Extensions and Applications of the Tail Pairwise Dependence Matrix

Matthew Pawley

October 28, 2024

Table of contents

List of Figures

List of Tables

Preface

Draft thesis of Matthew Pawley, created on October 28, 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 Background & literature review

2.1 Univariate extreme value theory

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

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

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

The exact distribution of M_n is given by

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

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

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

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

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

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

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

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

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

2.1.2 Threshold exceedances and the generalised Pareto distribution (GPD)

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

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

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

2.1.3 Non-stationary extremes

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

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

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

2.2 Multivariate extreme value theory

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

2.2.1 Componentwise maxima

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

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

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

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

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

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

Applying Theorem 2.1 to the marginal components reveals that the margins of G follow a univariate GEV distribution. The crucial difference to the univariate setting is that now the limit (joint) distribution G does *not* admit a parametric representation. The inherently challenging nature of MEVT largely stem from this fact. The problem of estimating/modelling G is usually split into two (sequential) steps. First, one models the margins to describe the extreme behaviour of each variable individually (using univariate EVT). Then, one standardises to common margins and models the extremal dependence structure, i.e. the inter-relationships between extremes across multiple variables. Copula theory provides a rigorous justification for this two-step process.

2.2.2 Copulae and marginal standardisation

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

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

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

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

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

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

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

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

2.2.3 The exponent measure and angular measure

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

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

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

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

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

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

Here

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

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

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

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

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

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

2.2.4 Parametric multivariate extreme value models

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

There are several drawbacks to the parametric approach. Working with a parametric model instead of the general class runs the risk of model misspecification. Generating valid models is a challenging endeavour due to the moment constraints, resulting in models that are either overly simplistic or have unwieldy distribution functions and parameter constraints. Striking a balance between flexibility and parsimony becomes especially in high dimensions (i.e. when d is large). For these reasons, parametric models are not a primary focus of this thesis. Nevertheless, we now review a small selection of models. These primarily feature as data-generating processes for our numerical experiments. Functionality for generating independent observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ or $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n \sim H$ based on the sampling algorithms formulated in Dombry et al. (2016) is provided in the R package `mev`.

2.2.4.1 Logistic-type models

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2.4.2 The Brown-Resnick process and Hüsler-Reiss distribution

The Brown-Resnick process of Brown and Resnick (1977) is a class of stochastic processes commonly used to model the extremal dependence structure of spatial phenomena, including rainfall (A. C. Davison et al. 2012), snow depths (Schellander and Hell 2018) and wind gusts (Oesting et al. 2017). It is naturally defined through a transformation of a Gaussian process – a formal construction can be found in CITE Kabluchko et al. (2009). Let

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

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

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

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

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

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

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

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

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

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

2.2.4.3 The max-linear model

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

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

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

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

2.2.5 Multivariate regular variation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

finite measure H on the simplex

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

such that

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

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

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

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

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

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

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

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

The angular measure is unique only with respect to a pre-specified norm $\|\cdot\|$ and lies on the corresponding unit simplex (2.28). As mentioned previously, we exclusively choose the

L_p-norm

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

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

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

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

2.2.6 Extremal dependence measures

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

2.2.6.1 The tail dependence coefficient

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

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

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

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

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

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

$\beta' \supseteq \beta$ such that

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

For example, consider the angular measures

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

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 vertices of the simplex. This corresponds to full asymptotic independence, since then $\chi_\beta = 0$ for all $\beta \subseteq \{1, \dots, d\}$ with cardinality at least equal to two. The angular measure $H^{(2)}$ concentrates at a single point at the centre of the simplex. This implies that $\chi_{\{1,\dots,d\}} > 0$ and consequently $\chi_\beta > 0$ for all subsets β.

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

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

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

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

Example 2.2. Let $\boldsymbol{X} = (X_1, \dots, X_d)$ 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

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

where Φ is the standard normal distribution function. Variables X_i and X_j are asymptotically dependent for all $\lambda_{ij} > 0$, with asymptotic independence in the limit as $\lambda_{ij} \to \infty$. *Refer back to this equation when discussing Hazra and Bose changepoint method – it gives* *one-to-one relationship between HR parameter and dependence strength, so testing for change in λ or χ are equivalent.*

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

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

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

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

The tail dependence coefficients under these models are

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

corresponding to complete dependence and asymptotic dependence, as expected.

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

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

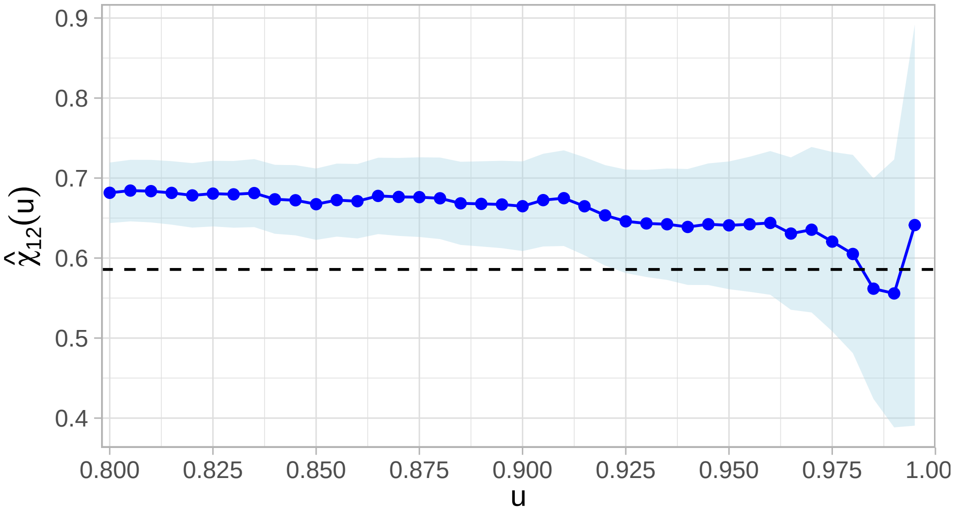

Figure 2.1: Empirical estimates $\hat{\chi}_{12}(u)$ of the tail dependence coefficient for bivariate symmetric logistic data with $\gamma = 0.5$ and $n = 5,000$ observations. The true coefficient $\chi_{12} = 2 - 2^{\gamma} \approx 0.59$ is marked by the dashed line. The shaded region represents the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

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

2.2.6.2 Extremal dependence measure

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

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

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

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

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

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

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

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

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

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

2.3 Inference

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

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

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

2.3.1 Framework and notation

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

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

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

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

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

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

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

2.3.2 Selecting the radial threshold or the number of exceedances

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

such that

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

These arise as sufficient conditions for proving various asymptotic properties (e.g. consistency, asymptotic normality) of estimators. The condition $k \to \infty$ ensures that the number of extremes – the effective sample size – grows arbitrarily large. The second condition $k/n \to 0$ requires that the proportion of threshold exceedances becomes vanishingly small, ensuring that inference is targeting the tail. In practice, n is fixed and selecting k requires striking a balance between these two aspects. Choosing k too small reduces the amount of available information and leads to unnecessarily high uncertainty. If k is too large, we risk using data that does not reflect the extremal dependence structure leading to bias. An appropriate choice depends on both the sample size and the underlying distribution of $\boldsymbol{X}$. If the convergence in (2.40) is rapid, then a low threshold may be adequate. Several threshold selection procedures have been proposed in univariate extremes (Section XX), but the literature on radial threshold selection is comparatively scant. By combining two sub-tests regarding (i) independence of the radial and angular components and (ii) regular variation of the radial component, Einmahl, Yang, et al. (2020) devise a formal procedure testing the validity of the MRV assumption. They suggest choosing the threshold by examining a plot of the sequence of p-values against k. The support-detection algorithm of Meyer and Wintenberger (2023) chooses k automatically via minimisation of a penalised log-likelihood. This procedure is specific to their setting and relies on additional technical assumptions. Most applied studies use a rule-of-thumb approach and/or produce a threshold stability plot checking the (in)sensitivity of some quantity to the choice of k – see Jiang et al. (2020), Szemkus and Friederichs (2024) and Russell and Hogan (2018) for examples.

2.3.3 The empirical angular measure

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

Definition 2.11. The empirical angular measure based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is the random measure on $\mathbb{S}_+^{d-1}$ defined as

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

Note that $\hat{H}$ does not enforce the moment constraints (2.12), so is not necessarily a valid angular measure. Einmahl and Segers (2009) construct an alternative non-parametric estimator that does enforce these restrictions, but it is limited to the bivariate setting. Proposition 3.3 in Janßen and Wan (2020) establishes consistency $\hat{H} \xrightarrow{p} H$ of the empirical angular measure provided the level k satisfies the rate conditions (2.43). Their result holds for general norms in arbitrary dimensions. 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 results hold with $\|\cdot\| = \|\cdot\|_p$ for $p \in [1,\infty]$. Since $\hat{H}$ is a discrete measure concentrating at k points, there exists a max-linear random vector $\boldsymbol{X}$ with parameter matrix

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

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

2.3.4 Non-parametric estimators

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

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

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

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

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

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

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

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

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

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

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

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

Then

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

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

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

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

Under the conditions of Theorem 2.3,

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

2.4 Tail pairwise dependence matrix (TPDM)

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

2.4.1 Definition and examples

Preamble.

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

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

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

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

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

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

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

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

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

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

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

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

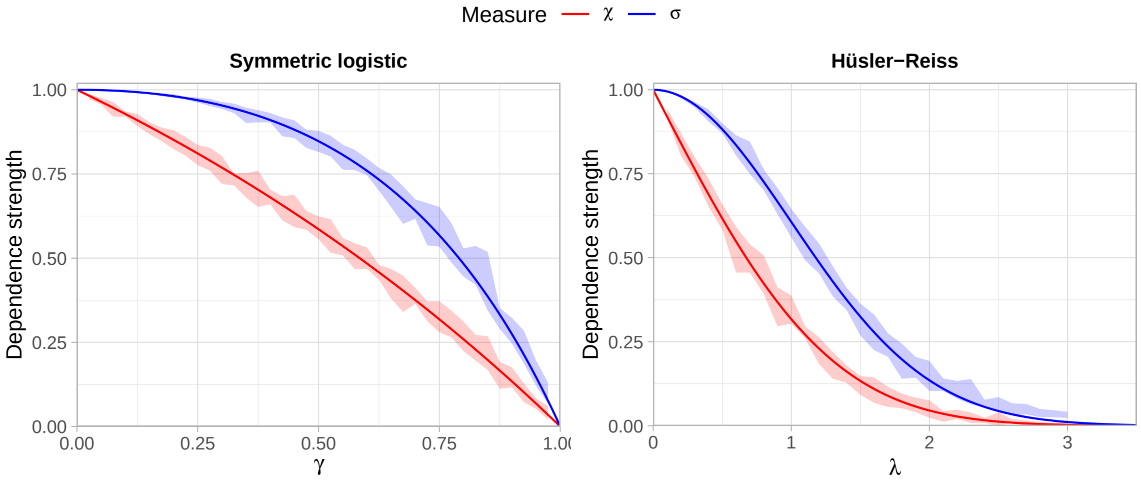

Figure 2.2: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate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

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

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

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

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

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

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

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

2.4.2 Interpretation of the TPDM entries

The definition of the TPDM

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

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

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

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

As discussed earlier, a well-defined notion of scale must fix either the sequence b_n or the mass of the angular measure in advance. In the above definition, the normalisation sequence is fixed and scaling information is contained in H. The scale is so-called because it yields information about the scale of the marginal distributions. Using (2.30), one can show that

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

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

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

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

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

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

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

and

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

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

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

2.4.3 Decompositions of the TPDM

The TPDM is useful as a summary statistic for quantifying pairwise dependencies, but what sets it apart from other pairwise dependence matrices (e.g. the TDM)? The TPDM admits two types of decomposition: eigendecomposition and the completely positive decomposition (Cooley and Thibaud 2019). These underpin the key statistical applications of the TPDM described in Section XX. The following results and proofs are reproduced from Kiriliouk and C. Zhou (2022).

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

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

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

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

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

□

By standard linear algebra results, the TPDM can be decomposed as $\Sigma = UDU^T$, where $D \in \mathbb{R}^{d\times d}$ is a diagonal matrix of eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ and $U \in \mathbb{R}^{d\times d}$ is an orthogonal matrix whose columns are the corresponding eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d \in \mathbb{R}^d$. The eigendecomposition potentially offers a low-rank representation of the TPDM expressed in terms of its eigenvalues and eigenvectors. In contrast, the completely positive decomposition represents the matrix as a product of a (potentially low-rank) non-negative matrix with its transpose.

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

Proposition 2.2. *The TPDM is completely positive.*

Proof. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with angular measure H and TPDM Σ. By Proposition 5 in Fougères et al. (2013), there exists a sequence of matrices $\{A_q \in \mathbb{R}_+^{d\times q} : q \geq 1\}$ such that $H_q \xrightarrow{v} H$, where H_q is the angular measure of the max-linear random vector $\boldsymbol{X}_q \in \mathcal{RV}_+^d(\alpha)$ parametrised by A_q. The TPDM of $\boldsymbol{X}_q$ is $\Sigma_q = A_q^{\alpha/2}(A_q^{\alpha/2})^T$ by Example 2.7. Thus, $\{\Sigma_q : q \geq 1\}$ is a sequence of completely positive matrices. The limit $\lim_{q\to\infty} \Sigma_q = \Sigma$ must also be completely positive because the set of completely positive matrices is closed (Haufmann 2011, Theorem 2.1.9).

□

In principle this provides a way to check whether a given matrix is a TPDM, but the membership problem for the completely positive cone is NP-hard (Dickinson and Gijben 2014). The following example illustrates how these two decompositions apply to the symmetric logistic model and hints towards their use for dimension reduction, to be formalised in Section XX.

Example 2.8. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is symmetric logistic with parameter $\gamma \in (0, 1]$. Then

$$\Sigma = (1 - \sigma)I_d + \sigma J_d,$$

where the constant σ depends on γ via the formula in Example 2.5. The eigenvalues of Σ are $\lambda_1 = 1+(d-1)\sigma$ and $\lambda_2 = \ldots = \lambda_d = 1-\sigma$. The principal eigenvector is $\boldsymbol{u}_1 = d^{-1/2}\mathbf{1}_d$ and the remaining eigenvectors $\boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ are orthogonal to $\boldsymbol{u}_1$. Rewriting the TPDM as

$$\Sigma = \sum_{i=1}^{d}(1-\sigma)\boldsymbol{e}_i\boldsymbol{e}_i^T + \sigma\mathbf{1}_d\mathbf{1}_d^T,$$

and using Example 2.7, the TPDM of $\boldsymbol{X}$ is identical to that of a max-linear random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d) \in \mathcal{RV}_+^d(\alpha)$ with parameter matrix

$$A = \begin{pmatrix} (1-\sigma)^{1/\alpha} & 0 & 0 & \cdots & 0 & \sigma^{1/\alpha} \\ 0 & (1-\sigma)^{1/\alpha} & 0 & \cdots & 0 & \sigma^{1/\alpha} \\ 0 & 0 & (1-\sigma)^{1/\alpha} & \cdots & 0 & \sigma^{1/\alpha} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & (1-\sigma)^{1/\alpha} & \sigma^{1/\alpha} \end{pmatrix}.$$

Consider the limiting case of complete dependence as $\gamma \to 0$, whereby the angular measure tends towards $H^{(2)}$ from (2.36) in the limit. The eigenvalues are $\lambda_1 \to d$, $\lambda_2, \ldots, \lambda_d \to 0$, indicating a single eigenvector $\boldsymbol{u}_1$ is sufficient to fully 'explain' the dependence structure. We also have $A \to (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{1}_d) = \mathbf{1}_d = A^{(2)}$. (This is an abuse of notation; we simply mean that zero columns have no effect and may be omitted.) Both perspectives point towards a low-rank representation of the dependence structure involving the vector directed towards the centre of the simplex. Indeed, this perfectly describes $H^{(2)} = d\delta_{\mathbf{1}_d/\|\mathbf{1}_d\|_\alpha}$.

2.4.4 The empirical TPDM

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

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

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

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

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

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

□

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

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

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

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

□

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

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

Proof. See Example 2.4.

□

If X_i and X_j are asymptotically independent ($\sigma_{ij} = 0$), then $\nu_{ij}^2 = 0$ and the limit distribution is degenerate. In this case, the above result only proves consistency, i.e. $\hat{\sigma}_{ij} \to 0$, and cannot be used to formally test for asymptotic independence (Lehtomaa and Resnick 2020).

Using asymptotic normality one may construct asymptotic confidence intervals

$$\lim_{n\to\infty} \mathbb{P}\left[|\sigma_{ij} - \hat{\sigma}_{ij}| < z_{\beta/2}\sqrt{\nu_{ij}^2/k}\right] = 1 - \beta, \tag{2.59}$$

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

Example 2.9. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is symmetric logistic with $\gamma = 0.6$. Using Example 2.5 and results in Appendix XX, $\sigma_{ij} \approx 0.760$ and $\nu_{ij}^2 \approx 0.065$ for all $i \neq j$. For sufficiently large n,

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

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

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

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

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

$$\boldsymbol{\sigma} := \mathrm{vecu}(\Sigma) := (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{1d}, \sigma_{23}, \ldots, \sigma_{2d}, \ldots, \sigma_{d-1,d}), \tag{2.60}$$

$$\hat{\boldsymbol{\sigma}} := \mathrm{vecu}(\hat{\Sigma}) := (\hat{\sigma}_{12}, \hat{\sigma}_{13}, \ldots, \hat{\sigma}_{1d}, \hat{\sigma}_{23}, \ldots, \hat{\sigma}_{2d}, \ldots, \hat{\sigma}_{d-1,d}). \tag{2.61}$$

Each vector contains

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

entries. This is justified because the matrices are symmetric and their diagonal entries are irrelevant. Components are indexed according to the sub-indices of the corresponding matrix entry, e.g. the first entry of $\boldsymbol{\sigma}$ is σ_{12} rather than σ_1.

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

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

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

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

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

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

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

Example 2.10. Suppose $\boldsymbol{X} = (X_1, \ldots, X_4) \in \mathcal{RV}_+^4(1)$ is max-linear with (randomly generated) parameter matrix $A \in \mathbb{R}_+^{4\times 12}$ as shown in Figure 2.3 (top). The TPDM $\Sigma = A^{1/2}(A^{1/2})^T$ is visualised in the bottom-left plot, with each cell's colour intensity representing the magnitude of the corresponding entry of Σ. All pairs of components exhibit strong dependence. The matrix in the bottom-right is the asymptotic covariance matrix V of $\hat{\boldsymbol{\sigma}}$, derived in Appendix XX. It has $\binom{4}{2} = 6$ rows and columns. *Any comments about the matrix itself?* We now run simulations verifying/illustrating Proposition 2.6 for this example. We generate $n = 10^4$ independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z}$ (see eq-max-linear-X) and compute the empirical TPDM using $k = \sqrt{n} = 100$ extremes. Repeating this process, we obtain 1,000 independent realisations of $\hat{\Sigma}$. After row-wise vectorisation, these estimates should be approximately $N(\boldsymbol{\sigma}, k^{-1}V)$ distributed. Figure 2.4 examines whether this is the case. First consider the diagonal panels. These show that the density function of an $N(\sigma_{ij}, \nu_{ij}^2/k)$ random variable (blue curve) provides a good fit for the empirical distribution of $\hat{\sigma}_{ij}$ (red histogram). Now consider the scatter plots in the lower triangular portion of the plot. The grey points represent 1,000 realisations of $(\hat{\sigma}_{ij}, \hat{\sigma}_{lm})$. The blue ellipses are the true asymptotic 95% data ellipses centred at $(\sigma_{ij}, \sigma_{lm})$ (blue crosses). Their orientation relates to the association $\rho_{ij,lm}$ between $\hat{\sigma}_{ij}$ and $\hat{\sigma}_{lm}$, while the lengths of the major and minor axes are dictated by the asymptotic variances ν_{ij}^2, ν_{lm}^2. The red ellipses and crosses are defined analogously but estimated from the data. They are generally in close agreement. The upper-triangular panels list the values of $\rho_{ij,lm}$ (blue) alongside empirical estimates (red) based on the sample covariance between $\hat{\sigma}_{ij}$ and $\hat{\sigma}_{lm}$.

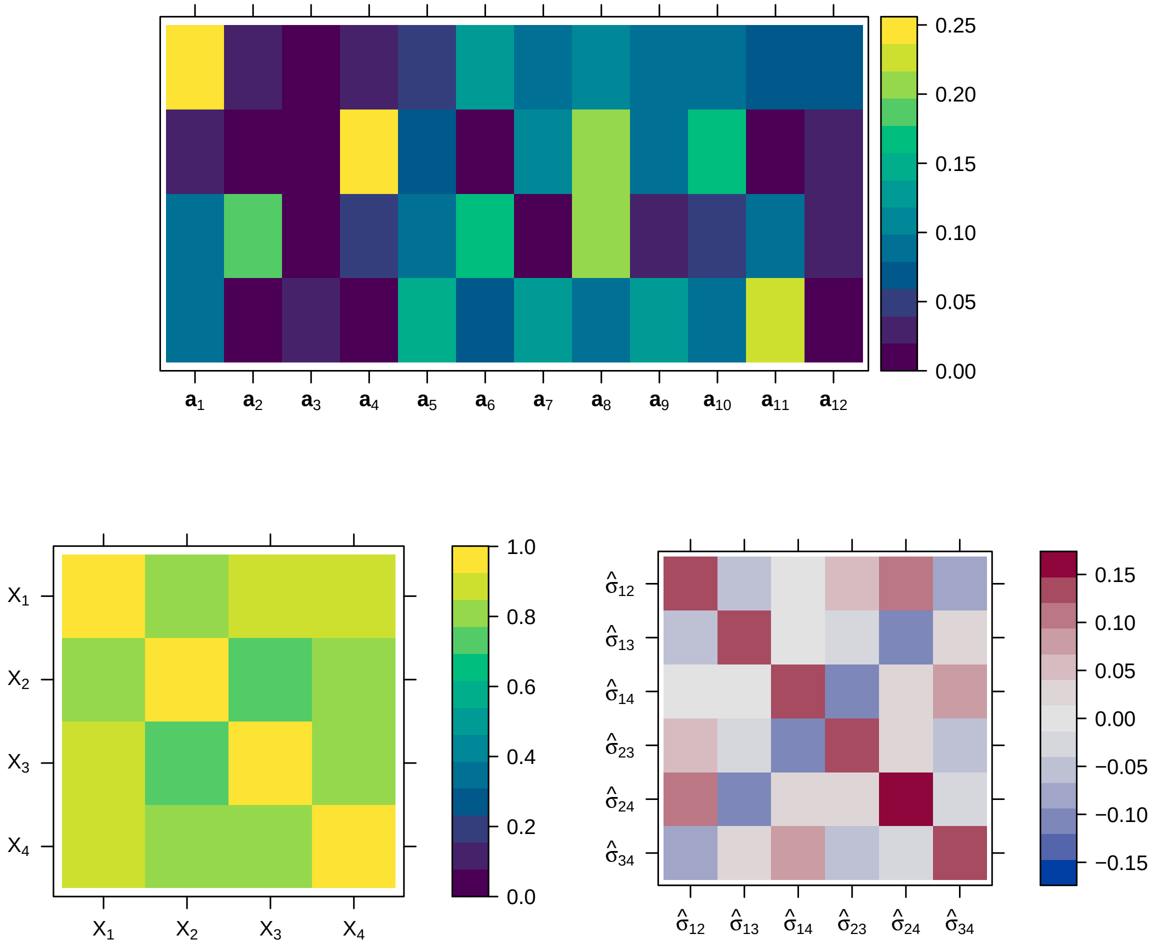

Figure 2.3: Visual representation of the matrices discussed in Example 2.10. Top: a randomly generated max-linear parameter matrix A with $d = 4$ and $q = 12$. Bottom left: the TPDM Σ of $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z}$. Bottom right: the asymptotic covariance matrix V of $\hat{\boldsymbol{\sigma}}$.

2.5 Existing applications and extensions of the TPDM

The general objective of this thesis is to develop novel statistical tools for analysing extremal dependence based on the TPDM. Before presenting these, we acquaint the reader with existing TPDM-based methods, selected according to their relevance to the thesis. Our survey divides the related literature into two main categories: principal components analysis (PCA) and inference for the max-linear model. Clustering features occasionally (e.g. in Chapter XX), but does not constitute an essential pillar of our research; a brief

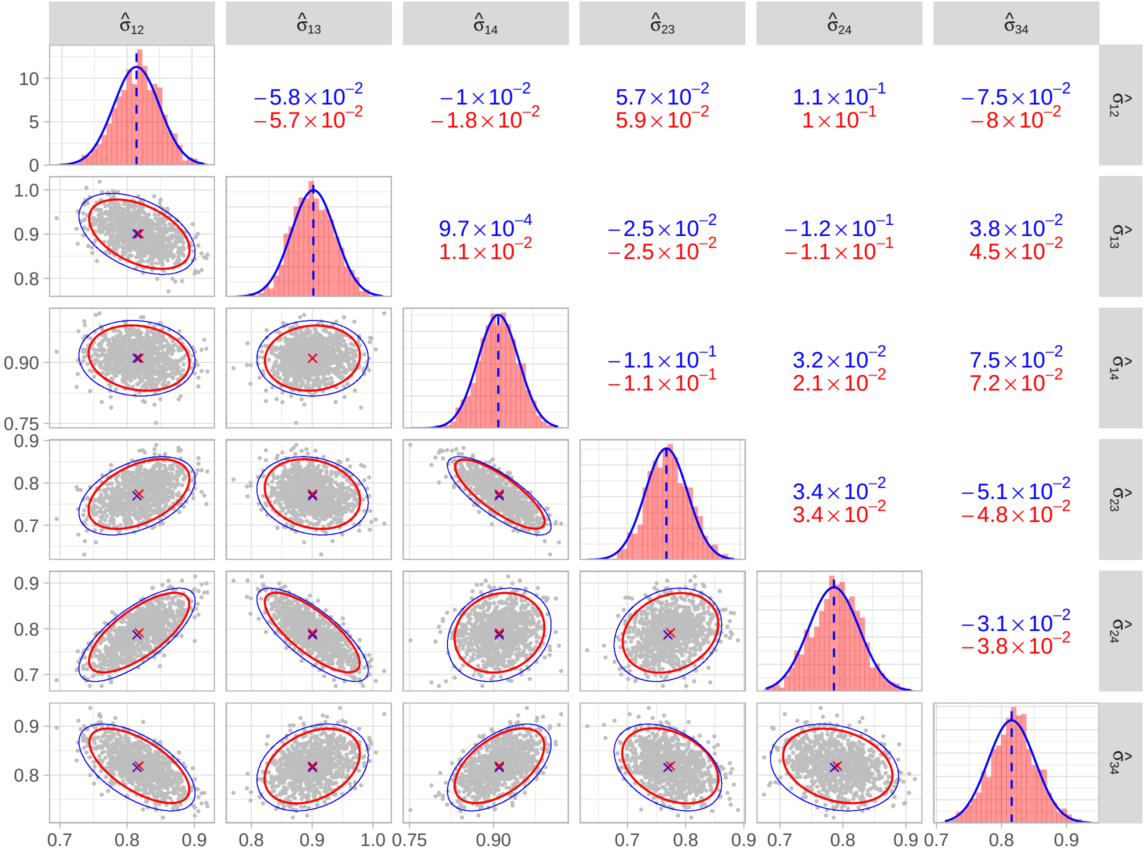

Figure 2.4: Pairs plot illustrating asymptotic normality of the empirical TPDM – see Example 2.10 for details. All panels: red represents the empirical quantity based on the 1,000 repeated simulations; blue represents the theoretical quantity based on asymptotic normality. Diagonal panels: the distribution (histogram or density function) of $\hat{\sigma}_{ij}$. Lower triangular panels: pairwise scatter plots of $(\hat{\sigma}_{ij}, \hat{\sigma}_{lm})$ (grey points) along with the mean (crosses) and the 95% data ellipse. Upper triangular panels: the entries $v_{ij,lm}$ of V.

overview of TPDM-based clustering algorithms (Fomichov and Ivanovs 2023; Richards et al. 2024) is contained in Appendix XX. Further interesting topics that are not covered include time series (Mhatre and Cooley 2021; Wixson and Cooley 2023) and graphical models (Gong et al. 2024; Lee and Cooley 2023). Throughout this section, $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ is a random vector on α-Fréchet margins with angular measure H with respect to $\|\cdot\| = \|\cdot\|_\alpha$ and $b_n = n^{1/\alpha}$, while $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are independent copies of $\boldsymbol{X}$.

2.5.1 Principal component analysis (PCA) for extremes

In classical multivariate statistics, principal component analysis (PCA) is the flagship method for reducing the dimension of a random vector by finding linear subspaces that minimise the distance between the data and its low-dimensional projections (Blanchard et al. 2007; Jolliffe 2002). The central idea is to transform the original set of correlated variables into a new set of uncorrelated variables – the principal components – which are ordered so that the first few capture most of the variability in the data. Computing these variables boils down to computing the eigendecomposition of a symmetric, positive semi-definite matrix. A more detailed review of the theory of PCA is given in Appendix XX.

In multivariate extremes, it is often assumed that the angular measure has a low-dimensional structure (Engelke and Jevgenijs Ivanovs 2021). For example, weather extremes typically exhibit spatial patterns related to geographical or topographical drivers, e.g. north/south, coastal/inland, high-lying/low-lying (Bernard et al. 2013; Jiang et al. 2020). These patterns permit a description of a process' extremal behaviour in terms of a smaller number of variables. This is the key objective of PCA for extremes.

Classical (non-extreme) PCA is not appropriate for this task for several reasons. The original variables $X_1, \dots, X_d$ are usually heavy-tailed, so the requirement on the existence second-order moments may be violated. (The variance of an α-regularly varying random variable is infinite if $\alpha < 2$.) Standard PCA reveals relationships between variables in the centre rather than the tail of the joint distribution, because it arises from the covariance matrix. Moreover, it captures dependence in both directions around the origin/mean, whereas we focus on a particular direction of interest. Finally, standard PCA fails to capitalise on the probabilistic structure inherent to MRV random vectors. The heavy-tailed, univariate radial component accounts for most of the variability in the data, but it is (asymptotically) independent of the angular component that actually contains the relevant information about the association between the variables. This suggests performing dimension reduction on the (empirical) angular measure via eigendecomposition of the (empirical) TPDM (Cooley and Thibaud 2019; Drees and Sabourin 2021).

Drees and Sabourin (2021) adopt a risk minimisation perspective aiming to minimise the mean-squared reconstruction error of $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$ with respect to the limit distribution H. They define the (asymptotic) risk of a subspace $\mathcal{S} \subset \mathbb{R}^d$ as

$$R(\mathcal{S}) = \mathbb{E}_H[\|\boldsymbol{\Theta} - \Pi_{\mathcal{S}}\boldsymbol{\Theta}\|_2^2],$$

where $\Pi_{\mathcal{S}}$ denotes orthogonal projection onto $\mathcal{S}$. The true risk cannot be minimised directly because H is unknown. Instead, they minimise the empirical risk

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

This is justified because, above a sufficiently high threshold, the extremal angles will lie in a neighbourhood of the target subspace. Let $\mathcal{V}_p$ denote the class of all linear subspaces of dimension $1 \leq p \leq d$ in $\mathbb{R}^d$. Minimisers of $\hat{R}$ are computed via eigendecomposition of the empirical TPDM. Let $(\hat{\boldsymbol{u}}_j, \hat{\lambda}_j)$ denote the (ordered) eigenpairs of $\hat{\Sigma}$ for $j = 1, \dots, d$. Then $\hat{\mathcal{S}}_p := \operatorname{span}\{\hat{\boldsymbol{u}}_1, \dots, \hat{\boldsymbol{u}}_d\}$ minimises R in $\mathcal{V}_p$ and $\hat{R}(\mathcal{S}_p) = \sum_{j>p} \hat{\lambda}_j$ (Drees and Sabourin 2021, Lemma 2.1). If the data exhibit a low-dimensional (linear) structure, then one can find $p \ll d$ such that the risk is acceptable small. It is recommended to plot $\hat{R}(\mathcal{S}_p)$ against p when choosing the number of principal components to retain. In terms of theoretical statistical guarantees, they prove that the learnt subspace converges to the optimal one as the sample size increases to infinity (Drees and Sabourin 2021, Theorem 2.4). Suppose there exists $p^\star < d$ and a linear subspace $\mathcal{S}^\star \in \mathcal{V}_{p^\star}$ such that $R(\mathcal{S}^\star) = 0$ and $R(\mathcal{S}) > 0$ for any $\mathcal{S} \in \cup_{p>p^\star}\mathcal{V}_p$. Then, provided $k(n)$ satisfies the rate conditions (2.43), $\hat{\mathcal{S}}_{p^\star} \to \mathcal{S}^\star$ in the sense that

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

Treating the angles $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$ as points in $\mathbb{R}^d$ rather than $\mathbb{S}_{+(\alpha)}^{d-1}$ simplifies the derivation of theoretical guarantees but creates interpretability issues. The rank-p reconstructions of the angles do not lie in the simplex, in general. Shifting/normalising the reconstructed vectors corrects this, but optimality properties will be not be preserved. One may also question the appropriateness of the Euclidean norm as a measure for the angular reconstruction error. In the context of clustering, Janßen and Wan (2020) argue that angular distances (e.g. the cosine dissimilarity) are a more natural choice. On a similar note, their working hypothesis that the low-dimensional structure of H is linear in $\mathbb{R}^d$ is restrictive. Avella-Medina et al. (2022) develop a kernel PCA method for extracting non-linear patterns. In Chapter XX, we propose our own PCA method, inspired by compositional PCA, that addresses all of these concerns: reconstructions are in $\mathbb{S}^{d-1}_{+(\alpha)}$, errors are defined using a simplicial metric, and non-linearity (curvature) in the data is captured.

Cooley and Thibaud (2019) propose an alternative approach based on the so-called transformed-linear inner product space on $\mathbb{R}^d_+$, the sample space of $\boldsymbol{X}$. It is grounded on the softplus transformation

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

This transformation is bijective with inverse function $\tau^{-1}(y) = \log[\exp(y) - 1]$ and, crucially, it is tail-preserving, i.e. $\lim_{x\to 1} \tau(x)/x = 1$. The role of τ is to provide a pathway between $\mathbb{R}^d$ and $\mathbb{R}^d_+$ that doesn't disturb the tails. The inner product space is constructed as follows. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d_+$ and $\alpha \in \mathbb{R}$, they define operations

$$\boldsymbol{x} \oplus \boldsymbol{y} = \tau[\tau^{-1}(\boldsymbol{x}) + \tau^{-1}(\boldsymbol{y})], \qquad \alpha \odot \boldsymbol{x} = \tau[a\tau^{-1}(\boldsymbol{x})].$$

and an inner product and norm

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_\tau = \left\langle \tau^{-1}(\boldsymbol{x}), \tau^{-1}(\boldsymbol{y}) \right\rangle, \qquad \|\boldsymbol{x}\|_\tau = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_\tau^{1/2} = \|\tau^{-1}(\boldsymbol{x})\|_2.$$

The PCA procedure may then be formulated in the transformed-linear space $\mathcal{H} = \mathbb{R}^d_+$ or in $\mathbb{R}^d$ under the transform/back-transform approach (see Appendix XX). As in Drees and Sabourin (2021), let $(\hat{\boldsymbol{u}}_j, \hat{\lambda}_j)$ be the ordered eigenpairs (in $\mathbb{R}^d$) of $\hat{\Sigma}$. Then $\{\hat{\boldsymbol{\omega}}_1, \dots, \hat{\boldsymbol{\omega}}_d\} = \{\tau(\hat{\boldsymbol{u}}_1), \dots, \tau(\hat{\boldsymbol{u}}_d)\}$ forms an orthonormal basis of $\mathbb{R}^d_+$. In this new basis, the random vector $\boldsymbol{X}$ may be decomposed as

$$\boldsymbol{X} = \bigoplus_{j=1}^{d} (\hat{V}_j \odot \hat{\boldsymbol{\omega}}_j) = \tau\left(\sum_{j=1}^{d} \hat{V}_j \hat{\boldsymbol{u}}_j\right), \tag{2.62}$$

where $\hat{V}_j = \langle \boldsymbol{X}, \hat{\boldsymbol{\omega}}_j \rangle_\tau$ for $j = 1, \dots, d$. Truncating the expansion (2.62) yields low-rank reconstructions of $\boldsymbol{X}$. The random variables $\hat{V}_1, \dots, \hat{V}_d$ are called the extremal principal components of $\boldsymbol{X}$. The MRV $\mathbb{R}^d$-valued random vector $\hat{\boldsymbol{V}}$ has the same dimensions as $\boldsymbol{X}$, but its components are ordered according to their contribution to the extremal behaviour of $\boldsymbol{X}$ in the sense that (Cooley and Thibaud 2019, Proposition 6)

$$\text{scale}(|\hat{V}_i|) = \hat{\lambda}_i^{1/\alpha}, \qquad (i = 1, \ldots, d),$$

Thus, the ith eigenvector $\hat{\boldsymbol{\omega}}_i$ represents the direction of maximum scale after accounting for information contained in the previous eigenvectors $\{\hat{\boldsymbol{\omega}}_j : j < i\}$.

Visualising/examining the TPDM eigenvectors a powerful tool for gaining insight into the extremal dependence structure. In a study of precipitation extremes in the United States, Jiang et al. (2020) relate the leading eigenvectors to the El-Niño Southern Oscillation (ENSO), a cyclical phenomenon that is known to be a key climatological driver. Low-rank reconstructions of Hurricane Floyd broadly capture the large-scale structure, but recreating localised features requires a large number of components. Russell and Hogan (2018) compare covariance matrix eigenvectors against TPDM eigenvectors to characterise performance differences between typical and elite-level National Football League (NFL) performers across a battery of physical tests. Szemkus and Friederichs (2024) apply PCA to the cross-TPDM, an extension to the TPDM that is analogous to the cross-covariance matrix, to analyse the dynamics of compound extreme weather events. For event detection and attribution purposes, they devise indices quantifying whether particular patterns of interest – those signified by the cross-TPDM's singular vectors – are highly pronounced. Rohrbeck and Cooley (2023) move beyond exploratory analysis and demonstrate how the framework can be used to generate synthetic extreme events. Hazard event sets are widely used in catastrophe modelling to assess exposure to extreme events (CITE). Their sampling algorithm exploits the fact that the leading components of $\hat{\boldsymbol{V}}$ account for a significant proportion of the extremal behaviour of $\boldsymbol{X}$. Roughly speaking, dependence between $\hat{V}_1, \ldots, \hat{V}_p$ is captured with a flexible model and a simple model is used to account for the remaining, relatively unimportant components. They use this model to generate samples of $\hat{\boldsymbol{V}}$, from which samples of $\boldsymbol{X}$ are produced via (2.62).

Results based on Cooley and Thibaud (2019) require accurate estimation of the TPDM so that the empirical eigenvectors reflect the true eigenvectors. However, in weak-dependence scenarios the empirical TPDM suffers from a positive bias (Section XX). This is problematic when the spatial extent of the study region is large relative to that of the modelled phenomenon. Jiang et al. (2020) ameliorate this using a 'pairwise-thresholded' estimator instead of (2.56), defined as

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

where $R_l^{ij} = \|(X_{li}, X_{lj})\|$ and $R_{(k+1)}^{ij}$ is the $(k+1)$th upper order statistic of $\{R_l^{ij} : l = 1, \ldots, n\}$. However, the resulting estimator $\hat{\Sigma}^{(p)}$ is not positive semi-definite. This may be resolved by projecting $\hat{\Sigma}^{(p)}$ onto the space of correlation matrices (Higham 2002), but this ad-hoc step does not address the fundamental problem. Partly motivated by this, Chapter XX proposes an improved estimator that is positive semi-definite.

2.5.2 Inference for the max-linear model

Estimating the parameter matrix A) of the max-linear model is a challenging task. The lack of an angular density function precludes the use of standard maximum likelihood procedures. Einmahl, Kiriliouk, et al. (2018) propose a procedure that minimises a weighted least-squares distance to some initial (non-parametric) estimator. Their procedure becomes computationally intensive when q is large. Janßen and Wan (2020) and Medina et al. (2021) cluster the angles of extreme observations and identify the normalised columns of A with the q cluster centres. The minimum-distance and clustering approaches assume q is fixed; Kiriliouk (2020) present a hypothesis test to assist with choosing q.

Recently, the TPDM has emerged as a promising tool for inference for the max-linear model (Fix et al. 2021; Kiriliouk and C. Zhou 2022). Recall from Example 2.7 that the TPDM of a max-linear random vector $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ is $\Sigma = \hat{A}^{\alpha/2}(\hat{A}^{\alpha/2})^T$. Now consider Proposition 2.2, which says that the TPDM is completely positive (Definition 2.15). Based on this connection, originally observed by Cooley and Thibaud (2019), any matrix belonging to the set

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

may be considered a reasonable estimate for A, in the sense that the pairwise dependencies of the fitted model conform with those implied by $\hat{\Sigma}$. The set $\mathcal{CP}(\hat{\Sigma})$ is in one-to-one correspondence with the set of completely positive (CP) factors of $\hat{\Sigma}$. We call $\hat{A} \in \mathcal{CP}(\hat{\Sigma})$

a CP-estimate of A. In general, a completely positive matrix may have many CP factorisations (Shaked-Monderer 2020). Among these, the simplest CP-estimate is the empirical estimate $\hat{A} \in \mathbb{R}_+^{d\times k}$ as defined in (2.45). Cooley and Thibaud (2019) describe the empirical estimate as 'naive' because it probably contains more columns than necessary. Kiriliouk and C. Zhou (2022) provide an algorithm for efficiently factorising $\hat{\Sigma}$ to obtain further. The performance of their CP-estimation procedure is assessed in simulation studies by computing tail event probability estimates under the true and fitted models using (2.23). The fitted models capture the dependence structure reasonably well, except for certain classes of failure regions. This is partly attributed to estimation error in the TPDM.

Fix et al. (2021) analyse the effect of TPDM estimation error for max-linear model fitting in more detail. Focussing on spatial extremes, they define the extremal spatial autoregressive (SAR) model, a special case of the max-linear model where $A = A(\rho)$ is determined by a single dependence parameter $\rho \in (0, 1/4)$. The model parameter ρ is estimated by minimising the discrepancy between $\hat{\Sigma}$ and the theoretical TPDM $\Sigma(\rho) := A(\rho)A(\rho)^T$ (assuming $\alpha = 2$):

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

They find that $\hat{\rho}$ has a positive bias when ρ is small (weak dependence). The proximate cause is that $\hat{\Sigma}$ overestimates weak dependencies, biasing the fitted model. This fundamental problem, and their proposed remedy, is the subject the following section.

2.6 Bias in the empirical TPDM in weak-dependence scenarios

The empirical TPDM is consistent and asymptotically unbiased (Proposition 2.6). This provides a guarantee that, with sufficient data, the empirical TPDM reflects the true pairwise dependence structure. The associated rate of convergence is $\mathcal{O}(k^{-1/2})$, where $k = k(n)$ represents the number of extreme observations and satisfies the rate conditions (2.43). However, in real-world applications, data are limited and extreme observations are scarce. For example, commonly available climate records typically span approximately 50 years (Boulaguiem et al. 2022). A study of summer heatwaves might then be based on, say, $n \approx 50 \times 100 = 5,000$ daily observations. The second condition in (2.43) requires that the effective sample size is some small fraction of n, resulting in a very limited number of extreme data points. Asymptotic guarantees are therefore of limited value for the sample sizes available in practice. This motivates an analysis of the empirical TPDM's finite-sample performance. As alluded to in the previous section, it will transpire that the TPDM is biased in scenarios where tail dependence is weak (Cooley and Thibaud 2019; Fix et al. 2021; Mhatre and Cooley 2021), herein referred to as the '(weak dependence) bias issue'. Chapter XX proposes bias-corrected estimators with superior finite-sample performance, but the bias issue will arise at various points in the preceding chapters, so we choose to highlight it now.

2.6.1 Bias in the TPDM and threshold-based estimators

The bias issue is not exclusive to the empirical TPDM. In fact, it applies more generally to threshold-based estimators in multivariate extremes. For example, Raphaël Huser et al. (2016) conduct simulation studies examining the finite-sample performance of estimators of γ, the dependence parameter of the symmetric logistic model. The results show that block-maxima based estimators have a small bias but very high variability. On the other hand, the estimator $\hat{\gamma}$ based on threshold exceedances tends to overestimate the dependence strength, that is $\text{Bias}(\hat{\gamma}) = \mathbb{E}[\hat{\gamma}] - \gamma < 0$. This discrepancy increases as dependence weakens; see the second column of Figure 3 in Raphaël Huser et al. (2016). Problems of a similar nature can be found across the multivariate extremes literature, for example in spatial modelling (Boulaguiem et al. 2022, Figure 6c) and lower-tail dependence modelling (Dobrić and Schmid 2005).

The empirical TPDM suffers from the same issue when dependence is weak. This phenomenon is illustrated in Figure 2.5. Like in Figure 2.2, the blue lines represent the true dependence strength for a given model parameter and the shaded regions indicate the minimum/maximum over a set of ten empirical estimates. The only difference is that here the underlying sample size is $n = 5 \times 10^3$, whereas in Figure 2.2 it was $n = 5 \times 10^5$. In both plots, the tuning parameter was set as $k = \sqrt{n}$. The plot definitively shows that the empirical TPDM overestimates the dependence as γ and λ approach their upper limits, or, equivalently, as $\sigma \to 0$. *Should I keep χ in the plot?* This can be summarised as

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

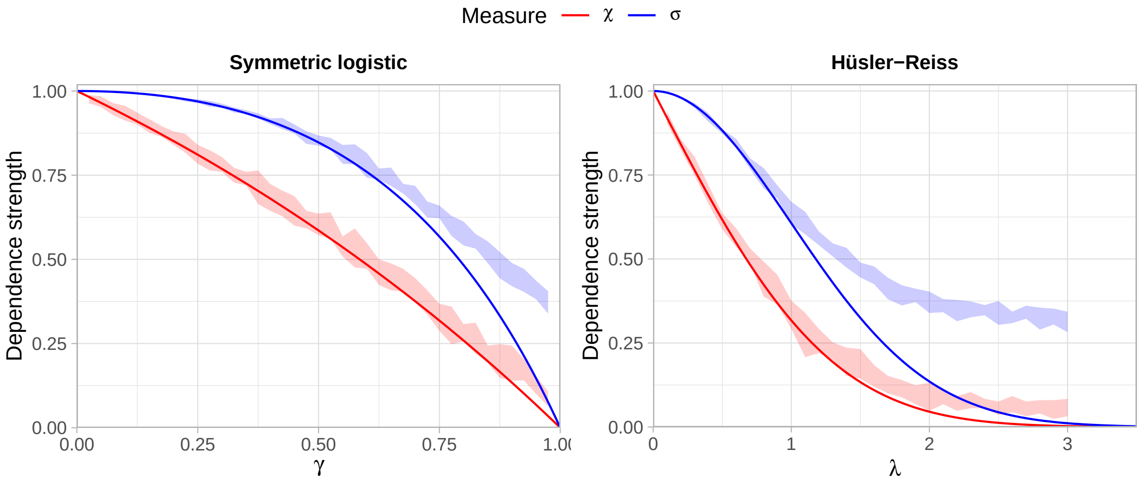

Figure 2.5: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate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

Note that overestimating the dependence strength corresponds to a positive bias in the TPDM estimate, so the inequality is reversed when compared to γ.

Estimation error in the empirical TPDM was first studied by Cooley and Thibaud (2019). Figure 3 in their Supplementary Material assesses the accuracy of the eigenvalues/eigenvectors of the empirical TPDM based on data generated from a Brown-Resnick model. The leading eigenvalue is consistently overestimated ($\hat{\lambda}_1 > \lambda_1$) and subsequent eigenvalues are underestimated ($\hat{\lambda}_j < \lambda_j$ for $j \geq 2$). The sample covariance matrix suffers a similar deficiency, especially when the sample size and dimension are comparable in magnitude (Mestre 2008). The magnitude of the bias depends on the sample size and the proportion k/n. Errors in the eigenvalues may influence the results of downstream PCA analysis, e.g. in deciding how many components are to be retained in the PCA.

2.6.2 Existing bias-correction approaches for the TPDM

The first strategy for tackling the bias issue is found in Mhatre and Cooley (2021). Working in a time series context, they study serial dependence in extremes using the tail pairwise dependence function (TPDF) $\sigma(h)$, which summarises the tail dependence between X_t and X_{t+h} for a tail stationary time series $\{X_t : t = 1 \ldots, n\}$. Simulation experiments reveal that the empirical TPDF $\hat{\sigma}(h)$ is biased at higher lags where the true dependence is close to zero. To counteract this, they subtract the mean from the time series in pre-processing. The rationale for this is described in terms of the position of extreme points in a lag plot (i.e. a scatter plot of (X_t, X_{t+h}) for fixed h). Subtracting the mean has negligible effect on the angles corresponding to joint extremes, but points near a coordinate axis are shifted even closer to the axis.

Fix et al. (2021) develop the first bias-corrected estimate of the TPDM. Recall from Section XX the problem of estimating the spatial dependence parameter ρ of the extremal SAR model (**??**). When the study domain is large or the modelled phenomenon is highly localised, the pairwise dependence between distant sites is weak and the empirical TPDM is prone to overestimation in the corresponding entries. This bias carries over to $\hat{\rho}$ defined in (2.63). Their bias-corrected estimate $\tilde{\Sigma}$ reduces the entries of $\hat{\Sigma}$ by element-wise application of the soft-thresholding operator (Rothman et al. 2009), that is

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

The threshold $\lambda \geq 0$ is selected by assuming that the pairwise tail dependence vanishes to zero as the distance between two sites increases. For $i \neq j$, let h_{ij} denote the (known) spatial distance between the sites corresponding to the variables X_i and X_j. Treating the empirical TPDM entries $\{\hat{\sigma}_{ij} : i \neq j\}$ as functions of distance, they model tail dependence strength against spatial distance via

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

The parameters $\beta_0, \beta_1, \beta_2$ are estimated from the data $\{(\hat{\sigma}_{ij}, h_{ij}) : 1 \leq i < j \leq d\}$ by non-linear least squares estimation, e.g. using `nls()`. Since $\hat{\sigma}(h) \to \beta_2$ as $h \to \infty$, the horizontal asymptote $\hat{\beta}_2$ of the fitted model is used as a proxy for the bias at large distances. It suggests itself to choose $\lambda = \hat{\beta}_2$. Clearly this procedure is only viable in spatial contexts where a notion of proximity exists.

The contrasting strategies of Mhatre and Cooley (2021) and Fix et al. (2021) point towards two qualitatively different ways of improving tail dependence estimation. The first approach acts directly on the data by moving (some of) the extremal angles $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(k)}$ closer to boundary of the simplex in some principled way. In other words, improved inference may be achieved by perturbing the empirical angular measure. This outlook is central to Chapter XX, where we employ sparse simplex projections (Meyer and Wintenberger 2021) to fit max-linear models. Under the second approach, bias-correction is undertaken as a post-processing step. Chapter XX pursues this idea in more detail. We propose a general class of shrinkage/thresholded TPDM estimators that includes (2.65). Unlike Fix et al. (2021), our tuning procedure for selecting the hyperparameter λ is purely data-driven and can be applied in general settings, not just spatial.

3 Testing for time-varying extremal dependence

Multivariate extreme value models typically assume that the data represent independent realisations from some fixed distribution. This requires that both the marginal distributions and the extremal dependence structure are constant throughout the observation period. As explained in Section XX with regards to univariate (marginal) modelling, this assumption is not always valid and non-stationary models are being developed to account for this. However, there is much less research on the topic of non-stationarity in the extremal dependence structure, even though the same problems apply. Anthropogenic climate change is driving changes in the spatial structure of climate extremes (S. Zhou et al. 2023) and regulatory changes can cause structural changes in the joint tail behaviour of financial asset prices (Poon et al. 2003). Thus, a crucial step in the modelling process is to determine whether it is reasonable to assume stationary dependence or not. In this chapter, we present a formal procedure for testing this assumption.

Before proceeding, we clarify an important distinction between *testing for* versus *modelling* non-stationary dependence. Both represent very challenging statistical problems: the underlying signal (e.g. climate change) may be very weak, perhaps only becoming apparent over very long observation periods. The latter task refers to the development of multivariate extreme value models that allow temporal non-stationarity in the dependence structure. For example, the regression model of Castro-Camilo, De Carvalho, et al. (2018) and the spectral density ratio model of De Carvalho and Anthony C. Davison (2014) can incorporate covariate effects, including time. These models rely on parametric assumptions and are restricted to a small number of dimensions. To the best of our knowledge, the only existing work on *testing* for changing dependence is Drees (2023). Roughly speaking, their procedure involves partitioning the observation period into temporal blocks and testing for deviations in $\hat{H}$ between blocks. This is very computationally intensive and thus is restricted to $d \leq 5$ in practice. Our contribution is to devise a procedure that instead tests for changes in $\hat{\Sigma}$, the empirical TPDM. Considering pairwise dependencies instead of the full angular measure eases the computational burden significantly and enables testing even in high dimensions. Our test achieves superior power in many realistic scenarios (Section XX). The trade-off is that neglecting higher-order dependencies necessarily incurs some information loss and we lose power in certain circumstances (Section XX).

3.1 Framework

Suppose $\{\boldsymbol{X}(t) = (X_1(t), \ldots, X_d(t)) : t \in [0,1]\}$ is an $\mathbb{R}_+^d$-valued, continuous time stochastic process with no serial dependence. For $t \in [0,1]$, assume that the random vector $\boldsymbol{X}(t)$ is MRV (Definition 2.8) with constant index of regular variation $\alpha(t) = \alpha$ and potentially time-varying angular measure $H(\cdot\,; t)$ on $\mathbb{S}_{+(\alpha)}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_\alpha = 1\}$. The underlying norm is $\|\cdot\|_\alpha$ and the normalising sequence is $b_n = n^{1/\alpha}$, so that $H(\mathbb{S}_{+(\alpha)}^{d-1}\,; t) = d$ for all $t \in [0,1]$. Denote the radial and angular components of $\boldsymbol{X}(t)$ by $R(t) := \|\boldsymbol{X}(t)\|_\alpha$ and $\boldsymbol{\Theta}(t) := \boldsymbol{X}(t)/\|\boldsymbol{X}(t)\|_\alpha$, respectively. In this time-dependent setting, the MRV property states that for all $z > 0$ and Borel sets $\mathcal{B} \subset \mathbb{S}_{+(\alpha)}^{d-1}$,

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

We assume $\boldsymbol{X}(t)$ is on stationary α-Fréchet margins, perhaps after a suitable marginal transformation. This pre-processing step may require removing marginal non-stationarity using the univariate techniques described in Section XX. Without loss of generality, we take $\alpha = 2$ throughout.

Following Drees (2023), our working null and alternative hypotheses are

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

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

Under the null hypothesis, the angular measure (extremal dependence structure) is constant/stationary. The alternative states that the angular measure is time-varying. The nature of the time-dependence is unspecified. This includes the possibility of instantaneous change-points, smooth gradual evolutions, or a mixture of both. Our goal is to devise a statistical procedure for testing (3.2) against (3.3) based on a discretised sample path of $\{\boldsymbol{X}(t) : t \in [0,1]\}$. This will be achieved by testing for deviations in a time-dependent version of the TPDM.

3.2 The local TPDM and integrated TPDM

Given non-stationary dependence as in (3.1), a time-dependent version of the TPDM is naturally defined by replacing H with the local angular measure $H(\cdot;t)$ in Definition 2.12.

Definition 3.1. The local TPDM of $\boldsymbol{X}(t)$ is the $d \times d$ matrix

$$\Sigma(t) = (\sigma_{ij}(t)), \qquad \sigma_{ij}(t) = \int_{\mathbb{S}^{d-1}_{+(2)}} \theta_i\theta_j \,\mathrm{d}H(\boldsymbol{\theta};t) = \mathbb{E}_{H(\cdot;t)}[\Theta_i\Theta_j]. \tag{3.4}$$

Since $H(\cdot\,;t)$ is a valid angular measure, the local TPDM possesses all the usual properties of a TPDM (Section XX). Its entries summarise the tail dependence strength between pairs of components of $\boldsymbol{X}(t)$. 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 (CITE). This strategy is employed by Drees (2023), whose test statistics are not directly based on the angular measure, but rather on the integrated angular measure,

$$\mathrm{IH}(\cdot;t) := \int_0^t H(\cdot;s)\,\mathrm{d}s.$$

We define the integrated TDPM analogously.

Definition 3.2. The integrated TPDM of $\{\boldsymbol{X}(t) : t \in [0,1]\}$ at a fixed time $t \in [0,1]$is the $d \times d$ matrix given by

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

The integrated TPDM can be equivalently expressed in terms of the integrated angular measure, since

$$\begin{aligned}\psi_{ij}(t) &= \int_0^t \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i\theta_j \,\mathrm{d}H(\boldsymbol{\theta};s)\,\mathrm{d}s \\ &= \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i\theta_j \int_0^t \mathrm{d}H(\boldsymbol{\theta};s)\,\mathrm{d}s \\ &= \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i\theta_j \,\mathrm{dIH}(\boldsymbol{\theta};t).\end{aligned}$$

The following result lists some useful properties of $\Psi(t)$.

Lemma 3.1. *Let $\Psi(t)$ be an integrated TPDM for some $t \in [0,1]$. Then:*

1. *For any $i \neq j$, the entries of $\Psi(t)$ satisfy $\psi_{ii}(t) = t$ and $\psi_{ij}(t) \in [0,t]$.*
2. *The entry $\psi_{ij}(t) = 0$ if and only if $X_i(s)$ and $X_j(s)$ are asymptotically independent for almost every $s \in [0,t]$.*
3. *$\Psi(t)$ is symmetric, positive semi-definite.*

Proof. Recall that the local TPDM possesses the properties of the TPDM.

1. From Section XX, we know that $\sigma_{ii}(s) = 1$ and $\sigma_{ij}(s) \in [0,1]$ for all $s \in [0,1]$. It immediately follows that

$$\begin{aligned}\psi_{ii}(t) &= \int_0^t \sigma_{ii}(s)\,\mathrm{d}s = \int_0^t \mathrm{d}s = t, \\ \psi_{ij}(t) &= \int_0^t \sigma_{ij}(s)\,\mathrm{d}s \leq \int_0^t \mathrm{d}s = t.\end{aligned}$$

2. For any (measurable) non-negative function $f : \mathbb{R} \to \mathbb{R}_+$, the condition

$$\int_a^b f(x)\,\mathrm{d}x = 0$$

holds if and only if $f(x) = 0$ for almost every $x \in [a,b]$ (*citation needed?*). Applying this fact to $f(t) = \sigma_{ij}(t)$ with $a = 0$ and $b = t$ yields the result.

3. Symmetry of $\Psi(t)$ is inherited from symmetry of the local TPDM. For any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, we have

$$\boldsymbol{y}^T \Psi(t)\boldsymbol{y} = \boldsymbol{y}^T \left(\int_0^t \Sigma(s)\,\mathrm{d}s\right)\boldsymbol{y} = \int_0^t \boldsymbol{y}^T \Sigma(s)\boldsymbol{y}\,\mathrm{d}s.$$

   Positive semi-definiteness of $\Sigma(t)$ guarantees the integrand is non-negative. Therefore $\boldsymbol{y}^T \Psi(t)\boldsymbol{y} \geq 0$.

□

Due to properties (1) and (3) we may focus on vectorised forms of $\Sigma(t)$ and $\Psi(t)$ defined analogously to $\boldsymbol{\sigma}$ in (2.60), i.e. by row-wise flattening of the strictly upper-triangular elements of $\Sigma(t)$ and $\Psi(t)$:

$$\boldsymbol{\sigma}(t) := \mathrm{vecu}(\Sigma(t)) = (\sigma_{12}(t), \sigma_{13}(t), \ldots, \sigma_{1d}(t), \sigma_{23}(t), \ldots, \sigma_{2d}(t), \ldots, \sigma_{d-1,d}(t)), \tag{3.5}$$

$$\boldsymbol{\psi}(t) := \mathrm{vecu}(\Psi(t)) = (\psi_{12}(t), \psi_{13}(t), \ldots, \psi_{1d}(t), \psi_{23}(t), \ldots, \psi_{2d}(t), \ldots, \psi_{d-1,d}(t)). \tag{3.6}$$

Each vector has dimension $\mathcal{D} = \binom{d}{2}$.

Using a simple example, we now sketch how the integrated TPDM will be used to test for non-stationary dependence. Suppose $\boldsymbol{X}(t)$ follows a symmetric logistic distribution with dependence parameter $\gamma(t)$. Consider the following two scenarios: $\gamma(t) = 0.7$ (constant dependence) and $\gamma(t) = 0.5 + |t - 0.5|$ (changing dependence). These cases correspond to the left- and right- plots in Figure 3.1, respectively, with $\gamma(t)$ represented by the blue line. The red and green lines depict the local TPDM $\sigma_{ij}(t)$ and integrated TPDM $\psi_{ij}(t)$, respectively, as functions of t. Under the symmetric logistic model all pairs are equivalent, so we may suppress the ij subscript. In the left-hand plot, constant dependence manifests as a horizontal line for $\sigma(t)$ and a straight line for $\psi(t)$. In the right-hand plot, the dependence strength $\sigma(t)$ is greatest near the centre of the time interval and $\psi(t)$ is the time-integral of this non-linear function. Intuitively, our test works by quantifying whether (estimates of) $\psi_{ij}(t)$ deviate from straight lines. Mathematically it will prove more convenient to instead consider deviations of (estimates of) $\psi_{ij}(t) - t\psi_{ij}(1)$ from zero. The function $\psi(t) - t\psi(1)$

is plotted in black. In the constant dependence case, $\sigma(t) = \sigma$ for all $t \in [0,1]$, so

$$\psi(t) - t\psi(1) = \int_0^t \sigma(s)\,\mathrm{d}s - t\int_0^1 \sigma(s)\,\mathrm{d}s = \sigma\left(\int_0^t \mathrm{d}s - t\int_0^1 \mathrm{d}s\right) = 0. \tag{3.7}$$

The following sections concern the main statistical challenges, namely (i) how to estimate $\boldsymbol{\sigma}(t)$ and $\boldsymbol{\psi}(t)$, and (ii) the construction of test statistics quantifying whether an estimate of $\{\boldsymbol{\psi}(t) - t\boldsymbol{\psi}(1) : t \in [0,1]\}$ is sufficiently different from zero.

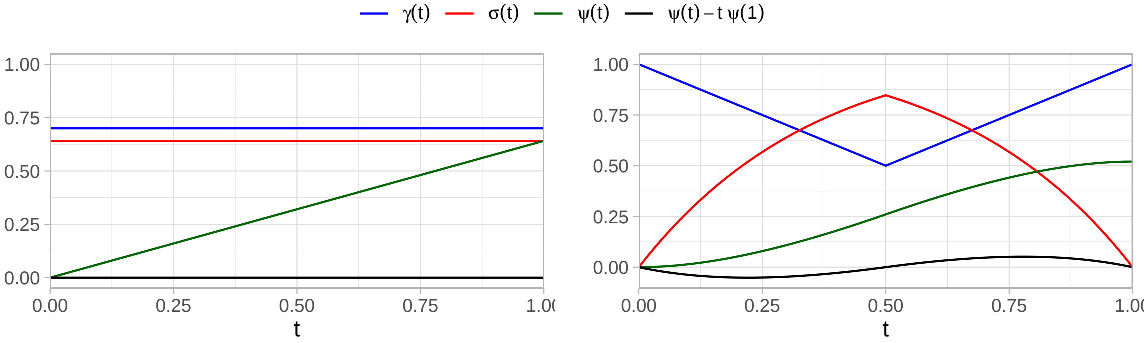

Figure 3.1: The quantities $\sigma_{ij}(t)$, $\psi_{ij}(t)$, and $\psi_{ij}(t) - t\psi_{ij}(1)$ as functions of t when $\boldsymbol{X}(t)$ is symmetric logistic with dependence parameter $\gamma(t)$. Left: constant dependence with $\gamma(t) = 0.7$. Right: time-varying dependence with $\gamma(t) = 0.5 + |t - 0.5|$.

3.3 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, corresponding to realisations of the independent random vectors $\{\boldsymbol{X}(i/n) : i = 1, \dots, n\}$.

3.3.1 The empirical local TPDM

Provided the tail distribution of $\boldsymbol{X}(t)$ varies sufficiently smoothly with t, we may infer the local dependence structure at time $t \in [0,1]$ using the most extreme observations lying within a small neighbourhood of t. For some positive bandwidth $h = h(n)$, define by

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

the index set of observations in a h-neighbourhood of t. Among the observations $\{\boldsymbol{X}(i/n) : i \in \mathcal{I}(t)\}$, only those whose norm exceeds some high radial threshold will enter into our estimator for $\Sigma(t)$. The threshold $\hat{u}(t)$ is set as the $k+1$ upper order statistic of $\{R(i/n) : i \in \mathcal{I}(t)\}$, resulting in exactly k radial threshold exceedances. Drees (2023) use the same setup to define the empirical local angular measure, which forms the basis of our estimator for the local TPDM.

Definition 3.3. For any $t \in [0,1]$, the empirical local angular measure is the random measure on $\mathbb{S}_{+(2)}^{d-1}$ defined by

$$\hat{H}(\cdot;t) := \frac{d}{k}\sum_{i\in\mathcal{I}(t)} \mathbf{1}\{R(i/n) > \hat{u}(t), \boldsymbol{\Theta}(i/n) \in \cdot\}. \tag{3.8}$$

Definition 3.4. For $t \in [0,1]$, the empirical local TPDM estimator is the $d \times d$ matrix $\hat{\Sigma}(t) = (\hat{\sigma}_{ij}(t))$, where

$$\hat{\sigma}_{ij}(t) := \int_{\mathbb{S}_{(2)+}^{d-1}} \theta_i\theta_j \,\mathrm{d}\hat{H}(\boldsymbol{\theta};t) = \frac{d}{k}\sum_{l\in\mathcal{I}(t)} \Theta_i(l/n)\Theta_j(l/n)\mathbf{1}\{R(l/n) > \hat{u}(t)\}. \tag{3.9}$$

We recognise $\hat{H}(\cdot;t$ and $\hat{\Sigma}(t)$ as simply time-localised versions of the empirical angular measure (Definition 2.11) and empirical TPDM (Definition 2.16). It is easy to see that the empirical local TPDM possesses the same finite-sample properties as the empirical TPDM, such as positive semi-definiteness and complete positivity. Unlike the empirical TPDM, the performance of $\hat{\Sigma}(t)$ depends not only on k but also on the additional tuning parameter h. Joint selection of k and h involves managing several trade-offs. A large effective sample size can be achieved by increasing k and/or h. As we know, increasing k risks bias due the inclusion of observations from the bulk. Increasing h introduces the possibility of another kind of bias due to the influence *non-local* extreme observations, i.e. data that are not representative of the tail distribution *at time* t. These trade-offs will appear via modified rate conditions when describing the asymptotic properties of $\hat{\Sigma}(t)$ in Section XX.

3.3.2 The empirical integrated TPDM

The best approach for estimating the integrated TPDM is slightly less obvious. For computational and mathematical reasons, we elect for a block-based approach that involves performing estimation in a piecewise constant fashion over disjoint time intervals of width $2h$. For any $t \in [0,1]$ and $h > 0$, by the additive property of the integral we have that

$$\begin{aligned}\psi_{ij}(t) &= \int_0^t \sigma_{ij}(s)\,\mathrm{d}s \\ &= \int_0^{2h} \sigma_{ij}(s)\,\mathrm{d}s + \int_{2h}^{4h} \sigma_{ij}(s)\,\mathrm{d}s + \ldots + \int_{2h\lfloor t/(2h)\rfloor}^{t} \sigma_{ij}(s)\,\mathrm{d}s \\ &= \sum_{l=1}^{\lfloor t/(2h)\rfloor} \int_{2h(l-1)}^{2hl} \sigma_{ij}(s)\,\mathrm{d}s + \int_{2h\lfloor t/(2h)\rfloor}^{t} \sigma_{ij}(s)\,\mathrm{d}s.\end{aligned}$$

The first term corresponds to the $\lfloor t/(2h)\rfloor$ whole blocks in $[0,t]$. The second term corresponds to the final partial block; this term vanishes if t is a multiple of $2h$. Our estimator for $\psi_{ij}(t)$ assumes that $\sigma_{ij}(s)$ is constant across each of the disjoint intervals and then estimates $\sigma_{ij}(s)$ empirically at the interval centres using h as the bandwidth. Henceforth, assume for simplicity that the number of blocks $1/(2h)$ is an integer.

Definition 3.5. For $t \in [0,1]$, the empirical integrated TPDM estimator is the $d \times d$ matrix $\hat{\Psi}(t) = (\hat{\psi}_{ij}(t))$, where

$$\begin{aligned}\hat{\psi}_{ij}(t) &:= \sum_{l=1}^{\lfloor t/(2h)\rfloor} \int_{2h(l-1)}^{2hl} \hat{\sigma}_{ij}((2l-1)h)\,\mathrm{d}s + \int_{2h\lfloor t/(2h)\rfloor}^{t} \hat{\sigma}_{ij}((2\lfloor t/(2h)\rfloor + 1)h)\,\mathrm{d}s \\ &= 2h \sum_{l=1}^{\lfloor t/(2h)\rfloor} \hat{\sigma}_{ij}((2l-1)h) + (t - 2h\lfloor t/(2h)\rfloor)\hat{\sigma}_{ij}((2\lfloor t/(2h)\rfloor + 1)h).\end{aligned}$$

This construction permits efficient computation of the entire process $\{\hat{\Psi}(t) : t \in [0,1]\}$. Note that the partition of the time-interval $[0,t]$ depends only on h, not on t. Hence, the estimates $\hat{\Sigma}(s)$ at the time points $s \in \{h, 3h, \ldots, 1-h\}$ are sufficient to estimate $\Psi(t)$ at any $t \in [0,1]$ via a simple weighted sum. The piecewise constant assumption on $\Sigma(t)$ implies that $\hat{\psi}_{ij}(t)$ is a piecewise linear function of t. Therefore, to compute the full path $\{\hat{\Psi}(t) : t \in [0,1]\}$ one need only compute $\hat{\Psi}(s)$ at the interval endpoints $s \in \{2h, 4h, \ldots, 1\}$ and fill in the intermediate time points by linear interpolation.

Mathematically, the appeal of a block-based construction is that $\hat{\Psi}(t)$ is a weighted sum of independent random matrices $\hat{\Sigma}(h), \hat{\Sigma}(3h), \dots, \hat{\Sigma}(1-h)$. Independence is due to the blocks being non-overlapping and our choosing the bandwidth as half the block width, so that dependence within each block is estimated using only observations contained in it. This independence is crucial in the elicitation of the asymptotic results to follow.

3.3.3 Asymptotic properties of $\hat{\Sigma}(t)$ and $\hat{\Psi}(t)$

We now formulate the asymptotic properties of the estimators $\hat{\boldsymbol{\sigma}}(t) := \mathrm{vecu}(\hat{\Sigma}(t))$ and $\hat{\boldsymbol{\psi}}(t) := \mathrm{vecu}(\hat{\Psi}(t))$ of $\boldsymbol{\sigma}(t)$ and $\boldsymbol{\psi}(t)$ defined in (3.5) and (3.6). The main result is asymptotic normality of $\hat{\Sigma}(t)$, cf. Proposition 2.6.

Proposition 3.1. *Assume $k(n)$ and $h(n)$ satisfy the rate conditions*

$$h(n) \to 0, \quad k(n) \to \infty, \quad nh(n) \to \infty, \quad \frac{k(n)}{nh(n)} \to 0 \tag{3.10}$$

as $n \to \infty$. Then, for any $t \in [0,1]$,

$$\sqrt{k}(\hat{\boldsymbol{\sigma}}(t) - \boldsymbol{\sigma}(t)) \to N(\mathbf{0}, V(t)) \tag{3.11}$$

as $n \to \infty$. The $\mathcal{D} \times \mathcal{D}$ asymptotic covariance matrix $V(t)$ has entries given by

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

where

$$\begin{aligned} \nu_{ij}^2(t) &:= \mathrm{Var}_{H(\cdot;t)}(\Theta_i\Theta_j) \\ \rho_{ij}(t) &:= \frac{1}{2}\left[\mathrm{Var}_{H(\cdot;t)}(\Theta_i\Theta_j + \Theta_l\Theta_m) - \nu_{ij}^2(t) - \nu_{lm}^2(t)\right] \end{aligned}$$

Proof. Proof here.

□

We will always assume that $V(t)$ is invertible. (This would not be permissible if we had defined the vecu operator to include diagonal entries.) The above result means that, roughly speaking, the entries of $\hat{\Sigma}(t)$ behave like (correlated) normal random variables when n is sufficiently large. Combining this with Definition 3.5, the implication is that each entry $\psi_{ij}(t)$ of $\hat{\Psi}(t)$ is the weighted sum of independent, asymptotically normal random variables. With this intuition, we can apply a functional central limit theorem type argument to derive the asymptotic behaviour of the stochastic process $\{\hat{\boldsymbol{\psi}}(t) : t \in [0,1]\}$.

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

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

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

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

Proof. Proof here.

□

The drift and diffusion coefficients associated with each univariate process $\{Y_{ij}(t) : t \in [0,1]\}$ are controlled by the cumulative dependence $\{\psi_{ij}(t) : t \in [0,1]\}$ and asymptotic variance $\{\nu_{ij}^2(t) : t \in [0,1]\}$, respectively. Meanwhile, the cross-correlation between $\{Y_{ij}(t) : t \in [0,1]\}$ and $\{Y_{lm}(t) : t \in [0,1]\}$ is determined by $\{\rho_{ij,lm}(t) : t \in [0,1]\}$.

Corollary 3.1. *Suppose the conditions of Proposition 3.2 hold and the null hypothesis (3.2) is true. Define*

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

Then $\{\hat{\boldsymbol{Z}}(t) : t \in [0,1]\}$ converges to $\{\boldsymbol{B}(t) : t \in [0,1]\}$, a standard $\mathcal{D}$-dimensional Brownian bridge.

Proof. Under the null, the asymptotic covariance matrix does not depend on t, so we may write $V = V(t)$. Using Proposition 3.2 and pre-multiplying the process (3.12) by $(2V)^{-1/2}$, it follows that

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

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

$$\mathrm{Cov}(Y_{ij}(s),Y_{lm}(t))=\int_0^{\min(s,t)}\mathrm{d}\tau=\min(s,t). \tag{3.15}$$

This is precisely the covariance function of a Brownian motion. By (3.7), $\boldsymbol{\psi}(t)=t\boldsymbol{\psi}(1)$ under the null and therefore

$$\begin{aligned}\hat{\boldsymbol{Z}}(t)&=\sqrt{\frac{k}{2h}}V^{-1/2}(\hat{\boldsymbol{\psi}}(t)-t\hat{\boldsymbol{\psi}}(1))\\&=\sqrt{\frac{k}{2h}}V^{-1/2}\left[\hat{\boldsymbol{\psi}}(t)-\boldsymbol{\psi}(t)-t(\hat{\boldsymbol{\psi}}(1)-\boldsymbol{\psi}(1))\right]\\&=\sqrt{\frac{k}{2h}}V^{-1/2}\left(\hat{\boldsymbol{\psi}}(t)-\boldsymbol{\psi}(t)\right)-t\left[\sqrt{\frac{k}{2h}}V^{-1/2}\left(\hat{\boldsymbol{\psi}}(t)-\boldsymbol{\psi}(1)\right)\right]\\&\to\boldsymbol{W}(t)-t\boldsymbol{W}(1).\end{aligned}$$

The process in the final line is equal in distribution to a Brownian bridge (*citation needed?*).

□

This main result provides the foundation for our test. The test statistics defined the following section will quantify whether the realised sample path of (3.14) is consistent with a Brownian bridge. We are not the first to use Brownian bridges in hypothesis testing in extremes; Gadeikis and Paulauskas (2005) use the same principle to test for changes in the tail index.

3.4 Test statistics and critical values

From the test process (3.14) we define Kolmogorov-Smirnov (KS) and Cramér-von-Mises (CM) type test statistics by

$$T^{(KS)} := \sup_{t\in[0,1]} \left\|\hat{\boldsymbol{Z}}(t)\right\|_\infty = \sup_{\substack{t\in[0,1]\\ i<j}} |Z_{ij}(t)|, \tag{3.16}$$

$$T^{(CM)} := \sup_{1\leq i<j\leq d} \left\|\hat{Z}_{ij}(t)\right\|^2_{L^2[0,1]} = \sup_{i<j} \int_0^1 |\hat{Z}_{ij}(t)|^2 \,\mathrm{d}t, \tag{3.17}$$

where $\|\boldsymbol{x}\|_\infty := \max\{|x_i| : i = 1,\dots,\mathcal{D}\}$ 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),*

$$T^{(KS)} \to \sup_{t\in[0,1]} \|\boldsymbol{B}(t)\|_\infty \overset{d}{=} \sup_{i<j} K_{ij}, \qquad T^{(CM)} \to \sup_{i<j} \|B_{ij}(t)\|^2_{L^2[0,1]}, \tag{3.18}$$

where $\boldsymbol{B}(t) = (B_{ij}(t) : i < j)$ denotes a standard $\mathcal{D}$-dimensional Brownian bridge and $\{K_{ij} : i < j\}$ are independent Kolmogorov random variables with distribution function

$$\mathbb{P}(K_{ij} < x) = F_K(x) = \begin{cases} 1 + 2\sum_{m=1}^{\infty}(-1)^m \exp(-2m^2x^2), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Proof. The asymptotic null distributions follows directly from Corollary 3.1. It is well known that $K_{ij} := \sup_{t\in[0,1]} |B_{ij}(t)|$ follows a Kolmogorov distribution (Henze 2024, p. 328).

□

Each test statistic $T^{(KS)}$ and $T^{(CM)}$ may be used to define an asymptotic test for constant dependence. For the KS test,

$$\mathbf{1}\{T^{(KS)} > c_\alpha\}, \qquad c_\alpha = F_K^{-1}((1-\alpha)^{1/\mathcal{D}}) \tag{3.19}$$

Table 3.1: Asymptotic 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

constitutes an asymptotic level α test. The critical value c_α represents the value for which a set of $\mathcal{D}$ independent one-dimensional Brownian bridges *all* remain in the region $(-c_\alpha, c_\alpha)$ with probability $1-\alpha$. Note that we avoid issues with multiple testing because the critical value implicitly accounts for the dimension d. Specifically, the critical value increases with d. This is intuitive because with a greater number of paths $\mathcal{D} = \mathcal{O}(d^2)$ there is a higher chance that at least one of them will exit a fixed interval $(-c, c)$. A CM-type test is constructed analogously. The only material difference is that 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 and compute their L^2-norms by numerical integration. Quantiles of the empirical distribution of these values are used to obtain approximate critical values. Critical values for selected dimensions and significance levels are listed in Table 3.1.

It is only feasible to produce a table of critical values due to the inclusion of the 'nuisance process' $\{V(t) : t \in [0,1]\}$ in (3.14). Its role is to standardise and remove cross-correlation in $\hat{\boldsymbol{Z}}(t)$, ensuring a convenient asymptotic null distribution for our test statistics. In contrast, when $d \geq 3$ the critical values in Drees (2023) depend on the class of sets $\mathcal{A}$ under consideration, so one is forced to resort to (intensive) simulations. To deal with the nuisance process, we simply estimate it from the data under the assumption of stationarity. Under the null, $\{V(t) : t \in [0,1]\}$ reduces to a single matrix, V, which we estimate as the sample covariance matrix of $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}(t)$ based on the set of radial threshold exceedances over all blocks. That is

$$\hat{v}_{ij,lm} := 2h \sum_{s=1}^{1/(2h)} \hat{v}_{ij,lm}((2s-1)h),$$
$$\hat{v}_{ij,lm}(t) := \frac{1}{k-1} \sum_{\tau \in \mathcal{I}(t)} \left[d\Theta_i\left(\frac{\tau}{n}\right)\Theta_j\left(\frac{\tau}{n}\right) - \hat{\sigma}_{ij}(t)\right]\left[d\Theta_l\left(\frac{\tau}{n}\right)\Theta_m\left(\frac{\tau}{n}\right) - \hat{\sigma}_{lm}(t)\right] \mathbf{1}\left\{R\left(\frac{\tau}{n}\right) > \hat{u}(t)\right\}.$$

Provided the rank condition

$$k_{\text{total}} := k/(2h) > \mathcal{D} \tag{3.20}$$

is satisfied, the estimator $\hat{V}$ is full-rank and therefore invertible. For a fixed sample size and choice of k and h, the rank condition imposes an upper limit on the dimension, roughly $d < \sqrt{2k_{\text{total}}} = \sqrt{k/h}$. The existence of this limit reflects the principle that reliable inference in high-dimensional settings requires commensurate data. For fixed n, we may increase the limit by increasing k and/or h, 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 the issue of invertibility altogether. This avenue is not explored on the basis that it doesn't seem sensible to proceed when the rank condition indicates there is insufficient data for the task at hand.

3.5 Simulation experiments

In this section, we present a series of numerical experiments demonstrating our method's performance and, where applicable, providing comparisons against Drees (2023).

3.5.1 Data generating processes

Suppose the extremal dependence structure of $\boldsymbol{X}(t) = (X_1(t), \dots, X_d(t))$ 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 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):** with $\gamma(t) = 1/\vartheta(t)$ and $\Omega = [1, \infty)$. With this parametrisation, asymptotic independence occurs when $\vartheta(t) = 1$ and complete asymptotic dependence as $\vartheta(t) \to \infty$.
2. **Hüsler-Reiss (HR):** with $\Lambda(t) = \vartheta(t)\Lambda_0$ and $\Omega = (0, \infty)$, where $\Lambda_0 \in \mathbb{R}_+^{d\times d}$ is a valid HR parameter matrix (see Section XX). 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 Λ_0. While not strictly necessary, we take $\vartheta_0 = 1$ so that $\Lambda(0) = \Lambda_0$. For each dimension d, the initial matrix Λ_0 is randomly generated using the procedure outlined in Appendix B1 in Fomichov and Ivanovs (2023).

The three dependence scenarios and two parametric distributions give six qualitatively different models. We refer to these as, e.g. SL-constant, HR-jump, and so on. When $d = 2$, the test of Drees (2023) is included as a comparator. Results pertaining to their test are based on our own implementation with $\mathcal{A} = \{A_y : y = 0.01, 0.02, \ldots, 0.99\}$, where

$$A_y := \{\boldsymbol{\theta} \in \mathbb{S}^1_{+(2)} : \theta_1 \leq y\} \subset \mathbb{S}^1_{+(2)}. \tag{3.21}$$

3.5.2 Large sample performance

In an idealised setting with infinite data, the null distribution of the test statistics is as described in Proposition 3.3. This may be empirically validated via large-sample simulations by checking whether the p-values are uniformly distributed.

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 containing $b := 2nh = 2,000$ observations, a tail sampling fraction $k/b = 2.5\%$, and an overall effective sample size of $k_{\text{total}} = 25,000$. Figure 3.2 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 (left), the theoretical quantiles in the QQ plots are computed using the Kolmogorov quantile function implemented in the CPAT package. The theoretical quantiles for the CM-type test (right) are estimated from the aforementioned simulated Brownian bridges. In each panel, the dimension and parametric distribution are represented by the line type and colour, respectively. In all cases, the p-values appear to be 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.

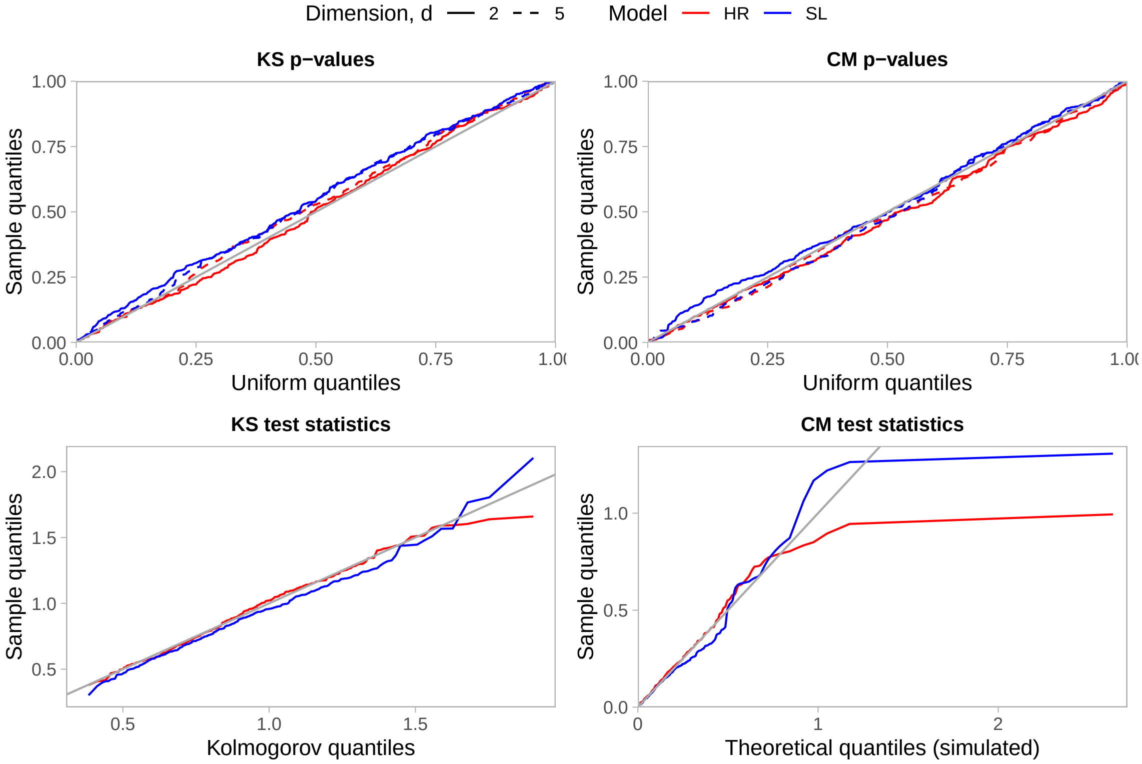

Figure 3.2: Large sample Q-Q plots for the p-values (top) and test statistics (bottom) associated with the KS test (left) and CM test (right). Based on 350 simulations from the SL- and HR-constant models with $n = 10^6$, $b = 2000$ and $k = 50$.

Next, we check that our procedure detects dependence changes with high probability when the number of samples is large. The experimental procedure is unchanged, except that data are now generated from the SL-jump model with $\vartheta_0 = 2$ and $\vartheta_1 = 2.5$. These values bring about relatively subtle shifts in the dependence strength, with $\sigma_{ij}(t) = 0.85$ for $t < 0.5$ and $\sigma_{ij}(t) = 0.91$ for $t \geq 0.5$. All p-values are less than 10^{-15}, meaning our method consistently and overwhelmingly identifies the dependence chang. Obviously we do not expect the test to maintain this performance level for samples of realistic size.

3.5.3 Small sample performance

In finite sample settings, the empirical size of an asymptotic test will generally differ from its nominal size. The only guarantee is that the correct level is attained as $n \to \infty$. The hope is that convergence occurs with sufficient rapidity that the difference between the prescribed and actual Type I error rates is acceptably small. To test this, we run simulations using the SL-constant ($\vartheta_0 = 2$) and HR-constant ($\vartheta_0 = 1$) models in dimensions $d \in \{2, 5, 10, 25\}$ with sample sizes $n \in \{2.5 \times 10^3, 5 \times 10^3, 10^4\}$. For each data set, we apply our method at the 5% level. 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 combination of tuning parameters were excluded because they violate the rank condition (3.20) or because $k \leq d$. Each value in the table is based on N repeated simulations, where $N = 10^3$ if $d \leq 5$ and $N = 300$ otherwise. Results from the large sample experiments in dimensions $d \in \{2, 5\}$ are included (bottom row) for completeness.

First we review the results for $d = 2$. Under our method and Drees' method, the rejection rate of the CM-based test is consistently around 5% for almost any choice of b and k, even when $n = 2,500$. The KS-based tests are universally more conservative than the CM-based tests, particularly for larger block sizes. Drees (2023) attribute this to the fact that a coarsely discretised path may only attain its supremum at a small number of time points (integer multiples of $2h$), whereas the corresponding critical values arise from suprema of continuous processes. However, for any value of n, there exists a pair of tuning parameters such that the discrepancy between the empirical and nominal size is at most 0.7%.

Now consider the columns for $d \geq 5$. It is not possible to include Drees (2023) as a comparator here because, by their own admission, the necessary computations become prohibitively expensive. We find that the KS-test remains rather conservative, so we shall focus on the CM-test instead. When $d = 5$, our procedure works well for both models, even when n is small. For $d = 10$ and $d = 25$, the rank conditions on $\hat{V}$ and $\hat{\Sigma}(t)$ take effect, drastically reducing the set of admissible tuning parameters when $n \leq 5,000$. Even in these high-dimensional settings, the test produces reasonable results, especially for symmetric logistic data. Performance deteriorates under the 25-dimensional Hüsler-Reiss

Table 3.2: Empirical Type I error rates (%) across repeated simulations. The number of simulations is $N = 1000$ if $d \leq 5$, or $N = 300$ otherwise. All tests have nominal size 5%. The parameters of the SL-constant and HR-constant models are $\vartheta_0 = 2$ and $\vartheta_0 = 1$, respectively.

(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				

model, suggesting there may be insufficient data for the asymptotic approximations to hold.

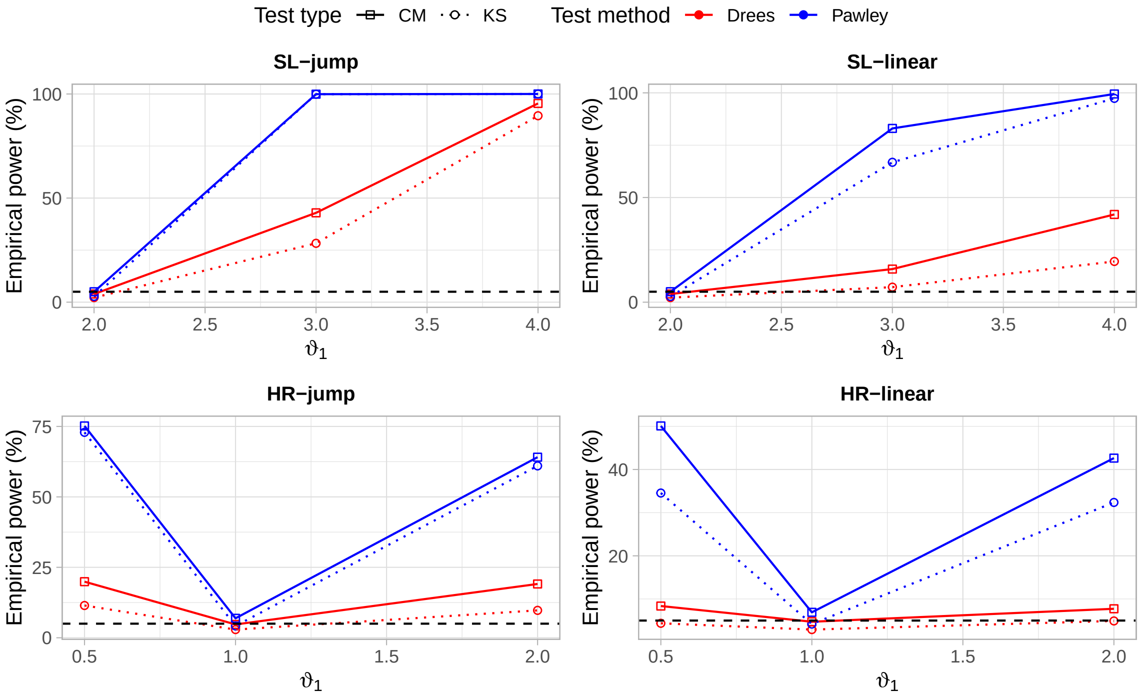

Figure 3.3: Empirical power (%) as a function of the dependence parameter ϑ_1. Based on 1000 simulations with $n = 2,500$ and $d = 2$. For the SL and HR models, $\vartheta_0 = 2$ and $\vartheta_0 = 1$, respectively. Tests are conducted at the 5% level (black dashed line).

Our next experiment assesses the empirical power under alternatives. For this, we generate data from the SL-jump ($\vartheta_0 = 2$), SL-linear ($\vartheta_0 = 2$), HR-jump ($\vartheta_0 = 1$) and HR-linear ($\vartheta_0 = 1$) models. The dependence ϑ_1 at time 1 is allowed to vary, allowing us to examine the relationship between the power and the magnitude of the dependence change. For each model and value of ϑ_1, we simulate 1,000 data sets with $d = 2$ and $n = 2,500$. All tests are conducted at the 5% level.

Figure 3.3 displays the results for $n/b = 25$ and $k/b = 0.1$; Figure F.1 in Appendix XX confirms that the conclusions are not overly sensitive to this choice. The plots show that our test achieves superior power compared to Drees (2023) in all scenarios. Note that when $\vartheta_1 = \vartheta_0$, the null is true so power reverts to 5%. All tests easily detect the largest SL-jump change, achieving near full power in this case. On the other hand, Drees' test is virtually powerless under the more challenging HR-linear scenario, while our procedure maintains a respectable level of performance. Here, the effect of focussing on bivariate summaries rather than the full angular measure becomes evident. Imposing additional structure/assumptions – namely that the TPDM provides an adequate summary of dependence – improves the signal-to-noise ratio, so that subtle dependence changes may be detected. In the case of both the symmetric logistic and Hüsler-Reiss models, this assumption is valid (and therefore helpful) due to the one-to-one correspondence between the model parameter and the TPDM. An example where this is not the case is given in Section XX. When dependence changes abruptly (SL-jump and HR-jump), the CM- and KS-based test perform equally well. Upon further investigation, we find that the path $\{\hat{Z}_{12}(t) : t \in [0,1]$ is a $\wedge$-shaped curve attaining its maximum at $t = 0.5$ (when the changepoint occurs). Thus $T^{KS} \approx \hat{Z}_{12}(0.5)$ and, upon approximating $\{|\hat{Z}_{12}(t)|^2 : t \in [0,1]$ by a triangle, $T^{CM} \approx \hat{Z}_{12}^2(0.5)/2$. Both test statistics are determined by $\hat{Z}_{12}(t)$, so they invariably reach the same outcome. For gradual changes the CM-based test is superior, cf. Figures 1 and 2 in Drees (2023).

Figure 3.4 shows how the power of the KS test evolves as more data is acquired. The experimental procedure is the same as above, except the sample size is allowed to vary. The Q-Q plots depict the distribution of the p-values across 1,000 repeated tests for the HR-jump (left) and HR-linear (right) models based on different values of ϑ_1 (line type) and n (line colour). If a test is highly-powered, then the associated curve will lie below the diagonal. In both scenarios, the power improves as n increases. The effect is more pronounced for the linear dependence change. Intuitively, for an instantaneous change the test only needs to learn the dependence strength before the change-point and after the change-point. This does not require a large amount of information. Detecting gradual changes places more emphasis on accurate local estimation, especially near the end-points of the time interval, so acquiring additional data has a greater effect.

3.5.4 Computation time

A key advantage of our approach is that the computations are considerably less intensive than those required by Drees (2023). As explained earlier, this permits its application in moderate to high-dimensional settings. Another benefit is reduced run times, which may be an important consideration if the test is to be performed repeatedly (e.g. see the data application in Section XX). This prompts us to analyse the computation times for the simulation experiments in Section XX. The left-hand plot in Figure 3.5 shows the

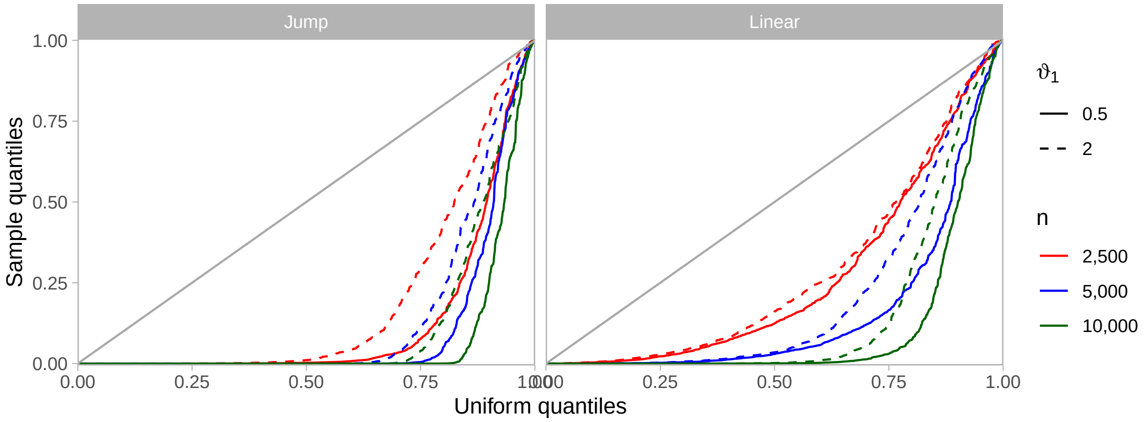

Figure 3.4: QQ-plots for the KS p-values with varying sample size. Based on 1,000 simulations from the HR-jump (left) and HR-linear (right) models with tuning parameters $n/b = 25$ and $k/b = 0.1$.

distribution of the total elapsed time (in seconds) against n. The number of blocks is $n/b = 50$; the number of extremes per block is indicated by the bar colour. Our procedure is faster than Drees', though the difference is only a few hundredths of a second. The test remains fast even when $n = 10^4$. This isn't particularly surprising, since discarding the non-extreme observations means the effective sample size is never actually very large. The right-hand plot shows average computation time as a function of k_{total} for different dimensions d. Dimension is clearly the key determinant of computation time. This is predominantly due to the inversion of the $\mathcal{D} \times \mathcal{D}$ matrix $\hat{V}$, which has $\mathcal{O}(\mathcal{D}^3) = \mathcal{O}(d^6)$ complexity. *More comments? Address it in future work e.g. ways to improve the efficiency of this step, or perhaps circumvent it altogether, would be welcome.*

3.6 Loss of power under TPDM-invariant dependence changes

Our test affords several advantages compared to Drees (2023), most notably the ability to conduct tests in high dimensions. However, the correspondence between dependence structures and TPDMs is many-to-one, so we forgo the ability to detect certain dependence changes where the TPDM is invariant. For this class of alternatives our test is inherently predisposed to commit Type II errors. In contrast, Drees' test is consistent under general alternatives (Drees 2023, Corollary 3.2(ii)). We exemplify this using a time-dependent generalisation of the max-linear model.

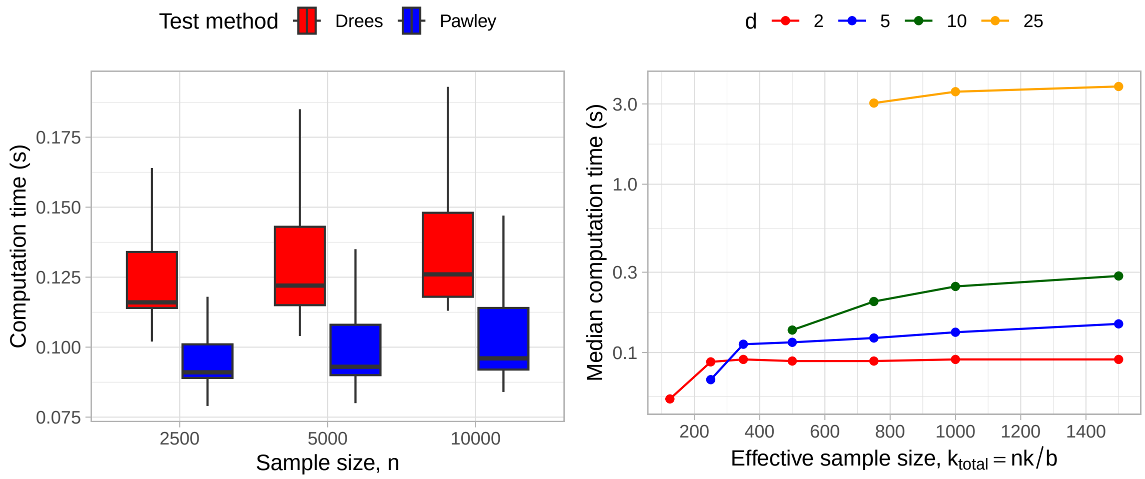

Figure 3.5: Analysis of computation times across the numerical experiments described in Section XX. Left: empirical distribution of run times when $d = 2$, $n/b = 50$, $k/b = 0.1$. Right: median computation time as a function of the total number of threshold exceedances.

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

$$\boldsymbol{X}(t) = A(t) \times_{\max} \boldsymbol{Z}(t), \qquad A(t) = A_0 \mathbf{1}\{t < 0.5\} + A_1 \mathbf{1}\{t \geq 0.5\}. \tag{3.22}$$

The stochastic innovations process $\{\boldsymbol{Z}(t) = (Z_1(t), \dots, Z_q(t)) : t \in [0,1]\}$ is a collection of independent random vectors such that, for any $t \in [0,1]$, the $q \geq 1$ components of $\boldsymbol{Z}(t)$ are independent 2-Fréchet random variables. 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 (3.22), $A(t)$ 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 complicated models can easily be conceived, whereby $A(t)$ evolves smoothly, perhaps even with a varying number of factors $q = q(t)$. The local angular measure associated with (3.22) 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

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

By Example 2.7, the local TPDM is given by $\Sigma(t) = A(t)A(t)^T$. A formula for the asymptotic asymptotic covariance $V(t)$ matrix is derived in Appendix XX.

Suppose $A_0 \neq A_1$ (including up to permutations of their columns) are such that $\Sigma(0) = \Sigma(1)$ and $V(0) = V(1)$. Then the alternative hypothesis (3.3) is true but the asymptotic distributions of $T^{(KS)}$ and $T^{(CM)}$ are the null distributions in (3.18). Clearly this presents an issue for our test.

To illustrate this problem empirically, we seek a pair of matrices A_0, A_1 with a common TPDM and asymptotic covariance. Constructing a non-trivial (i.e. $q > 2$) pair by hand would be extremely laborious. Instead, we generate a large set of valid candidate matrices $A \in \mathbb{R}_+^{d\times q}$ with $d = 2$ and $q = 20$ and search for a valid pair among these. This process yields the matrices shown in Figure 3.6. To emphasise that they parametrise different extreme value distributions, the matrices' columns are reordered so that $a_{11} < a_{12} < \ldots < a_{1q}$. Substituting these into (3.22) gives $\sigma_{12}(t) = 0.100$ and $\nu_{12}^2 = 0.060(t)$ for all $t \in [0, 1]$.

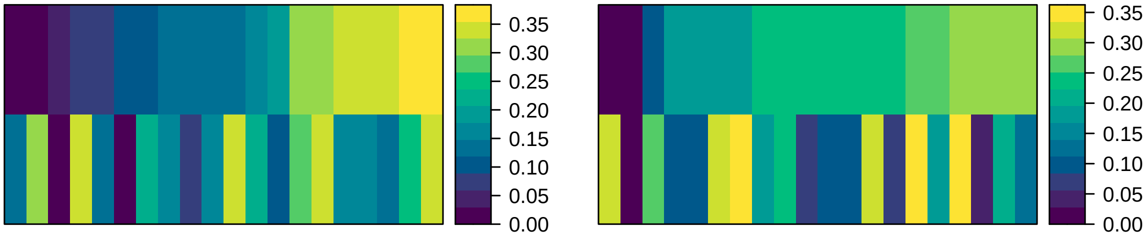

Figure 3.6: A pair of max-linear parameter matrices A_0 (left) and A_1 (right) such that the process (3.22) satisfies $\sigma_{12}(t) = 0.100$ and $\nu_{12}^2 = 0.060(t)$.

We generate 1,000 realisations of (3.22) with $n = 10,000$ and test for changing dependence using 25 blocks of size $b = 400$ and $k = 40$ extremes per block. The diagnostic plots in Figure 3.7 provide insight into what happens for one of these tests. The top-left plot shows that $\hat{\sigma}(t) \approx 0.8$ for all $t \in [0, 1]$, up to some random variation. Using (2.59) one can show that $\mathbb{P}(\hat{\sigma}_{12}(t) \in (0.724, 0.876)) \approx 0.95$. Indeed, the empirical coverage of this interval is 93.56%, based on all $1,000 \times 25 = 25,000$ estimates of $\sigma_{12}(t)$. The empirical integrated TPDM $\{\hat{\psi}_{12}(t) : t \in [0, 1]\}$ (top-right) is approximately a straight line and the test process $\{\hat{Z}_{12}(t) : t \in [0, 1]\}$ (bottom-left) resembles a typical Brownian bridge sample path. The bottom-right plot depicts $\int_0^t |\hat{Z}_{12}(s)|^2 \, \mathrm{d}s$ (upper sub-panel) and $\sup_{0\leq s\leq t} |\hat{Z}_{12}(s)|$ (lower sub-panel) as functions of t. The maxima of these processes are the CM and KS test statistics. Neither exceed the associated critical values at the 5% level, marked by the dashed lines. We conclude there is insufficient evidence to reject the null and commit a Type II error. The empirical Type II error rates across 1,000 repetitions of the experiment are 94.5% (CM) and 96.5% (KS). In other words, the rejection rate approximately equals the nominal size of the test, meaning the test has no power.

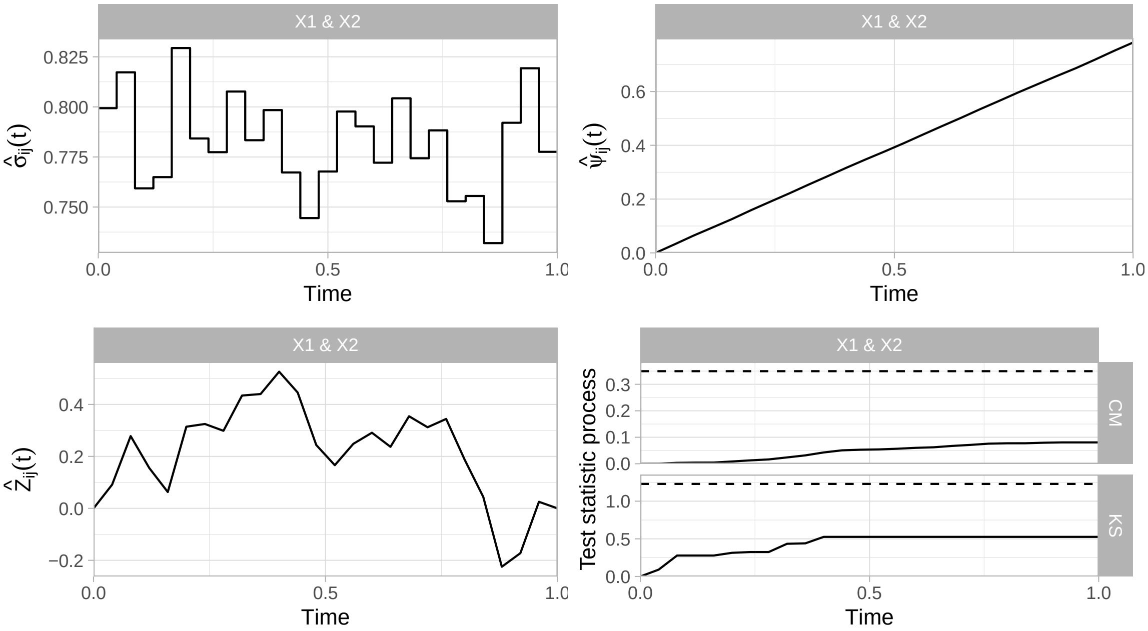

Figure 3.7: Diagnostic plots for our testing procedure based on one realisation of (3.22) with A_0 and A_1 as shown in Figure 3.6 and $n = 10,000$. The tuning parameters are $b = 400$ and $k = 40$. Top-left: the empirical local TPDM $\{\hat{\sigma}_{12}(t) : t \in [0,1]\}$. Top-right: the empirical integrated TPDM $\{\hat{\psi}_{12}(t) : t \in [0,1]\}$. Bottom-left: the test process $\{\hat{Z}_{12}(t) : t \in [0,1]\}$. Bottom-right: $\int_0^t |\hat{Z}_{12}(s)|^2 \,\mathrm{d}s$ (upper sub-panel) and $\sup_{0\leq s\leq t} |\hat{Z}_{12}(s)|$ (lower sub-panel) as functions of t along with the CM and KS critical values at the 5% level (dashed line).

Figure 3.8 presents analogous plots based on Drees' testing method applied to the same data. The left-hand plot shows the normalised empirical integrated angular measure $d^{-1}\hat{\mathrm{IH}}(A_y;t)$ as a function of t. Each curve corresponds to a particular set $A_y \in \mathcal{A}$ defined in (3.21) with darker colours indicating larger values of y. Close inspection of these curves reveals that some of them are slightly kinked at $t = 0.5$. The middle panel visualises the processes

$$\hat{Z}_y(t) := \sqrt{\frac{k}{2h}}(\widehat{\mathrm{IH}}(A_y;t) - t\widehat{\mathrm{IH}}(A_y;1)), \qquad (A_y \in \mathcal{A}). \tag{3.23}$$

These perform the same role as $\{\hat{Z}_{ij}(t) : t \in [0,1]\}$ but are less straightforward to interpret due to the presence of cross-correlation. The maxima (taken over all $A_y \in \mathcal{A}$) of the processes $\int_0^t |\hat{Z}_y(s)|^2 \,\mathrm{d}s$ (right, upper sub-panel) and $\sup_{0\leq s\leq t} |\hat{Z}_y(s)|$ (right, bottom sub-panel) are the CM and KS test statistics. These exceed the dashed lines marking the critical values (Drees 2023, Table 1). According to either test we (correctly) reject the null hypothesis at the 5% level. The empirical power based on 1,000 repeats is 100% (CM) and 99.8% (KS). Clearly the dependence change is easily detectable with the available data, emphasising that the deficiency of our test is purely methodological and not due to, say, a lack of data.

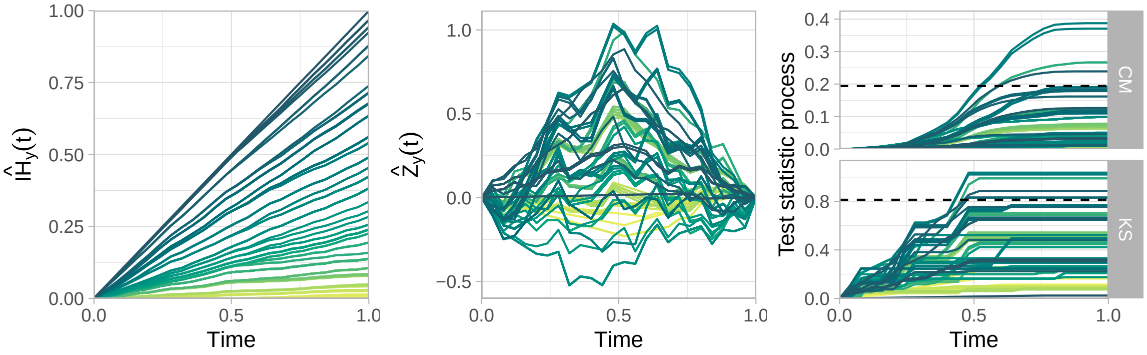

Figure 3.8: Diagnostic plots for Drees' testing procedure based on one realisation of (3.22) with A_0 and A_1 as shown in Figure 3.6 and $n = 10,000$. The tuning parameters are $b = 400$ and $k = 40$. Each curve represents a set A_y with darker colours indicating larger values of y. Left: the empirical integrated angular measure $\{\widehat{\mathrm{IH}}_{12}(A_y; t) : t \in [0, 1]\}$. Middle: the test process $\{\hat{Z}_y(t) : t \in [0, 1]\}$. Right: $\int_0^t |\hat{Z}_y(s)|^2 \,\mathrm{d}s$ (upper sub-panel) and $\sup_{0\leq s\leq t} |\hat{Z}_y(s)|$ (lower sub-panel) as functions of t along with the CM and KS critical values at the 5% level (dashed line).

3.7 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 (Castro-Camilo, Mhalla, et al. 2021; Rohrbeck, Emma S. Simpson, et al. 2021; Emma S. Simpson and Wadsworth 2021) having been the subject of the EVA (2019) Data Challenge (Raphaël Huser 2021). Further details about the data set and pre-processing can be found in Raphaël Huser (2021). *Mention removal of non-stationarity in the margins citing Castro-Camilo.*

Detecting changes in extremal dependence in the Red Sea is of significant practical importance. Increase dependence could lead to prolonged periods of elevated sea temperatures, exacerbating ecological issues such as coral bleaching and reducing the resilience of marine biodiversity. Red Sea surface temperatures are influenced by broader climate drivers, including the El Niño–Southern Oscillation (ENSO) (Karnauskas and Jones 2018). In a study of extreme precipitation, Jiang et al. (2020) found evidence for a positive temporal trend coefficients associated to principal eigenvectors related with ENSO. Moreover, Figure 3 in Kakampakou et al. (2024) shows increases in the tail dependence coefficient χ between pairs of sites based on data from the periods 1985-1989 and 2011-2015, especially in the north. These findings points towards the possibility of non-stationary tail dependence.

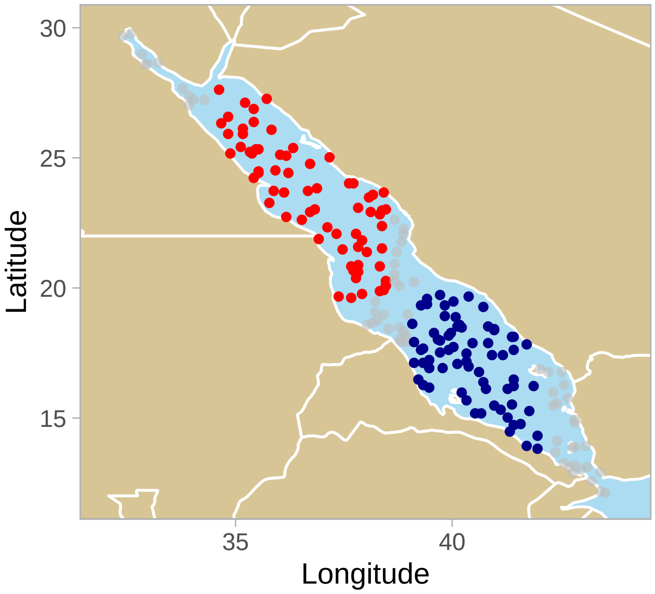

Figure 3.9: Locations of the 70 sites in each of the two sub-regions in the Red Sea.

Before applying our test, we divide the spatial domain into northerly and southerly sub-regions, each comprising 70 sites as shown in Figure 3.9. Emma S. Simpson and Wadsworth (2021) advise treating these areas separately because surface temperature extremes exhibit differing behaviour in the north and south. Additionally, they show that, at any particular location, high temperatures may persist across several days resulting in (temporal) clusters of extremes of daily maxima. To eliminate serial dependence we instead work with weekly maxima, yielding $n = 1605$ samples spanning approximately 31 years. Let $X_i^{(\text{north})}(t)$ and $X_i^{(\text{south})}(t)$ denote the surface temperature anomaly (on stationary 2-Fréchet margins) at site $i \in \{1, \ldots, 70\}$ and time $t \in [0, 1]$ in the two sub-regions. Our goal is to determine whether it is reasonable to assume stationary dependence for either/both of

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

To this end, we run the test using 15 blocks of size $b = 107$ and $k = 20$, yielding $k_{\text{total}} =$

$15 \times 20 = 300$. The rank condition (3.20) restricts us to testing up to 17 sites at a time, but even this seems excessive with only 20 extremes per block. Our strategy will be to repeatedly re-sample $2 \leq d \leq 17$ sites from each region and apply the test to these lower-dimensional data sets. The distribution of p-values across $1,000$ repeats with $d \in \{2, 5, 10\}$ are shown in Figure 3.10. The columns correspond to different numbers of re-sampled sites. The rows indicate the sub-region and the test type. The value printed at the top of each panel is the proportion of tests that are rejected at the 5% level. If dependence is constant (resp. time-varying) across the sub-region, then the distribution of p-values is expected to be approximately uniform (resp. positively skewed). The evidence for non-stationarity is fairly strong in the north (average rejection rate of 48.6%) and comparatively weaker in the south (20.1%). This aligns with the findings in Kakampakou et al. (2024). The CM test rejects the null much more frequently than the KS test. Our simulation studies suggested that CM tends to be superior when the dependence change is gradual, as is likely to be the case here. For all tests and regions, the rejection rate is highest when $d = 5$, and drops off when d is reduced or increased. We believe this reflects the trade-off between two factors. On the one hand, taking a large pool of sites increases the chance that among them there exists at least one pair with time-varying dependence. On the other hand, the local TPDM estimates become noisier, potentially masking any temporal trends.

3.8 Extensions and outlook

3.8.1 Change-point detection

In certain applications (e.g. finance), it may be more interesting to ask *when*, not if, dependence has changed. This is the realm of change-point detection. *How much should I say here? Mention that this is being actively pursued with colloborators?*

3.8.2 Mitigating the bias issue

The bias issue means that the empirical TPDM struggles to discriminate between weak and very weak dependence (Figure 2.5). Therefore power probably worse when dependence is weak. Fix this by using bias-corrected estimator.

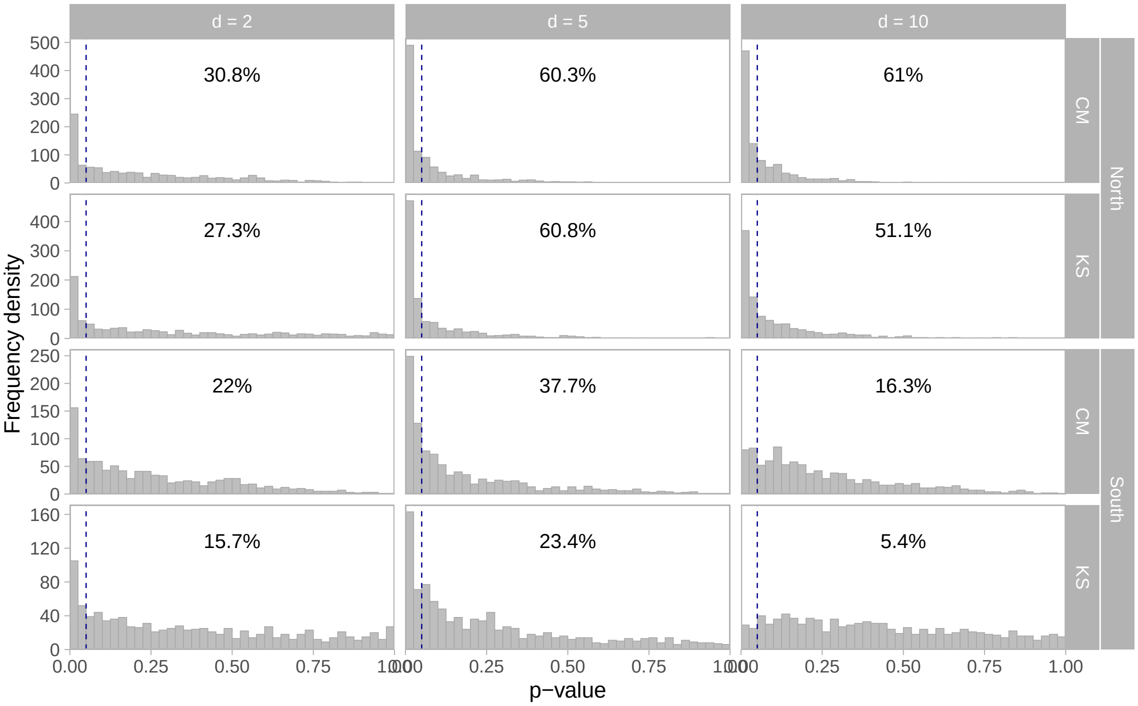

Figure 3.10: Empirical distributions of the p-values based on tests for changing dependence in the north and south regions of the Red Sea. The columns correspond to the number of re-sampled sites; the rows indicate the sub-region and the test type. The rejection rate at the 5% level is printed at the top of each panel.

3.8.3 Alternative dependence measures

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

$$\sigma_{ij}(t) = \mathbb{E}_H(\cdot;t)[f(\boldsymbol{\Theta}(t))], \tag{3.24}$$

where $f : \mathbb{S}_{+(2)}^{d-1} \to \mathbb{R}_+$ is set as $f(\boldsymbol{\theta}) = \theta_i\theta_j$. However, alternative dependence measures generated by some other function $g : \mathbb{S}_+^{d-1} \to \mathbb{R}_+$ could be used instead, provided they are continuous and bounded so that the necessary asymptotic theory holds. *Speculate as to what would happen?*

3.8.4 Robustness to change in serial dependence

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

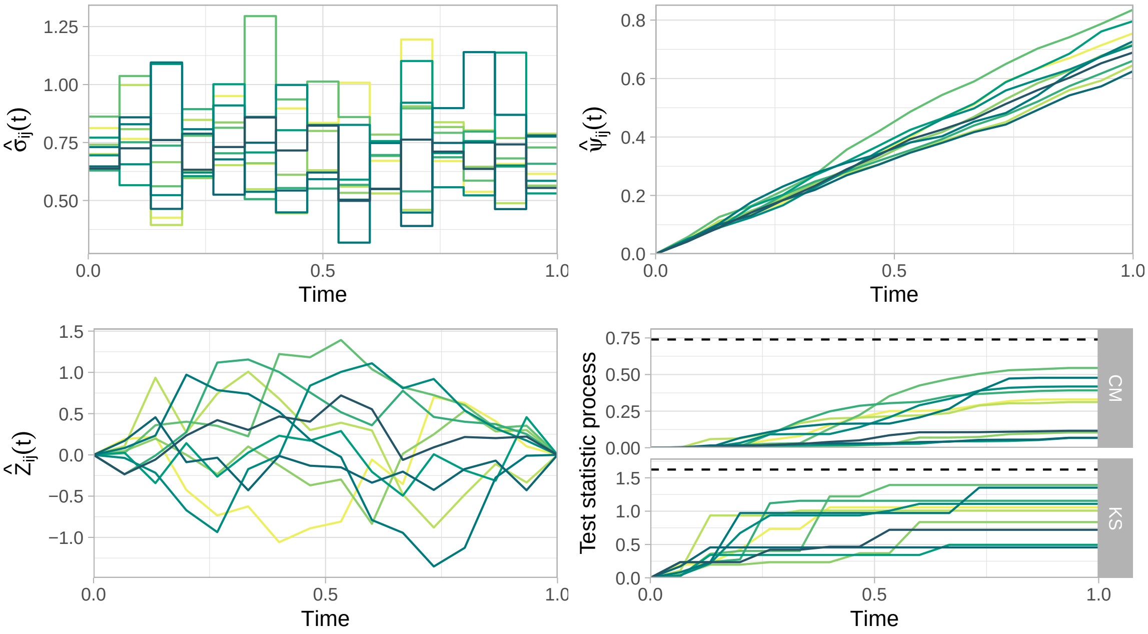

Figure 3.11: Diagnostic plots for our test, based $b = 107$ and $k = 20$, applied to data from $d = 5$ randomly selected northerly sites in the Red Sea. The interpretation of each plot is the same as in Figure 3.7, except now there are $\mathcal{D} = 10$ curves, one for each component pair. The variable pairs are coloured light to dark with respect to their lexicographical ordering. In this example, there is sufficient evidence to reject the null at the 5% level.

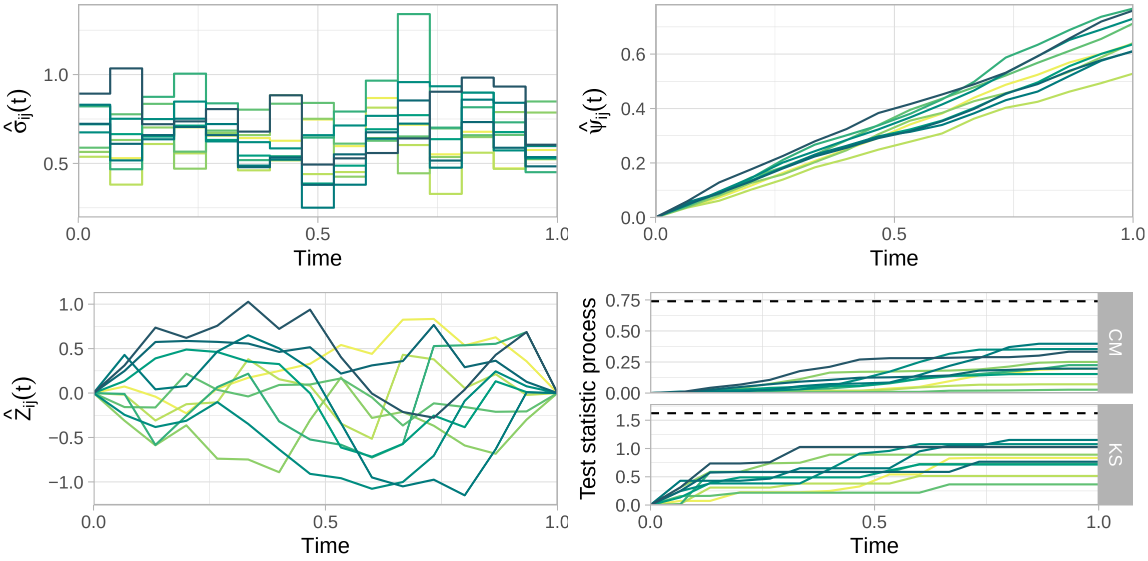

Figure 3.12: Diagnostic plots for our test, based $b = 107$ and $k = 20$, applied to data from $d = 5$ randomly selected southerly sites in the Red Sea. The interpretation of each plot is the same as in Figure 3.7, except now there are $\mathcal{D} = 10$ curves, one for each component pair. The variable pairs are coloured light to dark with respect to their lexicographical ordering. In this example, there is insufficient evidence to reject the null at the 5% level.

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

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

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

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

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

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

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

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

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

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

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

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

$$\operatorname{cen}_a(\boldsymbol{X}) := \underset{y \in \mathbb{S}_+^{d-1}}{\arg\min}\, \mathbb{E}[d_a^2(\boldsymbol{X}, \boldsymbol{y})] = \mathcal{C}(\exp(\mathbb{E}[\log(\boldsymbol{X})]),$$
$$\operatorname{totVar}_a(\boldsymbol{X}) := \mathbb{E}[d_a^2(\boldsymbol{X}, \operatorname{cen}_a(\boldsymbol{X}))] = \sum_{j=1}^{d} \operatorname{Var}([\operatorname{clr}(\boldsymbol{X})]_j).$$

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 $\boldsymbol{X} = (X_1, \ldots, X_d)$ denote a d-part centred random composition with finite second order moments, that is $\operatorname{cen}_a(\boldsymbol{X}) = \boldsymbol{e}_a$ and $\mathbb{E}[\|\boldsymbol{X}\|_a^2] < \infty$. Let $\Gamma = \operatorname{Cov}(\operatorname{clr}(\boldsymbol{X}))$ 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:

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

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

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 (B.2) on an unseen validation set $\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_{n^\star}^\star$ 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.

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

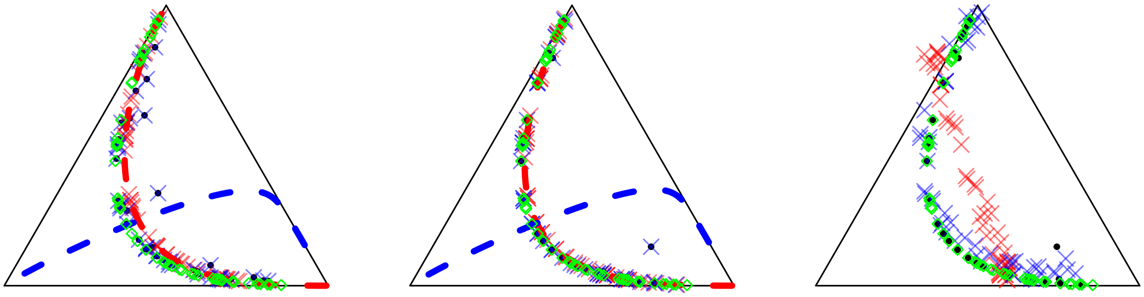

Figure 4.1: Examples of CoDA-PCA (left and middle) and DS-PCA (right) applied to RV- (left) and MS-max-linear (middle and right) data. The green diamonds represent the true angular measure. The black points are the angular components associated with the $k = 50$ largest observations among a sample of 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.

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. (B.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!

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

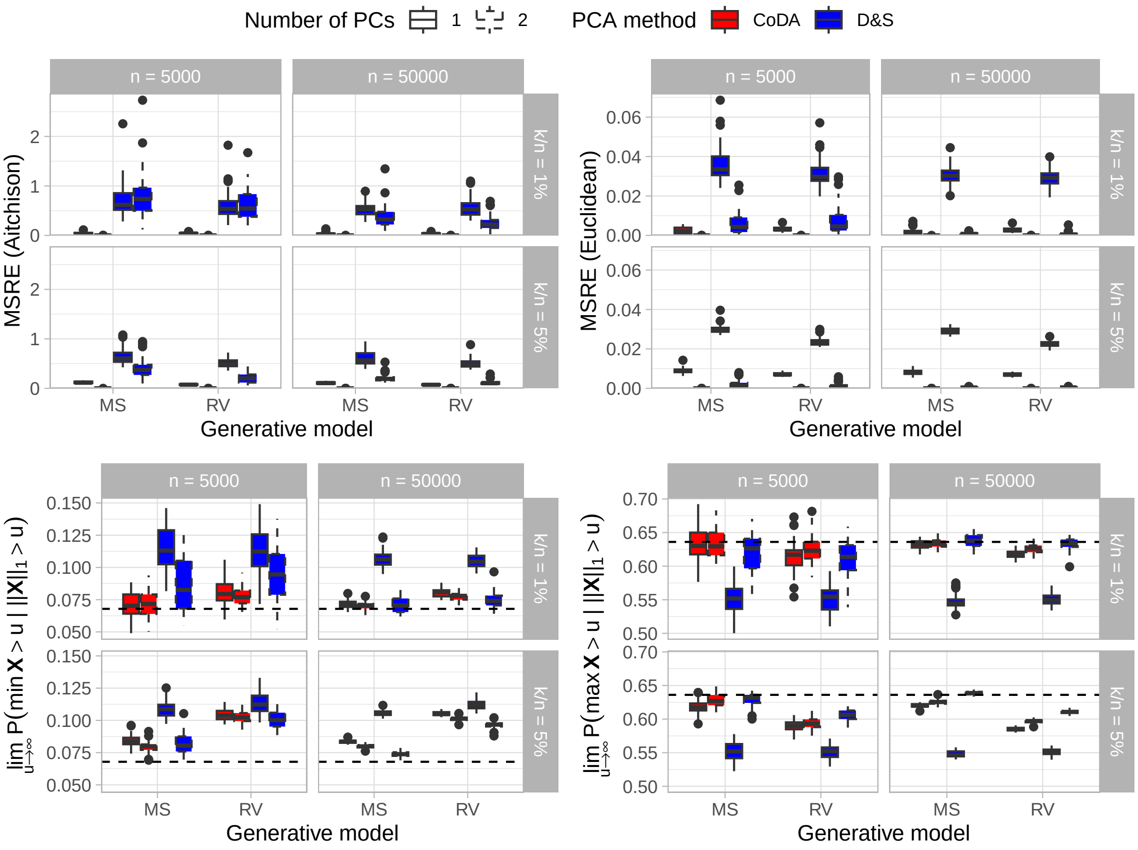

Figure 4.2: PCA performance metrics based on trivariate max-linear data.

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.

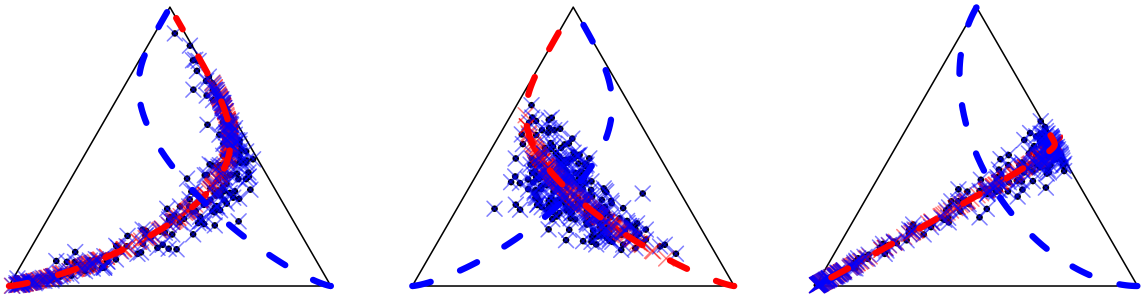

Figure 4.3: Example data from the three trivariate Hüsler-Reiss models. Based on $n = 10^4$ and $k = 250$.

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.

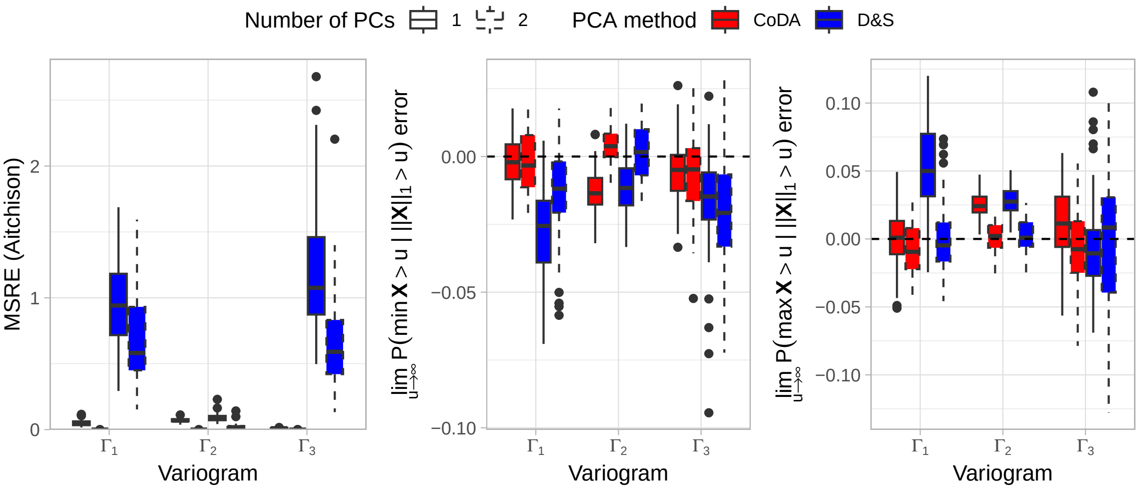

Figure 4.4: PCA performance metrics based on trivariate Hüsler-Reiss data.

4.2.3.3 Hüsler-Reiss model in high dimensions

For this experiment, the random vector $\boldsymbol{X} = (X_1, \ldots, X_{10})$ 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 $\{X_1, \ldots, X_4\}$ and $\{X_9, X_{10}\}$, while the pairwise dependence strengths between the components in $\{X_5, \ldots, X_8\}$ 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).

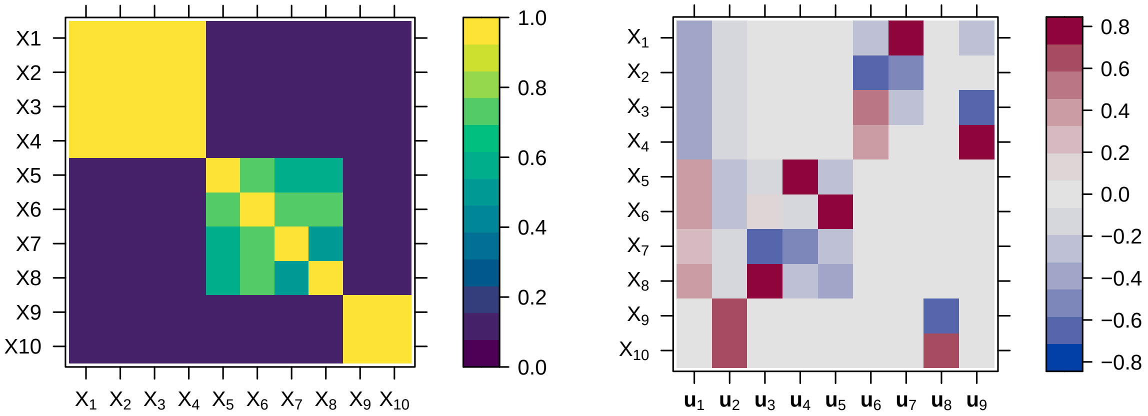

Figure 4.5: Matrix of tail dependence coefficients $\chi_{ij} = 2\bar{\Phi}(\sqrt{\Gamma_{ij}}/2)$ (left) and a matrix of CoDA-PCA sample eigenvectors (right).

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.

4.2.4 Discussion

Discuss conclusions of CoDA PCA stuff here.

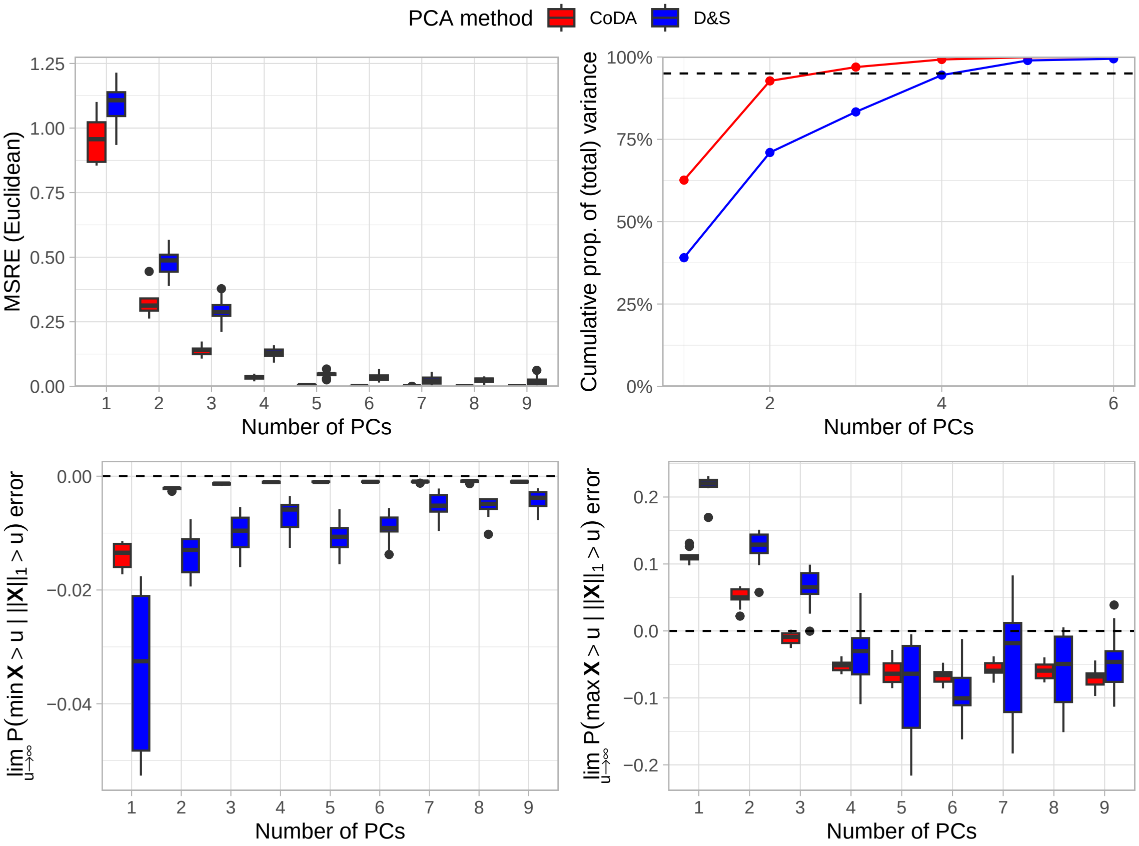

Figure 4.6: PCA performance metrics based on 10-dimensional Hüsler-Reiss data.

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

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

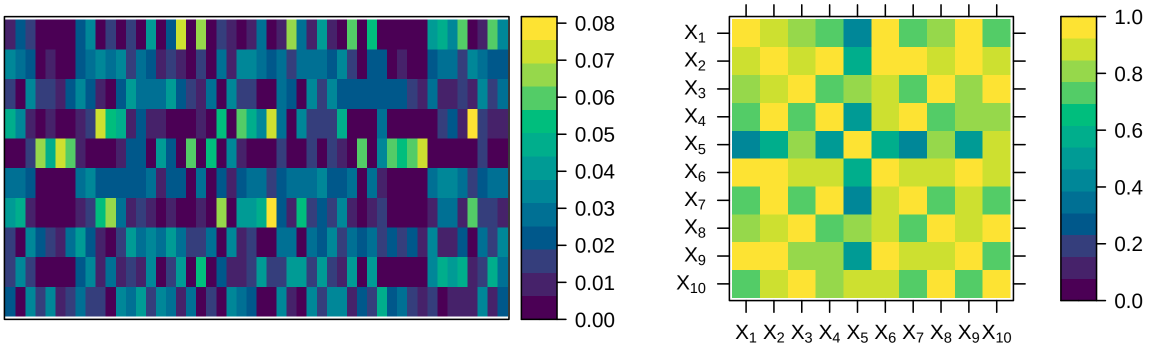

Figure 4.7: Parameter matrix A (left) and corresponding model TPDM (right) for 10-dimensional max-linear data.

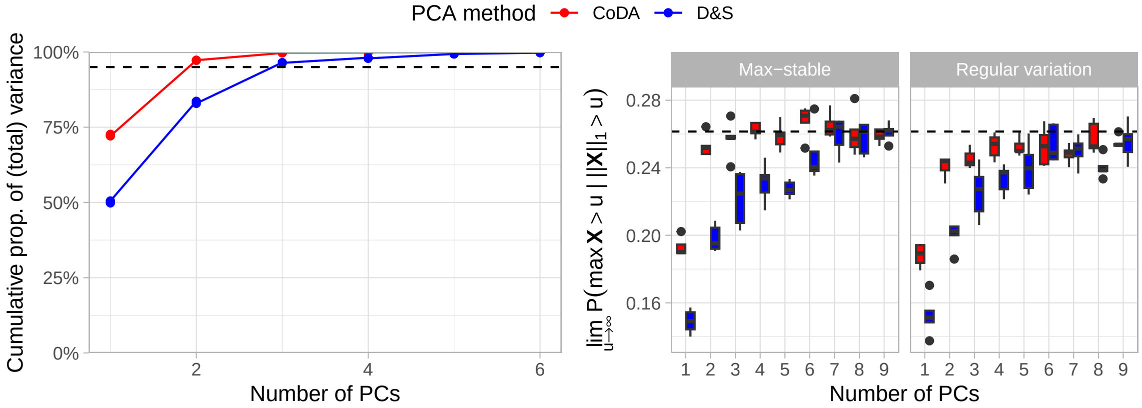

Figure 4.8: PCA performance metrics based on 10-dimensional max-linear data.

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.

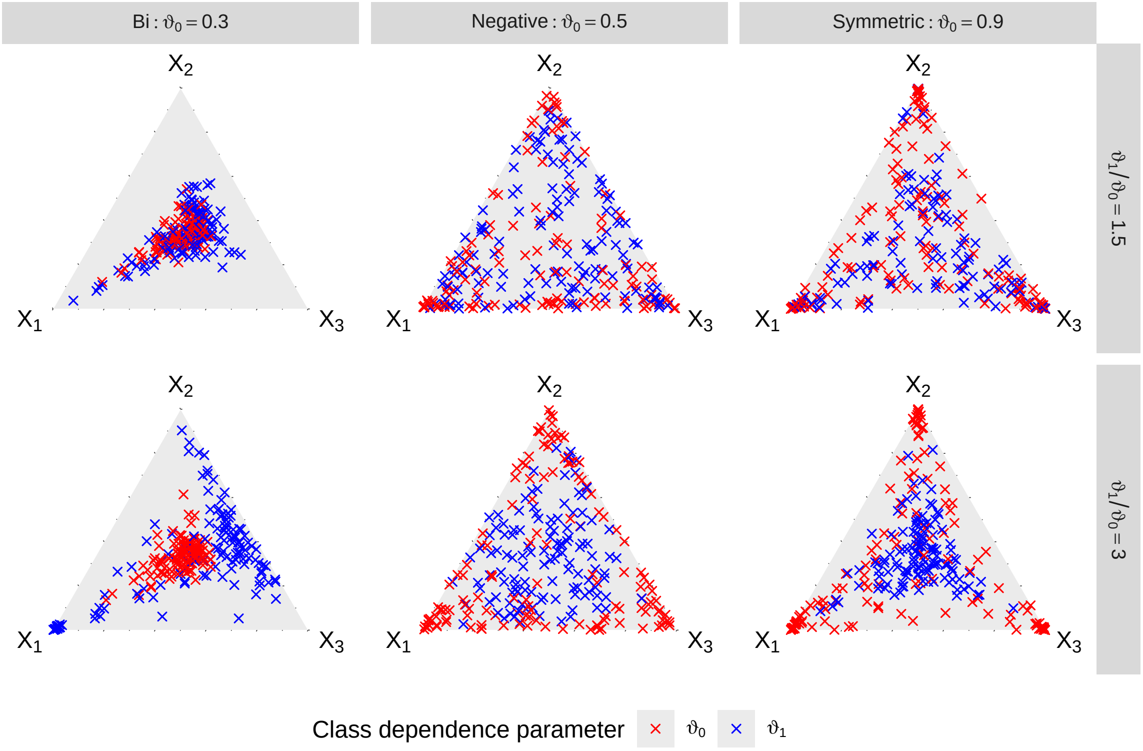

Figure 4.9: Blah.

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 $\alpha \in \{0, 0.1, \ldots, 0.9, 1\}$, 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 $\mathcal{G}_{\boldsymbol{\Theta}} = \{g_\psi : \psi \in \Psi\}$ be a family of simplicial classifiers $g_\psi : \mathbb{S}_+^{d-1} \to \{-1, +1\}$ parametrised by $\psi \in \Psi$. For example, if $\mathcal{G}_{\boldsymbol{\Theta}}$ represents the k-NN(α) class with fixed

α, then $\psi = k$ and $\Psi = \mathbb{N}$.

2. For a given training set $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$, let $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)}$ be the angular components of the k largest observations and $y_{(1)}, \ldots, y_{(k)}$ their associated classes. Let t denote the implicit threshold, i.e. the $k+1$ order statistic of $\{\|\boldsymbol{x}_i\|_1 : i = 1, \ldots, n\}$
3. Partition $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)}$ into $J = 10$ balanced folds by stratified sampling.
4. For $j = 1, \ldots, J$ 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.

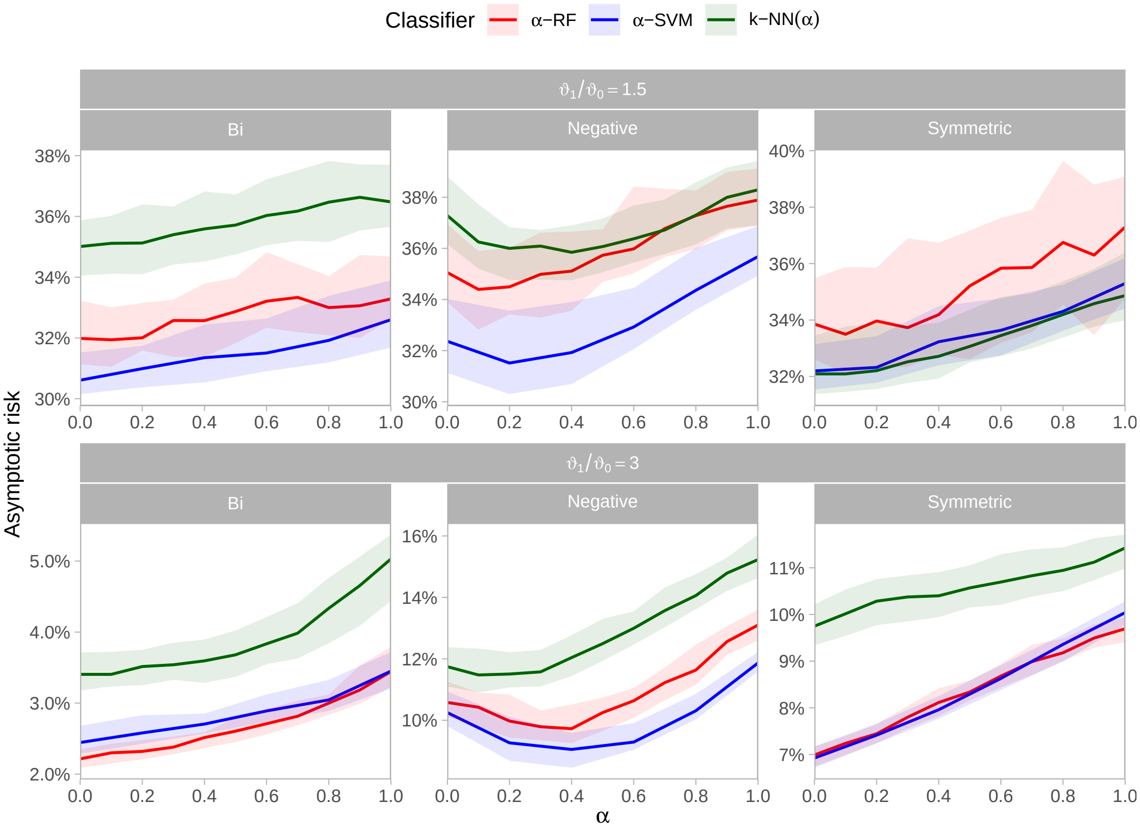

Figure 4.10: Blah.

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.

This hypothesis is borne out by Figure 4.12, which plots the median values of the 'training

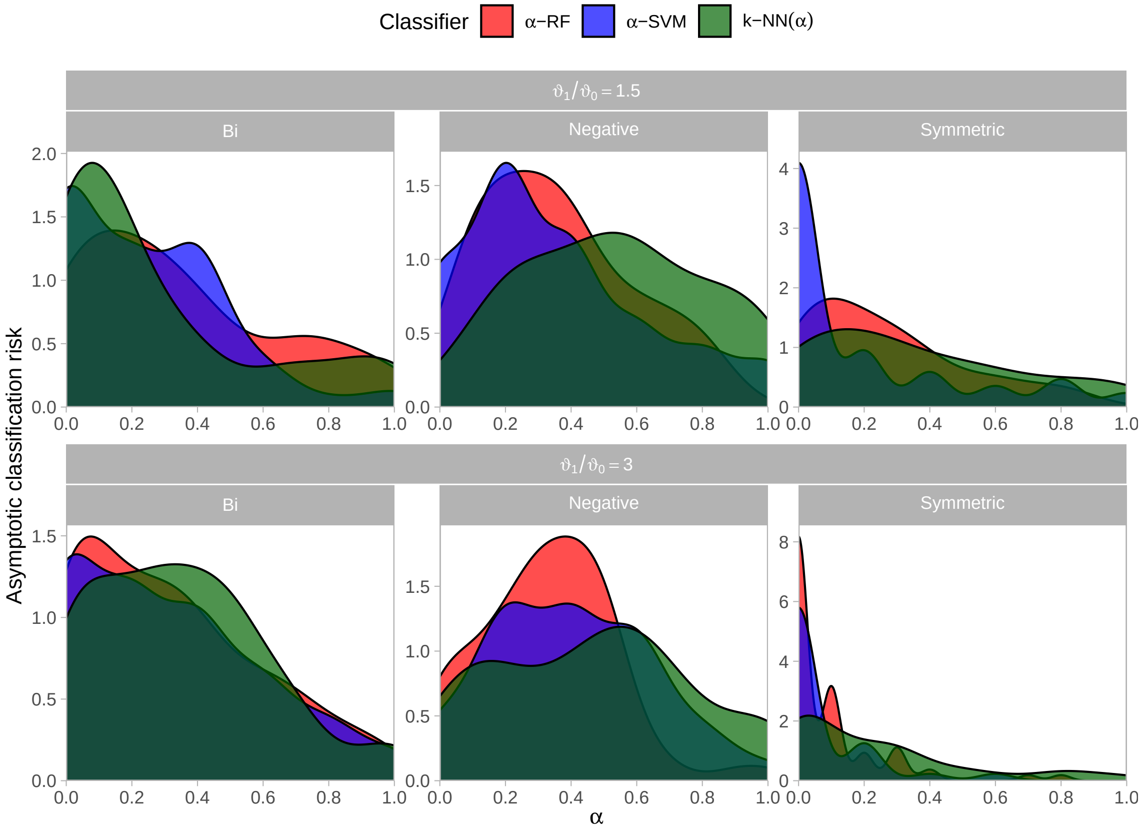

Figure 4.11: Blah.

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?*

4.3.3 Discussion

Discuss conclusions of CoDA classification stuff here.

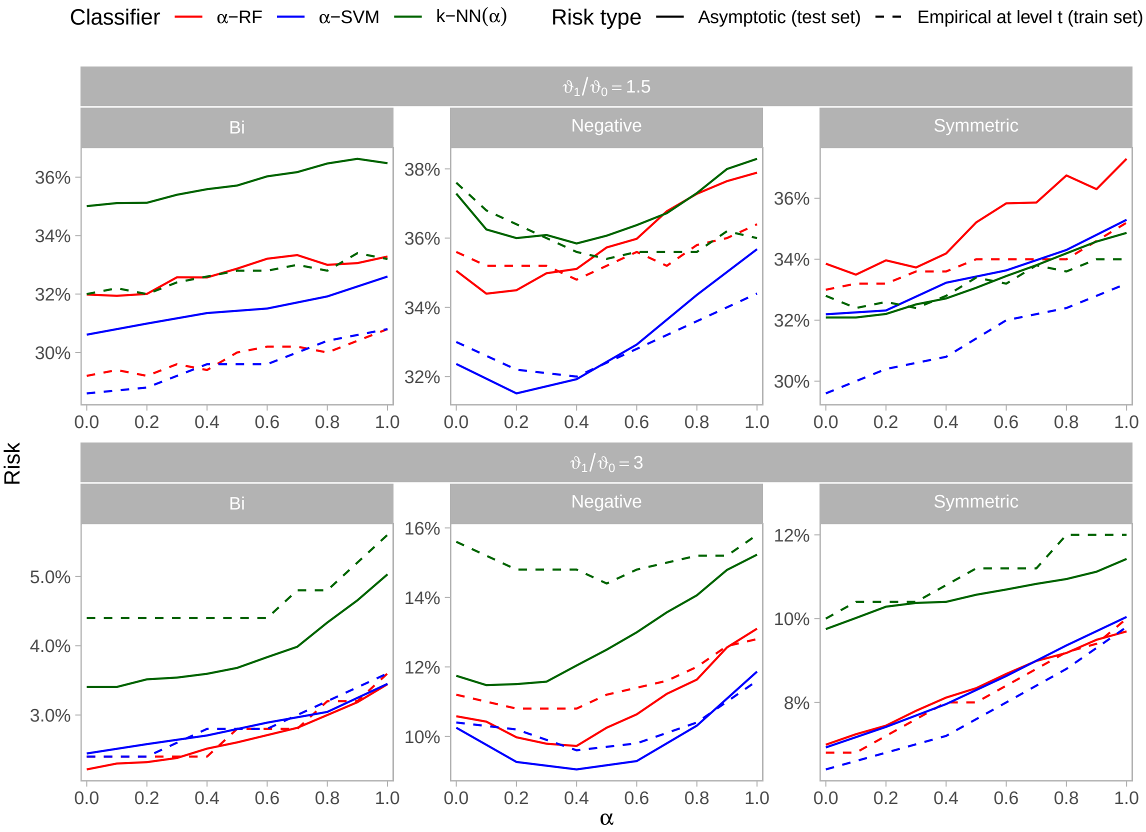

Figure 4.12: Blah.

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.*

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.

Avella-Medina, Marco, Richard A. Davis, and Gennady Samorodnitsky (2022). *Kernel PCA for Multivariate Extremes.* URL: http://arxiv.org/abs/2211.13172 (visited on 10/21/2024). Pre-published.

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

Blanchard, Gilles, Olivier Bousquet, and Laurent Zwald (2007). “Statistical Properties of Kernel Principal Component Analysis”. In: *Machine Learning* 66.2-3, pp. 259–294.

Boulaguiem, Younes et al. (2022). “Modeling and Simulating Spatial Extremes by Combining Extreme Value Theory with Generative Adversarial Networks”. In: *Environmental Data Science* 1, e5.

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

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

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.

Castro-Camilo, Daniela, Linda Mhalla, and Thomas Opitz (2021). “Bayesian Space-Time Gap Filling for Inference on Extreme Hot-Spots: An Application to Red Sea Surface Temperatures”. In: *Extremes* 24.1, pp. 105–128.

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

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

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

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.

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

De Carvalho, Miguel and Anthony C. Davison (2014). “Spectral Density Ratio Models for Multivariate Extremes”. In: *Journal of the American Statistical Association* 109.506, pp. 764–776.

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

Dobrić, Jadran and Friedrich Schmid (2005). “Nonparametric Estimation of the Lower Tail Dependence $_L$ in Bivariate Copulas”. In: *Journal of Applied Statistics* 32.4, pp. 387–407.

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

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

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

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

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

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

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

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.

Gadeikis, K. and V. Paulauskas (2005). “On the Estimation of a Changepoint in a Tail Index”. In: *Lithuanian Mathematical Journal* 45.3, pp. 272–283.

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

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

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

Goix, Nicolas, Anne Sabourin, and Stephan Clémençon (2017). “Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection”. In: *Journal of Multivariate Analysis* 161, pp. 12–31.

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

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.

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

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

Haufmann, Torkel A (2011). “On Completely Positive Matrices”. MA thesis. Oslo: University of Oslo.

Henze, Norbert (2024). *Asymptotic Stochastics: An Introduction with a View towards Statistics*. Vol. 10. Mathematics Study Resources. Berlin, Heidelberg: Springer Berlin Heidelberg.

Higham, N. J. (2002). “Computing the Nearest Correlation Matrix–a Problem from Finance”. In: *IMA Journal of Numerical Analysis* 22.3, pp. 329–343.

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

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

Huser, Raphaël (2021). “Editorial: EVA 2019 Data Competition on Spatio-Temporal Prediction of Red Sea Surface Temperature Extremes”. In: *Extremes* 24.1, pp. 91–104.

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

Hüsler, Jürg and Rolf-Dieter Reiss (1989). “Maxima of Normal Random Vectors: Between Independence and Complete Dependence”. In: *Statistics & Probability Letters* 7.4, pp. 283–286.

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

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

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

Jiang, Yujing, Daniel Cooley, and Michael F. Wehner (2020). “Principal Component Analysis for Extremes and Application to U.S. Precipitation”. In: *Journal of Climate* 33.15, pp. 6441–6451.

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

Jolliffe, Ian (2002). *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag.

Kakampakou, Lydia, Emma S. Simpson, and Jennifer L. Wadsworth (2024). *Spatial Extremal Modelling: A Case Study on the Interplay between Margins and Dependence*. URL: http://arxiv.org/abs/2409.16373 (visited on 10/26/2024). Pre-published.

Karnauskas, Kristopher B. and Burton H. Jones (2018). “The Interannual Variability of Sea Surface Temperature in the Red Sea From 35 Years of Satellite and In Situ Observations”. In: *Journal of Geophysical Research: Oceans* 123.8, pp. 5824–5841.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

Rohrbeck, Christian, Emma S. Simpson, and Ross P. Towe (2021). "A Spatio-Temporal Model for Red Sea Surface Temperature Anomalies". In: *Extremes* 24.1, pp. 129–144.

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.

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

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

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

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

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

Shaked-Monderer, Naomi (2020). *On the Number of CP Factorizations of a Completely Positive Matrix.* URL: http://arxiv.org/abs/2009.12290 (visited on 10/21/2024). Prepublished.

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

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

Simpson, Emma S. and Jennifer L. Wadsworth (2021). "Conditional Modelling of Spatio-Temporal Extremes for Red Sea Surface Temperatures". In: *Spatial Statistics* 41, p. 100482.

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

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

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

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

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.

Wixson, Troy P. and Daniel Cooley (2023). “Attribution of Seasonal Wildfire Risk to Changes in Climate: A Statistical Extremes Approach”. In: *Journal of Applied Meteorology and Climatology* 62.11, pp. 1511–1521.

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

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

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

A Properties of the TPDM

A.1 Equivalence of TPDM definitions

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

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

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

Then

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

□

Extra things to find a place for:

Symmetric logistic angular density:

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

Hüsler-Reiss angular density:

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

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

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

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

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

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

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

A.3 Proof of Proposition 2.6

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

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

where

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

The corresponding empirical estimator is

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

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

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

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

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

□

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

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

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

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

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

and therefore

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

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

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

B PCA in general finite-dimensional Hilbert spaces

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

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

We present PCA for random vectors in $\mathbb{R}^d$, with the understanding that the data may have undergone an isometric transformation in pre-processing and outputs may need to be back-transformed to lie in the original space. This transform/back-transform approach is equivalent to conducting the analysis in the original space with appropriately generalised notions of mean, variance, etc. (Pawlowsky-Glahn and Egozcue 2001).

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

$\mathcal{H}$	$\mathbb{R}^d$	$\mathbb{R}^d_+$	$\mathbb{S}^{d-1}_{+(1)}$
$h : \mathcal{H} \to \mathbb{R}^d$	$h(\boldsymbol{x}) = \boldsymbol{x}$	$h(\boldsymbol{x}) = \tau^{-1}(\boldsymbol{x}) = \log[\exp(\boldsymbol{x}) - 1]$	$h(\boldsymbol{x}) = \operatorname{clr}(\boldsymbol{x}) = \log[\boldsymbol{x}/\bar{g}(\boldsymbol{x})]$
$h^{-1} : \mathbb{R}^d \to \mathcal{H}$	$h^{-1}(\boldsymbol{y}) = \boldsymbol{y}$	$h^{-1}(\boldsymbol{y}) = \tau(\boldsymbol{y}) = \log[1 + \exp(\boldsymbol{y})]$	$h^{-1}(\boldsymbol{y}) = \operatorname{clr}^{-1}(\boldsymbol{y}) = \mathcal{C}\exp(\boldsymbol{y})$
$\boldsymbol{x} \oplus \boldsymbol{y}$	$\boldsymbol{x} + \boldsymbol{y}$	$\tau[\tau^{-1}(\boldsymbol{x}) + \tau^{-1}(\boldsymbol{y})]$	$\mathcal{C}(x_1y_1, \ldots, x_dy_d)$
$\alpha \odot \boldsymbol{x}$	$\alpha\boldsymbol{x}$	$\tau[\alpha\tau^{-1}(\boldsymbol{x})]$	$\mathcal{C}(x_1^\alpha, \ldots, x_d^\alpha)$
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{H}}$	$\sum_{i=1}^d x_iy_i$	$\sum_{i=1}^d \tau^{-1}(x_i)\tau^{-1}(y_i)$	$\sum_{i=1}^d \log[x_i/\bar{g}(\boldsymbol{x})] \log[y_i/\bar{g}(\boldsymbol{x})]$

is measured as

$$R(\mathcal{S}) := \mathbb{E}[\|\boldsymbol{Y} - \Pi_{\mathcal{S}}\boldsymbol{Y}\|_2^2] \tag{B.1}$$

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

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

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

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

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

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

produces the optimal p-dimensional projection of $\boldsymbol{Y}$. In other words, the subspace $\mathcal{S}_p = \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$ minimises the criterion (B.1) over $\mathcal{V}_p$, the set of all linear subspaces of dimension p of $\mathbb{R}^d$. It is the unique minimiser provided the multiplicity of λ_p is one. The corresponding risk is determined by the eigenvalues of the discarded components via $R(\mathcal{S}_p) = \sum_{j>p} \lambda_j$.

In practice, the covariance matrix is unknown so (B.1) cannot be minimised directly. Instead we resort to an empirical risk minimisation (ERM) approach, whereby the risk is replaced by

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

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

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

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

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

where $\hat{V}_{ij} := \langle \boldsymbol{Y}_i, \boldsymbol{u}_j \rangle$. The subspace $\hat{\mathcal{S}}_p = \mathrm{span}\{\hat{\boldsymbol{u}}_1, \dots, \hat{\boldsymbol{u}}_p\}$ minimises (B.2) in $\mathcal{V}_p$; the objective at the minimum is $\hat{R}(\hat{\mathcal{S}}_p) = \sum_{j>p} \hat{\lambda}_j$.

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

If $\boldsymbol{Y}$ is mean-zero (or the $n \times d$ data matrix is column-centred in pre-processing), then Σ is the covariance matrix of $\boldsymbol{Y}$ and the procedure is termed centred PCA. In this case, PCA can be equivalently reformulated in terms of finding low-dimensional projections that maximally preserve variance. In the non-centred case this interpretation is not valid, the projections merely maximise variability around the origin. A detailed comparison between centred PCA and non-centred PCA is conducted in Cadima and Jolliffe (2009). They obtain relationships between and bounds on the eigenvectors/eigenvalues of the non-centred and standard covariance matrices. Based on their theoretical analysis and a series of example, they conclude that both types of PCA generally produce similar results. In particular, the leading eigenvector (up to sign and scaling) of the non-centred covariance matrix is very often close to the vector of the column means of the data matrix. Thus the first non-centred principal component essentially relates to the centre of the data.

C Applications – original write up, can be removed

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

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

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

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

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

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

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

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

Rohrbeck and Cooley (2023) move beyond the use of the extremal PCA for purely exploratory purposes and demonstrate how it be used to generate synthetic extreme events. Hazard event sets are widely used in catastrophe modelling to assess exposure to extreme events. Imagine an insurance company insures against damage to a portfolio of properties, and wishes to gauge its exposure to claims caused by flooding. Given (i) the spatial locations of these properties, (ii) other relevant characteristics such as property value and construction standard, and (iii) a set of simulated flood events, one can derive a probabilistic loss distribution. If the exposure is unacceptably high, they might adjust their underwriting strategy or purchase reinsurance. Rohrbeck and Cooley (2023) show how to generate approximate samples from H, even in high-dimensions, by leveraging the PCA method of Cooley and Thibaud (2019). Their generative framework hinges on the fact that the leading components of $\boldsymbol{V}$ account for the greatest proportion of extremal behaviour of $\boldsymbol{X}$. Thus, efforts may be concentrated towards modelling the dependence structure of the sub-vector $(V_1,\ldots,V_p)$ for some appropriately chosen $p<d$. To achieve this, they use a spherical kernel density estimate to flexibly model the dependence between $V_1,\ldots,V_p$ and additionally between $(V_1,\ldots,V_p)$ and $(V_{p+1},\ldots,V_d)$. The dependence structure of $(V_{p+1},\ldots,V_d)$ is simply modelled by a nearest-neighbours approach. The number of components p entering into the complex model is selected by a leave-one-out cross validation procedure. This involves discarding an extreme observation $\boldsymbol{x}_{(i)}$, generating a large number of samples $\tilde{\boldsymbol{x}}_1^{[p]},\ldots,\tilde{\boldsymbol{x}}_N^{[p]}$ for a range of values p, and then assessing whether any of the generated samples resemble the discarded event using

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

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

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

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

D Review of clustering methods based on the TPDM

Within multivariate extremes, the umbrella term ‘clustering’ has many meanings. To avoid confusion, we briefly describe these and clarify which type we are referring to.

- **Prototypical events.** Assume that the angular measure concentrates at/near a small number of points in $\mathbb{S}_+^{d-1}$. Then one might wish to identify cluster centres $\boldsymbol{w}_1, \dots \boldsymbol{w}_K$ minimising some objective function of the form

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

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

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

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

  Only those groups (‘clusters’) of components indexed by $\beta_1, \dots, \beta_K$ may be simultaneously extreme. Identification of the support of the angular measure is notoriously challenging because the extremal angles $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$ lie (almost surely) in the interior of the simplex. Goix et al. (2017) and E S Simpson et al. (2020) identify clusters according to whether observations fall within appropriately sized rectangular/conic neighbourhoods of the corresponding axis in $\mathbb{R}_+^d$. Meyer and Wintenberger (2020) take a different approach, whereby the angular component is defined with respect to the Euclidean projection (Liu and Ye 2009) rather than usual projection based on self-normalisation. The geometry of the projection is such that the projected data lie on subfaces of the simplex. The price paid is that the limiting conditional distribution of the angles is related to, but not identical to, the angular measure.

- **Partitioning into AD/AI groups components.** This notion of clustering is related to the previous type. We assume that the variables $X_1, \ldots, X_d$ can be partitioned into K clusters, such that X_i and X_j are asymptotically dependent if and only if they belong to the same cluster. In other words, there exists $2 \leq K \leq d$ and a partition $\beta_1, \ldots, \beta_K$ of $\{1, \ldots, d\}$ such that the angular measure is supported on $C_{\beta_1}, \ldots, C_{\beta_K}$ or lower-dimensional subspaces thereof, i.e.

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

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

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

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

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

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

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

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

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

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

E Summary of Drees (2023)

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

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

Here $\hat{H}(A;s)$ denotes a non-parametric estimate of the angular measure $H(A;s)$ at time $s\in[0,1]$ – see (**??**) for a formal definition. The null is rejected if any paths in (E.1) 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 **?@sec-constant-tpdm**.

F Empirical power for different tuning parameters b and k

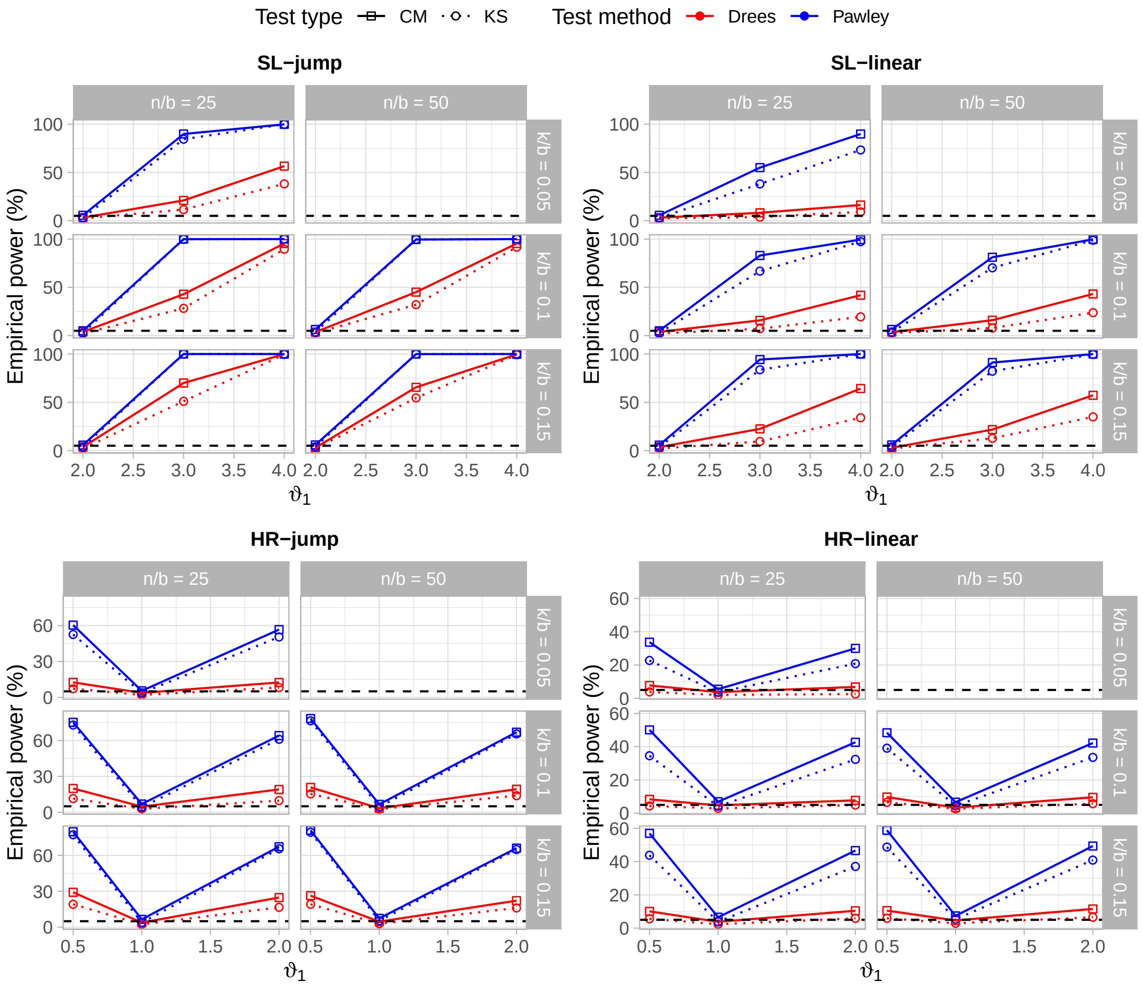

Figure F.1: Empirical power (%) as a function of the dependence parameter ϑ_1 for different combinations of tuning parameters b and k. Based on 1000 simulations with $n = 2,500$ and $d = 2$. For the SL and HR models, $\vartheta_0 = 2$ and $\vartheta_0 = 1$, respectively. Tests are conducted at the 5% level (black dashed line).