nearest neighbours (k-NN) algorithm is a simple, popular non-parametric classifier (Hastie et al. 2009). Suppose we observe training samples $(\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)$ of $(\boldsymbol{X}, Y)$, where $Y \in \{0, 1\}$ is the binary class label of $\boldsymbol{X}$. In the classification step, a new test observation $\boldsymbol{x}^\star$ is allocated to the majority class among its k nearest neighbours, with ties broken randomly or according to some other pre-specified criterion. The tuning parameter $k \geq 1$ determines the flexibility of the classification boundaries and is usually selected by a cross-validation procedure.

The notion of neighbours implicitly assumes an underlying metric, which is typically taken to be the Euclidean distance. However, if $\boldsymbol{X}$ is a compositional random vector, then the CoDA philosophy dictates that a simplicial distance measure should be preferred. The Aitchison metric (4.1) is an obvious choice. However, Greenacre (2024) argues that in the supervised setting, where an objective performance criterion (e.g. the out-of-sample classification error rate) is available, we should not be wedded to this choice. In this spirit, Tsagris et al. (2016) propose a compositional classification algorithm called α-transformed compositional k-nearest neighbours, henceforth denoted k-NN(α). The additional tuning parameter $\alpha \in \mathbb{R}$ relates to a Box-Cox-type data transformation upon which their proposed simplicial distance is based (Tsagris et al. 2011).

Definition 4.7. For $\alpha \neq 0$, the α-transformation of a composition $\boldsymbol{x} \in \mathbb{S}_+^{d-1}$ is

$$\boldsymbol{z}_\alpha : \mathbb{S}_+^{d-1} \to \mathbb{R}^d, \qquad \boldsymbol{x} \mapsto H \cdot \left(\frac{d(\alpha \odot \boldsymbol{x}) - \mathbf{1}_d}{\alpha} \right),$$

where H is any $(d-1) \times d$ real matrix with orthonormal rows. For $\alpha = 0$, we define $\boldsymbol{z}_0(\boldsymbol{x}) := \lim_{\alpha \downarrow 0} \boldsymbol{z}_\alpha(\boldsymbol{x})$.

Typically H is chosen as the Helmert matrix with its first row removed, but in any case k-NN(α) is invariant to this choice. The α-transformation induces a metric on $\mathbb{S}_+^{d-1}$ in a similar fashion to the CLR-transformation.

Definition 4.8. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}_+^{d-1}$ be closed compositions. For $\alpha \in \mathbb{R}$, the α-metric is defined as

$$d_\alpha(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{z}_\alpha(\boldsymbol{x}) - \boldsymbol{z}_\alpha(\boldsymbol{y})\|_e$$

The special cases $\alpha = 0$ and $\alpha = 1$ correspond to the Aitchison and Euclidean distances (up to a scalar multiple), respectively, that is $d_0(\boldsymbol{x}, \boldsymbol{y}) = d_a(\boldsymbol{x}, \boldsymbol{y})$ and $d_1(\boldsymbol{x}, \boldsymbol{y}) = d \cdot d_e(\boldsymbol{x}, \boldsymbol{y})$. This means that the family of k-NN(α) classifiers encompasses Euclidean- and Aitchison-based k-NN classifiers, but some alternative value of α may be selected if it achieves superior performance. One can easily devise analogues of other classifiers, such as α-transformed support vector machines (α-SVM) and α-transformed random forests (α-RF) (Tsagris et al. 2016).

4.2 Compositional PCA for extremes

4.2.1 Framework and motivation

Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathrm{RV}_+^d$ is an $\mathbb{R}_+^d$-valued random vector with tail index α. Let $H_{\boldsymbol{X}}$ denote the normalised angular measure of $\boldsymbol{X}$ when its radial and angular components are defined with respect to some norm $\|\cdot\|$, that is

$$\mathbb{P}(\boldsymbol{X}/\|\boldsymbol{X}\| \in \cdot \mid \|\boldsymbol{X}\| > t) \to H_{\boldsymbol{X}}(\cdot), \qquad (t \to \infty).$$

The goal is to find a (linear) subspace on which $H_{\boldsymbol{X}}$ is supported (or at least highly concentrated). Specifically, we assume that $H_{\boldsymbol{X}}$ is supported on a linear subspace $V^\star$ of dimension $p^\star < d$.

The standard technique for identifying supporting (linear) subspaces, given iid observations of $\boldsymbol{X}$, is principal components analysis (PCA). If $\|\cdot\| = \|\cdot\|_1$, then $H_{\boldsymbol{X}}$ is a distribution on the unit simplex and correspondingly $\boldsymbol{X}/\|\boldsymbol{X}\|_1$ is a random composition. This opens up the possibility of employing CoDA-PCA for this task. On the other hand, taking $\|\cdot\| = \|\cdot\|_2$ results in a random vector $\boldsymbol{X}/\|\boldsymbol{X}\|_2$ and limiting distribution $H_{\boldsymbol{X}}$ that are not compositional, but rather circular/spherical. They are also restricted to the non-negative orthant, precluding the use of techniques fro the field of directional/spherical statistics (Fisher et al. 1987). Drees and Sabourin (2021) sweep this difficulty under the rug: they ignore the unit-norm constraint, and apply standard PCA to the the pseudo-angles as though they were points in $\mathbb{R}^d$. By grounding their method on Euclidean norms and distances they are able to derive statistical guarantees concerning reconstruction errors, i.e. the error incurred in representing an observation by its projection into a principal subspace. However, the utility of such guarantees may be called into question. First, for compositional data on the simplex, arbitrarily small Euclidean reconstruction errors can be arbitrarily large in the Aitchison metric. The discrepancy between the two metrics is most pronounced near the simplex boundary (Park et al. 2022). Accurate modelling of the angular measure in such regions is critical for risk assessment. Second, constrained data often exhibit curvature that cannot be described by standard linear dimension reduction techniques – see Figure 1b in Aitchison (1983) for an illustration of this phenomenon. Ultimately, this limits the dimension reducing capability of the method, since additional basis vectors are needed to reproduce the curvature. This leads us to suspect that, while the subspace detected by Drees and Sabourin (2021) is optimal among a particular class of subspaces, it is not optimal with respect to a different (arguably more natural) class.

4.2.2 CoDA-PCA for extremes

Fix $\|\cdot\| = \|\cdot\|_1$ and let $R := \|\boldsymbol{X}\|_1$ and $\boldsymbol{\Theta} := \boldsymbol{X}/\|\boldsymbol{X}\|_1$ denote the radial and angular components of $\boldsymbol{X}$. In this section, algebraic-geometric terms (e.g. linear, orthogonal, dimension, etc.) are to be interpreted in the Aitchison sense, unless stated otherwise. For any linear subspace $V \subset \mathbb{S}_+^{d-1}$, let Π_V denote the orthogonal projection (matrix) onto V. In the spirit of Drees and Sabourin (2021), we define the risk associated with V by

$$R_\infty(V) := \mathbb{E}_{\boldsymbol{\Theta}\sim H_{\boldsymbol{X}}}[\|\boldsymbol{\Theta} \ominus \Pi_V \boldsymbol{\Theta}\|_a^2].$$

The risk $R_\infty(V)$ represents the expected squared reconstruction error under the limit model when $\boldsymbol{\Theta}$ is approximated by its projection $\Pi_V\boldsymbol{\Theta}$. Our working assumption is that there exists a $p^\star$-dimensional subspace $V^\star \subset \mathbb{S}_+^{d-1}$ such that $R_\infty(V^\star) = 0$ and $R_\infty(V) > 0$ for any subspace V with $\dim(V) < p^\star$. In applications this assumption is only likely to hold approximately, introducing the familiar trade-off between dimension reduction and reconstruction error. Of course, the limit model is unknown and we cannot access samples from it, so it is impossible to compute $V^\star$ by minimising R_∞ directly. Instead, we introduce the conditional risk of V at a finite threshold $t > 0$, defined by

$$R_t(V) := \mathbb{E}[\|\boldsymbol{\Theta} \ominus \Pi_V \boldsymbol{\Theta}\|_a^2 \mid R > t].$$

Since the angular measure represents the limiting conditional distribution of angles above high thresholds, we intuitively expect that a minimiser $V_t^\star$ of R_t is close to a minimiser of R_∞, provided t is sufficiently large. We estimate $V_t^\star$ by empirical risk minimisation. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent copies of $\boldsymbol{X}$ and define $R_i = \|\boldsymbol{X}_i\|_1$ and $\boldsymbol{\Theta}_i = \boldsymbol{X}_i/\|\boldsymbol{X}_i\|_1$ for $i = 1, \ldots, n$. The empirical (conditional) risk is defined by

$$\hat{R}_t(V) := \frac{1}{\sum_{i=1}^n \mathbf{1}\{R_i > t\}} \sum_{i=1}^n \|\boldsymbol{\Theta}_i \ominus \Pi_V \boldsymbol{\Theta}_i\|_a^2 \mathbf{1}\{R_i > t\}. \tag{4.2}$$

The following result explains how to compute minimisers of these risk functions.

Proposition 4.1.

Fix $t > 0$. Without loss of generality, assume that $\boldsymbol{\xi}_t := \mathrm{cen}_a(\boldsymbol{\Theta} \mid R > t) = \boldsymbol{e}_a$, i.e. $\boldsymbol{\Theta}$ is *conditionally compositionally centred given $R > t$. Finally, assume that $\mathbb{E}[\|\boldsymbol{\Theta}\|_a^2 \mid R > t] < \infty$.*

1. *Let $\Sigma_t = \mathrm{Cov}(\mathrm{clr}(\boldsymbol{\Theta}) \mid R > t)$ be the conditional CLR-covariance matrix of $\boldsymbol{\Theta}$ given $R > t$. Suppose Σ_t has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{d-1} \geq \lambda_d = 0$ and corresponding eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d \in \mathbb{R}^d$. Then $V_p = \mathrm{span}_a(\{\mathrm{clr}^{-1}(\boldsymbol{u}_j) : j = 1, \ldots, p\})$ minimises $R_t(V)$ among all linear subspaces V of dimension p. It is the unique minimiser if $\lambda_p > \lambda_{p+1}$.*

2. *Let $k = \sum_{i=1}^n \mathbf{1}\{R_i > t\}$ and suppose the empirical conditional CLR-covariance matrix $\hat{\Sigma}_t = \frac{1}{k}\sum_{i=1}^k \mathrm{clr}(\boldsymbol{\Theta}_i)\mathrm{clr}(\boldsymbol{\Theta}_i)^T \mathbf{1}\{R_i > t\}$ has eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{d-1} \geq \hat{\lambda}_d = 0$ and corresponding eigenvectors $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d \in \mathbb{R}^d$. Then $\hat{V}_p = \mathrm{span}_a(\{\mathrm{clr}^{-1}(\hat{\boldsymbol{u}}_j) : j = 1, \ldots, p\})$ minimises $\hat{R}_t(V)$ among all linear subspaces V of dimension p. It is the unique minimiser if $\hat{\lambda}_p > \hat{\lambda}_{p+1}$.*

Proof. Note that $\boldsymbol{Y}_t := \mathrm{clr}(\boldsymbol{\Theta}) \mid (R > t)$ is a random vector taking values in $\mathcal{T}^{d-1} \subset \mathbb{R}^d$ and

$$\mathbb{E}[\|\boldsymbol{Y}_t\|_2^2] = \mathbb{E}[\|\mathrm{clr}(\boldsymbol{\Theta})\|_2^2 \mid (R > t)] = \mathbb{E}[\|\boldsymbol{\Theta}\|_a^2 \mid R > t] < \infty,$$

by assumption. Thus $\boldsymbol{Y}_t$ satisfies the usual assumptions of unconstrained PCA in $\mathbb{R}^d$ and standard results, e.g. Theorem 5.3 in Seber (1984), concerning minimiser(s) of $V \mapsto \mathbb{E}[\|\Pi_V \boldsymbol{Y}_t - \boldsymbol{Y}_t\|_2^2]$ yield (i). The argument follows analogously for (ii), except now $\boldsymbol{Y}_t$ following the empirical distribution of the CLR-transformed angular components associated with radial threshold exceedances.

□

Minimising the expected angular reconstruction error is natural, but does not guarantee good performance in terms of estimation of the angular measure. Thus, we additionally consider the performance of the standard non-parametric estimator of the angular measure based on the compressed data versus the raw observations. For a given threshold $t > 0$, the empirical angular measure is given by

$$\hat{H}_{\boldsymbol{X}} = \frac{1}{\sum_{i=1}^n \mathbf{1}\{R_i > t\}} \sum_{i=1}^n \delta_{\boldsymbol{\Theta}_i} \mathbf{1}\{R_i > t\}. \tag{4.3}$$

If the pseudo-angles are first projected onto a subspace V, then we obtain an alternative estimator

$$\hat{H}_{\boldsymbol{X},V} = \frac{1}{\sum_{i=1}^n \mathbf{1}\{R_i > t\}} \sum_{i=1}^n \delta_{\Pi_V \boldsymbol{\Theta}_i} \mathbf{1}\{R_i > t\}. \tag{4.4}$$

The probabilities associated with certain joint tail events can be expressed in terms of the angular measure, for example:

$$\lim_{u\to\infty} \mathbb{P}(\min \boldsymbol{X} > u \mid \|\boldsymbol{X}\|_1 > u) = \int_{\mathbb{S}_+^{d-1}} \left(\min_{j=1,\dots,d} \theta_j\right)^{\alpha} H_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}), \tag{4.5}$$

$$\lim_{u\to\infty} \mathbb{P}(\max \boldsymbol{X} > u \mid \|\boldsymbol{X}\|_1 > u) = \int_{\mathbb{S}_+^{d-1}} \left(\max_{j=1,\dots,d} \theta_j\right)^{\alpha} H_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}). \tag{4.6}$$

These probabilities indicate how the components' contribution to extreme events are spread. If the underlying data-generating process is known, then the true values can be computed analytically or via Monte Carlo simulation. Replacing $H_{\boldsymbol{X}}$ with $\hat{H}_{\boldsymbol{X}}$ or $\hat{H}_{\boldsymbol{X},V}$ yields empirical estimates of these probabilities, the errors in which can be used to quantify the performance of our low-dimensional models. If the generative process is unknown, then we can still use the $\hat{H}_{\boldsymbol{X}}$-based estimate as a benchmark (based on all available information) against which to compare estimates obtained via $\hat{H}_{\boldsymbol{X},V}$.

4.2.3 Simulation experiments

We now perform a series of simulation experiments comparing CoDA-PCA against the procedure of Drees and Sabourin (2021), herein referred to as DS-PCA for brevity.

One complication that arises is that CoDA-PCA and DS-PCA define angles with respect to different norms, so the sets of reconstructions are not immediately comparable. We resolve this by performing DS-PCA in the usual way, but projecting all points onto the simplex via self-normalisation with respect to $\|\cdot\|_1$ before computing the performance metrics. This ensures that the metrics are well-defined, except for DS-PCA projections that lie outside of the positive orthant. In such instances, the projected vector cannot properly be called a composition and the Aitchison reconstruction error is undefined. One might consider projecting it to the nearest point on the simplex, but this would lie on the boundary resulting in infinite Aitchison distances. In the absence of better options, we elect to discard such points. Arguably we are being charitable to DS-PCA by not directly penalising its tendency to produce invalid angles.

To guard against overfitting, we compute the empirical risk $\hat{R}_t$ across an unseen validation set, so that we are actually measuring out-of-sample reconstruction error. This ensures that the dimension reducing capabilities of the PCA model generalise to future observations. Specifically, for a fixed threshold $t >$, we detect the set of principal subspaces by applying Proposition 4.1 based on independent training samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, but compute the empirical risk (4.2) on an unseen validation set $\boldsymbol{X}^\star_1, \dots, \boldsymbol{X}^\star_{n^\star}$ of independent observations using the same threshold. The threshold t is selected by specifying a desired number of extreme observations k and setting $t := R_{(k+1)}$, the $k+1$ order statistic of $\{R_1, \dots, R_n\}$. Since we work in a simulation setting, the number of threshold exceedances in the validation set, roughly $n^\star k/n$, can be made arbitrarily large by increasing $n^\star$ accordingly.

4.2.3.1 Max-linear model with compositionally colinear factors

Our first experiment is based on the max-linear model with a parameter matrix that is carefully constructed to favour CoDA-PCA. This example is somewhat contrived, but illustrates the many benefits of our proposed methodology. To facilitate visualisation we restrict ourselves to $d = 3$ dimensions, but the construction and our findings extend to higher dimensions.

Let $\boldsymbol{u}^\star \in \mathbb{S}_+^{d-1}$ be a composition and suppose $A = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_q) \in \mathbb{R}_+^{d\times q}$ is such that, for all $j = 1, \dots, d$, there exists $\beta_j \in \mathbb{R}$ such that $\mathcal{C}\boldsymbol{a}_j = \beta_j \odot \boldsymbol{u}^\star$. In other words, the $q \geq 1$ columns of A lie on the compositional straight line through $\boldsymbol{e}_a$ in the direction $\boldsymbol{u}^\star$. Let $\boldsymbol{Z} = (Z_1, \dots, Z_q)$ with $Z_1, \dots, Z_q$ independent standard Fréchet random variables. Suppose $\boldsymbol{X} = (X_1, \dots, X_d)$ is defined by either of the following:

$$\boldsymbol{X} = A \circ \boldsymbol{Z} := \tau(A\tau^{-1}(\boldsymbol{Z})), \tag{4.7}$$

$$\boldsymbol{X} = A \times_{\max} \boldsymbol{Z} := \left(\max_{j=1,\dots,q} a_{1j}Z_j, \dots, \max_{j=1,\dots,q} a_{dj}Z_j\right), \tag{4.8}$$

where $\tau : \mathbb{R} \to (0,\infty)$ is the softplus function defined by

$$\tau(y) := \log(1 + \exp(y)).$$

Based on the discussion in Section 3 of Cooley and Thibaud (2019), we refer to random vectors generated by (4.7) and (4.8) as RV-max-linear and MS-max-linear, respectively. In either case, $\boldsymbol{X}$ is multivariate regularly varying with tail index $\alpha = 1$ and common angular measure

$$H_{\boldsymbol{X}}(\cdot) = \frac{\sum_{j=1}^{q} \|\beta_j \odot \boldsymbol{u}^\star\|_1 \delta_{\beta_j \odot \boldsymbol{u}^\star}(\cdot)}{\sum_{j=1}^{q} \|\beta_j \odot \boldsymbol{u}^\star\|_1} = \frac{1}{q} \sum_{j=1}^{q} \delta_{\beta_j \odot \boldsymbol{u}^\star}(\cdot).$$

This means that

$$\operatorname{supp}(H_{\boldsymbol{X}}) = \{\beta_j \odot \boldsymbol{u}^\star : j = 1, \dots, q\} \subset \{\beta \odot \boldsymbol{u}^\star : \beta \in \mathbb{R}\} = \operatorname{span}_a(\boldsymbol{u}^\star),$$

a one-dimensional linear subspace of $\mathbb{S}_+^{d-1}$. The probabilities (4.5) and (4.6) are computed as

$$\lim_{u \to \infty} \mathbb{P}(\min \boldsymbol{X} > u \mid \|\boldsymbol{X}\|_1 > u) = \sum_{j=1}^{q} \min_{i=1,\dots,d} a_{ij} \approx 0.636, \tag{4.9}$$

$$\lim_{u \to \infty} \mathbb{P}(\max \boldsymbol{X} > u \mid \|\boldsymbol{X}\|_1 > u) = \sum_{j=1}^{q} \max_{i=1,\dots,d} a_{ij} \approx 0.0679. \tag{4.10}$$

The reason for introducing the two generative processes is that they differ in terms of their finite sample properties, allowing a more detailed exploration of the finite-sample performance of our methodology. The angular components associated with large realisations of the MS-max-linear process tend to lie exactly at the discrete locations in $\mathcal{C}\boldsymbol{a}_j$, whereas for the RV-type process extremal angles tend to lie close to, but not exactly, at these points.

All simulations are based on $d = 3$, $q = 50$, $\boldsymbol{u}^\star = (0.12, 0.58, 0.3)$ and fixed values $\beta_1, \dots \beta_{50}$ sampled uniformly between -4 and 4. We generate training sets of size $n \in \{5 \times 10^3, 5 \times 10^4\}$ and set $k/n \in \{1\%, 5\%\}$. Reconstruction errors are based on validation sets of size $n^\star = n$. For each parameter combination, simulations are repreated 50 times.

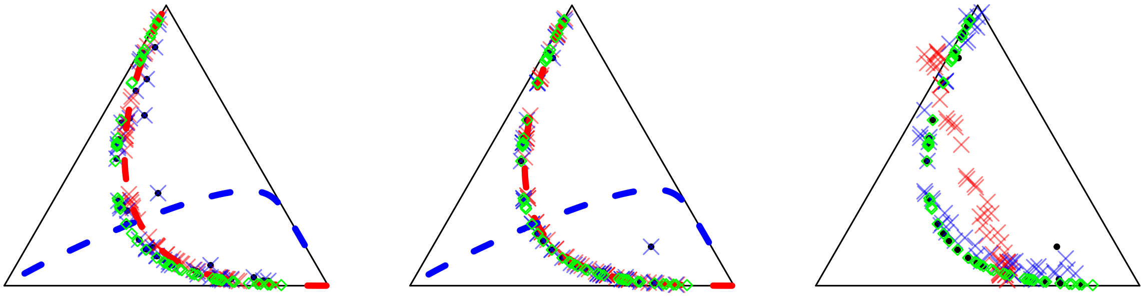

Figure 4.1: Examples of CoDA-PCA (left and middle) and DS-PCA (right) applied to RV- (left) and MS-max-linear (middle and right) data. The green diamonds represent the true angular measure. The black points are the angular components associated with the $k = 50$ largest observations among a sample of size $n = 5000$. The red and blue dashed lines represent the first and second principal axes, respectively. The red and blue crosses represent the projections onto the first and second principal subspaces, respectively.

Before diving in to the full simulation study, we invite the reader to examine Figure 4.1, which illustrates the example under consideration and will help provide some intuition for the results to follow. In each plot, the green diamonds represent the points $\mathcal{C}\boldsymbol{a}_j$ at which the angular measure places mass. These points follow a decidely curved pattern, but the essential structure is obviously one-dimensional. The black points represent the angular components of the threshold exceedances (here $k = 50$ and $n = 5000$). The left and middle ternary plots show the trends described by the first (red dashed line) and second (blue dashed line) compositional principal components. In the middle plot (MS-max-linear data), this red line follows the green points almost exactly, whereas the left-hand plot (RV-max-linear data) shows a degree of estimation error due to the weaker signal-to-noise ratio. The red crosses represent the rank-one reconstructions obtained by projecting the black points onto the red line. The second principal component describes all remaining variability in the data. In theory, one component is sufficient to describe the target distribution, but in finite sample settings an additional component is required (due to noise and the samples not coming directly from the limit model). The right-hand plot shows the results of DS-PCA applied to the MS-max-linear data. Now the first principal component is unable to capture the curvature of the data, yielding poor first-order projections (red crosses). (Note: these points lie on straight line in $\mathbb{R}^3$, but the process of visualising them on a ternary plot distorts the line slightly.) As discussed earlier, there are even a handful of projected points that lie outside of the ternary plot. Adding a second component improves the reconstructions significantly (blue crosses). Unlike CoDA-PCA, the two-dimensional approximations are still imperfect, because DS-PCA treats the angles as points in $\mathbb{R}^3$.

With this example in mind, we consider the methods' performance over repeated simulations, shown in Figure 4.2. Within each sub-plot, we plot the empirical distribution of a given performance metric for each PCA method (bar colour) with varying number of principal components (bar outline). The maximum number of components is two, as including a third component does not add any information to the plot (*does this need explaining?*). The panels within each sub-plot correspond to the different combinations of n and k/n. First, consider the Aitchison MSREs in the top-left plot. As expected, CoDA-PCA is able to reconstruct the data almost perfect with a single principal component. In contrast, the DS-PCA projections are relatively poor, even with the inclusion of a second component. This is primarily caused by imperfect reconstructions near the simplex boundary (see Figure 4.1, right), which are heavily penalised by the Aitchison metric. The top-right plot shows the Euclidean reconstruction error, i.e. (4.2) with $\|\cdot\|_a$ replaced with $\|\cdot\|_2$. The main difference with the previous sub-plot is that now the two-dimensional DS-PCA are judged much more favourably. This shows how measuring performance using Euclidean distances can mask errors. Nevertheless, the CoDA method is vastly superior. This does not contradict the fact DS-PCA produces optimal subspaces (Drees and Sabourin 2021, Lemma 2.1(iii)). That optimality pertains to the class of linear subspaces in $\mathbb{R}^d$, which does not preclude the existence of better subspaces outside of this class. The bottom sub-plots show the empirical estimates of (4.9) and (4.10) obtained via the models (4.4), where V is the one/two principal subspace detected by each algorithm. With MS-max-linear data, 1D CoDA-PCA yields almost perfect estimates of both probabilities. When the data are generated from the RV-type process, the min and max probabilities tend to be overestimated/underestimated. This is because the sub-asymptotic distribution of the data is such that the first sample eigenvector $\hat{\boldsymbol{u}}_1$ is slightly biased for $\boldsymbol{u}^\star$. This is evidenced in Figure 4.1 (left), where the red line gets pulled to the right of the green diamonds. Retaining a second component helps correct this, and the remaining bias in the probability estimates can be attributed to the noisy samples. Generally speaking, increasing the sample size reduces variance but does not eliminate biases. This means that the errors (or a lack thereof) can be attributed to the statistical methodology, rather than a lack of available data. Even with infinite samples, DS-PCA will fit a straight line to curved data!

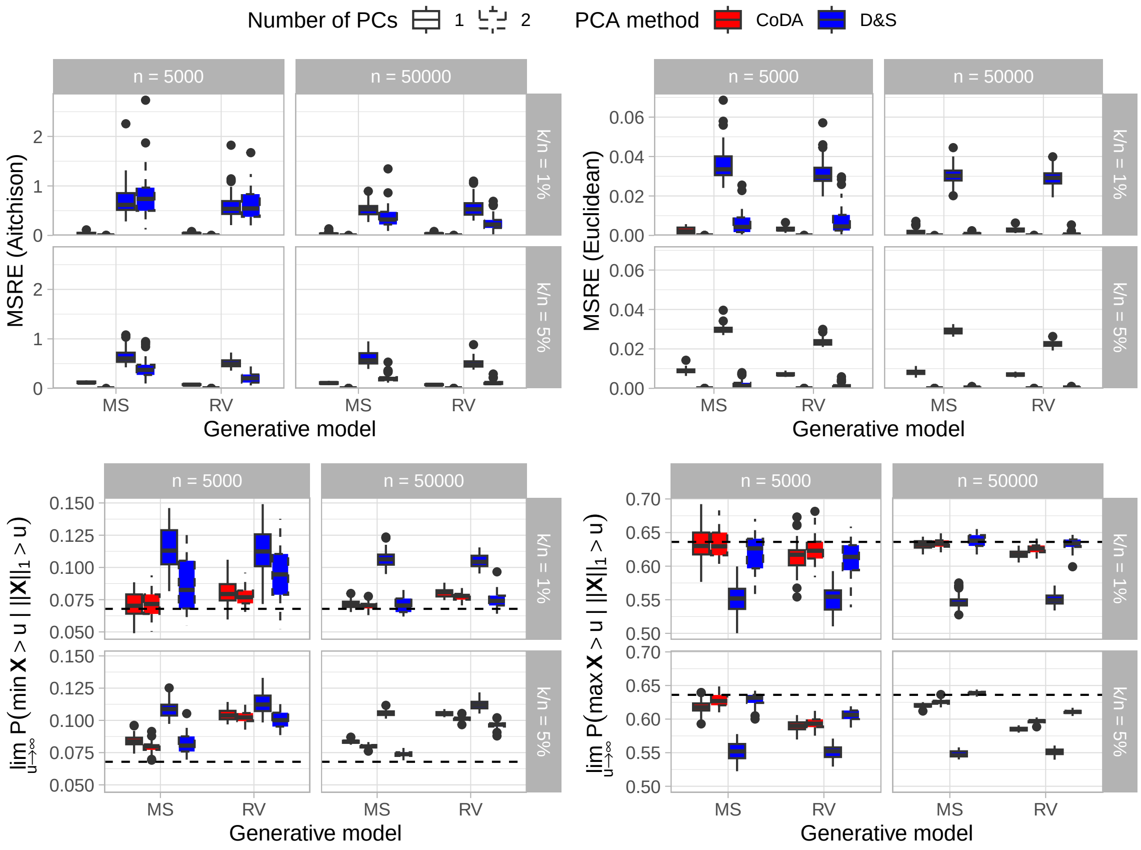

Figure 4.2: PCA performance metrics based on trivariate max-linear data.

4.2.3.2 Hüsler-Reiss model in low dimensions

Now we consider examples where the data are generated from a Hüsler-Reiss model. To start with we shall stay in three dimensions to facilitate visualisation. The relevance of this experiment is that, unlike the previous example, the true limit model is not designed to favour a particular PCA method.

Data are produced from three Hüsler-Reiss models parametrised by the following variograms (entries rounded to two decimal places):

$$\Gamma_1 = \begin{pmatrix} 0.00 & 0.10 & 1.24 \\ 0.10 & 0.00 & 0.67 \\ 1.24 & 0.67 & 0.00 \end{pmatrix}, \qquad \Gamma_2 = \begin{pmatrix} 0.00 & 0.14 & 0.04 \\ 0.14 & 0.00 & 0.10 \\ 0.04 & 0.10 & 0.00 \end{pmatrix}, \qquad \Gamma_3 = \begin{pmatrix} 0.00 & 0.01 & 1.29 \\ 0.01 & 0.00 & 1.24 \\ 1.29 & 1.24 & 0.00 \end{pmatrix}.$$

These randomly generated variograms induce qualitatively different dependence structures, as shown in Figure 4.3. The left plot (Γ_1) exhibits a curved trend with little variability in the direction orthogonal to this curve. This is a similar paradigm to the previous example. The extremes in the middle plot (Γ_2) is concentrated in near the barycentre of the simplex, implying strong asymptotic dependence between the three variables. The (empirical) angular measure exhibits a very slight curvature but is two-dimensional. Under the model parametrised by Γ_3 (right), extremes tend to occur in X_1 and X_2 jointly or X_3 singly. The extremal angles concentrate along a straight line joining the edge and vertex associated with these groups of components.

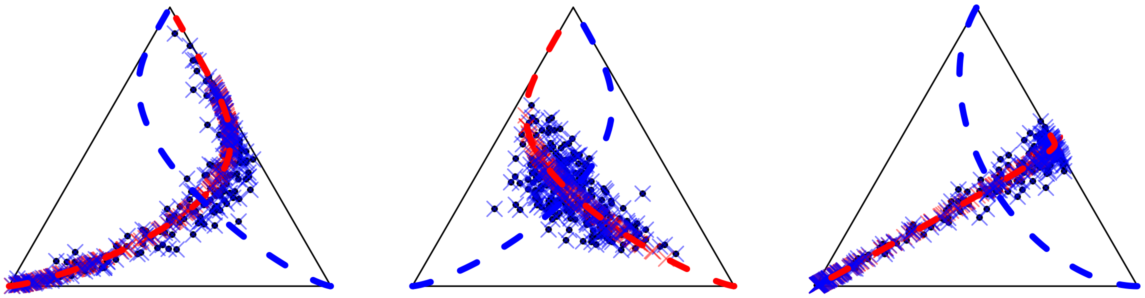

Figure 4.3: Example data from the three trivariate Hüsler-Reiss models. Based on $n = 10^4$ and $k = 250$.

Similar to before, we repeatedly generate samples ($n = 5000$) from each model and perform PCA based on the largest $k = 50$ observations in norm. The probabilities (4.5) and (4.5) are computed empirically from samples of size $n = 10^6$ using $u = 100$. To three decimal places, the true values of (4.5) (resp. (4.6)) under the three sub-models are found to be 0.089, 0.228, 0.081 (resp. 0.603, 0.456, 0.586). For each simulated data set, we compute the MSRE and the error in the probability estimates obtained using low-rank reconstructions via (4.3) and (4.4). The results are displayed in Figure 4.4. For Γ_1, one compositional principal component is sufficient to explain the data and produce good estimates. Drees and Sabourin (2021) requires at least two components due to the non-linear trend. The data generated by Γ_2 is approximately linear and distinctly two-dimensional. Thus, both PCA procedures perform similarly and there is no scope for dimension reduction. The angular measure associated with Γ_3 concentrates along a straight line, which both methods are able to captured relatively well with a single eigenvector. The upshot is that CoDA-PCA performs at least as well as DS-PCA across a range of scenarios and outperforms it by a significant margin in some cases. Whether there is a difference in the methods depends on whether the low-dimensional target subspace $V^\star$ can be well-approximated by a linear subspace. In low dimensions this can be gauged by simply inspecting the data. Of course this is not generally feasible in high-dimensional applications, which represent the typical use case of such techniques.

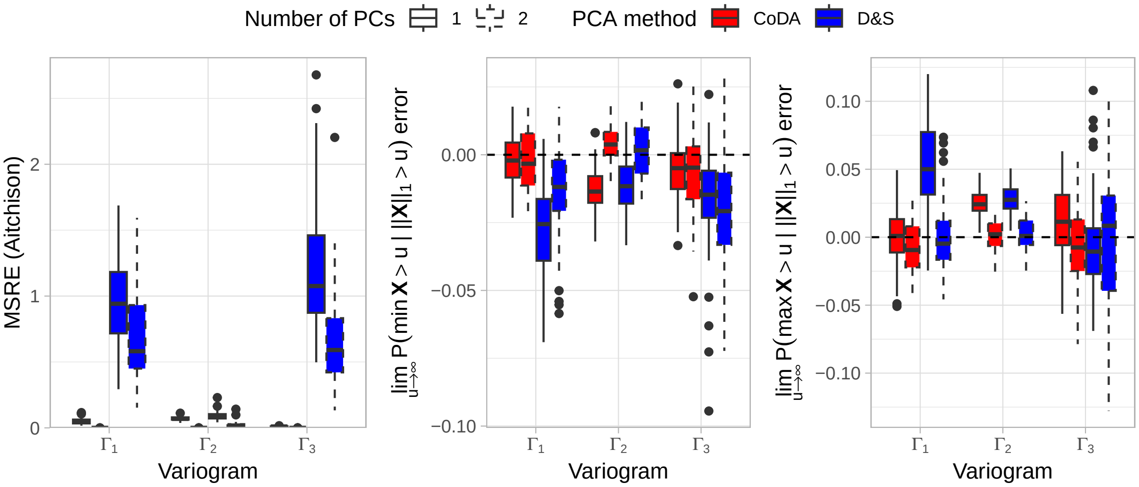

Figure 4.4: PCA performance metrics based on trivariate Hüsler-Reiss data.

4.2.3.3 Hüsler-Reiss model in high dimensions

For this experiment, the random vector $\boldsymbol{X} = (X_1, \dots, X_{10})$ follows a Hüsler-Reiss distribution where the variogram $\Gamma \in \mathbb{R}_+^{10\times 10}$ is randomly generated according to the procedure of Fomichov and Ivanovs (2023) (see Assumption A and Appendix B1 therein) with three clusters. The matrix of tail dependence coefficients associated with this variogram is given by $\chi_{ij} = 2\bar{\Phi}(\sqrt{\Gamma_{ij}}/2)$. Figure 4.5 (left) provides a visual representation of this matrix. Recall that $\chi_{ij} = 0$ if and only if X_i and X_j are asymptotically independent, and the magnitude of χ_{ij} indicates the extremal dependence strength between the corresponding pair of variables. We observe three groups/clusters and asymptotically dependent variables. Dependence is very strong among $\{X_1, \dots, X_4\}$ and $\{X_9, X_{10}\}$, while the pairwise dependence strengths between the components in $\{X_5, \dots, X_8\}$ is moderate and more variable. Figure 4.5 (right) shows the eigenvectors of the CLR-covariance matrix, estimated from a sample of size $n = 10^6$ with $k = 200$. The leading eigenvectors describe the extremal dependence structure with increasing resolution. The eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2$ determine which cluster an extreme belong to, $\boldsymbol{u}_3, \boldsymbol{u}_4, \boldsymbol{u}_5$ capture the fine-scale behaviour within the second cluster. Subsequent eigenvectors account for patterns within the first and third clusters. We would speculate that retaining three to five principal components will be sufficient, depending on the complexity of the dependence in cluster two. This is borne out in Figure 4.6 (top-right), which shows that first three components account for at least 95% of the total variability. In contrast, satisfying the same criterion (albeit with respect to a different notion of variance) under DS-PCA requires five components. Compressing the data to three-dimensions yields no discernible deterioration in the probability estimates (bottom sub-plots).

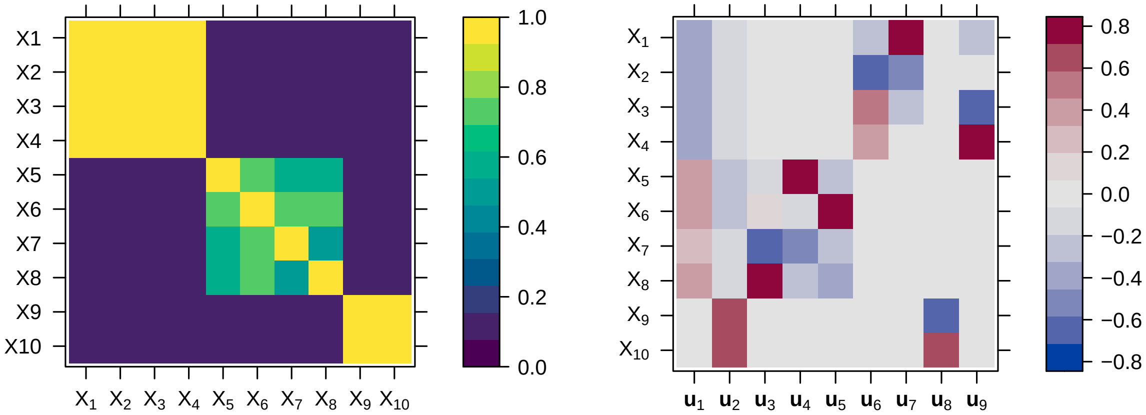

Figure 4.5: Matrix of tail dependence coefficients $\chi_{ij} = 2\bar{\Phi}(\sqrt{\Gamma_{ij}}/2)$ (left) and a matrix of CoDA-PCA sample eigenvectors (right).

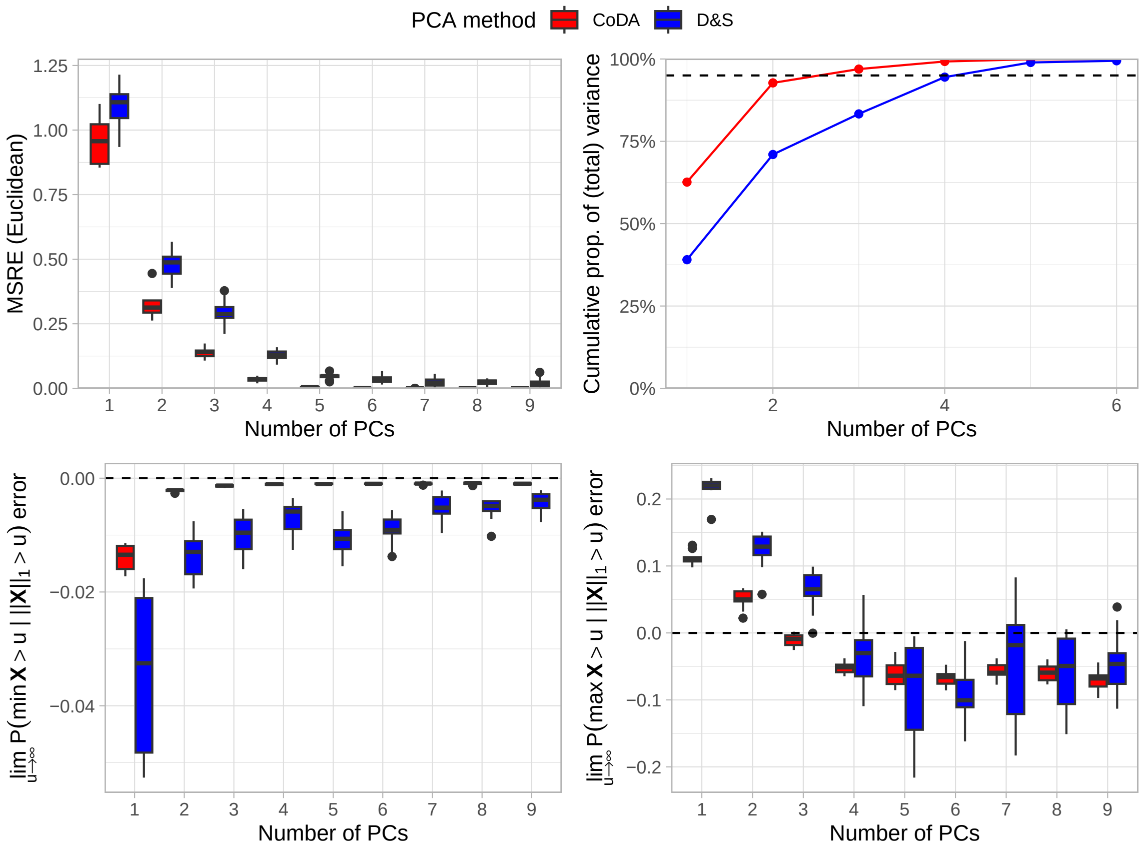

Figure 4.6: PCA performance metrics based on 10-dimensional Hüsler-Reiss data.

4.2.3.4 Max-linear model in high dimensions

- Same structure as earlier example, but now $d = 10$, $q = 25$, and $\mathcal{C}\boldsymbol{a}_j = \beta_{j1} \odot \boldsymbol{u}_1^\star \oplus \beta_{j2} \odot \boldsymbol{u}_2^\star$
- Figure 4.8 (left): Two/three components explain approx 95% of the total variance.
- Figure 4.8 (right): For MS data, two/three components gives good estimates. For RV data, four components gives good performance (corrects for error in previous eigenvectors). In each case, DS-PCA needs at least six components.

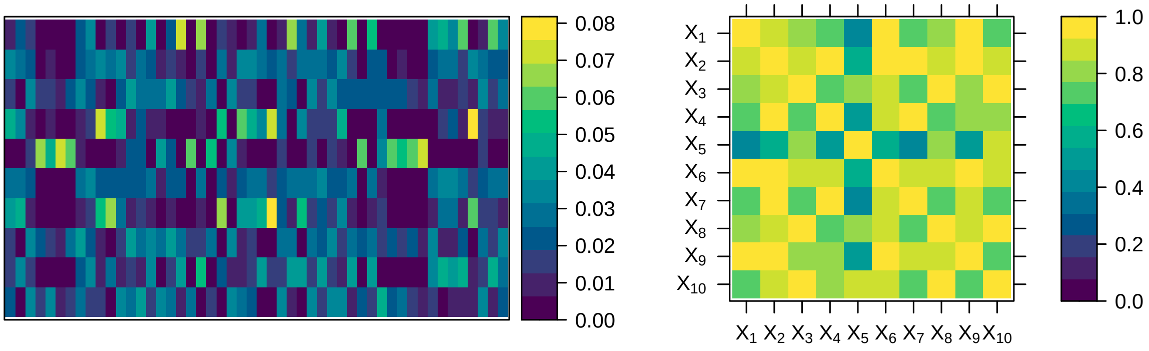

Figure 4.7: Parameter matrix A (left) and corresponding model TPDM (right) for 10-dimensional max-linear data.

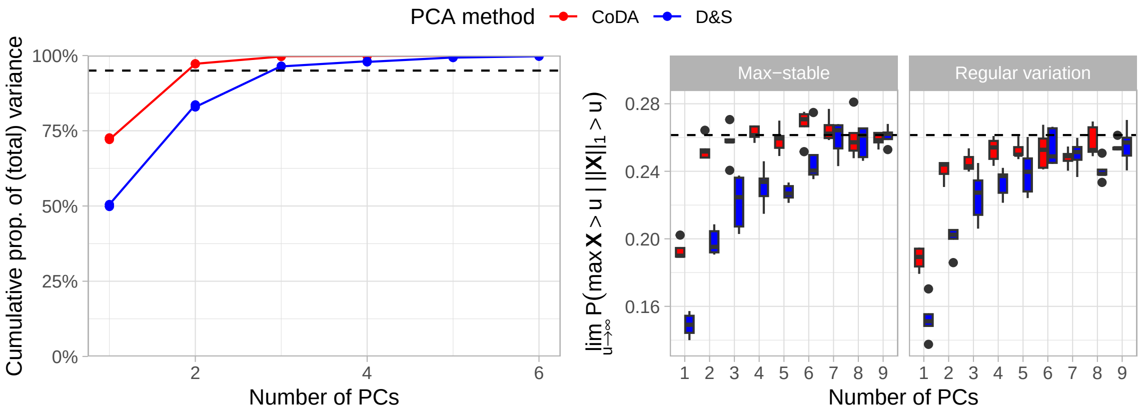

Figure 4.8: PCA performance metrics based on 10-dimensional max-linear data.

4.2.4 Discussion

Discuss conclusions of CoDA PCA stuff here.

4.3 Compositional classification for extremes

4.3.1 Framework/motivation

Let $(\boldsymbol{X}, Y)$ be a random pair with unknown joint distribution $F_{(\boldsymbol{X},Y)}$, where $Y \in \{-1, +1\}$ is a binary class label and $\boldsymbol{X} = (X_1, \ldots, X_d)$ is an $\mathbb{R}_+^d$-valued random vector containing covariate information that is presumed to be useful for predicting Y. For $\sigma \in \{-, +\}$, assume $\boldsymbol{X} \mid Y = \sigma 1$ is multivariate regularly varying with tail index $\alpha = 1$ and angular measure H_σ (with respect to a fixed norm $\|\cdot\|$ on $\mathbb{R}^d$). Given a labelled training sample

$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ of independent copies of $(\boldsymbol{X}, Y)$, the goal is to find a classifier $g : \mathbb{R}^d \to \{-1, +1\}$ that minimises the expected classification error rate

$$\begin{aligned}\mathcal{L}(g) &:= \mathbb{P}(Y \neq g(\boldsymbol{X})) \\ &= \mathbb{P}(Y \neq g(\boldsymbol{X}) \mid \|\boldsymbol{X}\| \leq t)\mathbb{P}(\|\boldsymbol{X}\| \leq t) + \mathbb{P}(Y \neq g(\boldsymbol{X}) \mid \|\boldsymbol{X}\| > t)\mathbb{P}(\|\boldsymbol{X}\| > t).\end{aligned}$$

Since $F_{(\boldsymbol{X},Y)}$ is unknown, the standard approach to this task is estimate the statistical risk $\mathcal{L}(g)$ by the empirical risk

$$\hat{\mathcal{L}}(g) = (1/n)\sum_{i=1}^{n} \mathbf{1}\{Y_i \neq g(\boldsymbol{X}_i)\}$$

and choose $\hat{g} = \arg\min_{g\in\mathcal{G}} \hat{L}(g)$ over a suitable class $\mathcal{G}$. However, Jalalzai et al. (2018) point out that the optimal classifier need not perform well in extreme regions of the predictor space, e.g. $\{\|\boldsymbol{X}\| > t\}$ for $t > 0$ large, since such regions exert a negligible influence over the global prediction error. This motivates the idea of building a classifier that minimises the asymptotic risk in the extremes, defined as

$$\mathcal{L}_\infty(g) := \lim_{t\to\infty} \mathbb{P}(Y \neq g(\boldsymbol{X}) \mid \|\boldsymbol{X}\| > t).$$

They prove that, under certain assumptions, the minimiser $g^\star := \arg\min_{g\in\mathcal{G}} \mathcal{L}_\infty(g)$ is of the form $g^\star(\boldsymbol{x}) = g^\star(\boldsymbol{x}/\|\boldsymbol{x}\|)$ (Jalalzai et al. 2018, Theorem 1). In practice, this suggests finding solutions of the minimisation problem

$$\min_{g\in\mathcal{G}_{\boldsymbol{\Theta}}} \hat{L}_t(g), \qquad \hat{\mathcal{L}}_t(g) = \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{\|\boldsymbol{X}\| > t\}} \sum_{i=1}^{n} \mathbf{1}\left\{Y_i \neq g\left(\frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}\right), \|\boldsymbol{X}\| > t\right\}. \tag{4.11}$$

for some high threshold $t > 0$, where $\mathcal{G}_{\boldsymbol{\Theta}}$ denotes a family of classifiers $g : \mathbb{S}_+^{d-1} \to \{-1 + 1\}$.

The remainder of their paper is devoted to providing theoretical guarantees for this learning principle, leaving aside “the practical issue of designing efficient algorithms for solving (4.11)”. This is exemplified in their numerical experiments, where they simply resort to popular, general-purpose classifiers such as k-NN and random forests. These algorithms disregard the unit-norm constraint imposed on the input data. By now, we hope the reader is persuaded that this mismatch can have significant practical ramifications. Upon taking

$\|\cdot\| = \|\cdot\|_1$, (4.11) becomes a compositional binary classification problem. The CoDA community develops bespoke algorithms for such tasks. Implementations of these algorithms are readily available in packages such as `Compositional` and `CompositionalML`.

4.3.2 Simulation experiments

For our simulation experiments, we generate realisations of $\boldsymbol{X} \mid Y = y$ on standard Fréchet margins from one of three MEV models: symmetric logistic, asymmetric logistic, and bilogistic. The negative ($y = -1$) and positive ($y = +1$) class instances are generated using different (scalar) dependence parameters, denoted ϑ_0 and ϑ_1, respectively. The classes are balanced globally and asymptotically, meaning

$$p = \mathbb{P}(Y = +1) = 0.5, \qquad p_\infty := \lim_{t\to\infty} \mathbb{P}(Y = +1 \mid \|\boldsymbol{X}\|_1 > t) = 0.5.$$

From each model, we simulate a labelled training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ of size $n = 5\times 10^3$. Each tail classifier is trained using the $k = 500$ largest observations (in L_1-norm) among this set. Figure 4.9 illustrates one realisation of this subset for each model. Each plot gives an indication of the two classes' tail dependence structures and the degree of difficulty in classifying them. The performance of each classifier will be assessed by its asymptotic risk. This is estimated empirically using a validation set comprising 10^5 samples from the limit model, i.e. angles sampled from the angular measures H_- and H_+ using the `rmevspec` function from the `mev` package. In practical scenarios where we cannot access unlimited samples from the limit model, we would instead assess the extrapolation capacity of the classifier by compute the empirical risk at a sequence of increasing thresholds on a hold-out validation set. All reported results are based on 100 repeated simulations.

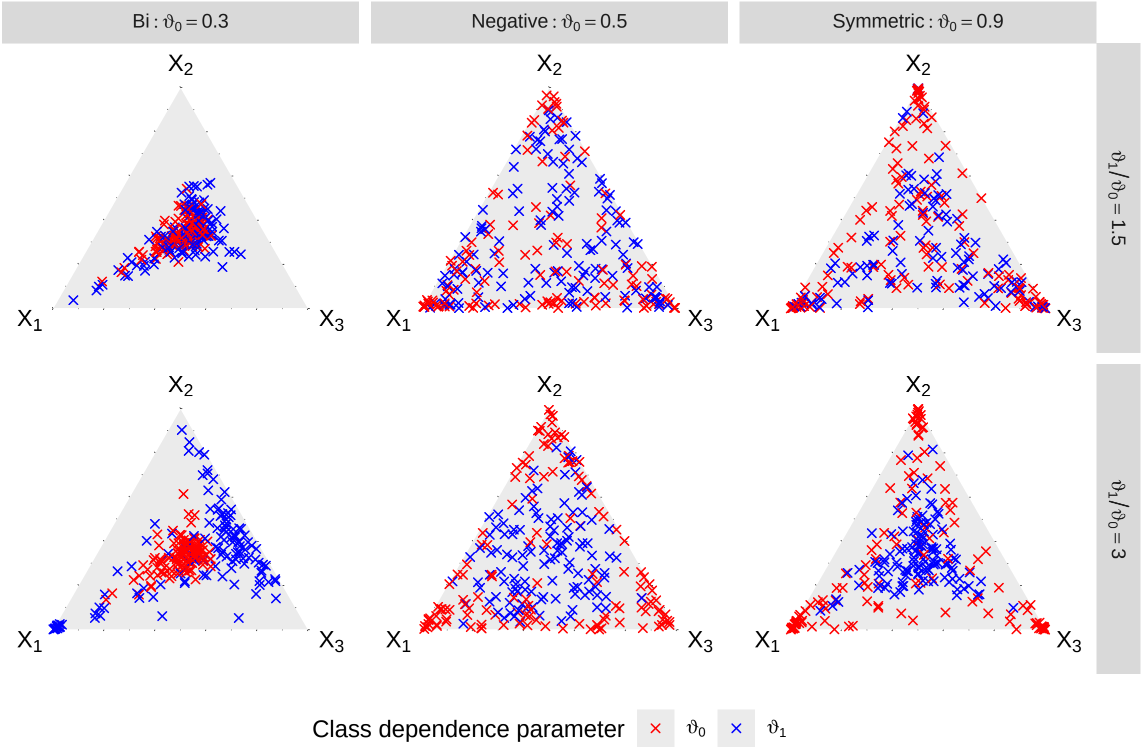

Figure 4.9: Blah.

The catalogue of classification algorithms is virtually limitless. For simplicity, we restrict ourselves to three types: k-nearest neighbours, support vector machines, and random forests. Since our primary goal is to compare CoDA-based classifiers to standard ones, we shall employ the α-transformed classifiers described earlier. For fixed $\alpha \in \{0, 0.1, \dots, 0.9, 1\}$, the classifiers' tuning parameters are selected by ten-fold cross-validation on the training set. The hyperparameters of the three classifiers considered here are described in XXX(A). The tuning procedure is summarised as follows:

1. Let $\mathcal{G}_{\mathbf{\Theta}} = \{g_\psi : \psi \in \Psi\}$ be a family of simplicial classifiers $g_\psi : \mathbb{S}_+^{d-1} \to \{-1, +1\}$ parametrised by $\psi \in \Psi$. For example, if $\mathcal{G}_{\mathbf{\Theta}}$ represents the k-NN(α) class with fixed α, then $\psi = k$ and $\Psi = \mathbb{N}$.
2. For a given training set $\{(\boldsymbol{x}_i, y_i) : i = 1, \dots, n\}$, let $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(k)}$ be the angular components of the k largest observations and $y_{(1)}, \dots, y_{(k)}$ their associated classes. Let t denote the implicit threshold, i.e. the $k+1$ order statistic of $\{\|\boldsymbol{x}_i\|_1 : i = 1, \dots, n\}$
3. Partition $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(k)}$ into $J = 10$ balanced folds by stratified sampling.

4. For $j = 1, \dots, J$ and $\psi \in \Psi$, predict the classes of the elements of fold j by applying the classification rule g_ψ trained on all the (labelled) data not in fold j. Denote the resulting classification error rate by $\hat{\mathcal{L}}_t^{(j)}(g_\psi)$.
5. Estimate the risk of g_ψ at level t as

$$\hat{\mathcal{L}}_t(g_\psi) = \frac{1}{J} \sum_{j=1}^{J} \hat{\mathcal{L}}_t^{(j)}(g_\psi).$$

6. Select ψ to minimise the empirical risk, that is $\hat{\psi} = \arg\min_{\psi \in \Psi} \hat{\mathcal{L}}_t(g_\psi)$. The tuned classifier is $\hat{g} = g_{\hat{\psi}}$.
7. Compute the asymptotic risk $\mathcal{L}_\infty(\hat{g})$ of $\hat{g}$ as the empirical classification error rate based on the samples from the true limit model. By taking sufficiently many Monte Carlo samples, the asymptotic risk can be computed to arbitrary precision.

Figure 4.10 presents the results of our experiment. Each sub-panel corresponds to a generative process, that is, the MEV model and the ratio between the dependence parameters. Within each sub-panel, we plot the asymptotic risk as a function of the data-transformation parameter α. The solid lines represent the median risk (across the 100 repeats) while the bands depict the interquartile range. The colours indicate the classifier type. The ratio of the dependence parameters dictates the difficulty level of the learning task. The error rates are between 30-40% when $\vartheta_1/\vartheta_0 = 1.5$ and between 2-16% when $\vartheta_1/\vartheta_0 = 3$. Universally, the statistical risk is maximised when the underlying geometry is Euclidean ($\alpha = 1$). For the bilogistic and symmetric logistic models, $\alpha = 0$ appears optimal. For the negative logistic data, the minimal risk is attained at some intermediate value, say $\alpha \approx 0.3$. Thus the optimal classifiers fall somewhere under the CoDA umbrella, corresponding to either the Aitchison metric (“quintessential CoDA”) or the α-metric with $\alpha \neq 1$ (“modern CoDA”), respectively. The choice of classifier is obviously a key determinant of performance, with k-NN(α) typically being the worst-performing and α-SVM the best. Notwithstanding this, we highlight that k-NN($\alpha = 0$) is usually fairly competitive against the Euclidean SVM/RF. This shows that a simple classifier in the ‘correct’ geometry can be as good as a sophisticated classifier in the ‘wrong’ geometry.

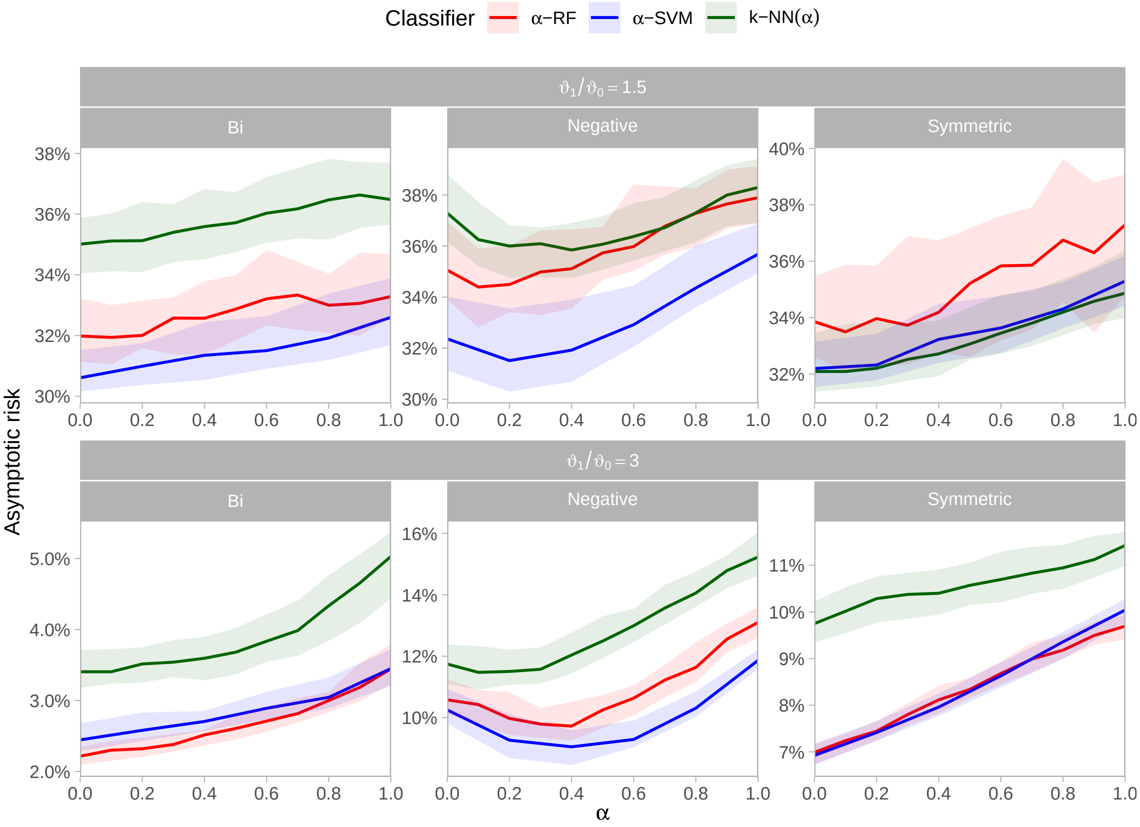

Figure 4.10: Blah.

In practice, the asymptotic risk cannot form the basis for our choice of α, since it is a non-observable quantity. Then, α becomes an additional hyperparameter in the model, which may be selected by cross-validation along with the other tuning parameters. The results of this approach are presented in Figure 4.11. Roughly speaking, the selected values of α accord with our earlier findings, with $\alpha \approx 0.3$ being favoured in the negative logistic case and $\alpha = 0$ in the other two. However, there is significant variation in the selected values. Moreover, $\alpha = 1$ is chosen in a non-negligible proportion of runs, despite this being the worst choice according to the asymptotic risk criterion. This suggests that $\hat{L}_t(\cdot)$ and $\hat{L}_\infty(\cdot)$, when viewed as functions of α, can exhibit differing profiles.

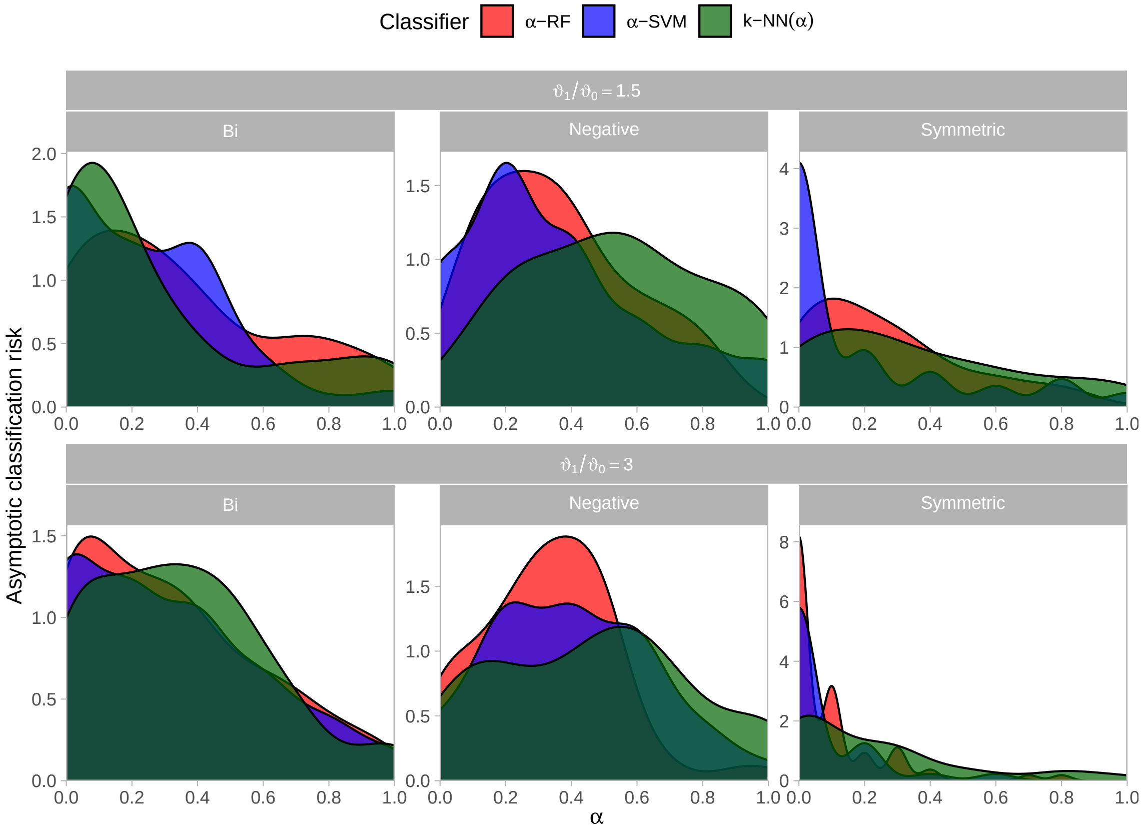

Figure 4.11: Blah.

This hypothesis is borne out by Figure 4.12, which plots the median values of the 'training loss' $\hat{L}_t$ (dashed lines) and 'test loss' $\hat{L}_\infty$ (solid lines) against α. Indeed, for the negative logistic model with $\vartheta_1/\vartheta_0 = 3$, the training loss of k-NN(α) is almost flat, while the test loss exhibits a definite positive gradient. This explains why the optimal α values are almost uniformly distributed – see the green kernel density estimate in bottom-middle panel of Figure 4.11. *Speculate as to what is going on here, and give some concluding remarks. Influence of n and k? More work needed on estimating asymptotic risk?*

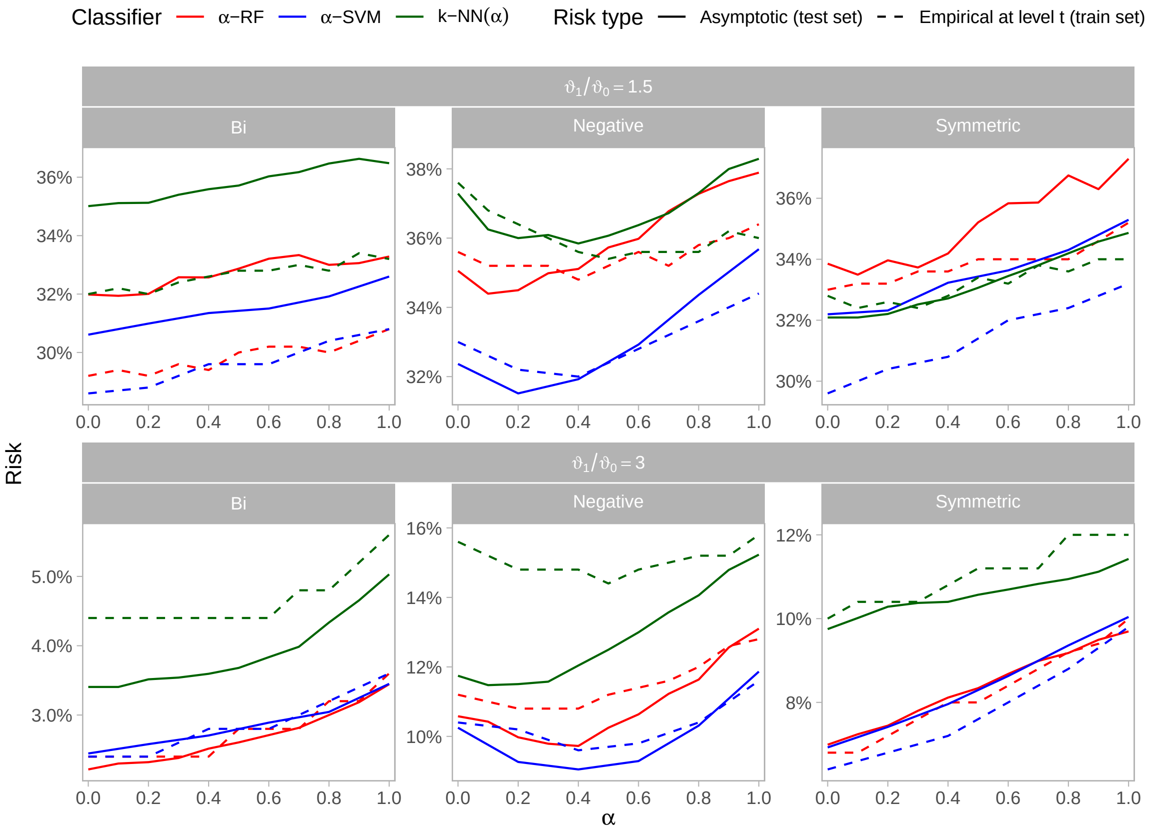

Figure 4.12: Blah.

4.3.3 Discussion

Discuss conclusions of CoDA classification stuff here.

4.4 Appendix material

4.4.1 (A) Computational details regarding the k-NN(α), α-SVM and α-RF classifiers

List the hyperparameters of each method, describe what they mean, list the ranges of values used, and give any relevant computational details. Refer to `Compositional` *and* `CompositionalML` *packages.*

5 EVA 2023 Data Challenge

5.1 Introduction

5.1.1 General

5.1.2 Multivariate challenges

In Challenges 3 and 4 we are presented with a d-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_d) \sim F_{\boldsymbol{X}}$, representing the value of an environmental variable at d sites in Utopia, where $F_{\boldsymbol{X}}$ is an unknown joint distribution function. Our goal is to estimate the probability that $\boldsymbol{X}$ lies in a given extreme 'failure region' based on a sample of independent observations of $\boldsymbol{X}$. The failure regions of interest are such that certain components of $\boldsymbol{X}$ are simultaneously large (while all remaining components, if any, are of lower order). The inherent difficulty of the task stems from the fact that the events' return periods are of similar order or even significantly longer than the observation period over which the data are collected; empirical methods based solely on the relative frequency of event occurrences are fruitless. Instead, we use the observed data to infer an estimate for the probabilistic structure of the joint tail of $\boldsymbol{X}$ and subsequently compute tail event probabilities under this model. This encapsulates the philosophy of multivariate extreme value statistics.

In the absence of prior knowledge about the physical processes driving the environment of Utopia, we are compelled to recourse to data-driven, statistical learning methods for multivariate extremes. The particular tools we choose will depend on the particular nature and difficulties of the task at hand. In Challenge 3, the failure regions are defined by *different* subsets of variables being large, thereby placing emphasis on accurately modelling the so-called extremal directions of $\boldsymbol{X}$. The salient characteristic of Challenge 4 is its high dimensionality, which calls for the utilisation of dimension reduction techniques, such as clustering, in order to overcome the curse of dimensionality inherent to tail dependence estimation.

5.2 Background

Our methodology relies on the parametric family of max-linear combinations of regularly varying random variables (Fougères et al. 2013; Kiriliouk and C. Zhou 2022) and the closely related tail pairwise dependence matrix (Cooley and Thibaud 2019; Larsson and Resnick 2012). In this section, we review the requisite background theory and introduce a novel approach to inference for max-linear models using sparsity-inducing simplex projections (Meyer and Wintenberger 2021).

5.2.1 Multivariate regular variation and the angular measure

Let $\boldsymbol{X}$ denote a random vector that takes values in the positive orthant $\mathbb{R}_+^d := [0,\infty)^d$. As is commonly done in multivariate extremes, we work in the framework of multivariate regular variation (MRV).

Definition 5.1. We say that $\boldsymbol{X}$ is *multivariate regularly varying* with index $\alpha > 0$, denoted $\boldsymbol{X} \in \mathrm{RV}_+^d(\alpha)$, if it satisfies the following (equivalent) definitions (Resnick 2007):

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

$$n\mathbb{P}(b_n^{-1}\boldsymbol{X} \in \cdot) \xrightarrow{\mathrm{v}} \nu_{\boldsymbol{X}}(\cdot), \qquad (n \to \infty), \tag{5.1}$$

   where $\xrightarrow{\mathrm{v}}$ denotes vague convergence in the space of non-negative Radon measures on $\mathbb{E}_0$. The *exponent measure* $\nu_{\boldsymbol{X}}$ is homogeneous of order $-\alpha$, that is $\nu_{\boldsymbol{X}}(s\,\cdot) = s^{-\alpha}\nu_{\boldsymbol{X}}(\cdot)$ for any $s > 0$.

2. For any norm $\|\cdot\|$ on $\mathbb{R}^d$, there exists a sequence $b_n \to \infty$ and a finite *angular measure* $H_{\boldsymbol{X}}$ on $\mathbb{S}_+^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\}$ such that for $(R, \boldsymbol{\Theta}) := (\|\boldsymbol{X}\|, \boldsymbol{X}/\|\boldsymbol{X}\|)$,

$$n\mathbb{P}((b_n^{-1}R, \boldsymbol{\Theta}) \in \cdot) \xrightarrow{\mathrm{v}} \nu_\alpha \times H_{\boldsymbol{X}}(\cdot), \qquad (n \to \infty), \tag{5.2}$$

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 $\nu_{\boldsymbol{X}}$ and $H_{\boldsymbol{X}}$ are related via

$$\nu_{\boldsymbol{X}}(\{\boldsymbol{x} \in \mathbb{E}_0 : \|\boldsymbol{x}\| > s, \boldsymbol{x}/\|\boldsymbol{x}\| \in \cdot\}) = s^{-\alpha} H_{\boldsymbol{X}}(\cdot),$$
$$\nu_{\boldsymbol{X}}(\mathrm{d}r \times \mathrm{d}\boldsymbol{\theta}) = \alpha r^{\alpha-1} \mathrm{d}r \, \mathrm{d}H_{\boldsymbol{X}}(\boldsymbol{\theta}).$$

The probabilistic tail of $\boldsymbol{X}$ decomposes into a univariate α-regularly varying radial component that is asymptotically independent of the angular component (Resnick 2007). The angular measure represents the limiting distribution of the angular component and encodes all information about the tail dependence structure.

Henceforth, assume that the marginal components of $\boldsymbol{X} \in \mathrm{RV}_+^d(\alpha)$ are identically distributed with a Fréchet distribution with shape parameter α, that is $\mathbb{P}(X_i < x) = \Psi_\alpha(x) := \exp(-x^{-\alpha})$ for $x > 0$ and $i = 1, \ldots, d$. (The data for Challenges 3 and 4 are on Gumbel margins, but a marginal transformation will be applied.) Moreover, we choose $\|\cdot\| = \|\cdot\|_\alpha$, the L_α-norm on $\mathbb{R}^d$, and specify that the normalising sequence in (5.2) is $b_n = n^{1/\alpha}$. With these particular choices the marginal variables have unit scale (Klüppelberg and Krali 2021, Definition 4) and $H_{\boldsymbol{X}}(\mathbb{S}_+^{d-1}) = d$.

The problem of modelling the angular measure has attracted considerable attention in recent years – a survey of related literature can be found in Engelke and Jevgenijs Ivanovs (2021). One research avenue concerns learning which sets of variables that may be concurrently extreme; this can be posed as a support detection problem (Goix et al. 2017; Simpson et al. 2020). Let $\mathcal{P}_d^\star := \mathcal{P}(\mathbb{V}(d)) \setminus \emptyset$ denote the power set of the index set $\mathbb{V}(d) := \{1, \ldots, d\}$ excluding the empty set. A set $\beta \in \mathcal{P}_d^\star$ is termed an extremal direction of $\boldsymbol{X} \in \mathrm{RV}_+^d(\alpha)$ if the subspace

$$C_\beta = \{\boldsymbol{w} \in \mathbb{S}_+^{d-1} : w_i > 0 \iff i \in \beta\} \subseteq \mathbb{S}_+^{d-1}$$

has non-zero $H_{\boldsymbol{X}}$-mass. Another branch of research aims at developing dimension reduction techniques for analysing the angular measure in high dimensions. To this end, one often considers a summary of the full dependence structure encoded in a matrix of pairwise extremal dependence metrics. One such matrix, originating from Larsson and Resnick

(2012) and later popularised by Cooley and Thibaud (2019), is the tail pairwise dependence matrix (TPDM). The TPDM of $\boldsymbol{X} \in \mathrm{RV}_+^d(2)$ is the $d \times d$ matrix $\Sigma = (\sigma_{ij})$ with entries

$$\sigma_{ij} = \int_{\mathbb{S}_+^{d-1}} \theta_i \theta_j \, \mathrm{d}H_{\boldsymbol{X}}(\boldsymbol{\theta}), \qquad (i, j = 1, \ldots, d). \tag{5.3}$$

The diagonal entries are the squared margin scales, i.e. $\sigma_{ii} = 1$ for $i = 1, \ldots, d$. The off-diagonal entries measure extremal dependence between the associated pairs of variables. In particular, $\sigma_{ij} = 0$ if and only if X_i and X_j are asymptotically independent. For asymptotically dependent variables the magnitude of σ_{ij} represents the dependence strength. An important property of the TPDM is its complete positivity (Cooley and Thibaud 2019, Proposition 5).

Definition 5.2. A matrix S is *completely positive* if there exists a matrix A with non-negative entries such that $S = AA^T$. We call A (resp. AA^T) a *CP-factor* (resp. *CP-decomposition*) of S.

This property connects the TPDM to the model class introduced in the following section.

5.2.2 The max-linear model for multivariate extremes

Our proposed methods for Challenges 3 and 4 employ the *max-linear model*, a parametric model based on the class of random vectors constructed by max-linear combinations of independent Fréchet random variables (Fougères et al. 2013). This model is appealing for several reasons. First, it is flexible, in the sense that any regularly varying random vector can be arbitrarily well-approximated by a max-linear model with sufficiently many parameters (Fougères et al. 2013). Since neither Challenge 3 nor Challenge 4 provide any prior information about the underlying data-generating processes, it is preferable to avoid imposing overly restrictive assumptions on the tail dependence structure. Secondly, although the number of parameters grows rapidly – at least $\mathcal{O}(d)$ but often even $\mathcal{O}(d^2)$ – efficient inference procedures are available even in high dimensions. Scalability is critical for Challenge 4. Finally, extremal directions and failure probabilities can be straightforwardly identified and computed directly from the model parameter (Kiriliouk and C. Zhou 2022).

Definition 5.3. For some $q \geq 1$ and $\alpha > 0$, let $\boldsymbol{Z} = (Z_1, \dots, Z_q)$ be a random vector with independent components $Z_1, \dots, Z_q \sim \Psi_\alpha$ and $A = (a_{ij}) \in \mathbb{R}_+^{d \times q}$ a deterministic matrix. If

$$X_i := \bigvee_{j=1}^{q} a_{ij} Z_j, \qquad (i = 1, \dots, d),$$

then $\boldsymbol{X} = (X_1, \dots, X_d)$ is said to be *max-linear* with *noise coefficient matrix* A, denoted $\boldsymbol{X} \sim \text{MaxLinear}(A; \alpha)$ and we write $\boldsymbol{X} = A \circ \boldsymbol{Z}$.

Cooley and Thibaud (2019) show that $\boldsymbol{X} = A \circ \boldsymbol{Z} \in \text{RV}_+^d(\alpha)$ with angular measure

$$H_{\boldsymbol{X}}(\cdot) = \sum_{j=1}^{q} \|\boldsymbol{a}_j\|_\alpha^\alpha \delta_{\boldsymbol{a}_j / \|\boldsymbol{a}_j\|_\alpha}(\cdot), \tag{5.4}$$

where δ is the Dirac mass function. The angles along which extremes can occur (in the limit) are precisely the q self-normalised columns of A. Therefore $\beta \in \mathcal{P}_d^\star$ is an extremal direction of $\boldsymbol{X}$ if and only if there exists $j \in \{1, \dots, q\}$ such that $\boldsymbol{a}_j / \|\boldsymbol{a}_j\| \in C_\beta$. The testing procedure of Kiriliouk (2020) can provide guidance for choosing q; for our purposes, q either represents a tuning parameter (Challenge 3) or takes a fixed value owing to computational/algorithmic restrictions (Challenge 4). Substituting (5.4) into (5.3) shows the TPDM of $\boldsymbol{X} \sim \text{MaxLinear}(A; \alpha = 2)$ is $\Sigma = AA^T$. In other words, the noise coefficient matrix is a CP-factor of the model TPDM. Conversely, given an arbitrary random vector $\boldsymbol{X} \in \text{RV}_+^d(2)$ with TPDM Σ, any CP-factor A of Σ parametrises a max-linear model with identical pairwise tail dependence metrics to $\boldsymbol{X}$.

Kiriliouk and C. Zhou (2022) present classes of tail events $\mathcal{C} \subset \mathbb{E}_0$ for which $\mathbb{P}(\boldsymbol{X} \in \mathcal{C})$ is approximately some function of the parameter A. With a view to Challenges 3 and 4, we focus on tail events where $\boldsymbol{X}$ is large in the $s \leq d$ components indexed by $\beta = \{\beta_1, \dots, \beta_s\} \in \mathcal{P}_d^\star$, while all $d - s$ remaining components are of lower order. For $\boldsymbol{u} = (u_1, \dots, u_s) \in \mathbb{R}_+^s$ a vector of high thresholds and $\boldsymbol{l} \in \mathbb{R}_+^{d-s}$ a vector of comparatively low (upper) thresholds, then we write

$$\mathcal{C}_{\beta, \boldsymbol{u}, \boldsymbol{l}} := \{\boldsymbol{x} \in \mathbb{E}_0 : \boldsymbol{x}_\beta > \boldsymbol{u}, \boldsymbol{x}_{-\beta} < \boldsymbol{l}\}, \qquad \mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\beta, \boldsymbol{u}, \boldsymbol{l}}) = \mathbb{P}(\boldsymbol{X}_\beta > \boldsymbol{u}, \boldsymbol{X}_{-\beta} < \boldsymbol{l}),$$

When $\beta = \mathbb{V}(d)$ the threshold vector $\boldsymbol{l}$ is superfluous and may be omitted. In Section 2.3, Kiriliouk and C. Zhou (2022) specify an approximate formula for $\mathbb{P}(A \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d), \boldsymbol{u}})$ in terms of A. We derive a more general formula for $\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}})$ for arbitrary $\beta \in \mathcal{P}_d^\star$: if $\boldsymbol{X} \sim \mathrm{MaxLinear}(A;\alpha)$, then

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}}) \approx \hat{\mathbb{P}}(\boldsymbol{X} \in \mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}}) := \sum_{j:\frac{\boldsymbol{a}_j}{\|\boldsymbol{a}_j\|_\alpha} \in C_\beta} \min_{i=1,\ldots,s} \left(\frac{a_{\beta_i,j}}{u_i}\right)^\alpha. \tag{5.5}$$

The formula should be interpreted as being valid in the limit as the thresholds $u_i \to \infty$ while $\boldsymbol{l}$ is held fixed. If $\boldsymbol{u} = u\mathbf{1}_s$ for some large scalar $u > 0$, then we write $\mathcal{C}_{\beta,u} := \mathcal{C}_{\beta,u\mathbf{1}_s}$. We derive the formula for this simpler case but the steps are easily modified for general $\boldsymbol{u}$. From (5.1) we have that, provided u is sufficiently large,

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\beta,u}) = \frac{1}{n}\left[n\mathbb{P}\left(\frac{\boldsymbol{X}}{n^{1/\alpha}} \in \frac{\mathcal{C}_{\beta,u}}{n^{1/\alpha}}\right)\right] \approx \frac{1}{n}\nu_{\boldsymbol{X}}\left(\frac{\mathcal{C}_{\beta,u}}{n^{1/\alpha}}\right) = \nu_{\boldsymbol{X}}(\mathcal{C}_{\beta,u}),$$

where the last step follows from homogeneity of the exponent measure. Transforming to pseudo-polar coordinates we have

$$\nu_{\boldsymbol{X}}(\mathcal{C}_{\beta,u}) = \int_{\left\{(r,\boldsymbol{w})\,:\,r\boldsymbol{w}_\beta > u\mathbf{1}_s,\, r\boldsymbol{w}_{\beta^c} < l\mathbf{1}_{d-s}\right\}} \alpha r^{-\alpha-1}\,\mathrm{d}r\,\mathrm{d}H_{\boldsymbol{X}}(\boldsymbol{w}). \tag{5.6}$$

Based on (5.4), there are only q possible angles along which extremes can occur, namely $\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha$ for $j = 1,\ldots,q$. However, only those angles $\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha \in C_\beta$ will contribute to the integral. If $\boldsymbol{w} = \boldsymbol{a}/\|\boldsymbol{a}\|_\alpha \notin C_\beta$, then there exists $i \in \{1,\ldots,s\}$ such that $w_{\beta_i} = 0$ and hence $rw_{\beta_i} = 0 < u$ for all $0 < r < \infty$. On the other hand, if $\boldsymbol{w} = \boldsymbol{a}/\|\boldsymbol{a}\|_\alpha \in C_\beta$, then $\boldsymbol{w}_{\beta^c} = \mathbf{0} \in \mathbb{R}^{d-s}$ and so

$$r\boldsymbol{w}_\beta > u\mathbf{1}_s,\, r\boldsymbol{w}_{\beta^c} < l\mathbf{1}_{d-s} \iff r\boldsymbol{w}_\beta > u\mathbf{1}_s \iff r > \max_{i=1,\ldots,s}\left(\frac{\|\boldsymbol{a}\|_\alpha u}{a_{\beta_i}}\right).$$

Thus

$$\nu_{\boldsymbol{X}}(\mathcal{C}_{\beta,u}) = \sum_{j:\frac{\boldsymbol{a}_j}{\|\boldsymbol{a}_j\|_\alpha} \in C_\beta} \|\boldsymbol{a}_j\|_\alpha^\alpha \int_{\max_i\left(\frac{\|\boldsymbol{a}_j\|_\alpha u}{a_{\beta_i,j}}\right)}^\infty \alpha r^{-\alpha-1}\,\mathrm{d}r = \sum_{j:\frac{\boldsymbol{a}_j}{\|\boldsymbol{a}_j\|_\alpha} \in C_\beta} \min_{i=1,\ldots,s}\left(\frac{a_{\beta_i,j}}{u}\right)^\alpha.$$

For $\alpha = 1$ we can establish the upper bound $\hat{\mathbb{P}}(\boldsymbol{X} \in \mathcal{C}_{\beta,u}) \leq H_{\boldsymbol{X}}(C_\beta)/(su)$ with equality attained if and only if the angular measure places all of its mass on C_β at the centroid $\boldsymbol{e}(\beta)/|\beta|$, where $\boldsymbol{e}(\beta) := (\mathbf{1}\{i \in \beta\} : i = 1,\ldots,d) \in \{0,1\}^d$. This bound represents the limiting probability of the angle lying in the relevant subspace multiplied by the survivor function of a Pareto($\alpha = 1$) random variable evaluated at the effective radial threshold $\|u\mathbf{1}_s\|_1 = su$.

5.2.3 Existing approaches to inference for max-linear models

Suppose $\boldsymbol{X} \sim \text{MaxLinear}(A; \alpha)$, where α is known and A is to be estimated from a collection $\{\boldsymbol{x}_t = (r_t, \boldsymbol{\theta}_t) : t = 1, \ldots, n\}$ of independent observations of $\boldsymbol{X} = (R, \boldsymbol{\Theta})$. For $j = 1, \ldots, n$, let $r_{(j)}$ denote the jth upper order statistic of $\{r_1, \ldots, r_n\}$ and denote by $\boldsymbol{x}_{(j)}$ and $\boldsymbol{\theta}_{(j)}$ the corresponding observation vector and angular component, respectively. We consider two existing approaches for inferring A: an empirical estimate and a CP-decomposition based estimate (Cooley and Thibaud 2019; Kiriliouk and C. Zhou 2022). The former is a natural estimate in view of (5.4); the latter exploits the connection with CP-factors of the TPDM described in the ensuant discussion.

Definition 5.4. For $1 \leq k < n$, the *empirical estimate* of A based on is $\hat{A} = (\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_k) \in \mathbb{R}_+^{d \times k}$, where $\hat{\boldsymbol{a}}_j := (d/k)^{1/\alpha} \boldsymbol{\theta}_{(j)}$ for $j = 1, \ldots, k$.

The quantity k is the customary tuning parameter representing the number of ‘extremes’ – observations with norm not less than the implied radial threshold $r_{(k)}$ – that enter into the estimator. The associated angular measure (for any α) and TPDM (for $\alpha = 2$) are given by

$$\hat{H}_{\boldsymbol{X}}(\cdot) := H_{\hat{A} \circ \boldsymbol{Z}}(\cdot) = \frac{d}{k} \sum_{t=1}^{n} \mathbf{1}\{\boldsymbol{\theta}_t \in \cdot, r_t \geq r_{(k)}\} \tag{5.7}$$

$$\hat{\Sigma} := \hat{A}\hat{A}^T. \tag{5.8}$$

These are the empirical angular measure (Einmahl and Segers 2009) and empirical TPDM (Cooley and Thibaud 2019), respectively. Due to non-uniqueness of CP-factors, further estimates of A can be obtained by CP-decomposition of (5.8).

Definition 5.5. Any CP-factor of $\hat{\Sigma}$ is called a *CP-estimate* of A, denoted $\tilde{A}$.

In general, $\tilde{A}$ is distinct from $\hat{A}$ and induces a different angular measure $\tilde{H}_{\boldsymbol{X}} := H_{\tilde{A} \circ \boldsymbol{Z}}$. By construction, these angular measures give rise to identical tail pairwise dependence measures, since $\tilde{A}\tilde{A}^T = \hat{A}\hat{A}^T$. Note that $\tilde{A}$ implicitly depends on the same tuning parameter k as $\hat{A}$ via the empirical TPDM. All our CP-estimates are obtained using the procedure of Kiriliouk and C. Zhou (2022), which can efficiently factorise high-dimensional TPDMs. Their algorithm takes as input a TPDM with strictly positive entries and a permutation $(i_1, \ldots, i_d)$ of $\mathbb{V}(d)$ and returns a square CP-factor $\tilde{A} \in \mathbb{R}_+^{d\times d}$ whose columns satisfy $\tilde{\boldsymbol{a}}_j/\|\tilde{\boldsymbol{a}}_j\|_2 \in C_{\mathbb{V}(d)\setminus\{i_l\,:\,l<j\}}$ for $j = 1, \ldots, d$. (Note that not all input permutations will result in a valid decomposition.)

5.2.4 Inference for max-linear models based on sparse projections

A limitation of $\hat{A}$ and $\tilde{A}$ is that they do not capture the extremal directions of $\boldsymbol{X}$. For the empirical estimate, this stems from the angular component $\boldsymbol{X}/\|\boldsymbol{X}\|$ lying in the simplex interior $C_{\mathbb{V}(d)}$ almost surely. Consequently $\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_k \in\in C_{\mathbb{V}(d)}$ and $\hat{\mathbb{P}}(\hat{A} \circ \boldsymbol{Z} \in \mathcal{C}_{\beta,\boldsymbol{u}}) = 0$ for any $\beta \neq \mathbb{V}(d)$. On the other hand, the d extremal directions $\mathbb{V}(d), \mathbb{V}(d) \setminus \{i_1\}, \mathbb{V}(d) \setminus \{i_1, i_2\}, \ldots, \mathbb{V}(d)\setminus\{i_1, \ldots, i_{d-1}\}$ of $\tilde{A}\circ\boldsymbol{Z}$ are fully determined by the user-defined input path $(i_1, \ldots, i_d)$. To address this shortcoming in the existing estimates, we propose augmenting the empirical estimate with an alternative notion of angle based on Euclidean projections onto the L_1-simplex (Meyer and Wintenberger 2021).

Definition 5.6. The Euclidean projection onto the L_1-simplex is defined by

$$\pi : \mathbb{R}_+^d \to \mathbb{S}_+^{d-1}, \qquad \pi(\boldsymbol{v}) = \underset{\boldsymbol{w}\in\mathbb{S}_+^{d-1}}{\arg\min} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2.$$

This projection is useful because $\pi(\boldsymbol{v})$ may lie on the simplex boundary even when $\boldsymbol{v}/\|\boldsymbol{v}\|_1$ does not. Assume now that $\alpha = 1$.

Definition 5.7. For $1 \leq k < n$, the *sparse empirical estimate* of A is $\hat{A}^\star = (\hat{\boldsymbol{a}}_1^\star, \ldots, \hat{\boldsymbol{a}}_k^\star) \in \mathbb{R}_+^{d\times k}$, where $\hat{\boldsymbol{a}}_j^\star = (d/k)\pi(\boldsymbol{x}_{(j)}/r_{(k+1)})$ for $j = 1, \ldots, k$.

The corresponding angular measure

$$\hat{H}_{\boldsymbol{X}}^\star(\cdot) := H_{\hat{A}^\star\circ\boldsymbol{Z}}(\cdot) = \frac{d}{k}\sum_{j=1}^{k} \mathbf{1}\{\pi(\boldsymbol{x}_{(j)}/r_{(k+1)}) \in \cdot\}$$

spreads mass across the subspaces $C_\beta \subseteq \mathbb{S}_+^{d-1}$ in which the projected data lie and hence $\hat{\mathbb{P}}(\hat{A}^\star \circ \boldsymbol{Z} \in \mathcal{C}_{\beta,\boldsymbol{u}}) \neq 0$ for all corresponding β. A full study of the theoretical properties of $\hat{A}^\star$ has not been conducted. Having introduced our estimator and all the requisite theory, we are ready to present our methods for the multivariate challenges.

5.3 Challenge 3

5.3.1 Data

Challenge 3 considers a trivariate random vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ on standard Gumbel margins, i.e. $\mathbb{P}(Y_i < y) = G(y) := \exp(-\exp(-y))$ for $y \in \mathbb{R}$ and $i = 1, 2, 3$. It entails estimating

$$p_1 := \mathbb{P}(Y_1 > 6, Y_2 > 6, Y_3 > 6), \qquad p_2 := \mathbb{P}(Y_1 > 7, Y_2 > 7, Y_3 < -\log(\log(2))).$$

The data comprise $n = 21,000$ independent observations $\{\boldsymbol{y}_t = (y_{t1}, y_{t2}, y_{t3}) : t = 1, \ldots, n\}$ of $\boldsymbol{Y}$. Additional covariate information is available but not leveraged by our method.

5.3.2 Methodology

Let $\boldsymbol{X} = (X_1, X_2, X_3)$ denote the random vector obtained by transforming $\boldsymbol{Y}$ to Fréchet margins with shape parameter 1, i.e. $X_i = \Psi_1^{-1}(G(Y_i)) = \exp(Y_i) \sim \Psi_1$ for $i = 1, 2, 3$. The above probabilities can be expressed as

$$p_1 = \mathbb{P}(X_1 > e^6, X_2 > e^6, X_3 > e^6), \qquad p_2 = \mathbb{P}(X_1 > e^7, X_2 > e^7, X_3 < 1/\log(2)). \tag{5.9}$$

The threshold values e^6, e^7, and $1/\log(2)$ correspond approximately to the 99.8%, 99.9% and 50% quantiles of Ψ_1, respectively. Our solution models $\boldsymbol{X}$ as max-linear and infers the noise coefficient matrix using the sparse empirical estimator; this step involves fixing the hyperparameter k. Then we estimate p_1 and p_2 under this model via (5.5), that is

$$\hat{p}_1 = \hat{\mathbb{P}}(\hat{A}^\star \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d),\exp(6)}), \qquad \hat{p}_2 = \hat{\mathbb{P}}(\hat{A}^\star \circ \boldsymbol{Z} \in \mathcal{C}_{\{1,2\},\exp(7),1/\log(2)}). \tag{5.10}$$

Inference is based on the transformed data $\{\boldsymbol{x}_t = (x_{t1}, x_{t2}, x_{t3}) : t = 1, \dots, n\}$, where $x_{ti} := \exp(y_{ti})$ for $t = 1, \dots, n$ and $i = 1, 2, 3$.

5.3.3 Results

The results presented are based on $k = 500 \approx 2.5\% \times n$. The results' sensitivity to this choice will be examined later.

The ternary plots in Figure 5.1 depict the angular components associated with the k largest observations in norm, i.e. exceedances of the radial threshold $r_{(k+1)} \approx 138.77$. Those in the left-hand plot represent the self-normalised vectors $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(k)}$. We find points lying near the centre of the ternary plot as well as in the neighbourhood of each of three edges and three vertices. This suggests that the angular measure spreads mass across all seven subspaces C_β, $\beta \in \mathcal{P}_3^\star$. However, we reiterate that no points lie *exactly* on the boundaryIn contrast, the sparse angles $\{\pi(\boldsymbol{x}_t / r_{(k+1)}) : r_t > r_{(k+1)}\}$ in the right-hand plot lie in the interior (black, 40 points), along the edges (red, 139 points), and on the vertices (blue, 321 points) of the closed simplex. Only the 40 vectors in $C_{\mathbb{V}(3)}$ and 23 vectors in $C_{\{1,2\}}$ will enter into the estimates of (5.9).

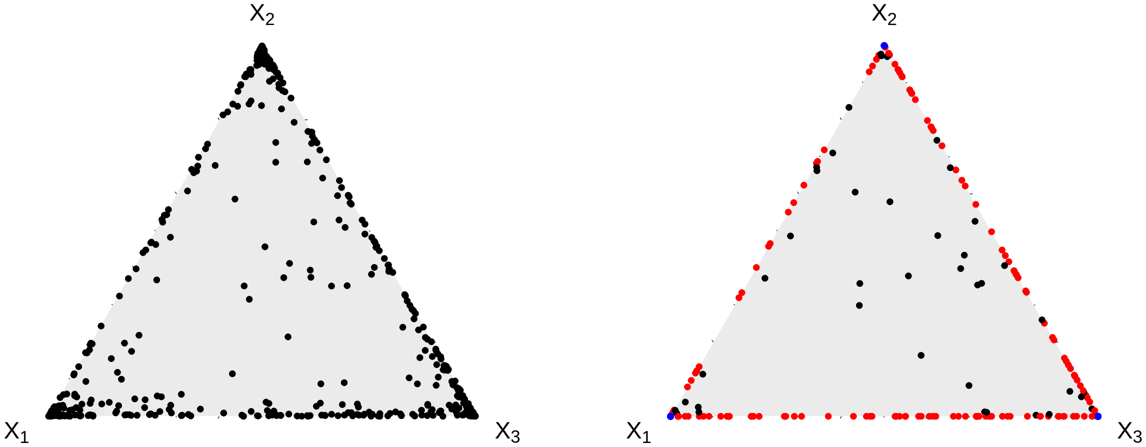

Figure 5.1: The angles $\boldsymbol{\theta}_{(j)} \in \mathbb{S}_+^2$ (left) and Euclidean projections $\pi(\boldsymbol{x}_{(j)}/r_{(k+1)}) \in \mathbb{S}_+^2$ (right) for $j = 1, \dots, 500$ based on the Challenge 3 data. Points are coloured according to whether they lie in the interior (black), on an edge (red) or on a vertex (blue).

The projected vectors are collated to form the 3×500 matrix $\hat{A}^\star$. The first 100 columns of the matrices $\hat{A}$ and $\hat{A}^\star$ are represented visually in Figure 5.2. As expected, $\hat{A}$ is dense while $\hat{A}^\star$ exhibits a high degree of sparsity, in the sense that most of its columns satisfy

$\|\hat{\boldsymbol{a}}_j^\star\|_0 < \|\hat{\boldsymbol{a}}_j\|_0 = 3$. The duplicate columns in $\hat{A}^\star$ can be merged to produce a compressed estimate $\hat{A}^\star_{\text{comp}}$ comprising $q_{\text{comp}} = 40 + 139 + 3 = 182$ unique columns. One might prefer $\hat{A}_{\text{comp}}$ over $\hat{A}^\star$ for model parsimony reasons, but ultimately they parametrise identical models since $H_{\hat{A}^\star \circ \boldsymbol{Z}} = H_{\hat{A}^\star_{\text{comp}} \circ \boldsymbol{Z}}$.

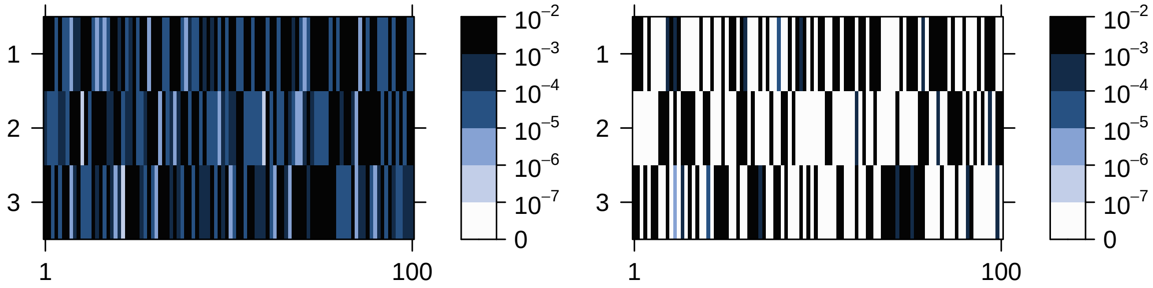

Figure 5.2: The first 100 columns of $\hat{A}$ (left) and $\hat{A}^\star$ (right) in Challenge 3. The cells' colour intensities represent the magnitude of the corresponding matrix entries.

Substituting $\hat{A}^\star$ (or $\hat{A}^\star_{\text{comp}}$) into (5.9) yields our final point estimates $\hat{p}_1 = 3.36 \times 10^{-5}$ and $\hat{p}_2 = 2.76 \times 10^{-5}$, to three significant figures. We are pleased to find these are very close to the true values given in **Rohr23**. Moreover, Figure 5.3 shows that the estimates are fairly stable with respect to k in the range $1.5\% < k/n < 6\%$.

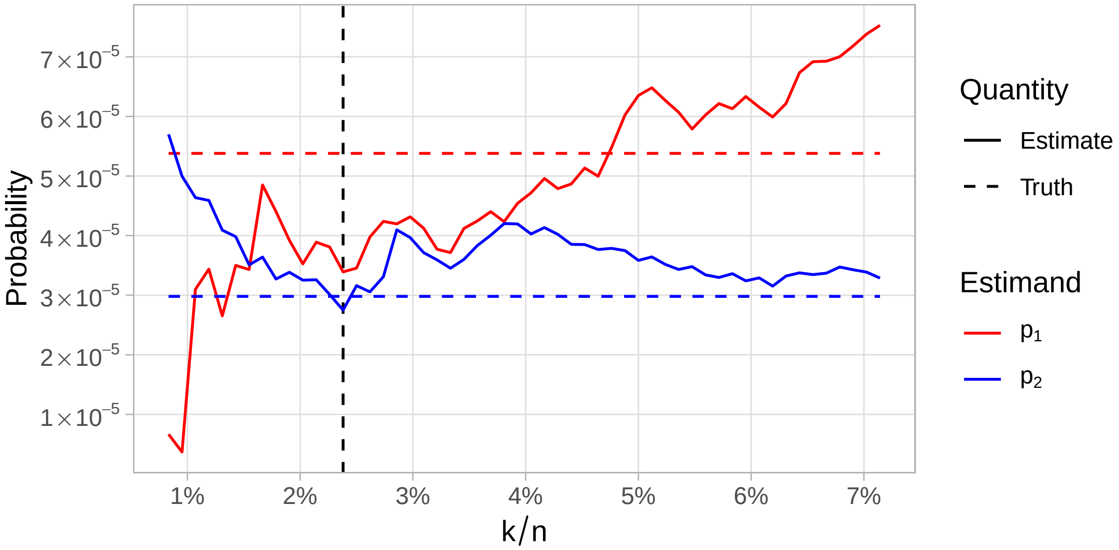

Figure 5.3: Estimates obtained using our Challenge 3 methodology with different choices for the tuning parameter, k. The horizontal dashed lines indicate the estimands' true values. Our submitted solution was based on $k = 500$ (vertical dashed line).

5.4 Challenge 4

5.4.1 Data

Challenge 4 regards a $d = 50$ dimensional random vector $\boldsymbol{Y}$ on standard Gumbel margins. The components of $\boldsymbol{Y}$ are random variables $Y_{i,j}$ for $i = 1, \ldots, 25$ and $j = 1, 2$, representing the value of an environmental variable at the ith site in the administrative area of government area j. The joint exceedance probabilities

$$\begin{aligned} p_1 &:= \mathbb{P}(Y_{i,j} > G^{-1}(1 - \phi_j) : i = 1, \ldots, 25,\, j = 1, 2), \\ p_2 &:= \mathbb{P}(Y_{i,j} > G^{-1}(1 - \phi_1) : i = 1, \ldots, 25,\, j = 1, 2), \end{aligned}$$

where $\phi_1 = 1/300$ and $\phi_2 = 12/300$, are to be estimated from $n = 10,000$ independent observations $\{\boldsymbol{y}_t = (y_{t,i,j} : i = 1, \ldots, 25,\, j = 1, 2) : t = 1, \ldots, n\}$ of $\boldsymbol{Y}$.

5.4.2 Methodology

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ denote the random vector obtained from $\boldsymbol{Y}$ by re-indexing its variables and transforming to Fréchet margins with shape parameter $\alpha = 2$, i.e.

$$X_{i+25(j-1)} := \Psi_2^{-1}(G(Y_{i,j})) = \exp(Y_{i,j}/2) \sim \Psi_2, \qquad (i = 1, \ldots, 25,\, j = 1, 2).$$

The choice of $\alpha = 2$ will be justified later. The data are transformed in an identical way yielding $\{\boldsymbol{x}_t = (x_{t1}, \ldots, x_{td}) : t = 1, \ldots, n\}$, where $x_{t,i+25(j-1)} := \exp(y_{t,i,j}/2)$ for $t = 1, \ldots, n$, $i = 1, \ldots, 25$ and $j = 1, 2$. The estimands can now be expressed as

$$\begin{aligned} p_1 &= \mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d), \boldsymbol{u}_1}), \qquad [\boldsymbol{u}_1]_i = \begin{cases} \Psi_2^{-1}(1 - \phi_1), & \text{if } 1 \leq i \leq 25 \\ \Psi_2^{-1}(1 - \phi_2), & \text{if } 26 \leq i \leq 50 \end{cases}, \\ p_2 &= \mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d), \boldsymbol{u}_2}), \qquad \boldsymbol{u}_2 = \Psi_2^{-1}(1 - \phi_1)\mathbf{1}_{50}. \end{aligned}$$

At a high level, our method proceeds in a similar vein to Challenge 3: we will model $\boldsymbol{X}$ as max-linear and compute estimates for p_1 and p_2 based on (5.5). However, our exploratory analysis revealed that not all components of $\boldsymbol{X}$ are asymptotically dependent.

This implies that $H_{\boldsymbol{X}}(\mathbb{V}(d)) = 0$ and has important ramifications for how we construct our solution. On the one hand, the empirical or CP-estimates of A will assert that $C_{\mathbb{V}(d)}$ is an extremal direction, resulting in a misspecified model. On the other hand, a model that correctly identifies $C_{\mathbb{V}(d)}$ as a $H_{\boldsymbol{X}}$-null set is also of little value, since the joint exceedance probabilities are known to be non-zero. How do we resolve this apparent contradiction? The key is to identify clusters of asymptotically dependent variables prior to model fitting; a joint exceedance in all clusters equates to a joint exceedance across all variables. Our working hypothesis – that the marginal variables can be partitioned such that asymptotic independence is present between clusters but not within them – can be formalised as follows:

Assumption 1. There exists $2 \leq K \leq d$ and a partition $\beta_1, \ldots, \beta_K$ of $\mathbb{V}(d)$ such that the angular measure is supported on the closed subspaces $\bar{C}_{\beta_1}, \ldots, \bar{C}_{\beta_K} \subset \mathbb{S}_+^{d-1}$, where $\bar{C}_\beta := \{C_{\beta'} : \beta' \subseteq \beta\}$ for any $\beta \in \mathcal{P}_d^\star$. That is, $H_{\boldsymbol{X}}(\cup_{l=1}^K \bar{C}_{\beta_l}) = H_{\boldsymbol{X}}(\mathbb{S}_+^{d-1})$.

This scenario has already been explored in extremes, cf. Assumption 1 in Fomichov and Ivanovs (2023). If $\boldsymbol{X}$ is max-linear with parameter $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q) \in \mathbb{R}_+^{d \times q}$, then the assumption can be equivalently restated as follows.

Assumption 2. There exist permutations $\pi : \mathbb{V}(d) \to \mathbb{V}(d)$ and $\phi : \mathbb{V}(q) \to \mathbb{V}(q)$ such that $\boldsymbol{X}_\pi := (X_{\pi(1)}, \ldots, X_{\pi(d)}) \sim \text{MaxLinear}(A_\phi; \alpha)$, where $A_\phi := (\boldsymbol{a}_{\phi(1)}, \ldots, \boldsymbol{a}_{\phi(q)}) \in \mathbb{R}_+^{d \times q}$ is block-diagonal with $2 \leq K \leq d$ blocks. For $l = 1, \ldots, K$, the lth block matrix $A_\phi^{(l)}$ has $d_l = |\beta_l|$ rows, $1 \leq q_l < q$ columns, and is such that the $q_l \times q_l$ matrix $A_\phi^{(l)}(A_\phi^{(l)})^T$ has strictly positive entries. The blocks' dimensions satisfy $\sum_{l=1}^K d_l = d$ and $\sum_{l=1}^K q_l = q$.

Under this framework, $\boldsymbol{X}$ divides into random sub-vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$, where $\boldsymbol{X}^{(l)} := (X_j : j \in \beta_l) \sim \text{MaxLinear}(A_\phi^{(l)}; \alpha)$ for $l = 1, \ldots, K$. (Henceforth, assume for notational convenience that the columns of A are already ordered, so that $A = A_\phi$.) While the clustering assumption is simplistic and cannot be expected to hold in general applications, it reflects what we suspect to be the true dependence structure for Challenge 4. Moreover, it is dimension reducing because the original d-dimensional problem is transformed to a set of K independent problems with dimensions $d_1, \ldots, d_K < d$. This ameliorates the curse of dimensionality to some extent.

In general, a joint exceedance event can be decomposed into concurrent joint exceedances in all K clusters as $\{\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d),u}\} = \cap_{l=1}^{K}\{\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}\}$, where each threshold subvector $\boldsymbol{u}^{(l)}$ is defined from $\boldsymbol{u}$ analogously to $\boldsymbol{X}^{(l)}$. Since we consider variables in different clusters to be asymptotically independent, we assume that for large, finite thresholds, joint exceedances in each cluster are approximately independent events, so that

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}}) = \mathbb{P}\left(\bigcap_{l=1}^{K}\{\boldsymbol{X}^{(l)} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}\}\right) \approx \prod_{l=1}^{K}\mathbb{P}(\boldsymbol{X}^{(l)} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}).$$

Assuming $\boldsymbol{X}^{(l)} \sim \text{MaxLinear}(A^{(l)};\alpha)$ for $l = 1,\ldots,K$, each term in the product can be estimated using (5.5), so that

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}}) \approx \prod_{l=1}^{K}\hat{\mathbb{P}}(A^{(l)} \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}). \tag{5.11}$$

The final step is to replace $A^{(1)},\ldots,A^{(K)}$ with suitably estimated counterparts. We opted for CP-estimates for two reasons: (i) they are rooted in the TPDM, which is geared towards high-dimensional settings, and (ii) their non-uniqueness enables us to compute numerous parameter/probability estimates, whose variation reflects the model uncertainty that arises from summarising dependence via the TPDM and thereby overlooking higher-order dependencies between components. The use of CP-estimates justifies our choosing $\alpha = 2$ in the pre-processing step and throughout.

5.4.3 Results

We now present our results for Challenge 4. First, the variables $X_1,\ldots,X_d$ are partitioned into K groups based on asymptotic (in)dependence using the clustering algorithm of Bernard et al. (2013). This entails constructing a distance matrix $\mathcal{D} = (\hat{d}_{ij})$, where $\hat{d}_{ij}$ denotes a non-parametric estimate of the F-madogram distance between variables X_i and X_j. The distance metric is connected to the strength of extremal dependence between X_i and X_j, with $\hat{d}_{ij} \approx 0$ implying strong asymptotic dependence and $\hat{d}_{ij} = 1/6$ in the case of asymptotic independence. The partition around medoids (PAM) clustering algorithm (Kaufman and Rousseeuw 1990) returns a partition $\beta_1,\ldots,\beta_K$ of $\mathbb{V}(d)$ based on $\mathcal{D}$. The number of clusters K is a pre-specified tuning parameter; we identify $K = 5$ clusters whose sizes are

Table 5.1: Summary statistics for the Challenge 4 clusters and their empirical TPDMs.

(a) Challenge 4: cluster summary statistics.

				$\{\sigma_{ij} : i \neq j\}$		
Cluster	Size	U1 sites	U2 sites	Min.	Median	Max.
1	9	7	2	0.30	0.33	0.40
2	8	5	3	0.64	0.68	0.74
3	8	3	5	0.62	0.67	0.74
4	12	7	5	0.27	0.33	0.39
5	13	3	10	0.43	0.50	0.58

given in Table 5.1. Defining cluster membership variables $\mathcal{M}_1,\ldots,\mathcal{M}_d \in \{1,\ldots,K\}$ by $\mathcal{M}_i = l \iff i \in \beta_l$ for $i=1,\ldots,d$, we find that $\max\{\hat{d}_{ij} : \mathcal{M}_i = \mathcal{M}_j\} = 0.113 < 1/6$ and $\min\{\hat{d}_{ij} : \mathcal{M}_i \neq \mathcal{M}_j\} = 0.164 \approx 1/6$. These summary statistics are consistent with Assumptions 1/2.

Next, we compute the empirical TPDMs of $\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(K)}$. For $l=1,\ldots,K$ and $t=1,\ldots,n$, define the observational sub-vector $\boldsymbol{x}_t^{(l)} = (x_{ti} : i \in \beta_l)$ and its radial and angular components $r_t^{(l)} = \|\boldsymbol{x}_t^{(l)}\|_2$, $\boldsymbol{\theta}_t^{(l)} = \boldsymbol{x}_t^{(l)}/\|\boldsymbol{x}_t^{(l)}\|_2$, respectively. Let $\boldsymbol{x}_{(j)}^{(l)}$, $r_{(j)}^{(l)}$ and $\boldsymbol{\theta}_{(j)}^{(l)}$ denote the vector, radius, and angle associated with the jth largest observation in norm among $\boldsymbol{x}_1^{(l)},\ldots,\boldsymbol{x}_n^{(l)}$. Choose a tuning parameter $1 \leq k_l \leq n$ representing the number of extreme observations that enter into the estimate for cluster l. Then

$$\hat{A}^{(l)} = \left(\frac{d_l}{k_l}\boldsymbol{\theta}_{(1)}^{(l)},\ldots,\frac{d_l}{k_l}\boldsymbol{\theta}_{(k_l)}^{(l)}\right), \qquad \hat{\Sigma}_{\boldsymbol{X}^{(l)}} = \hat{A}^{(l)}(\hat{A}^{(l)})^T.$$

We set $k_1 = \ldots = k_K =: k = 250$, corresponding to a sampling fraction of $k/n = 2.5\%$ for each cluster. The empirical TPDMs for the first two clusters are displayed in Figure 5.4; summary statistics for all clusters' TPDMs are listed in Table 5.1. Asymptotic dependence is strongest in clusters 2 and 3 and weakest in cluster 1 and 4.

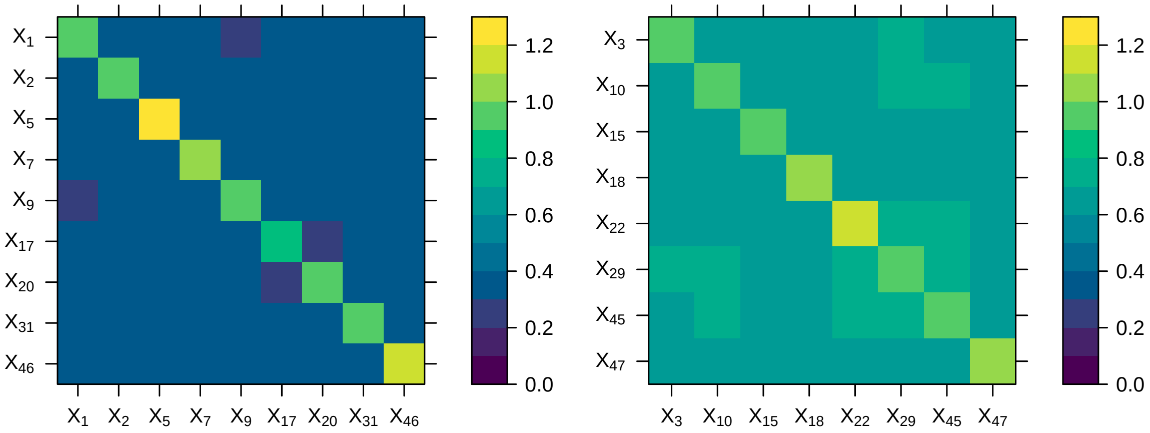

Figure 5.4: The empirical TPDMs for the first (left) and second (right) clusters in Challenge 4, based on the $k = 250$ largest observations.

By repeated application of the CP-decomposition algorithm of Kiriliouk and C. Zhou (2022) with randomly chosen inputs $(i_1, \ldots, i_{d_l})$, we obtain $N_{\text{cp}} = 50$ CP-estimates of each matrix $A^{(l)}$. The resulting CP-factors are denoted by $\tilde{A}_1^{(l)}, \ldots, \tilde{A}_{N_{\text{cp}}}^{(l)}$. Note that among $\tilde{A}_1^{(l)}, \ldots, \tilde{A}_{N_{\text{cp}}}^{(l)}$ there are at most d_l distinct leading columns, because there are only d_l unique ways to initialise the (deterministic) algorithm. But $\hat{\mathbb{P}}(\tilde{A} \times \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}})$ is fully determined by $\tilde{\boldsymbol{a}}_1$, so in fact $\{\hat{\mathbb{P}}(\tilde{A}_r^{(l)} \times \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}) : r = 1, \ldots, N_{\text{cp}}\}$ contains at most d_l distinct values. These values are represented by black points in the left-hand plot in **?@fig-c4-prob-estimates**. Clusters 2 and 3 are deemed most likely to experience a joint extreme event, because they contain a small number ($d_2 = d_3 = 8$) of strongly dependent variables. The relatively low risk in clusters 1 and 4 can be attributed to weak dependence between their components. The effect of changing the threshold from $\boldsymbol{u}_1$ to $\boldsymbol{u}_2$ is most pronounced in cluster 5, since it is primarily composed of sites in area U2.

Substituting each CP-estimate into (5.11) and enumerating over all possible combinations, we produce sets of estimates of p_i given by

$$\tilde{P}_i := \{\hat{\mathbb{P}}(\tilde{A}_{r_1}^{(1)} \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d_1),\boldsymbol{u}_i^{(1)}}) \times \ldots \times \hat{\mathbb{P}}(\tilde{A}_{r_K}^{(K)} \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d_K),\boldsymbol{u}_i^{(K)}}) : r_1, \ldots, r_K = 1, \ldots, N_{\text{cp}}\},$$

for $i = 1, 2$. Each set has size $N_{\text{cp}}^K \approx 3 \times 10^8$ (including repeated values) and contains $\prod_{l=1}^K d_l = 89,856$ distinct values. The distributions of the estimates are represented by the black points in the right-hand panel of **?@fig-c4-prob-estimates**. Our final point estimates are taken as the median values $\tilde{p}_1 := \text{median}(\tilde{P}_1) = 1.4 \times 10^{-16}$ and $\tilde{p}_2 :=$

median$(\tilde{P}_2) = 1.3 \times 10^{-16}$, respectively, to two significant figures.

5.4.4 Improving performance using sparse empirical estimates

Unfortunately, it transpires that our method dramatically overestimated the true probabilities. In hindsight, this could have been anticipated in view of the simulation study in Section 5.1 in Kiriliouk and C. Zhou (2022), where the authors remark that failure regions of the the type $\mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}}$ are poorly summarised by the TPDM. This prompts us to investigate whether using empirical or sparse empirical estimates instead of a CP-based approach would have improved our performance. For the former, this simply involves replacing $A^{(l)}$ with the precomputed matrix $\hat{A}^{(l)}$ in (5.11). The approach based on sparse empirical estimates proceeds analogously except that we must revert to the $\alpha = 1$ setting and transform the data and thresholds accordingly. The resulting estimates for the joint exceedance probabilities are represented by the red and blue points in **?@fig-c4-prob-estimates**. The values are lower in every cluster and consequently the final estimates are closer to the true probabilities (purple points). In fact, using sparse empirical estimates yields $\hat{p}_1^\star = 5.1 \times 10^{-23}$ and $\hat{p}_2^\star = 5.0 \times 10^{-24}$, which are remarkably close to the correct solutions $p_1 = 8.4 \times 10^{-23}$ and $p_2 = 5.4 \times 10^{-25}$. Had we submitted these values would have significantly improved our ranking for this sub-challenge.

5.5 Conclusion

Our performance in the multivariate challenges demonstrates that the max-linear model provides a good framework for estimating tail event probabilities. By connecting this model with sparse simplex projections, one can achieve exceptional performance on both Challenges 3 and 4. Given these results, further research on the theoretical properties of the sparse empirical estimator $\hat{A}^\star$ is warranted. An obvious shortcoming of our Challenge 3 methodology is that it ignores the available covariate information. Presently, to the best of our knowledge, there is no way to incorporate covariates into the max-linear/TPDM framework. Our methods switch between $\alpha = 1$ and $\alpha = 2$ in a way that is somewhat cumbersome and unsatisfactory. This could have been avoided by employing a generalised version of the TPDM (Kiriliouk and C. Zhou 2022, Equation 4), thereby allowing us to fix $\alpha = 1$ throughout.

6 Bias-corrected estimation of the TPDM

6.1 Introduction and motivation

6.2 Regularised TPDM estimators

In high-dimensional settings, where the number of variables is comparable to the number of observations, the empirical covariance matrix is noisy and poorly conditioned. Regularisation may be used to obtain more stable and accurate estimates. Shrinkage estimation achieves this by shrinking the empirical estimate towards a biased but highly structured target (Ledoit and Wolf 2003). Thresholded estimators enforce sparsity by shrinking small values towards zero (Rothman et al. 2009). This improves stability and interpretability in cases where many variables are weakly correlated. We review these techniques before applying them for TPDM estimation.

6.2.1 Thresholding

6.2.2 Ledoit-Wolf shrinkage

6.3 Selecting the regularisation parameter

6.3.1 Frobenius risk minimisation

6.3.2 Why are standard approaches for covariance matrices not viable?

6.3.3 The asymptotically optimal Ledoit-Wolf shrinkage

6.4 Simulation experiments

6.4.1 Symmetric logistic

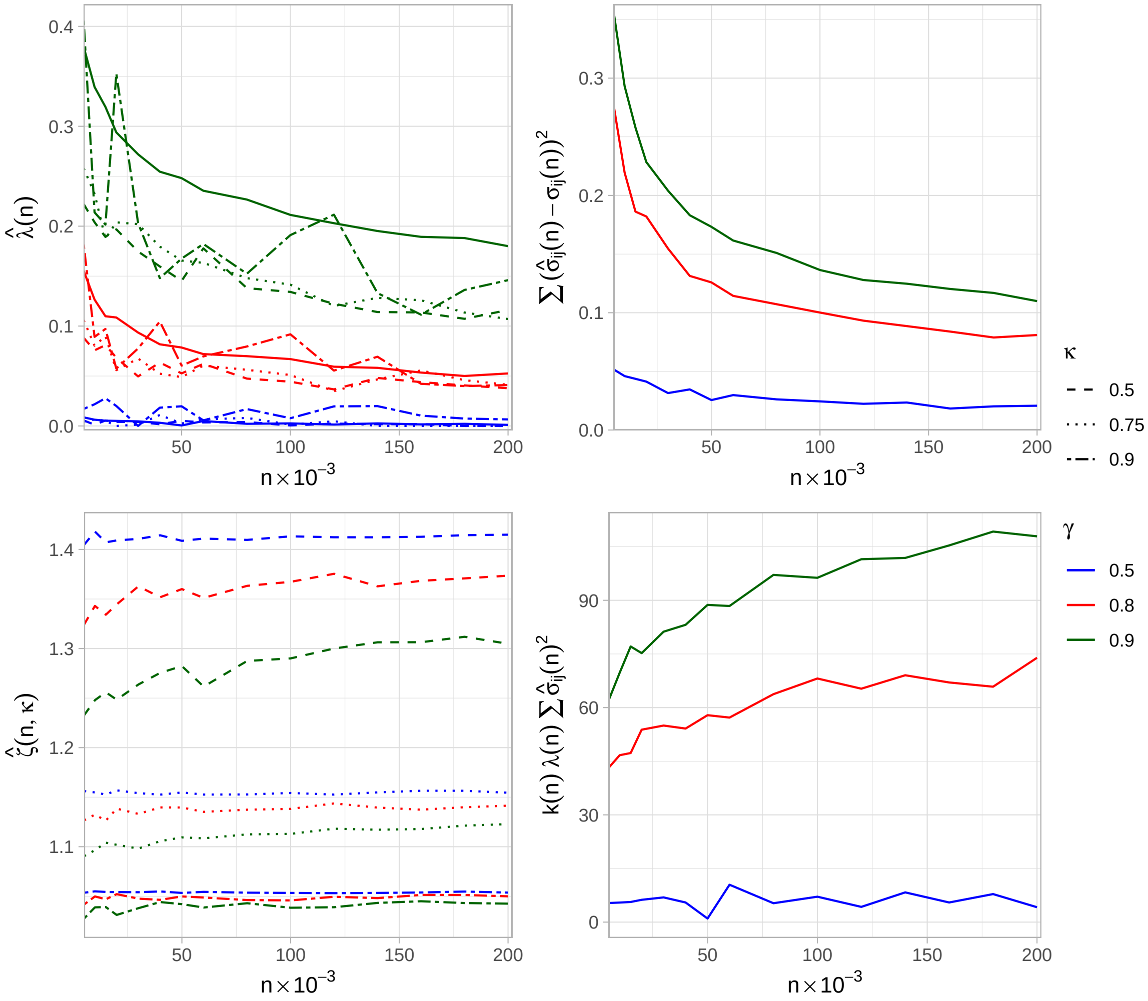

Figure 6.1: Blah

6.4.2 EVA (2023) Data Challenge 4

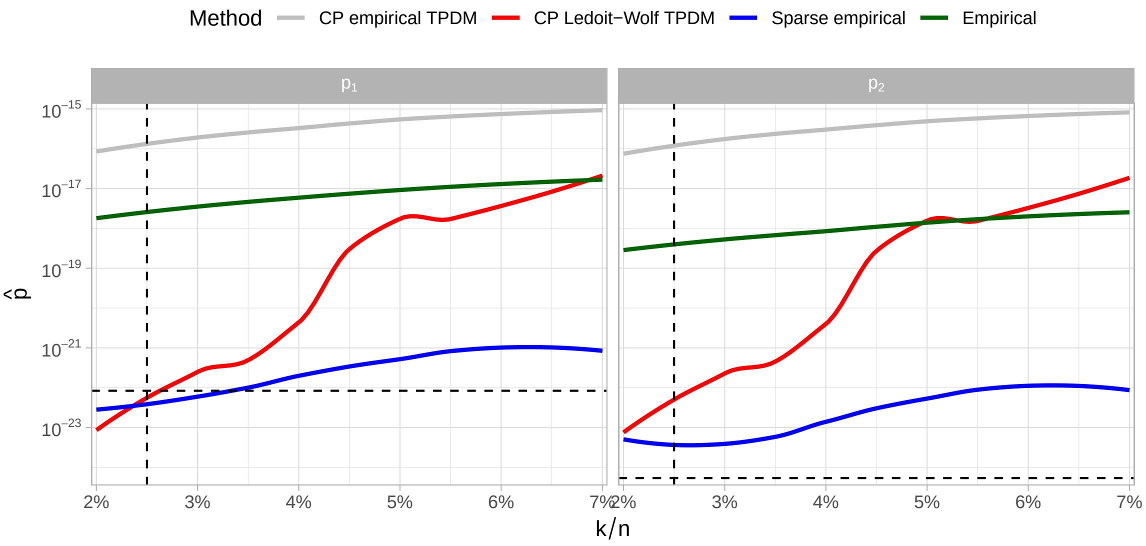

Figure 6.2: Blah

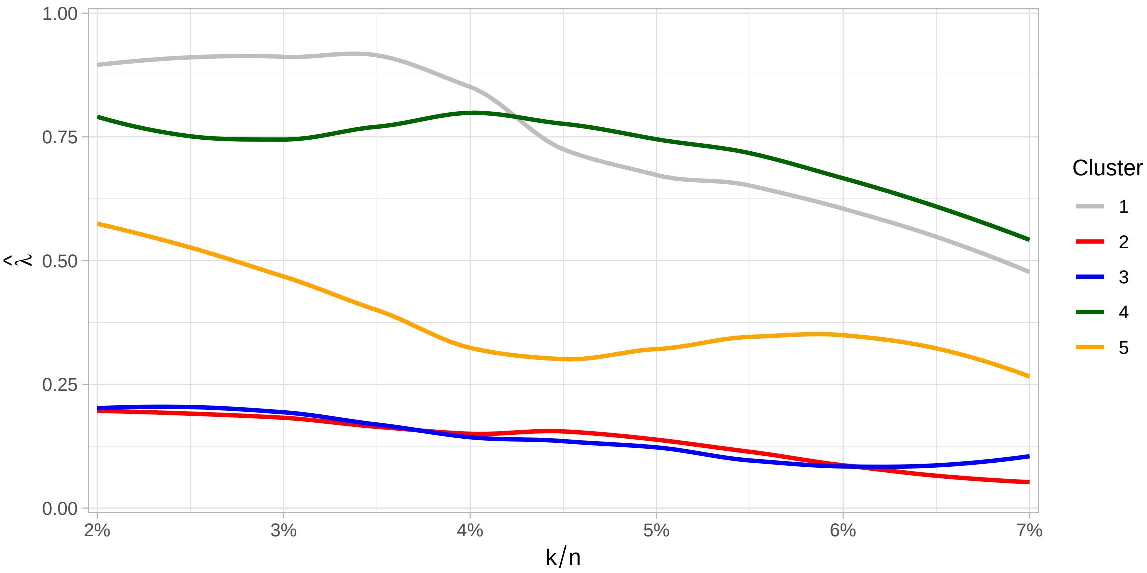

Figure 6.3: Blah

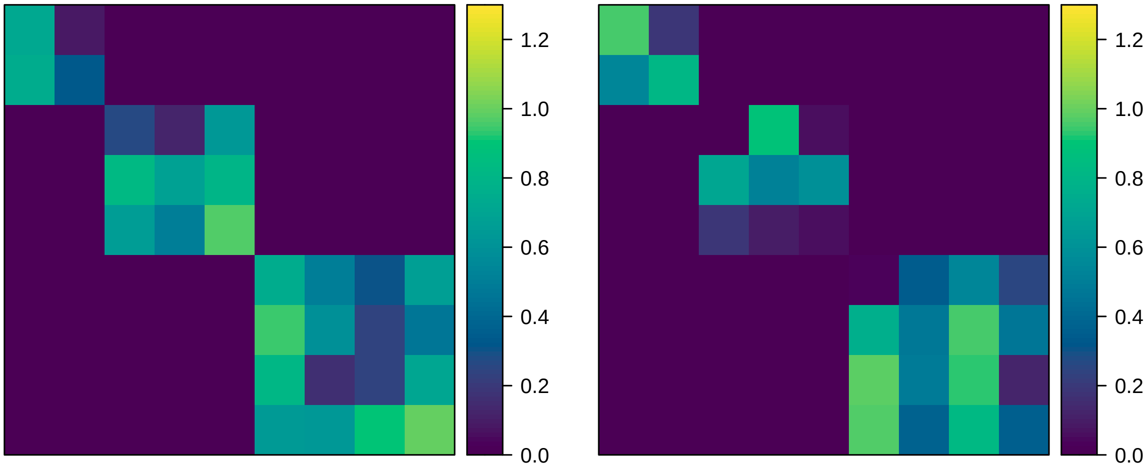

Figure 6.4: Blah

6.4.3 Extremal SAR model

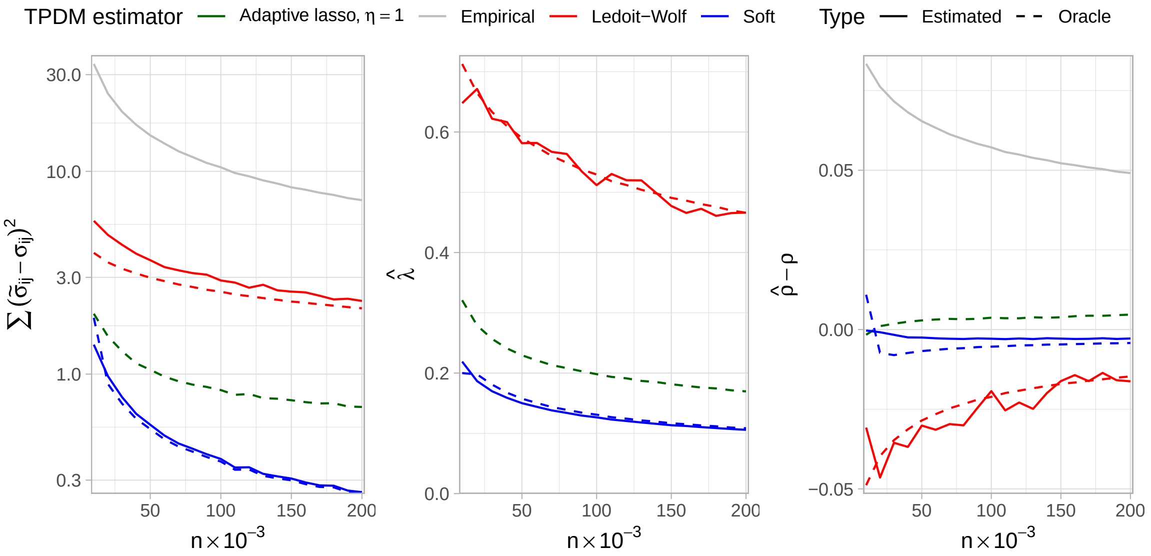

Figure 6.5: Blah

6.5 Conclusions and outlook

References

Aitchison, J. (1982). “The Statistical Analysis of Compositional Data”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 139–160.

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

– (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability.* Chapman and Hall.

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

Castro-Camilo, Daniela, Miguel De Carvalho, and Jennifer Wadsworth (2018). “Time-Varying Extreme Value Dependence with Application to Leading European Stock Markets”. In: *The Annals of Applied Statistics* 12.1.

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

Clémençon, Stephan, Nathan Huet, and Anne Sabourin (2024). “Regular Variation in Hilbert Spaces and Principal Component Analysis for Functional Extremes”. In: *Stochastic Processes and their Applications* 174, p. 104375.

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

Coles, Stuart G. and Jonathan A. Tawn (1991). “Modelling Extreme Multivariate Events”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 53.2, pp. 377–392.

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 (2023). “Statistical Inference on a Changing Extreme Value Dependence Structure”. In: *The Annals of Statistics* 51.4, pp. 1824–1849.

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.

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

Fisher, N. I., T. Lewis, and B. J. J. Embleton (1987). *Statistical Analysis of Spherical Data*. 1st ed. Cambridge University Press.

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.

Greenacre, Michael (2024). *The chiPower Transformation: A Valid Alternative to Logratio Transformations in Compositional Data Analysis*. URL: http://arxiv.org/abs/2211.06755 (visited on 07/09/2024). Pre-published.

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.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York.

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

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.

Jalalzai, Hamid, Stephan Clemencon, and Anne Sabourin (2018). "On Binary Classification in Extreme Regions". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18, pp. 3096–3104.

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.

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

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

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

Kiriliouk, Anna and 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.

Ledoit, Olivier and Michael Wolf (2003). "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection". In: *Journal of Empirical Finance* 10.5, pp. 603–621.

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.

Meyer, Nicolas and Olivier Wintenberger (2021). “Sparse Regular Variation”. In: *Advances in Applied Probability* 53.4, pp. 1115–1148.

Park, Junyoung et al. (2022). “Kernel Methods for Radial Transformed Compositional Data with Many Zeros”. In: *Proceedings of the 39th International Conference on Machine Learning*. Proceedings of Machine Learning Research, pp. 17458–17472.

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.

Pearson, Karl (1897). “Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs”. In: *Proceedings of the Royal Society of London* 60.359-367, pp. 489–498.

Poon, Ser-Huang, Michael Michael Rockinger, and Jonathan Tawn (2003). “Modelling Extreme-Value Dependence in International Stock Markets”. In: *Statistica Sinica* 13.4, pp. 929–953.

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.

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

Rothman, Adam J., Elizaveta Levina, and Ji Zhu (2009). “Generalized Thresholding of Large Covariance Matrices”. In: *Journal of the American Statistical Association* 104.485, pp. 177–186.

Seber, G. A. F. (1984). *Multivariate Observations*. 1st ed. Wiley Series in Probability and Statistics. Wiley.

Serrano, Javier Fernández (2022). *Semiparametric Bivariate Extreme-Value Copulas*. URL: http://arxiv.org/abs/2109.11307 (visited on 07/03/2024). preprint.

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.

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

Tsagris, Michail, Simon Preston, and Andrew T. A. Wood (2011). “A Data-Based Power Transformation for Compositional Data”. In: *Proceedings of the 4th International Workshop on Compositional Data Analysis*. CODAWORK 2011. Barcelona: CIMNE.

– (2016). “Improved Classification for Compositional Data Using the -Transformation”. In: *Journal of Classification* 33.2, pp. 243–261.

Wang, Huiwen et al. (2015). “Principal Component Analysis for Compositional Data Vectors”. In: *Computational Statistics* 30.4, pp. 1079–1096.

Zhou, Sha, Bofu Yu, and Yao Zhang (2023). “Global Concurrent Climate Extremes Exacerbated by Anthropogenic Climate Change”. In: *Science Advances* 9.10, eabo1638.