Extensions and Applications of the Tail Pairwise Dependence Matrix

Matthew Pawley

November 17, 2024

Table of contents

List of Figures

List of Tables

Preface

Draft thesis of Matthew Pawley, created on November 17, 2024.

1 Background & literature review

1.1 Univariate extreme value theory

1.1.1 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

$$M_n := \max(X_1, \ldots, X_n) = \bigvee_{i=1}^{n} X_i. \tag{1.1}$$

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

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} G(x), \qquad (n \to \infty). \tag{1.2}$$

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

When (1.2) 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

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \tag{1.3}$$

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 (1.2) 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 1.1. 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.

1.1.2 Threshold exceedances and the generalised Pareto distribution (GPD)

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

$$\lim_{u\to\infty} \mathbb{P}(X - u > x \mid X > u) = \left[1 + \frac{\xi x}{\tilde{\sigma}}\right]_+^{-1/\xi}, \qquad (x > 0), \tag{1.4}$$

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).

1.1.3 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. *Mention and cite evgam.*

1.2 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.

1.2.1 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

$$\mathbb{P}\left(\frac{\boldsymbol{M}_n - \boldsymbol{b}_n}{\boldsymbol{a}_n} \leq \boldsymbol{x} \right) \xrightarrow{d} G(\boldsymbol{x}), \qquad (n \to \infty). \tag{1.5}$$

Applying Theorem 1.1 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.

1.2.2 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 1.2. *Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ has joint distribution function F and continuous marginal distributions $X_i \sim F_i$ for $i = 1, \ldots, d$. Then there exists a unique copula C such that*

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{1.6}$$

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 (**resnickHeavytailPhenomenaProbabilistic2007**).

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

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

$$X_i \mapsto F_\star^{-1}(F_i(X_i)) \sim F_\star, \qquad (i = 1, \ldots, d). \tag{1.7}$$

If the marginal distributions are unknown, as is usually the case, then F_i is replaced with some estimate $\hat{F}_i$ in (1.7). 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 **russellAnalyzingDependenceMatrices2018** and **rohrbeckSimulatingFloodEvent2023**, respectively. Throughout this thesis, uncertainty arising from estimation of the marginal distributions shall be neglected. Relaxing this assumption, as in **clemenconConcentrationBoundsEmpirical2023**, represents an avenue for future work.

1.2.3 The exponent measure and angular measure

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

$$\mathbb{P}(X_i < x) = \exp(-1/x), \qquad (x > 0), \tag{1.8}$$

for $i = 1, \ldots, d$. This corresponds to a GEV distribution with $\mu = \sigma = \xi = 1$. The joint distribution G in (1.5) may be rewritten in the form

$$G(\boldsymbol{x}) = \exp(-V(\boldsymbol{x})), \tag{1.9}$$

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

$$V(\boldsymbol{x}) = d \int_{\mathbb{S}_{+(1)}^{d-1}} \bigvee_{i=1}^{d} \left(\frac{\theta_i}{x_i} \right) \mathrm{d}H(\boldsymbol{\theta}). \tag{1.10}$$

Here

$$\mathbb{S}_{+(p)}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_p = 1\} \tag{1.11}$$

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

$$\int_{\mathbb{S}^{d-1}_{+(1)}} \theta_i \, \mathrm{d}H(\boldsymbol{\theta}) = 1/d, \qquad (i = 1, \ldots, d). \tag{1.12}$$

Our notation for the simplex is borrowed from **fixSimultaneousAutoregressiveModels2021**. 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 (1.9) 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

$$h\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}\right) = -\frac{\|\boldsymbol{x}\|_1^{d+1}}{d} \frac{\partial^d}{\partial x_1 \cdots \partial x_d} V(\boldsymbol{x}). \tag{1.13}$$

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 (1.10) and subsequently G via (1.9). Modelling the angular measure now becomes our primary focus.

1.2.4 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 **gudendorfExtremeValueCopulas2010**.

There are several drawbacks to the parametric approach. Working with a parametric model instead of the general class runs the risk of model misspecification. Generating valid models is a challenging endeavour due to the moment constraints, resulting in models that are

either overly simplistic or have unwieldy distribution functions and parameter constraints. Striking a balance between flexibility and parsimony becomes especially in high dimensions (i.e. when d is large). For these reasons, parametric models are not a primary focus of this thesis. Nevertheless, we now review a small selection of models. These primarily feature as data-generating processes for our numerical experiments. Functionality for generating independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ or $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim H$ based on the sampling algorithms formulated in **dombryExactSimulationMaxstable2016** is provided in the R package `mev`.

1.2.4.1 Logistic-type models

One of the oldest and simplest multivariate extreme value models is the symmetric logistic distribution (**gumbelBivariateExponentialDistributions1960**).

Definition 1.1. The exponent measure of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ following the symmetric logistic distribution is

$$V(\boldsymbol{x}) = \left(\sum_{i=1}^{d} x_i^{-1/\gamma} \right)^{\gamma}, \qquad \gamma \in (0, 1]. \tag{1.14}$$

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 **tawnModellingMultivariateExtreme1990**. Greater control over the dependence structure is achieved by increasing the number of parameters.

Definition 1.2. The exponent measure of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ following

the asymmetric logistic distribution is of the form

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} \left[\sum_{i \in \beta} \left(\frac{\theta_{i,\beta}}{x_i} \right)^{1/\gamma_\beta} \right]^{\gamma_\beta}, \qquad \begin{cases} \gamma_\beta \in (0,1], & \\ \theta_{i,\beta} \in [0,1], & \text{if } i \in \beta, \\ \theta_{i,\beta} = 0, & \text{if } i \notin \beta, \\ \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} \theta_{i,\beta} = 1, & \end{cases} \tag{1.15}$$

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

The set of parameters $\{\gamma_\beta : \beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset\}$ control the dependence strength among the corresponding variables $\{X_i : i \in \beta\}$ in a similar way to the symmetric logistic model. The model's complexity arises from the set of asymmetry parameters $\boldsymbol{\theta}_\beta = (\theta_{i,\beta} : i \in \beta)$, which dictate the direction/composition of extreme events involving the variables $\{X_i : i \in \beta\}$. Further models can be generated by 'inverting' the logistic and asymmetric models. **The purpose of inverting is...**. When applied to the models described above, inversion yields the negative symmetric logistic model (**galambosOrderStatisticsSamples1975**) and the negative asymmetric logistic model (**joeFamiliesMinstableMultivariate1990**), respectively.

Definition 1.3. The exponent measure of a random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$ following the negative symmetric logistic distribution is

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} (-1)^{|\beta|+1} \left(\sum_{i \in \beta} x_i^\gamma \right)^{-1/\gamma}, \qquad \gamma > 0. \tag{1.16}$$

Definition 1.4. The exponent measure of a random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$ following the negative asymmetric logistic distribution is

$$V(\boldsymbol{x}) = \sum_{\beta \in \mathcal{P}(\{1,\ldots,d\}) \setminus \emptyset} (-1)^{|\beta|+1} \left(\sum_{i \in \beta} x_i^\gamma \right)^{-1/\gamma}, \qquad \gamma > 0. \tag{1.17}$$

Other logistic-type models include the bilogistic **smithStatisticsMultivariateExtremes1990**] and negative bilogistic (**colesStatisticalMethodsMultivariate1994**).

1.2.4.2 The Brown-Resnick process and Hüsler-Reiss distribution

The Brown-Resnick process of **brownExtremeValuesIndependent1977** is a class of stochastic processes commonly used to model the extremal dependence structure of spatial phenomena, including rainfall (**davisonStatisticalModelingSpatial2012**), snow depths (**schellanderModelingSnowDepth2018**) and wind gusts (**oestingStatisticalPostprocessingForecas**
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

$$\gamma(\boldsymbol{s}, \boldsymbol{s}') = (\|\boldsymbol{s} - \boldsymbol{s}'\|_2/\rho)^\kappa, \qquad \rho > 0, \kappa \in (0, 2]. \tag{1.18}$$

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 (**engelkeEstimationHuslerReissDistributions2015**). 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 (1.18) 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 (**huserCompositeLikelihoodEstimation2013**)

$$V(x_i, x_j) = \frac{1}{x_i}\Phi\left(\frac{a_{ij}}{2} + \frac{1}{a_{ij}}\log\frac{x_j}{x_i}\right) + \frac{1}{x_j}\Phi\left(\frac{a_{ij}}{2} + \frac{1}{a_{ij}}\log\frac{x_i}{x_j}\right), \tag{1.19}$$

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 **huslerMaximaNormalRandom1989**. The Hüsler-Reiss distribution is of fundamental importance in multivariate extremes: it has been labelled the Gaussian distribution for extremes (**engelkeGraphicalModelsExtremes2019**). In $d \geq 2$ dimensions the distribution is parametrised by a matrix $\Lambda = (\lambda_{ij}^2)_{1 \leq i,j \leq d}$ belonging to the class of symmetric, strictly conditionally negative definite matrices

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

The class of Hüsler-Reiss distributions is closed in the sense that if $\boldsymbol{X} = (X_1, \ldots, X_d)$ follows a Hüsler-Reiss distribution with parameter matrix Λ, then any random sub-vector (X_i, X_j) is also Hüsler-Reiss distributed with parameter λ_{ij}^2. This permits very flexible control over the pairwise dependence structure. The dependence between any pair of variables X_i and X_j can be adjusted by modifying the corresponding parameter λ_{ij}, subject to the constraint $\Lambda \in \mathcal{D}$. The finite-dimensional distribution of a Brown-Resnick process at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d$ is precisely the Hüsler-Reiss distribution with $\Lambda = (\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)/4)_{1 \leq i,j \leq d}$ (**engelkeEstimationHuslerReissDistributions2015**). Due to this link, the exponent measure of (X_i, X_j) is given by (1.19) with a_{ij} replaced by $2\lambda_{ij}$.

1.2.4.3 The max-linear model

The final parametric model we consider is the max-linear (factor) model (**einmahlMestimatorTailDepen**
fougeresDenseClassesMultivariate2013; **yuenCRPSMestimationMaxstable2014**).
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 (**yuenCRPSMestimationMaxstable2014**). Due to its flexibility and theoretical properties, the max-linear model has enjoyed widespread use across several areas of extremes, including clustering (**janssenKmeansClusteringExtremes2020**; **medinaSpectralLearningMultivariate2021**), graphical modelling for causal inference (**gissiblIdentifiabilityEstimationRecursive2019**; **gissiblMaxlinearModelsDirected2018**; **tranCausalDiscoveryRiver2021**) and tail event probability estimation (**kiriliouk EstimatingProbabi**

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

$$H(\cdot) = \frac{1}{\sum_{j=1}^q \|\boldsymbol{a}_j\|_\alpha^\alpha} \sum_{j=1}^q \|\boldsymbol{a}_j\|_\alpha^\alpha \delta_{\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha}(\cdot) \tag{1.20}$$

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 (1.20) is a valid angular measure: for any $i = 1, \ldots, d$,

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

The number of free parameters is $d \times (q-1)$ and the order of the columns of A is inconsequential. The factors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q$ correspond to the possible directions that extremal observations may take. The column norms $\|\boldsymbol{a}_1\|_\alpha, \ldots, \|\boldsymbol{a}_q\|_\alpha$ determine the respective weights assigned to these directions. There is a direct correspondence between the class of discrete angular measure placing mass on $q < \infty$ points and the class of max-linear random vectors with q factors (**yuenCRPSMestimationMaxstable2014**). Moreover, the class of angular measures (1.20) is dense in the class of valid angular measures (**fougeresDenseClassesMultivariate2013**). In other words, any extremal dependence structure can be arbitrarily well-approximated by that of a max-linear model with sufficiently many factors. This makes max-linear modelling a versatile and powerful framework, despite its simplicity.

There are several ways to construct a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with angular measure (1.20). This thesis uses two constructions. Let $Z_1, \ldots, Z_q$ be independent Fréchet

random variables with unit scale and shape parameter α, and set $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$. The two constructions are

$$\boldsymbol{X} = A \times_{\max} \boldsymbol{Z} := \left(\bigvee_{j=1}^{q} a_{1j} Z_j, \ldots, \bigvee_{j=1}^{q} a_{dj} Z_j \right) \tag{1.21}$$

and

$$\boldsymbol{X} = A \otimes \boldsymbol{Z} := \bigoplus_{j=1}^{q} (\boldsymbol{a}_j \odot Z_j). \tag{1.22}$$

Adopting the terminology of **cooleyDecompositionsDependenceHighdimensional2019**, 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 **cooleyDecompositionsDependenceHighdimensional2019**, 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 **cooleyDecompositionsDependenceHighdimensional2019**. 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

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{R}_f(x)) \approx \sum_{j=1}^{q} \frac{\|\boldsymbol{a}_j\|_\alpha^\alpha}{r_\star(\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha)^\alpha}, \tag{1.23}$$

where $r_\star = r_\star(\boldsymbol{\theta})$ is such that $f(r_\star \boldsymbol{\theta}) = x$ (**cooleyDecompositionsDependenceHighdimensional2019** **kirilioukEstimatingProbabilitiesMultivariate2022**). The formulae corresponding to

some popular failure regions are listed below:

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

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 (**kirilioukClimateExtremeEvent2020**) or quantifying the Value-at-Risk of an asset portfolio (**yuenUpperBoundsValuerisk2014**). 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.

1.2.5 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 1.6. A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index $\alpha \in \mathbb{R}$ if, for all $x > 0$,

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^{\alpha}. \tag{1.24}$$

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 1.7. 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

$$\mathbb{P}(X > x) = x^{-\alpha}L(x) \tag{1.25}$$

for some slowly-varying function L (**jessenRegularlyVaryingFunctions2006**). 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, (1.25) 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 **resnickHeavytailPhenomenaPr**
– 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 1.8. 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 (**resnickHeavytailPhenomenaProbabilistic2007**):

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

$$n\mathbb{P}(b_n^{-1}\boldsymbol{X} \in \cdot) \xrightarrow{\text{v}} \nu(\cdot), \qquad (n \to \infty), \tag{1.26}$$

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 $s > 0$,

$$\nu(s\,\cdot) = s^{-\alpha}\nu(\cdot). \tag{1.27}$$

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

$$\mathbb{S}_+^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\} \tag{1.28}$$

such that

$$n\mathbb{P}((b_n^{-1}R, \boldsymbol{\Theta}) \in \cdot) \xrightarrow{\text{v}} \nu_\alpha \times H(\cdot), \qquad (n \to \infty), \tag{1.29}$$

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 (1.26) and (1.29) are related via

$$\nu(\{\boldsymbol{x} \in \mathbb{E}_0 : \|\boldsymbol{x}\| > s, \boldsymbol{x}/\|\boldsymbol{x}\| \in \cdot\}) = s^{-\alpha}H(\cdot), \qquad \nu(\mathrm{d}r \times \mathrm{d}\boldsymbol{\theta}) = \alpha r^{-\alpha-1}\mathrm{d}r\,\mathrm{d}H(\boldsymbol{\theta}). \tag{1.30}$$

The attractive feature of MRV is best represented by its pseudo-polar formulation (1.29). 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 (1.29) 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

$$\mathbb{P}(X_i < x) = \exp(-x^{-\alpha}), \qquad (x > 0). \tag{1.31}$$

An MRV random vector on α-Fréchet margins (1.31) 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 (1.28). As mentioned previously, we exclusively choose the L_p-norm

$$\|\cdot\|_p : \mathbb{R}^d \to \mathbb{R}, \qquad \|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{d} x_i^p\right)^{1/p} \tag{1.32}$$

with (1.11) 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 (1.29). 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 **fougeresDenseClassesMultivariate2013**, 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:

$$\int_{\mathbb{S}_+^{d-1}} \theta_i \, \mathrm{d}H(\boldsymbol{\theta}) = \mu > 0, \qquad (i = 1, \ldots, d). \tag{1.33}$$

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 (1.12). If $\|\cdot\| = \|\cdot\|_2$, then $m/d \leq \mu \leq m/\sqrt{d}$ according to Lemma 2.1 in **fomichovSphericalClusteringDetection2023**. 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.

1.2.6 Extremal dependence measures

The extremal dependence structure of a random vector $\boldsymbol{X}$ can be quantified and classified using a plethora of summary measures (**colesDependenceMeasuresExtreme1999**). We focus on the tail dependence coefficient and the extremal dependence measure.

1.2.6.1 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 1.9. 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)

$$\chi_\beta = \lim_{u \to 1} \chi_\beta(u) = \lim_{u \to 1} \frac{\mathbb{P}(F_i(X_i) > u : i \in \beta)}{1 - u}. \tag{1.34}$$

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

$$H(\{\boldsymbol{\theta} \in \mathbb{S}_+^{d-1} : \theta_i > 0 \iff i \in \beta'\}) > 0. \tag{1.35}$$

For example, consider the angular measures

$$H^{(1)} = \frac{m}{d} \sum_{i=1}^{d} \delta_{\boldsymbol{e}_i}, \qquad H^{(2)} = m\delta_{\mathbf{1}_d/\|\mathbf{1}_d\|}, \tag{1.36}$$

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

If the bivariate exponent measure V_{ij} of (X_i, X_j) is known, then the tail dependence coefficient χ_{ij} may be computed using the relation $\chi_{ij} = 2 - V_{ij}(1, 1)$ (**colesDependenceMeasuresExtreme1999**). The following examples illustrate this for selected parametric models.

Example 1.1. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be symmetric logistic distributed with dependence parameter $\gamma \in (0, 1]$. For any $i \neq j$, let V_{ij} denote the bivariate exponent measure of (X_i, X_j). Then

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

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

Example 1.2. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be Hüsler-Reiss distributed with parameter matrix $\Lambda = (\lambda_{ij}^2)$. For any $i \neq j$, let V_{ij} denote the bivariate exponent measure of (X_i, X_j). Then

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

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

Example 1.3. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d)$ is max-linear with parameter matrix $A \in \mathbb{R}_+^{d \times q}$. Substituting (1.20) into (1.10) yields

$$\chi_{ij} = 2 - V_{12}(1,1) = 2 - 2\int_{\mathbb{S}^1_{+(1)}} (\theta_1 \vee \theta_2)\, \mathrm{d}H(\boldsymbol{\theta}) = 2 - \sum_{l=1}^{q} (a_{il} \vee a_{jl}). \tag{1.37}$$

Consider two max-linear random vectors with discrete angular measures $H^{(1)}$ and $H^{(2)}$ as in (1.36). 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 **reissStatisticalAnalysisExtreme2007** and produces a diagnostic plot as shown in Figure 1.1. 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.

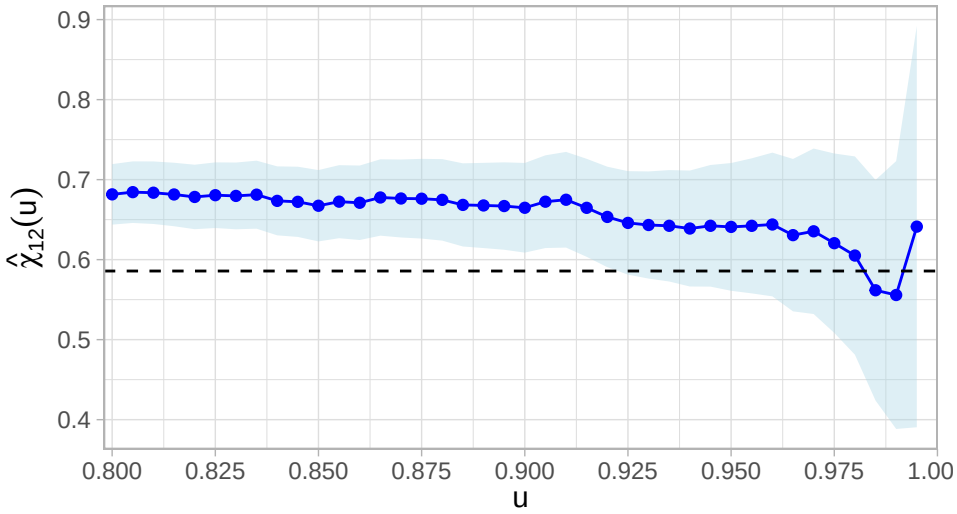

Figure 1.1: Empirical estimates $\hat{\chi}_{12}(u)$ of the tail dependence coefficient for bivariate symmetric logistic data with $\gamma = 0.5$ and $n = 5,000$ observations. The true coefficient $\chi_{12} = 2 - 2^{\gamma} \approx 0.59$ is marked by the dashed line. The shaded region represents the 95% Wald confidence interval.

Estimation of χ_β for $|\beta| > 2$ is more complicated and is related to the task of determining the support of the angular measure (**goixSparseRepresentationMultivariate2017**; **meyerMultivariateSparseClustering2023**; **simpsonDeterminingDependenceStructure2020**). This thesis primarily concerns dependence at the pairwise level, so we direct the reader to the aforementioned papers and the review **engelkeSparseStructuresMultivariate2021** for further details.

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 (**huangNewExploratoryTools2019**) and considered as a tool for clustering (**fomichovSphericalClusteringDetection2023**). Other works focus on its theoretical properties. **shyamalkumarTailDependenceMatrices2020** conjecture that the 'realisation problem' – determining whether a given matrix is a valid TDM – is NP-complete; this was recently proved by **janssenTaildependenceExceedanceSets2023**. By establishing a correspondence between the class of TDMs and a metric space, **janssenTaildependenceExceedanceSets2023** 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).

1.2.6.2 Extremal dependence measure

The extremal dependence measure (EDM) is a pairwise summary measure similar to χ_{ij}. It was originally proposed **resnickExtremalDependenceMeasure2004** and later generalised by **larssonExtremalDependenceMeasure2012**.

Definition 1.10. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ with angular measure H. The EDM between X_i and X_j is

$$\text{EDM}_{ij} := \int_{\mathbb{S}_+^{d-1}} \theta_i \theta_j \, \mathrm{d}H(\boldsymbol{\theta}). \tag{1.38}$$

The EDM depends on the choice of norm via the angular measure, but **larssonExtremalDependenceMe** show that EDMs under different norms are equivalent in a certain sense. The EDM was originally defined by **resnickExtremalDependenceMeasure2004** for bivariate random vectors $\boldsymbol{X} = (X_1, X_2)$. In their definition, the integrand is

$$\left(\frac{4}{\pi}\right)^2 \arctan\left(\frac{\theta_2}{\theta_1}\right)\left[\frac{\pi}{2} - \arctan\left(\frac{\theta_2}{\theta_1}\right)\right]. \tag{1.39}$$

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 (1.35) that variables X_i and X_j are asymptotically independent if and only if $H(\{\boldsymbol{\theta} : \theta_i, \theta_j > 0\}) = 0$. Then

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

The EDM is maximal when X_i and X_j are perfectly asymptotically dependent. The maximal value depends on the choice of norm and the mass of the angular measure. When

$d = 2$ and $\|\cdot\| = \|\cdot\|_p$ we have $\mathrm{EDM}_{ij} \leq 2^{-2/p} m$ 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.

1.3 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: (1.29) implies that

$$\boldsymbol{\Theta} \mid (R > t) \xrightarrow{d} H(\cdot), \qquad (t \to \infty). \tag{1.40}$$

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.

1.3.1 Framework and notation

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

$$R_i := \|\boldsymbol{X}_i\|, \qquad \boldsymbol{\Theta}_i := (\Theta_{i1}, \ldots, \Theta_{id}) = \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}, \tag{1.41}$$

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

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

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

$$R_{(i)} = \|\boldsymbol{X}_{(i)}\|, \qquad \boldsymbol{\Theta}_{(i)} = (\Theta_{(i),1}, \ldots, \Theta_{(i),d}) = \frac{\boldsymbol{X}_{(i)}}{\|\boldsymbol{X}_{(i)}\|}. \tag{1.42}$$

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 $t = R_{(k+1)}$.

1.3.2 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

$$\lim_{n\to\infty} k(n) = \infty, \qquad \lim_{n\to\infty} \frac{k(n)}{n} = 0. \tag{1.43}$$

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 (1.40) 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, **einmahlTestingMultivariateRegular2020** 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 **meyerMultivariateSparseClustering2023** 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 **jiangPrincipalComponentAnalysis2020**, **szemkusSpatialPatternsIndices2024** and **russellAnalyzingDependenceMatrices2018** for examples.

1.3.3 The empirical angular measure

Once the tuning parameter k has been chosen, attention turns towards the extremal angles $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$. In view of (1.40), the empirical distribution of $\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}$ is the natural non-parametric estimator for the angular measure.

Definition 1.11. The empirical angular measure based on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is the random measure on $\mathbb{S}_+^{d-1}$ defined as

$$\hat{H}(\cdot) := \frac{m}{k} \sum_{i=1}^{n} \delta_{\boldsymbol{\Theta}_i}(\cdot) \mathbf{1}\{R_i > R_{(k+1)}\} = \frac{m}{k} \sum_{i=1}^{k} \delta_{\boldsymbol{\Theta}_{(i)}}(\cdot). \tag{1.44}$$

Note that $\hat{H}$ does not enforce the moment constraints (1.12), so is not necessarily a valid angular measure. **einmahlMaximumEmpiricalLikelihood2009** construct an alternative non-parametric estimator that does enforce these restrictions, but it is limited to the bivariate setting. Proposition 3.3 in **janssenKmeansClusteringExtremes2020** establishes consistency $\hat{H} \xrightarrow{p} H$ of the empirical angular measure provided the level k satisfies the rate conditions (1.43). Their result holds for general norms in arbitrary dimensions. **clemenconConcentrationBoundsEmpirical2023** 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

$$\hat{A} := \left(\frac{m}{k}\right)^{1/\alpha} \left(\boldsymbol{\Theta}_{(1)}, \dots, \boldsymbol{\Theta}_{(k)}\right) \in \mathbb{R}_+^{d \times k}. \tag{1.45}$$

whose angular measure is $\hat{H}$. Estimates of tail event probabilities under the empirical model $\hat{H}$ may then be computed using the formula (1.23).

1.3.4 Non-parametric estimators

larssonExtremalDependenceMeasure2012 remark that analysing extremal dependence often involves quantities of the form

$$\mathbb{E}_H[f(\boldsymbol{\Theta})] := \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_{m^{-1}H}[mf(\boldsymbol{\Theta})], \tag{1.46}$$

where $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$. We have already seen an example of this in Definition 1.10: the EDM between X_i and X_j is defined as (1.46) 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:

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

kluppelbergEstimatingExtremeBayesian2021 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 (1.46), we simply replace H with the empirical angular measure $\hat{H}$, yielding (**kluppelbergEstimatingExtremeBayesian2021**)

$$\hat{\mathbb{E}}_H[f(\boldsymbol{\Theta})] := \mathbb{E}_{\hat{H}}[f(\boldsymbol{\Theta})] = \int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}\hat{H}(\boldsymbol{\theta}) = \frac{m}{k}\sum_{i=1}^{k} f(\boldsymbol{\Theta}_{(i)}). \tag{1.47}$$

kluppelbergEstimatingExtremeBayesian2021 prove asymptotic normality of these estimators by generalising a result in **larssonExtremalDependenceMeasure2012**.

Theorem 1.3. *Let $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$ be continuous and assume k satisfies the rate conditions (1.43). Moreover, suppose that*

$$\lim_{n\to\infty} \sqrt{k}\left[\frac{n}{k}\mathbb{E}[f(\boldsymbol{\Theta}_1)\mathbf{1}\{R_1 \geq b_{\lfloor n/k\rfloor}t^{-1/\alpha}\}] - \mathbb{E}_H[f(\boldsymbol{\Theta})]\frac{n}{k}\bar{F}_R(b_{\lfloor n/k\rfloor}t^{-1/\alpha})\right] = 0 \tag{1.48}$$

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*

$$\nu^2 := \mathrm{Var}_H(f(\boldsymbol{\Theta})) > 0. \tag{1.49}$$

Then

$$\sqrt{k}\left[\hat{\mathbb{E}}_H[f(\boldsymbol{\Theta})] - \mathbb{E}_H[f(\boldsymbol{\Theta})]\right] \to N(0, \nu^2), \qquad (n \to \infty). \tag{1.50}$$

The rate condition (1.48) requires that the dependence between the radius and angle decays sufficiently quickly. This condition is non-observable and must be assumed.

Example 1.4. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent copies of $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$. The estimator for the EDM between X_i and X_j is

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

Under the conditions of Theorem 1.3,

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

1.4 Tail pairwise dependence matrix (TPDM)

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

1.4.1 Definition and examples

Preamble.

Definition 1.12. 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

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i \theta_j \, \mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_H[\Theta_i \Theta_j]. \tag{1.51}$$

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 **kirilioukEstimatingProbabilitiesMultivariate2022** to permit general α.

Definition 1.13. 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

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \int_{\mathbb{S}_{+(\alpha)}^{d-1}} \theta_i^{\alpha/2} \theta_j^{\alpha/2} \, \mathrm{d}H(\boldsymbol{\theta}) = \mathbb{E}_H[\Theta_i^{\alpha/2} \Theta_j^{\alpha/2}]. \tag{1.52}$$

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 **kiriliioukEstimatingProbabilitiesMultivariate2022** provide no direct rationale for why (1.52) is the natural generalisation of (1.51). Appendix XX provides a series of results shedding light on this matter. After generalising a result in **fixSimultaneousAutoregressiveModels2021** (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 **semadeniInferenceAngularDistribution2020** 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 1.5. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ follows the symmetric logistic distribution with dependence parameter $\gamma \in (0, 1)$. For any $i \neq j$,

$$\sigma_{ij} = \frac{1-\gamma}{\gamma} \int_0^1 [u(1-u)]^{\frac{1}{\gamma} - \frac{3}{2}} [(1-u)^{1/\gamma} + u^{1/\gamma}]^{\gamma - 2} \, \mathrm{d}u. \tag{1.53}$$

Example 1.6. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ follows the Hüsler-Reiss distribution with parameter matrix $\Lambda = (\lambda_{ij}^2)$. For any $i \neq j$,

$$\sigma_{ij} = \int_0^1 \frac{\exp(-\lambda_{ij}/4)}{2\lambda_{ij} u(1-u)} \phi\left(\frac{1}{2\lambda_{ij}} \log\left(\frac{u}{1-u}\right)\right) \mathrm{d}u. \tag{1.54}$$

The blue lines in Figure 1.2 plot (1.53) and (1.54) against the model parameter. For comparison, we also include the tail dependence coefficients (red lines) computed using Example 1.1 and Example 1.2. 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 **cooleyDecompositionsDependenceHighdimensional2019**. The figure reveals that for a Brown-Resnick process with semi-variogram (1.18) 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 1.2. Further verification of our expressions are provided by the shaded regions in Figure 1.2. 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.

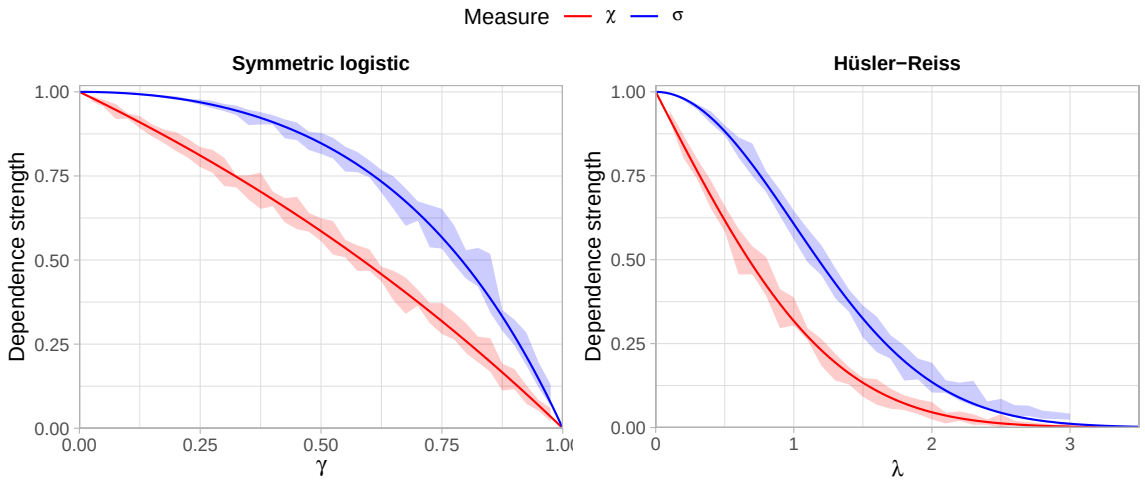

Figure 1.2: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate 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 (**cooleyDecompositionsDependenceHighdimensional2019**; **kirilioukEstimatingProbabilitiesMultivariate2022**).

Example 1.7. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is max-linear with parameter matrix A. Then for any $i \neq j$,

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

Therefore $\Sigma = A^{\alpha/2}(A^{\alpha/2})^T$. Taking A to be $A^{(1)}$ and $A^{(2)}$ as defined in Example 1.3, 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?*

1.4.2 Interpretation of the TPDM entries

The definition of the TPDM

$$\Sigma = \mathbb{E}_H\left[\boldsymbol{\Theta}^{\alpha/2}(\boldsymbol{\Theta}^{\alpha/2})^T\right], \tag{1.55}$$

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 1.14. 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 (**kluppelbergEstimatingExtremeBayesian2021**)

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

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 (1.30), one can show that

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

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

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

Comparing Definition 1.14 against Definition 1.13, the diagonal entries of the TPDM are related to the marginal scales via $\operatorname{scale}(X_i) = \sigma_{ii}^{1/\alpha}$. Consequently, if the marginal distributions are standardised to have unit scales, then all diagonal entries of the TPDM are equal to one. Moreover, when $b_n = n^{1/\alpha}$ and $\|\cdot\| = \|\cdot\|_\alpha$, the mass of the angular measure

relates to the marginal scales via

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

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

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

and

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

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

As mentioned earlier, the TPDM's off-diagonal entries are simply pairwise EDMs. Thus the interpretation of σ_{ij} is inherited from the EDM: X_i and X_j are asymptotically independent if and only $\sigma_{ij} = 0$, and the magnitude of $\sigma_{ij} > 0$ reveals the strength of tail dependence between X_i and X_j. Like a correlation matrix, σ_{ij} attains its maximal value (one) when X_i and X_j are completely dependent (Example 1.7).

1.4.3 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 (**cooleyDecompositionsDependenceHighdimensional2019**). These underpin the key statistical applications of the TPDM described in Section XX. The following results and proofs are reproduced from **kirilioukEstimatingProbabilitiesMultivariate2022**.

Proposition 1.1. *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 1.15. 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 1.2. *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 **fougeresDenseClassesMultivariate2013**, 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 1.7. 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 (**haufmannCompletelyPositiveMatrices2011**).

□

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 (**dickinsonComputationalComplexityMe**
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 1.8. 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 1.5. 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 1.7, 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 (1.36) 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 $H^{(2)} = d\delta_{\mathbf{1}_d/\|\mathbf{1}_d\|_\alpha}$.

1.4.4 The empirical TPDM

Definition 1.16. Let $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ on Fréchet margins (1.31) 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

$$\hat{\Sigma} = (\hat{\sigma}_{ij}), \qquad \hat{\sigma}_{ij} := \hat{E}_H[\Theta_i^{\alpha/2}\Theta_j^{\alpha/2}] = \frac{d}{k}\sum_{l=1}^{k} \Theta_{(l),i}^{\alpha/2}\Theta_{(l),j}^{\alpha/2}. \tag{1.56}$$

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 1.3. *The empirical TPDM is completely positive.*

Proof. Let $A = \hat{A}$, the $d \times k$ matrix with non-negative entries defined in (1.45). 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 1.4. *The empirical TPDM is symmetric and positive semi-definite.*

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

$$\boldsymbol{y}^T\hat{\Sigma}\boldsymbol{y} = \boldsymbol{y}^T AA^T\boldsymbol{y} = \|A^T\boldsymbol{y}\|_2^2 \geq 0. \tag{1.57}$$

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

□

Proposition 1.5. *Under the conditions of Theorem 1.3, the entries of $\hat{\Sigma}$ are consistent and asymptotically normal, that is, for any $i, j = 1, \ldots, d$,*

$$\sqrt{k}(\hat{\sigma}_{ij} - \sigma_{ij}) \to \mathrm{N}(0, \nu_{ij}^2), \qquad \nu_{ij}^2 := \operatorname{Var}_H(\Theta_i^{\alpha/2}\Theta_j^{\alpha/2}). \tag{1.58}$$

Proof. See Example 1.4.

□

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 (**lehtomaaAsymptoticIndependenceSupport2020**).

Using asymptotic normality one may construct asymptotic confidence intervals

$$\lim_{n\to\infty} \mathbb{P}\left[|\sigma_{ij} - \hat{\sigma}_{ij}| < z_{\beta/2}\sqrt{\nu_{ij}^2/k}\right] = 1 - \beta, \tag{1.59}$$

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 1.9. Suppose $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathcal{RV}_+^d(\alpha)$ is symmetric logistic with $\gamma = 0.6$. Using Example 1.5 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 (**leePartialTailCorrelation2023**)

$$\hat{\nu}_{ij}^2 := \frac{1}{k-1}\sum_{l=1}^{k}\left(d\Theta_{(l),i}\Theta_{(l),j} - \hat{\sigma}_{ij}\right)^2. \tag{1.60}$$

The following result, proved by **kraliCausalityEstimationMultivariate2018** 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

$$\boldsymbol{\sigma} := \text{vecu}(\Sigma) := (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{1d}, \sigma_{23}, \ldots, \sigma_{2d}, \ldots, \sigma_{d-1,d}), \tag{1.61}$$

$$\hat{\boldsymbol{\sigma}} := \text{vecu}(\hat{\Sigma}) := (\hat{\sigma}_{12}, \hat{\sigma}_{13}, \ldots, \hat{\sigma}_{1d}, \hat{\sigma}_{23}, \ldots, \hat{\sigma}_{2d}, \ldots, \hat{\sigma}_{d-1,d}). \tag{1.62}$$

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 1.6. *Under the conditions of Theorem 1.3, 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 1.5 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 **kraliCausalityEstimationMultivariate2018** to permit general α. The following example illustrates an application of Proposition 1.6 to the max-linear model.

Example 1.10. 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 1.3 (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 1.6 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 1.4 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 $\hat{\sigma}_{lm}$.

1.5 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 overview of TPDM-based clustering algorithms (**fomichovSphericalClusteringDetection2023**; **richardsModernExtremeValue2024**) is contained in Appendix XX. Further interesting topics that are not covered include time series (**mhatreTransformedLinearModelsTime2021**; **wixsonAttributionSeasonalWildfire2023**) and graphical models (**gongPartialTailCorrelationCoef**; **leePartialTailCorrelation2023**). 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

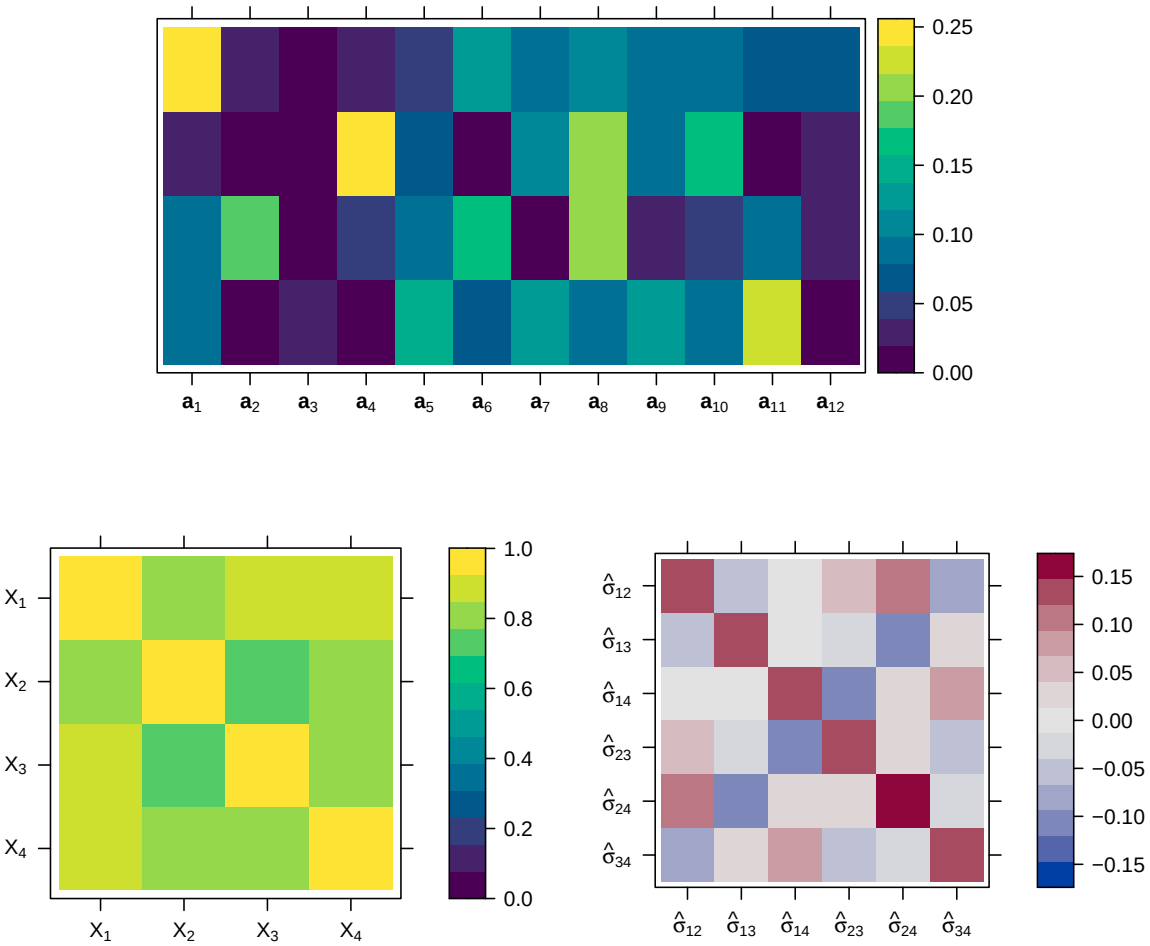

Figure 1.3: Visual representation of the matrices discussed in Example 1.10. Top: a randomly generated max-linear parameter matrix A with $d = 4$ and $q = 12$. Bottom left: the TPDM Σ of $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z}$. Bottom right: the asymptotic covariance matrix V of $\hat{\boldsymbol{\sigma}}$.

$b_n = n^{1/\alpha}$, while $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent copies of $\boldsymbol{X}$.

1.5.1 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 (**blanchardStatisticalPropertiesKernel2007**; **jolliffePrincipalComponentAnalysis2002**). 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

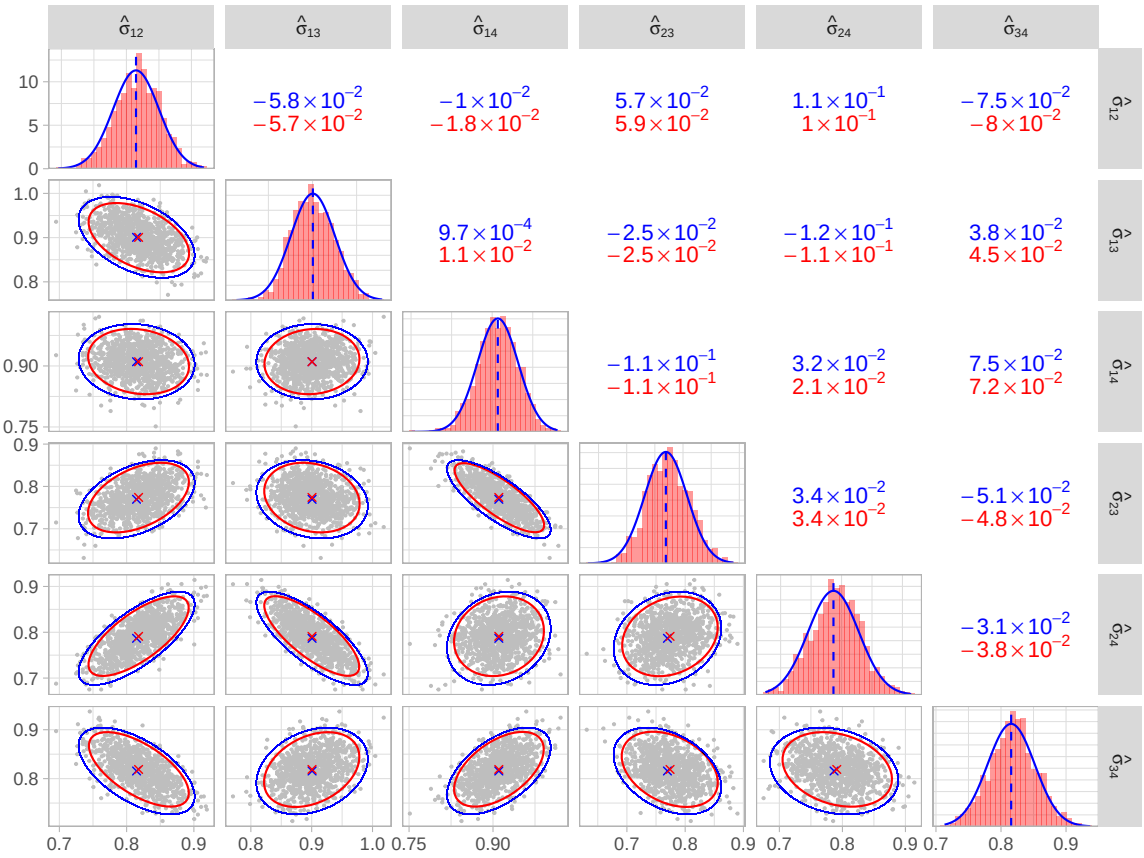

Figure 1.4: Pairs plot illustrating asymptotic normality of the empirical TPDM – see Example 1.10 for details. All panels: red represents the empirical quantity based on the 1,000 repeated simulations; blue represents the theoretical quantity based on asymptotic normality. Diagonal panels: the distribution (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.

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 (**engelkeSparseStructuresMultivariate2021**). For example, weather extremes typically exhibit spatial patterns related to geographical or topographical drivers, e.g. north/south, coastal/inland, high-lying/low-lying (**bernardClusteringMaximaSpatial2013**; **jiangPrincipalComponentAnalysis2020**). 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 $X_1, \ldots, X_d$ 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 (**cooleyDecompositionsDependenceHighdimensional2019**; **dreesPrincipalComponentAnalysis2021**).

dreesPrincipalComponentAnalysis2021 adopt a risk minimisation perspective aiming to minimise the mean-squared reconstruction error of $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ 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 $j = 1, \ldots, d$. Then $\hat{\mathcal{S}}_p := \operatorname{span}\{\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_d\}$ minimises R in $\mathcal{V}_p$ and $\hat{R}(\mathcal{S}_p) = \sum_{j>p} \hat{\lambda}_j$ (**dreesPrincipalComponentAnalysis2021**). 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 com-

ponents 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 (**dreesPrincipalComponentAnalysis2021**). 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 (1.43), $\hat{\mathcal{S}}_{p^\star} \to \mathcal{S}^\star$ in the sense that

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

Treating the angles $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ as points in $\mathbb{R}^d$ rather than $\mathbb{S}^{d-1}_{+(\alpha)}$ 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, **janssenKmeansClusteringExtremes2020** 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-medinaKernelPCAMultivariate2022** 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.

cooleyDecompositionsDependenceHighdimensional2019 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 **dreesPrincipalComponentAnalysis2021**, let $(\hat{\boldsymbol{u}}_j, \hat{\lambda}_j)$ be the ordered eigenpairs (in $\mathbb{R}^d$) of $\hat{\Sigma}$. Then $\{\hat{\boldsymbol{\omega}}_1, \ldots, \hat{\boldsymbol{\omega}}_d\} = \{\tau(\hat{\boldsymbol{u}}_1), \ldots, \tau(\hat{\boldsymbol{u}}_d)\}$ forms an orthonormal basis of $\mathbb{R}_+^d$. In this new basis, the random vector $\boldsymbol{X}$ may be decomposed as

$$\boldsymbol{X} = \bigoplus_{j=1}^{d} (\hat{V}_j \odot \hat{\boldsymbol{\omega}}_j) = \tau\left(\sum_{j=1}^{d} \hat{V}_j \hat{\boldsymbol{u}}_j \right), \tag{1.63}$$

where $\hat{V}_j = \langle \boldsymbol{X}, \hat{\boldsymbol{\omega}}_j \rangle_\tau$ for $j = 1, \ldots, d$. Truncating the expansion (1.63) yields low-rank reconstructions of $\boldsymbol{X}$. The random variables $\hat{V}_1, \ldots, \hat{V}_d$ 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 (**cooleyDecompositionsDependenceHighdimensional2019**)

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

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, **jiangPrincipalComponentAnalysis2020** 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. **russellAnalyzingDependenceMatrices2018** 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. **szemkusSpatialPatternsIndices2024** 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. **rohrbeckSimulatingFloodEvent2023** 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 (1.63).

Results based on **cooleyDecompositionsDependenceHighdimensional2019** 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. **jiangPrincipalComponentAnalysis2020** ameliorate this using a 'pairwise-thresholded' estimator instead of (1.56), 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 (**highamComputingNearestCorrelation2002**), 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.

1.5.2 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. **einmahlContinuousUpdatingWeighted2018** 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. **janssenKmeansClusteringExtremes2020** and **medinaSpectralLearningMultivariate2021** 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; **kirilioukHypothesisTestingTail2020** 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 (**fixSimultaneousAutoregressiveModels2021**; **kirilioukEstimatingProbabilitiesMultivariate202**. Recall from Example 1.7 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 1.2, which says that the TPDM is completely positive (Definition 1.15). Based on this connection, originally observed by **cooleyDecompositionsDependenceHighdimensional2019**, 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-mondererNumberCPFactorizations2020**). Among these, the simplest CP-estimate is the empirical estimate $\hat{A} \in \mathbb{R}_+^{d\times k}$ as defined in (1.45). **cooleyDecompositionsDependenceHighdimensional2019** describe the empirical estimate as 'naive' because it probably contains more columns than necessary. **kirilioukEstimatingProbabilitiesMultivariate2022** 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 (1.23). 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.

fixSimultaneousAutoregressiveModels2021 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 $\alpha = 2$):

$$\hat{\rho} = \underset{\rho \in (0,1/4)}{\arg\min} \|\Sigma(\rho) - \hat{\Sigma}\|_F^2. \tag{1.64}$$

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.

1.6 Bias in the empirical TPDM in weak-dependence scenarios

The empirical TPDM is consistent and asymptotically unbiased (Proposition 1.6). 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 (1.43). However, in real-world applications, data are limited and extreme observations are scarce. For example, commonly available climate records typically span approximately 50 years (**boulaguiemModelingSimulatingSpatial2022**). A study of summer heatwaves might then be based on, say, $n \approx 50 \times 100 = 5,000$ daily observations. The second condition in (1.43) 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 (**cooleyDecompositionsDependenceHighdimensional2019**;

fixSimultaneousAutoregressiveModels2021; **mhatreTransformedLinearModelsTime2021**), 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.

1.6.1 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, **huserLikelihoodEstimatorsMultivariate2016** 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 **huserLikelihoodEstimatorsMultivariate2016**. Problems of a similar nature can be found across the multivariate extremes literature, for example in spatial modelling (**boulaguiemModelingSimulatingSpatial2022**) and lower-tail dependence modelling (**dobricNonparametricEstimationLower2005**).

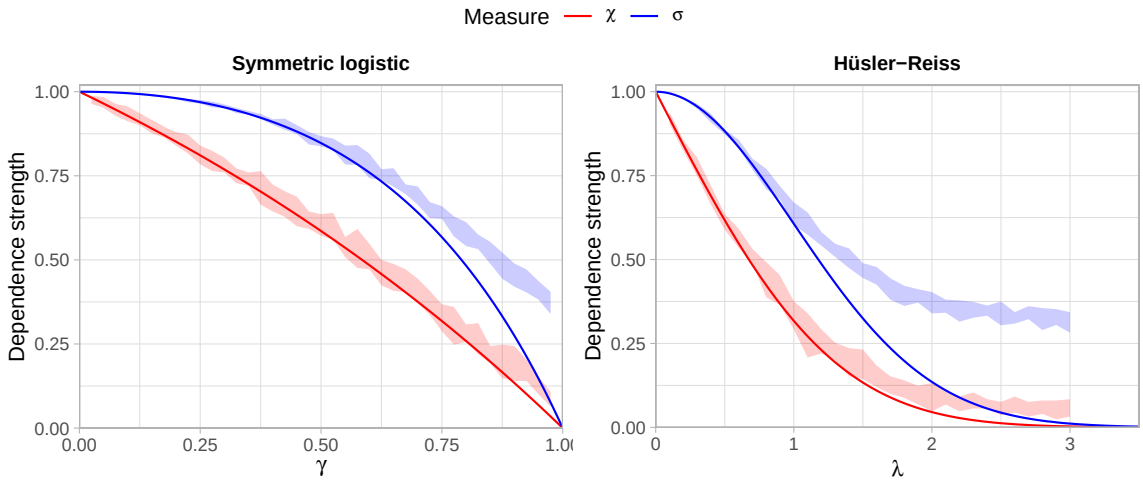

Figure 1.5: True dependence strengths for the symmetric logistic (left) and Hüsler-Reiss (right) models, measured using the tail dependence coefficient (red line) and TPDM (blue line). The shaded regions represent the minimum/maximum values of empirical estimates over 10 repeated simulations using bivariate samples of size $n = 5 \times 10^3$.

The empirical TPDM suffers from the same issue when dependence is weak. This phenomenon is illustrated in Figure 1.5. Like in Figure 1.2, 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 1.2 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

$$\sigma_{ij} \ll 1 \implies \text{Bias}(\hat{\sigma}_{ij}) = \mathbb{E}[\hat{\sigma}_{ij}] - \sigma_{ij} > 0. \tag{1.65}$$

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 **cooleyDecompositionsDependenceHigh** 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 (**mestreImprovedEstimationEigenvalues2008**). 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.

1.6.2 Existing bias-correction approaches for the TPDM

The first strategy for tackling the bias issue is found in **mhatreTransformedLinearModelsTime2021**. 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.

fixSimultaneousAutoregressiveModels2021 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 (1.64). Their bias-corrected estimate $\tilde{\Sigma}$ reduces the entries of $\hat{\Sigma}$ by element-wise application of the soft-thresholding operator (**rothmanGeneralizedThresholdingLarge2009**), that is

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij}), \qquad \tilde{\sigma}_{ij} = \begin{cases} \hat{\sigma}_{ij}, & i = j, \\ (\hat{\sigma}_{ij} - \lambda)_+, & i \neq j. \end{cases} \tag{1.66}$$

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 **mhatreTransformedLinearModelsTime2021** and **fixSimultaneousAutoregressiveModels2021** 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 $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)}$ 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 (**meyerSparseRegularVariation2021**) 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 (1.66). Unlike **fixSimultaneousAutoregressiveModels2021**, our tuning procedure for selecting the hyperparameter λ is purely data-driven and can be applied in general settings, not just spatial.

2 Testing for time-varying extremal dependence

2.1 Introduction

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 (**zhouGlobalConcurrentClimate2023**) and regulatory changes can cause structural changes in the joint tail behaviour of financial asset prices (**poonModellingExtremeValueDependence2003**). 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-camiloTimevaryingExtremeValue2018** and the spectral density ratio model of **decarvalhoSpectralDensityRatio2014**

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 **dreesStatisticalInferenceChanging2023**. 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).

2.2 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 1.8) 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 $\mathcal{B} \subset \mathbb{S}_{+(\alpha)}^{d-1}$,

$$\lim_{u \to \infty} \frac{\mathbb{P}(R(t) > zu, \boldsymbol{\Theta}(t) \in \mathcal{B})}{\mathbb{P}(R(t) > u)} = z^{-\alpha(t)} H(\mathcal{B}; t). \tag{2.1}$$

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 **dreesStatisticalInferenceChanging2023**, our working null and alternative

hypotheses are

$$\mathrm{H}_0 \ : \ \forall t \in [0,1],\, H(\cdot\,;t) = H(\cdot\,;1), \tag{2.2}$$

$$\mathrm{H}_1 \ : \ \exists t,\, H(\cdot\,;t) \neq H(\cdot\,;1). \tag{2.3}$$

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 (2.2) against (2.3) 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. In particular, we will work an integrated version of the TPDM. Since we work with an integrated quantity, the alternative hypothesis might be revised to instead say that there exists a set $\mathcal{T} \in [0,1]$ of non-zero measure such that $H(\cdot\,;t) \neq H(\cdot\,;1)$ for all $t \in \mathcal{T}$.

2.3 The local TPDM and integrated TPDM

Given non-stationary dependence as in (2.1), a time-dependent version of the TPDM is naturally defined by replacing H with the local angular measure $H(\cdot;t)$ in Definition 1.12.

Definition 2.1. The local TPDM of $\boldsymbol{X}(t)$ is the $d \times d$ matrix

$$\Sigma(t) = (\sigma_{ij}(t)), \qquad \sigma_{ij}(t) = \int_{\mathbb{S}_{+(2)}^{d-1}} \theta_i \theta_j \,\mathrm{d}H(\boldsymbol{\theta};t) = \mathbb{E}_{H(\cdot;t)}[\Theta_i \Theta_j]. \tag{2.4}$$

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 **dreesStatisticalInferenceChanging2023**, 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 2.2. 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 2.1. *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

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

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 $x \in [a, b]$. 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 $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, 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 in Lemma 2.1 we may focus on vectorised forms of $\Sigma(t)$ and $\Psi(t)$ defined analogously to $\boldsymbol{\sigma}$ in (1.61), i.e. by row-wise flattening of the strictly upper-triangular elements of $\Sigma(t)$ and $\Psi(t)$:

$$\boldsymbol{\sigma}(t) := \mathrm{vecu}(\Sigma(t)) = (\sigma_{12}(t), \sigma_{13}(t), \ldots, \sigma_{1d}(t), \sigma_{23}(t), \ldots, \sigma_{2d}(t), \ldots, \sigma_{d-1,d}(t)), \tag{2.5}$$
$$\boldsymbol{\psi}(t) := \mathrm{vecu}(\Psi(t)) = (\psi_{12}(t), \psi_{13}(t), \ldots, \psi_{1d}(t), \psi_{23}(t), \ldots, \psi_{2d}(t), \ldots, \psi_{d-1,d}(t)). \tag{2.6}$$

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 2.1, 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

$$\psi(t) - t\psi(1) = \int_0^t \sigma(s)\,\mathrm{d}s - t\int_0^1 \sigma(s)\,\mathrm{d}s = \sigma\left(\int_0^t \mathrm{d}s - t\int_0^1 \mathrm{d}s\right) = 0. \tag{2.7}$$

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.

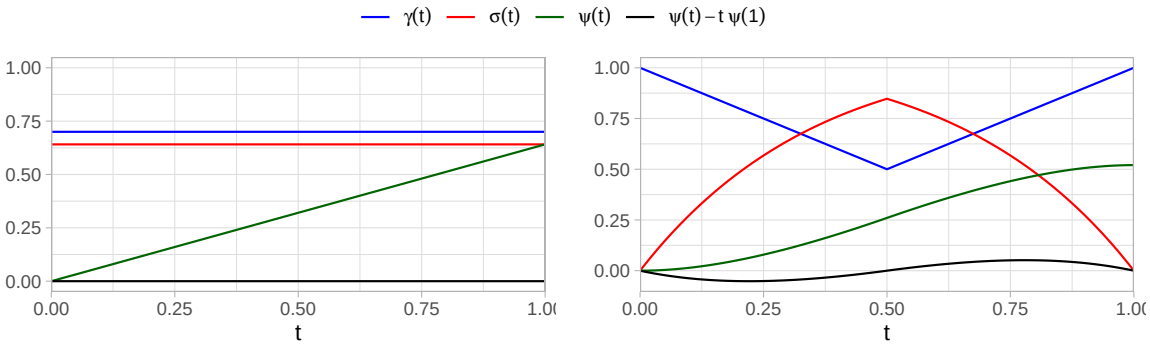

Figure 2.1: The quantities $\sigma_{ij}(t)$, $\psi_{ij}(t)$, and $\psi_{ij}(t) - t\psi_{ij}(1)$ as functions of t when $\boldsymbol{X}(t)$ is symmetric logistic with dependence parameter $\gamma(t)$. Left: constant dependence with $\gamma(t) = 0.7$. Right: time-varying dependence with $\gamma(t) = 0.5 + |t - 0.5|$.

2.4 Inference

Suppose we observe a sample path of $\{\boldsymbol{X}(t) : t \in [0, 1]\}$ along n discrete time-points according to an equidistant sampling scheme, corresponding to realisations of the independent random vectors $\{\boldsymbol{X}(i/n) : i = 1, \ldots, n\}$.

2.4.1 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

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

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. **dreesStatisticalInferenceChanging2023** use the same setup to define the empirical local angular measure, which forms the basis of our estimator for the local TPDM.

Definition 2.3. For any $t \in [0,1]$, the empirical local angular measure is the random measure on $\mathbb{S}^{d-1}_{+(2)}$ defined by

$$\hat{H}(\cdot; t) := \frac{d}{k} \sum_{l \in \mathcal{I}(t)} \mathbf{1}\{R(l/n) > \hat{u}(t), \boldsymbol{\Theta}(l/n) \in \cdot\}. \tag{2.8}$$

Definition 2.4. For $t \in [0,1]$, the empirical local TPDM estimator is the $d \times d$ matrix $\hat{\Sigma}(t) = (\hat{\sigma}_{ij}(t))$, where

$$\hat{\sigma}_{ij}(t) := \int_{\mathbb{S}^{d-1}_{(2)+}} \theta_i \theta_j \, \mathrm{d}\hat{H}(\boldsymbol{\theta}; t) = \frac{d}{k} \sum_{l \in \mathcal{I}(t)} \Theta_i(l/n)\Theta_j(l/n)\mathbf{1}\{R(l/n) > \hat{u}(t)\}. \tag{2.9}$$

We recognise $\hat{H}(\cdot; t$ and $\hat{\Sigma}(t)$ as simply time-localised versions of the empirical angular measure (Definition 1.11) and empirical TPDM (Definition 1.16). 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 of 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.

2.4.2 The empirical integrated TPDM

The best approach for estimating the integrated TPDM is slightly less obvious because $\Psi(t)$ depends on the entire history $\{\Sigma(s) : s \in [0,t]\}$. 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 2.5. 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 $\hat{\Sigma}(h), \hat{\Sigma}(3h), \ldots, \hat{\Sigma}(1-h)$. Independence is due to the blocks being non-overlapping and 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.

2.4.3 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 (2.5) and (2.6). The main result is asymptotic normality of $\hat{\Sigma}(t)$.

Lemma 2.2. *Assume $k(n)$ and $h(n)$ satisfy the rate conditions*

$$h(n) \to 0, \quad k(n) \to \infty, \quad nh(n) \to \infty, \quad \frac{k(n)}{nh(n)} \to 0 \tag{2.10}$$

as $n \to \infty$ and the null hypothesis (2.2) is true. Then, for any $t \in [0, 1]$,

$$\sqrt{k}(\hat{\boldsymbol{\sigma}}(t) - \boldsymbol{\sigma}(t)) \to N(\mathbf{0}, V(t)) \tag{2.11}$$

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

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

Proof. Fix $t \in [0,1]$. The random vectors $\{\boldsymbol{X}(l/n) : l \in \mathcal{I}(t)\}$ are independent and identically distributed with distribution $H(\cdot;t)$. The modified (2.10) ensure that the original rate conditions (1.43) are satisfied on the sub-interval $(t-h, t+h]$: the number of local observations $2nh(n) \to \infty$ and the extreme sampling fraction $k(n)/2nh(n) \to 0$ as $n \to \infty$. Thus, we may invoke asymptotic normality of the empirical TPDM (**?@prp-empirical-local-tpdm-normality**) and the result follows immediately.

□

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. 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 $\{\hat{\boldsymbol{\psi}}(t) : t \in [0,1]\}$.

Lemma 2.3. *Under the conditions of Lemma 2.2, the $\mathcal{D}$-dimensional, continuous-time stochastic process*

$$\left\{\sqrt{\frac{k}{2h}}\left(\hat{\boldsymbol{\psi}}(t) - \boldsymbol{\psi}(t)\right) : t \in [0,1]\right\}, \tag{2.12}$$

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

$$\mathrm{Cov}(\boldsymbol{Y}(s), \boldsymbol{Y}(t)) = \min(s,t)V \tag{2.13}$$

Proof. Since dependence is constant, we may denote by $\boldsymbol{\sigma}$ and V the true local TPDM and asymptotic covariance matrix for all $t \in [0,1]$. Let $N = 1/(2h)$ be the number of blocks. For $i = 1, \ldots, N$, let

$$\boldsymbol{Y}_{i/N} := \sqrt{k}(\hat{\boldsymbol{\sigma}}((2i-1)h) - \boldsymbol{\sigma}).$$

Then $\{\boldsymbol{Y}_{i/N} : i = 1, \ldots, N\}$ are independent, asymptotically normal random vectors with $\boldsymbol{Y}_{i/N} \to N(\boldsymbol{0}, V)$ as $n \to \infty$. (Note that $N \to \infty$ as $n \to \infty$ due to the rate conditions.) Donsker's invariance principle states that the empirical process

$$\{\boldsymbol{Y}_N(t) : t \in [0,1]\}, \qquad \boldsymbol{Y}_N(t) := \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} \boldsymbol{Y}_{i/N},$$

converges to the process $\{\boldsymbol{Y}(t) : t \in [0,1]\}$ defined in the statement of the lemma. To complete the proof, observe that

$$\begin{aligned} \boldsymbol{Y}_N(t) &= \sqrt{2h} \sum_{i=1}^{\lfloor t/(2h) \rfloor} \sqrt{k}(\hat{\boldsymbol{\sigma}}((2i-1)h) - \boldsymbol{\sigma}) \\ &= \sqrt{2kh} \left(\frac{\hat{\psi}(t)}{2h} - \frac{t\boldsymbol{\sigma}}{2h} \right) \\ &= \sqrt{\frac{k}{2h}} \left(\hat{\boldsymbol{\psi}}(t) - \boldsymbol{\psi}(t) \right). \end{aligned}$$

□

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

Theorem 2.1. *Suppose the conditions of Lemma 2.2 hold. Define*

$$\hat{\boldsymbol{Z}}(t) := \sqrt{\frac{k}{2h}} V(t)^{-1/2} (\hat{\boldsymbol{\psi}}(t) - t\hat{\boldsymbol{\psi}}(1)). \tag{2.14}$$

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. Using Lemma 2.3 and pre-multiplying the process (2.12) by $V^{-1/2}$, 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

$$\text{Cov}(\boldsymbol{Y}(s), \boldsymbol{Y}(t)) = \min(s,t), \tag{2.15}$$

i.e. a $\mathcal{D}$-dimensional standard Brownian motion. By (2.7), $\boldsymbol{\psi}(t) = t\boldsymbol{\psi}(1)$ under the null and therefore

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

which is equivalent to a $\mathcal{D}$-dimensional standard Brownian bridge.

□

This main result provides the foundation for our test. The test statistics defined in the following section will quantify whether the realised sample path of (2.14) is consistent with a Brownian bridge. We are not the first to use Brownian bridges in hypothesis testing in extremes; **gadeikisEstimationChangepointTail2005** use the same principle to test for changes in the tail index.

2.5 Test statistics and critical values

From the test process (2.14) we define Kolmogorov-Smirnov (KS) and Cramér-von-Mises (CM) type test statistics by

$$T^{(KS)} := \sup_{t\in[0,1]} \left\|\hat{\boldsymbol{Z}}(t)\right\|_\infty = \sup_{\substack{t\in[0,1] \\ i<j}} |Z_{ij}(t)|, \tag{2.16}$$

$$T^{(CM)} := \sup_{1\le i<j\le d} \left\|\hat{Z}_{ij}(t)\right\|^2_{L^2[0,1]} = \sup_{i<j} \int_0^1 |\hat{Z}_{ij}(t)|^2 \,\mathrm{d}t, \tag{2.17}$$

where $\|\boldsymbol{x}\|_\infty := \max\{|x_i| : i = 1, \dots, \mathcal{D}\}$ denotes the sup-norm in $\mathbb{R}^{\mathcal{D}}$, $\mathcal{D} = \binom{d}{2}$ 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 2.1. *Under the null hypothesis* (2.2)*,*

$$T^{(KS)} \to \sup_{t\in[0,1]} \|\boldsymbol{B}(t)\|_\infty \overset{d}{=} \sup_{i<j} K_{ij}, \qquad T^{(CM)} \to \sup_{i<j} \|B_{ij}(t)\|^2_{L^2[0,1]}, \tag{2.18}$$

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 Theorem 2.1. It is well known that $K_{ij} := \sup_{t\in[0,1]} |B_{ij}(t)|$ follows a Kolmogorov distribution (**henzeAsymptoticStochasticsIntroduction2024**).

□

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

$$\mathbf{1}\{T^{(KS)} > c_\alpha\}, \qquad c_\alpha = F_K^{-1}((1-\alpha)^{1/\mathcal{D}}) \tag{2.19}$$

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

Table 2.1: 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

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 2.1.

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 (2.14). 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 **dreesStatisticalInferenceChanging2023** 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

$$k_{\text{total}} := k/(2h) > \mathcal{D} \tag{2.20}$$

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.

2.6 Simulation experiments

In this section, we present a series of numerical experiments demonstrating our method's performance and, where applicable, providing comparisons against **dreesStatisticalInferenceChanging**

2.6.1 Data generating processes

Suppose the extremal dependence structure of $\boldsymbol{X}(t) = (X_1(t), \ldots, X_d(t))$ 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 **fomichovSphericalClusteringDetection2023**.

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 **dreesStatisticalInferenceChanging2023** is included as a comparator. Results pertaining to their test are based on our own implementation based on a family of Borel subsets $\mathcal{A} = \{A_y : y = 0.01, 0.02, \ldots, 0.99\}$, where

$$A_y := \{\boldsymbol{\theta} \in \mathbb{S}^1_{+(2)} : \theta_1 \leq y\} \subset \mathbb{S}^1_{+(2)}. \tag{2.21}$$

For further details about the meaning and role of $\mathcal{A}$, see Appendix XX.

2.6.2 Large sample performance

In an idealised setting with infinite data, the null distribution of the test statistics is as described in Proposition 2.1. 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 2.2 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 **dreesStatisticalInferenceChanging2023** method can be found in Figure 7 within their Supplementary Material.

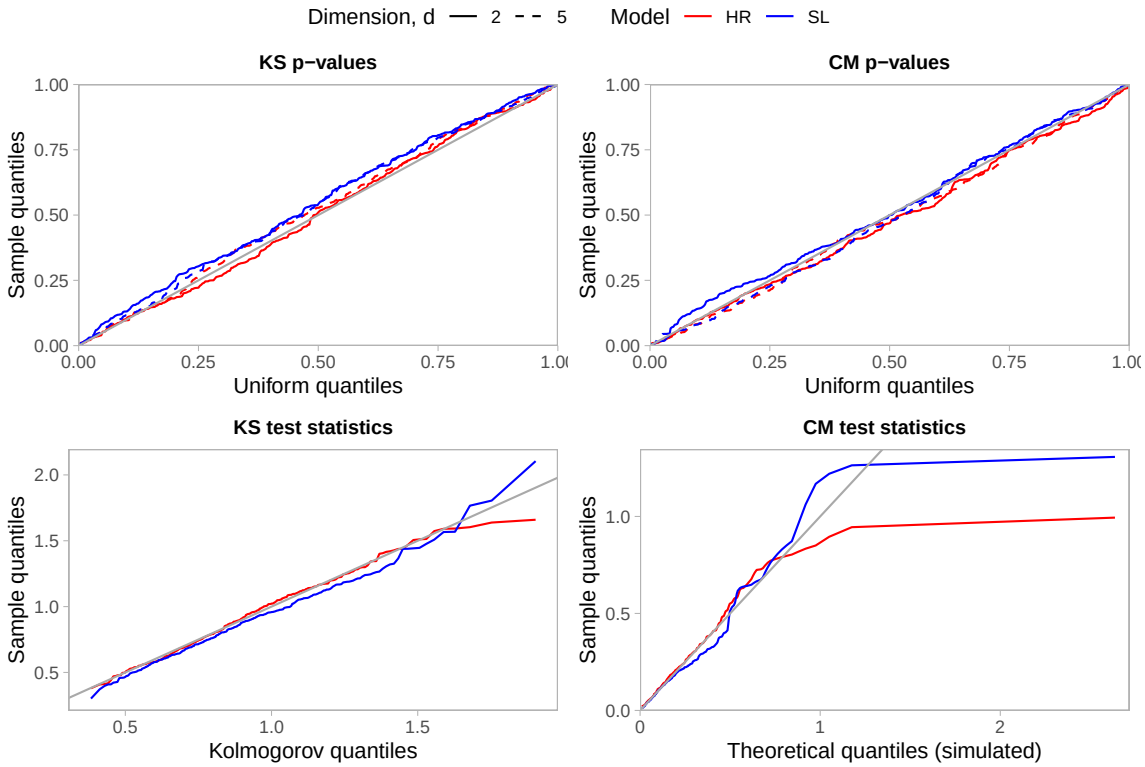

Figure 2.2: Large sample Q-Q plots for the p-values (top) and test statistics (bottom) associated with the KS test (left) and CM test (right). Based on 350 simulations from the SL- and HR-constant models with $n = 10^6$, $b = 2000$ and $k = 50$.

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 change. To fully explore the asymptotic properties of the proposed methods, in the next sections we investigate how they perform for more moderate sample sizes.

2.6.3 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 so 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 2.2 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 (2.20) 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. **dreesStatisticalInferenceChanging2023** 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 **dreesStatisticalInferenceChanging20** 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 $n \leq 5,000$. Nevertheless, even in these high-dimensional settings, the test produces reasonable results, especially for symmetric logistic data. Performance deteriorates under the 25-dimensional Hüsler-Reiss model, suggesting there may be insufficient data for the asymptotic approximations to hold.

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 parameter ϑ_1 at time 1 is allowed to vary, permitting an examination

Table 2.2: 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				

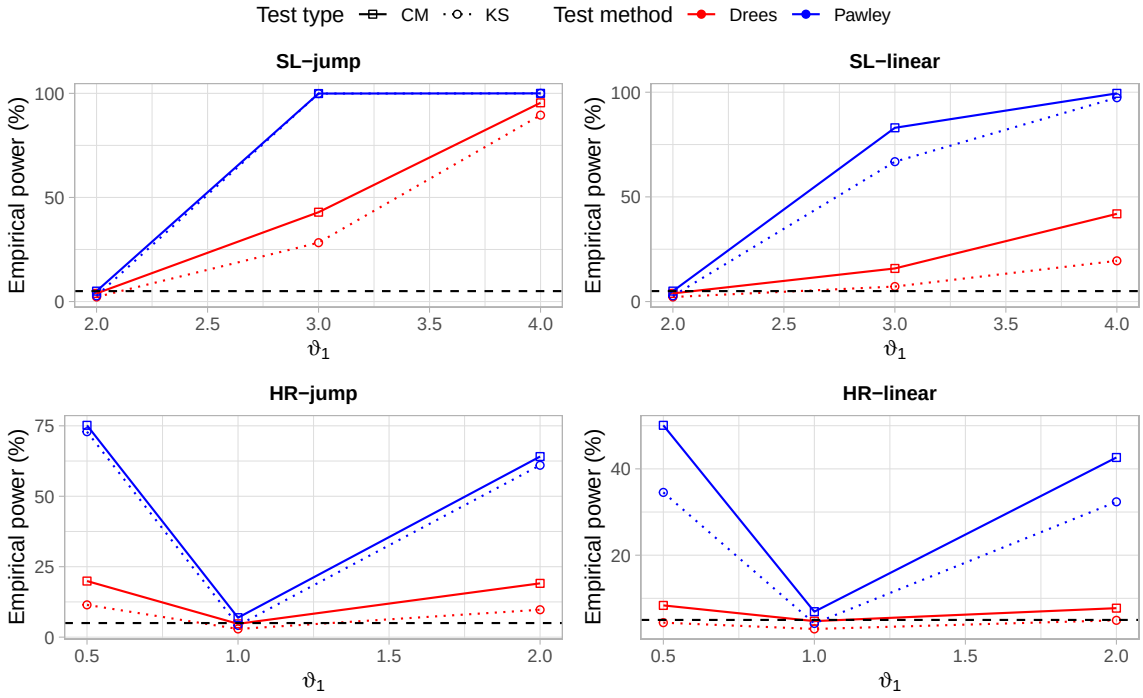

Figure 2.3: Empirical power (%) as a function of the dependence parameter ϑ_1. Based on 1000 simulations with $n = 2,500$ and $d = 2$. For the SL and HR models, $\vartheta_0 = 2$ and $\vartheta_0 = 1$, respectively. Tests are conducted at the 5% level (black dashed line).

of 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 2.3 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 **dreesStatisticalInferenceChanging2023** in all scenarios. When the null is true, $\vartheta_1 = \vartheta_0$, the power of the test reverts to 5%. 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 focusing 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, $T^{CM} \approx \hat{Z}^2_{12}(0.5)/2$. 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 **dreesStatisticalInferenceChanging2023**.

Figure 2.4 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.

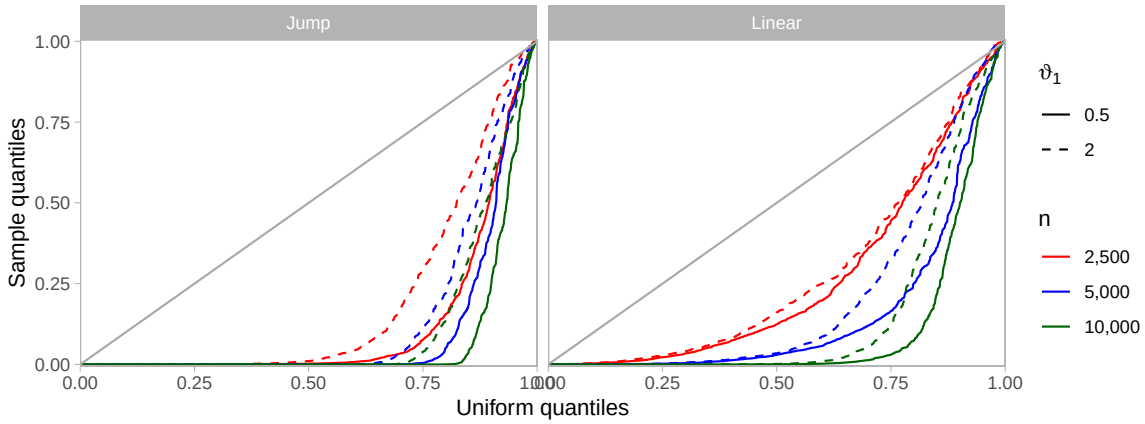

Figure 2.4: QQ-plots for the KS p-values with varying sample size. Based on 1,000 simulations from the HR-jump (left) and HR-linear (right) models with tuning parameters $n/b = 25$ and $k/b = 0.1$.

2.6.4 Computation time

A key advantage of our approach is that the computations are considerably less intensive than those required by **dreesStatisticalInferenceChanging2023**. 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 2.5 shows the 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 to run the test 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.

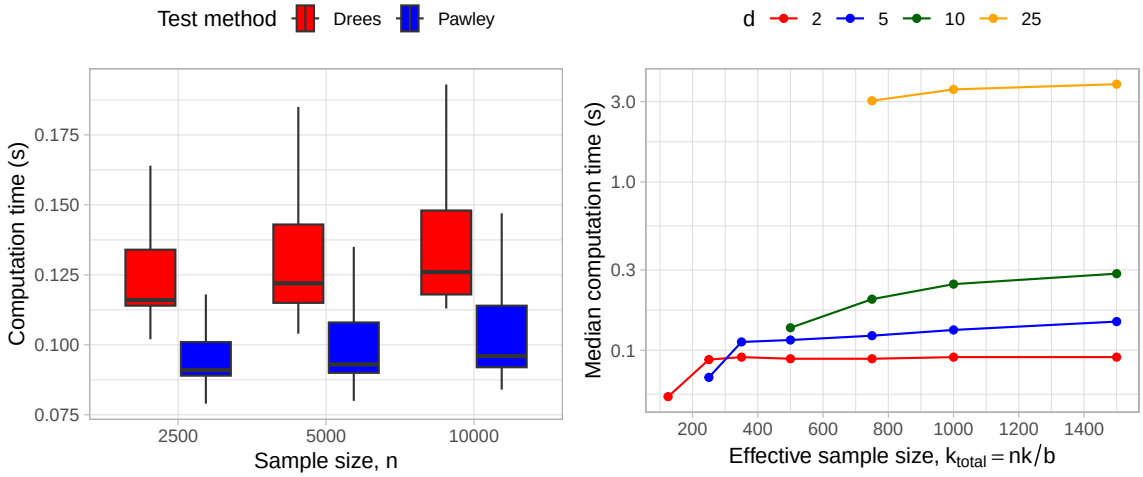

Figure 2.5: Analysis of computation times across the numerical experiments described in Section XX. Left: empirical distribution of run times when $d = 2$, $n/b = 50$, $k/b = 0.1$. Right: median computation time for the test as a function of the total number of threshold exceedances.

2.7 Loss of power under TPDM-invariant dependence changes

Our test affords several advantages compared to **dreesStatisticalInferenceChanging2023**, 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 (**dreesStatisticalInferenceChanging2023**). We exemplify this using a time-dependent generalisation of the max-linear model.

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

$$\boldsymbol{X}(t) = A(t) \times_{\max} \boldsymbol{Z}(t), \qquad A(t) = A_0 \mathbf{1}\{t < 0.5\} + A_1 \mathbf{1}\{t \geq 0.5\}. \tag{2.22}$$

The stochastic innovations process $\{\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_q(t)) : t \in [0,1]\}$ is a collection of independent random vectors 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 (2.22), $A(t)$ 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 (2.22) can be expressed in terms of the columns $\boldsymbol{a}_1(t), \ldots, \boldsymbol{a}_q(t) \in \mathbb{R}_+^d$ 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 1.7, 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 and A_1, with $A_0 \neq A_1$ including up to permutations of their columns, are such that $\Sigma(0) = \Sigma(1)$ and $V(0) = V(1)$. Then the alternative hypothesis (2.3) is true but the asymptotic distributions of $T^{(KS)}$ and $T^{(CM)}$ are the null distributions in (2.18). 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 suitable pair among these. This process yields the matrices shown in Figure 2.6. 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 (2.22) gives $\sigma_{12}(t) = 0.100$ and $\nu_{12}^2 = 0.060(t)$ for all $t \in [0, 1]$.

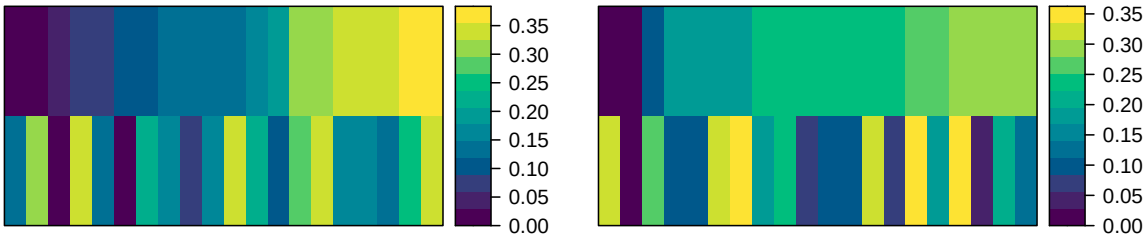

Figure 2.6: A pair of max-linear parameter matrices A_0 (left) and A_1 (right) such that the process (2.22) satisfies $\sigma_{12}(t) = 0.100$ and $\nu_{12}^2 = 0.060(t)$.

We generate 1,000 realisations of (2.22) 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 2.7 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 (1.59) 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 $\sup_{0\le s\le t} |\hat{Z}_{12}(s)|$ (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 2.8 presents analogous plots based on Drees' testing method applied to the same data. The left-hand plot shows the normalised empirical integrated angular measure $d^{-1}\hat{\mathrm{I}}\hat{\mathrm{H}}(A_y; t)$ as a function of t. Each curve corresponds to a particular set $A_y \in \mathcal{A}$ defined in (2.21) with darker colours indicating larger values of y. Close inspection of these

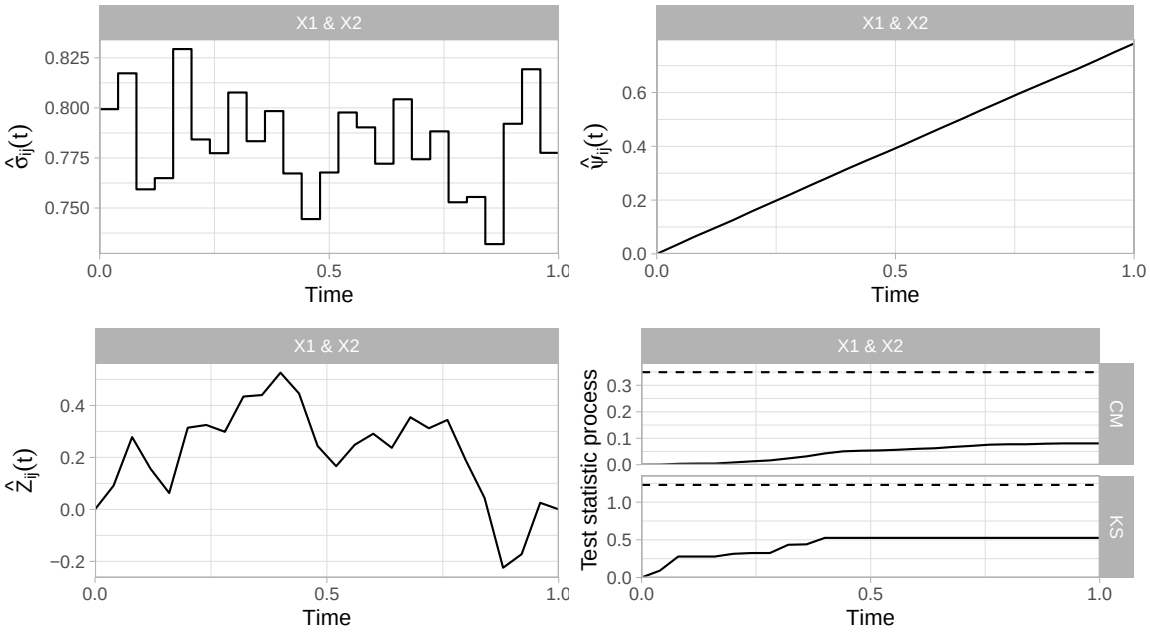

Figure 2.7: Diagnostic plots for our testing procedure based on one realisation of (2.22) with A_0 and A_1 as shown in Figure 2.6 and $n = 10,000$. The tuning parameters are $b = 400$ and $k = 40$. Top-left: the empirical local TPDM $\{\hat{\sigma}_{12}(t) : t \in [0,1]\}$. Top-right: the empirical integrated 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).

curves reveals that some of them are slightly kinked at $t = 0.5$. The middle panel visualises the processes

$$\hat{Z}_y(t) := \sqrt{\frac{k}{2h}}(\widehat{\mathrm{IH}}(A_y; t) - t\widehat{\mathrm{IH}}(A_y; 1)), \qquad (A_y \in \mathcal{A}). \tag{2.23}$$

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 (**dreesStatisticalInferenceChanging2023**). 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.

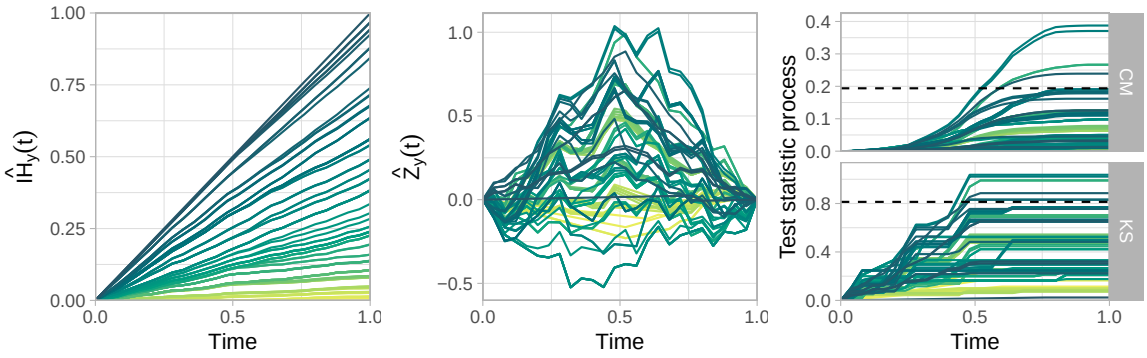

Figure 2.8: Diagnostic plots for Drees' testing procedure based on one realisation of (2.22) with A_0 and A_1 as shown in Figure 2.6 and $n = 10,000$. The tuning parameters are $b = 400$ and $k = 40$. Each curve represents a set A_y with darker colours indicating larger values of y. Left: the empirical 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 $\sup_{0 \le s \le t} |\hat{Z}_y(s)|$ (lower sub-panel) as functions of t along with the CM and KS critical values at the 5% level (dashed line).

2.8 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-camiloBayesianSpacetimeGap2021**; **rohrbeckSpatiotemporalModelRed2021**; **simpsonConditionalModellingSpatiotemporal2021**) having been the subject of the EVA (2019) Data Challenge (**huserEditorialEVA20192021**). Further details about the data set and pre-processing can be found in **huserEditorialEVA20192021**. non-stationarity in the marginal distributions was handled using the approach in **castro-camiloBayesianSpacetimeGap2021**.

Detecting changes in extremal dependence in the Red Sea is of significant practical importance. Increased spatial 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) (**karnauskasInterannualVariabilitySea2018**). In a study of extreme precipitation, **jiangPrincipalComponentAnalysis2020** found evidence for a positive temporal trend coefficients associated to principal eigenvectors related with ENSO. Moreover, Figure 3 in **kakampakouSpatialExtremalModelling2024** 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 point towards the possibility of non-stationary tail dependence.

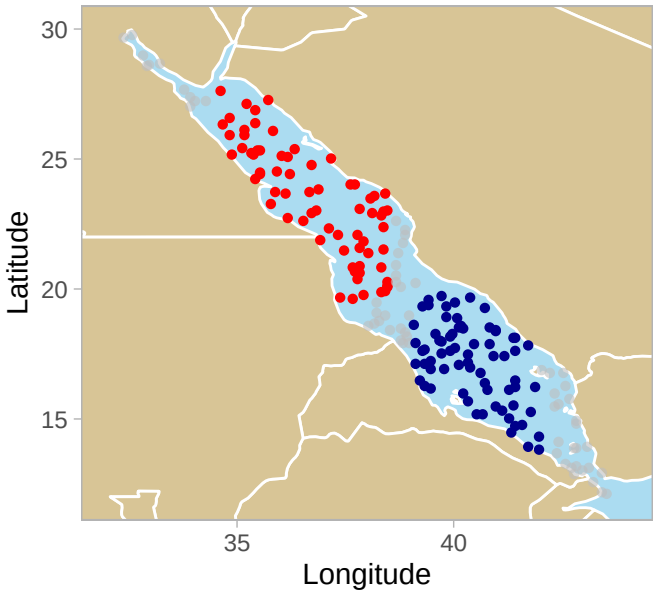

Figure 2.9: Locations of the 70 sites in each of the two sub-regions in the Red Sea.

Before applying our test, we divide the spatial domain into northerly and southerly sub-regions, each comprising 70 sites as shown in Figure 2.9. **simpsonConditionalModellingSpatiotempor** 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 (2.20) 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 2.10. 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 **kakampakouSpatialExtremalModelling2024**. 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.

2.9 Future work

2.9.1 Change-point detection

In certain applications (e.g. finance), one is not only interested in whether dependence has changed, but also *when* dependence has changed. This is the realm of change-point detection.

2.9.2 Mitigating the bias issue

Christian: did you mean moving this to the future work section of the thesis, i.e. Chapter 7? Or move it to the future work section of Chapter 6 where the new estimator is defined?

The bias issue means that the empirical TPDM struggles to discriminate between differences in dependence strengths at weak levels (Figure 1.5). As a result, the power of our

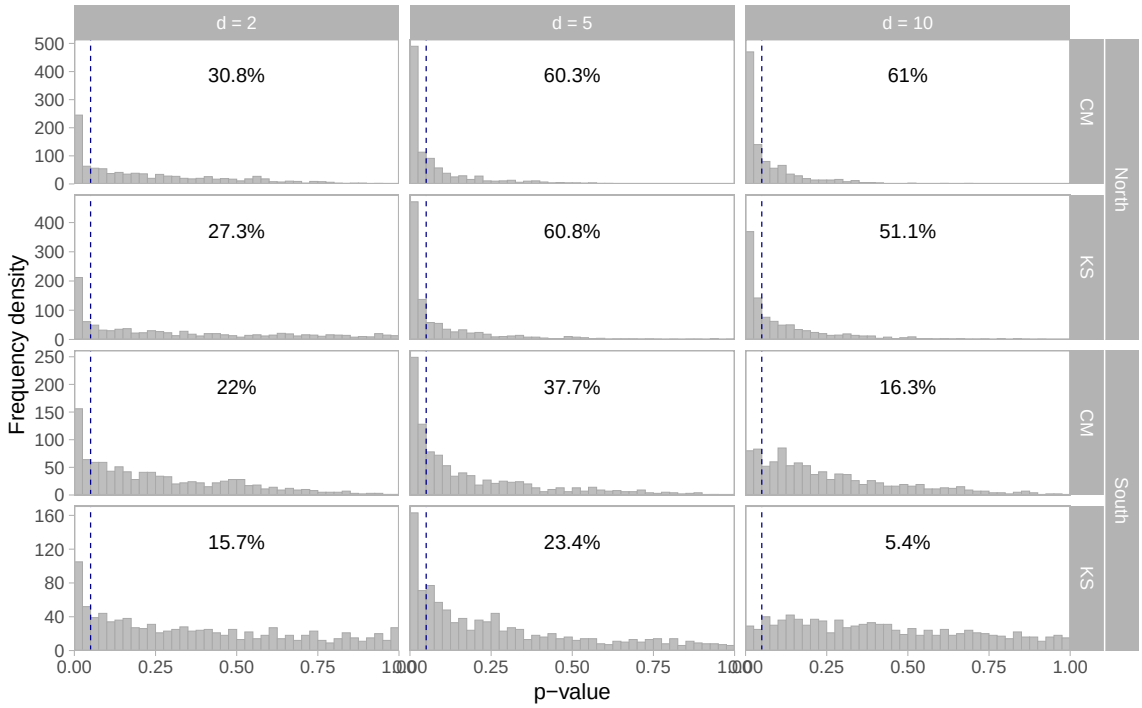

Figure 2.10: Empirical distributions of the p-values based on tests for changing dependence in the north and south regions of the Red Sea. The columns correspond to the number of re-sampled sites; the rows indicate the sub-region and the test type. The rejection rate at the 5% level is printed at the top of each panel.

test is likely dependent on the underlying dependence strength. This phenomenon is illustrated in Figure 2.13. The plots concern data generated from the SL-linear model with $d = 3$. In the top plot, the dependence parameter γ evolves linearly from $\gamma(0) = 0.06$ ($\sigma_{ij}(0) = 0.998$) to $\gamma(1) = 0.10$ ($\sigma_{ij}(0) = 0.996$). In the bottom plot, the corresponding values are $\gamma(0) = 0.95$ ($\sigma_{ij}(0) = 0.147$) to $\gamma(1) = 0.99$ ($\sigma_{ij}(0) = 0.031$). One might assume that the second dependence change is easier to detect, because the change in the TPDM is greater. However, the deficiencies of the empirical TPDM mean this is not the case. Indeed, the bottom-right sub-panels reveal that the null is only rejected for the first test. While we have not conducted a full study of the power, we expect this to replicate over repeated simulations. To address this shortcoming, one might consider employing a TPDM estimator with better finite-sample performance in weak dependence settings. Such an estimator is proposed in Chapter XX. In order to use this estimator in our test, one needs to undertake an asymptotic analysis analogous to Section XX. We conjecture that the results will follow identically, so in effect the new estimator may simply be plugged in to our testing framework.

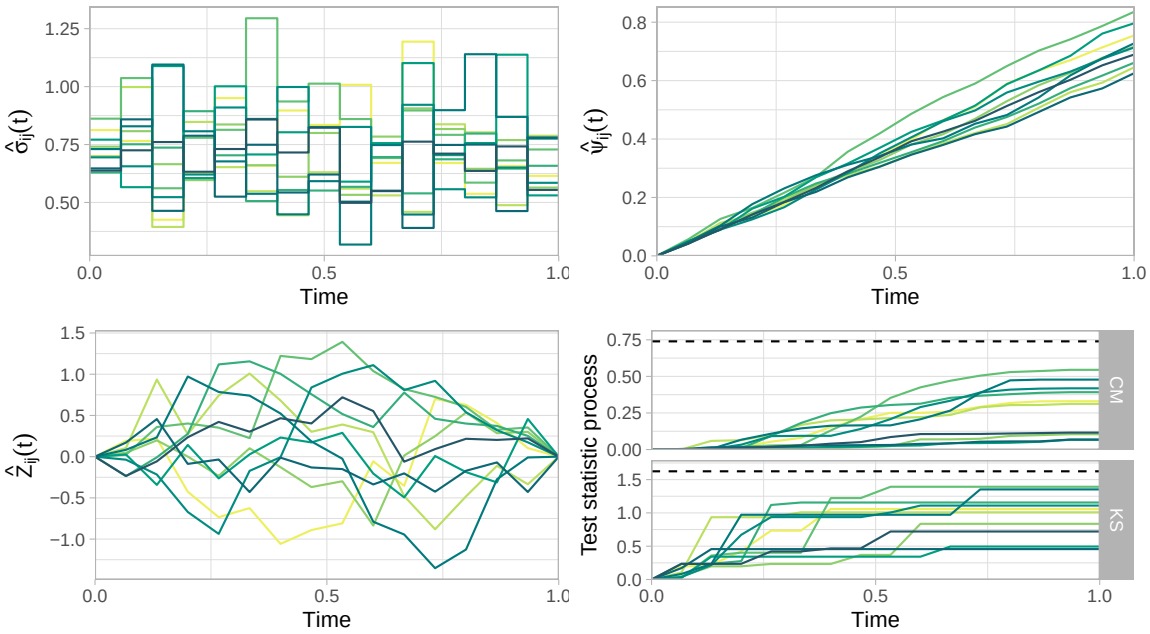

Figure 2.11: Diagnostic plots for our test, based $b = 107$ and $k = 20$, applied to data from $d = 5$ randomly selected northerly sites in the Red Sea. The interpretation of each plot is the same as in Figure 2.7, except now there are $\mathcal{D} = 10$ curves, one for each component pair. The variable 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.

Therefore power probably worse when dependence is weak. Fix this by using bias-corrected estimator.

2.9.3 Improving computational complexity

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

$$\sigma_{ij}(t) = \mathbb{E}_H(\cdot; t)[f(\boldsymbol{\Theta}(t))], \tag{2.24}$$

where $f : \mathbb{S}^{d-1}_{+(2)} \to \mathbb{R}_+$ 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?*

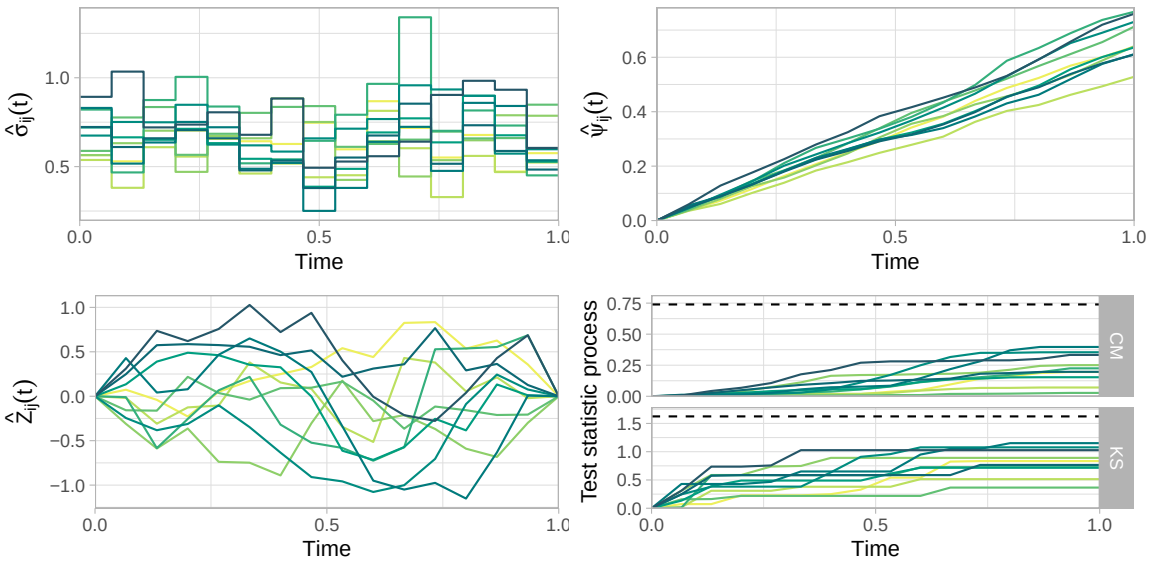

Figure 2.12: Diagnostic plots for our test, based $b = 107$ and $k = 20$, applied to data from $d = 5$ randomly selected southerly sites in the Red Sea. The interpretation of each plot is the same as in Figure 2.7, except now there are $\mathcal{D} = 10$ curves, one for each component pair. The variable 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.

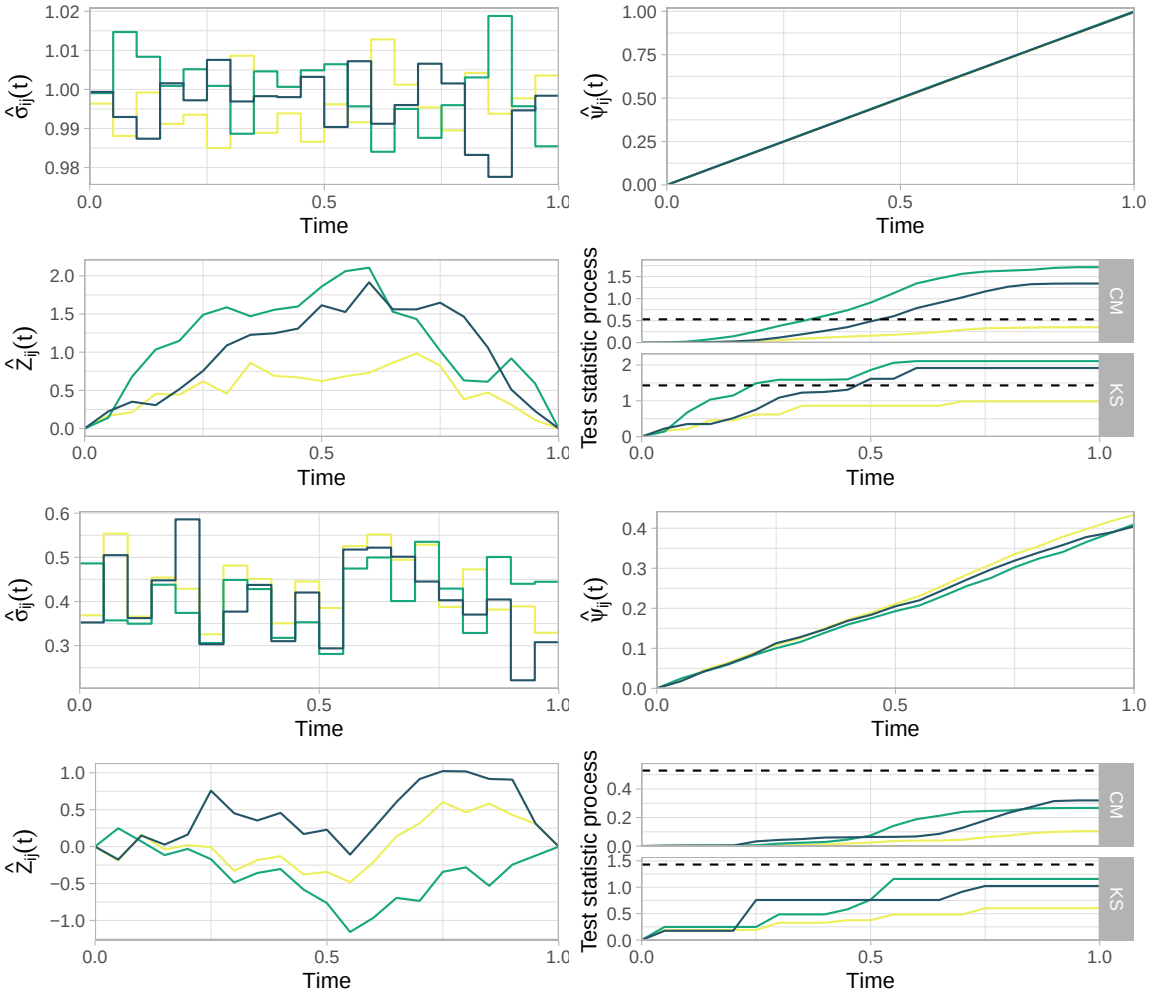

Figure 2.13: Diagnostic plots for our test based on $n = 5,000$ samples of SL-linear data in $d=3 dimensions. For the interpretation of the plot, see the caption of fig-counterexample-maxlinear-pawley. Top: a strong dependence scenario where $\gamma(0) = 0.06$ and $\gamma(1) = 0.10$. Bottom: a weak dependence scenario where $\gamma(0) = 0.95$ and $\gamma(1) = 0.99$.

3 A compositional perspective on multivariate extremes

3.1 Motivation

In multivariate extreme value statistics, the analysis of angular distributions is central to understanding the tail dependence structure. The angular measure describes the direction of large observations, focussing on the relative proportions of variables at extreme levels rather than their absolute values. Compositional data analysis (CoDA) provides tools to handle data that carry only relative information, enabling a more robust inference by accounting for the constraints imposed by their simplex structure (**aitchisonStatisticalAnalysisCompositional1986**). EVA and CoDA have been connected on several occasions: **colesModellingExtremeMultivariate1991** remark on the similarity between parametric models for CoDA and models for the angular measure; **decarvalhoStatisticsExtremesChallenges2016** sketch a possible approach to extremal PCA that involves applying "PCA for compositional data to the pseudo-angles" to "disentangle dependence into components of practical interest"; **serranoSemiparametricBivariateExtremevalue2022** construct bivariate extreme value copulas using compositional splines. Apart from these sporadic cases, the connection between these two disciplines remains unexplored in the current literature. This chapter aims to demonstrate this untapped potential.

First, we will review the core CoDA principles and concepts and draw high-level connections to MEV analysis. Then, we provide some tangible examples that illustrate our point more concretely. Specifically, we apply a CoDA lens to two statistical learning problems within multivariate extremes: tail dimension reduc-

tion via principal components analysis (**clemenconRegularVariationHilbert2024**; **cooleyDecompositionsDependenceHighdimensional2019**; **dreesPrincipalComponentAnalysis**) and binary classification in extreme regions (**jalalzaiBinaryClassificationExtreme2018**). We demonstrate that off-the-shelf CoDA methods are readily applicable to these problems and are competitive against the current state-of-the-art.

3.2 Compositional data analysis

3.2.1 Compositions and the Aitchison geometry

Compositional data analysis is a relatively modern statistical discipline that has been adopted across various fields including economics, geology, biology, political science, and many others (**alenaziReviewCompositionalData2023**). Its widespread applicability stems from the ubiquity of compositional data, i.e. data whose components represent parts of a whole and convey only relative information. For example, a geologist analysing the chemical composition of rock samples is interested in the relative abundances of its constituent elements rather than the absolute amounts (which ultimately depend on the size of the rock sample). Two 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 $\mathbb{R}^d_+$. For $\boldsymbol{x} \in \mathbb{R}^d_+$, the compositional class $[\boldsymbol{x}] := \{c\boldsymbol{x} : c > 0\}$ may be represented by a closed composition on the d-part unitary simplex,

$$\mathcal{C}\boldsymbol{x} := \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1} \in \mathbb{S}^{d-1}_{+(1)}.$$

Thus the natural sample space of compositional random vectors is the L_1-simplex $\mathbb{S}^{d-1}_{+(1)}$. In his seminal paper, **aitchisonStatisticalAnalysisCompositional1982** contends that standard multivariate analysis techniques are inappropriate for modelling compositions because they are intended for unconstrained data. Neglecting the compositional constraint causes an array of difficulties, including detecting spurious correlations (**pearsonMathematicalContributionsTheory1897**; **aitchisonStatisticalAnalysisCompositional1982**), difficulties representing the structure of the data (**aitchisonPrincipalComponentAnalysis1983**), and contradictory

conclusions between analyses (**aitchisonStatisticalAnalysisCompositional1986**). The consensus solution is to work in a statistical framework that satisfies the core CoDA principles laid out by **aitchisonStatisticalAnalysisCompositional1986**:

1. **(Scale invariance)** Analyses based on the raw observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d_+$ and on the closed compositions $\mathcal{C}\boldsymbol{x}_1, \ldots, \mathcal{C}\boldsymbol{x}_n \in \mathbb{S}^{d-1}_{+(1)}$ should yield the same results. In other words, the magnitudes of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are not relevant for inference.
2. **(Subcompositional coherence)** Analyses based on the full composition and any subcomposition should yield consistent conclusions.
3. **(Permutation invariance)** Analyses should yield equivalent results when parts in the composition are permuted.

Adherence to the foundational CoDA principles is achieved by discarding Euclidean geometry and instead working in the so-called Aitchison geometry (**aitchisonStatisticalAnalysisCompositior** Formally, this involves constructing a Hilbert space structure on the interior of $\mathbb{S}^{d-1}_{+(1)}$ (**pawlowsky-glahnGeometricApproachStatistical2001**). Suppose $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}_{+(1)} \setminus \partial\mathbb{S}^{d-1}_{+(1)}$, where

$$\partial\mathbb{S}^{d-1}_{+(1)} := \{\boldsymbol{x} \in \mathbb{S}^{d-1}_{+(1)} : \exists i \, x_i = 0\}$$

is the simplex boundary. For $\alpha \in \mathbb{R}$, the perturbation and power operations are defined by

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

The real vector space $(\mathbb{S}^{d-1}_{+(1)}, \odot, \oplus)$ has additive identity $\boldsymbol{e}_a := \mathcal{C}(1, \ldots, 1)$ and inverse elements $\ominus\boldsymbol{x} := \mathcal{C}(x_1^{-1}, \ldots, x_d^{-1})$. Central to the Aitchison geometry is the centred log-ratio (CLR) transformation

$$\operatorname{clr} : \mathbb{S}^{d-1}_{+(1)} \to \mathbb{R}^d, \qquad \boldsymbol{x} \mapsto \log\left(\frac{\boldsymbol{x}}{\bar{g}(\boldsymbol{x})}\right),$$

where $\bar{g}(\boldsymbol{x}) := (\prod_{i=1}^d x_i)^{1/d}$ denotes the geometric mean of the components of $\boldsymbol{x}$. The inverse transformation is $\operatorname{clr}^{-1}(\boldsymbol{v}) = \mathcal{C}\exp(\boldsymbol{v})$. With this transformation, the inner product may be defined in terms of $\langle\cdot,\cdot\rangle_e$, the Euclidean inner product in $\mathbb{R}^d$, as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a := \langle \operatorname{clr}(\boldsymbol{x}), \operatorname{clr}(\boldsymbol{y}) \rangle_e,$$

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

$$\|\boldsymbol{x}\|_a := \langle \boldsymbol{x}, \boldsymbol{x} \rangle_a^{1/2} = \|\mathrm{clr}(\boldsymbol{x})\|_2, \qquad d_a(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \|\mathrm{clr}(\boldsymbol{x}) - \mathrm{clr}(\boldsymbol{y})\|_2. \tag{3.1}$$

The CLR transformation is an isometry between the Aitchison simplex and the Euclidean hyperplane

$$\mathcal{T}^{d-1} := \{\boldsymbol{y} \in \mathbb{R}^d : y_1 + \ldots + y_d = 0\} \subset \mathbb{R}^d$$

passing through the origin and parallel to the simplex in $\mathbb{R}^d$. Therefore CoDA methods can be formulated in two equivalent ways: in terms of compositions in the Aitchison geometry (the in-the-simplex approach) or based on the CLR-transformed vectors in the Euclidean geometry (the out-of-the-simplex approach). If one chooses the in-the-simplex route, statistical and probabilistic concepts must be appropriately generalised. For example, **pawlowsky-glahnGeometricApproachStatistical2001** define measures of central tendency and variability for compositions as

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

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

3.2.2 Connections between CoDA and multivariate extremes

Suppose $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ and let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent copies of $\boldsymbol{X}$. Inference for H is typically based on the extremal angles $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ associated with radial threshold exceedances. This process is rigorously justified by the MRV property. Extreme events may be represented by their angles with zero loss of information since the radii $R_{(1)}, \ldots, R_{(k)}$ are uninformative for H, except insofar as they must exceed a high threshold. Extremal dependence modelling therefore proceeds under an assumption of scale invariance, the first axiom of CoDA. Moreover, choosing the L_1-norm implies that H is a distribution on $\mathbb{S}_{+(1)}^{d-1}$ and $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ are random compositions in $\mathbb{S}_{+(1)}^{d-1} \setminus \partial\mathbb{S}_{+(1)}^{d-1}$ almost surely. Thus CoDA techniques are readily applicable to the kind of data that arise in such analyses.

3.2.3 Comparison between Euclidean and Aitchison distances on the simplex

The Aitchison distance is a simplicial metric (i.e. satisfies the CoDA axioms) and is therefore an appropriate measure of distance between compositions (**aitchisonCriteriaMeasuresCompositio**
In general, the Aitchison dissimilarity $\|\boldsymbol{x} \ominus \boldsymbol{y}\|_a$ between two compositions $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}_{+(1)} \setminus \partial\mathbb{S}^{d-1}_{+(1)}$ behaves very differently to $\|\boldsymbol{x} - \boldsymbol{y}\|_2$. Figure 3.1 provides further insight into this. Each ternary plot graphs the distance between $d_\bullet(\boldsymbol{x}, \boldsymbol{y})$ as a function of $\boldsymbol{y}$ with respect to some fixed point $\boldsymbol{x}$, where $\bullet \in \{a, e\}$. The fixed points are $\boldsymbol{x} = (1/3, 1/3, 1/3)$ (top row) and $\boldsymbol{x} = (0.1, 0.45, 0.45)$ (bottom row). The left- and right-hand plots correspond to Euclidean and Aitchison distances, respectively. Consider the plots in the top row. The Euclidean distance is bounded by $\sqrt{2/3} \approx 0.82$, the distance between the centroid and any vertex, since the simplex is a bounded subspace of $\mathbb{R}^d$. In contrast, the Aitchison distance $d_a(\boldsymbol{x}, \boldsymbol{y})$ is unbounded. CLR transforms are based on log-ratios and $\log y$ diverges to $\pm\infty$ as $y \to 0$ or $y \to \infty$. Thus, points close to the boundary should be regarded as being close to infinity (**parkKernelMethodsRadial2022**). In the top-left plot, distance decays in a concentric fashion. Consequently, the point labelled A, which may be interpreted as being arbitrarily close to its neighbouring edge, is deemed closer to the centre than point B. On the other hand, the Aitchison contours are decidedly non-concentric and the points C and D are approximately equidistant to the centre. In the bottom plots, the fixed point $\boldsymbol{x} = (0.1, 0.45, 0.45)$ is located at the 'hotspot' near the right-hand edge. Intuitively, the Euclidean metric (bottom left) says that point E is very close to $\boldsymbol{x}$, while F is much further away. However, in the Aitchison geometry (bottom right) the distance between G and $\boldsymbol{x}$ becomes arbitrarily large as G is nudged towards the boundary, so that eventually H is actually closer to $\boldsymbol{x}$ than G!

These examples demonstrate that, when analysing data on the simplex, the choice of distance metric is of huge practical importance and should not dismissed as merely a mathematical technicality.

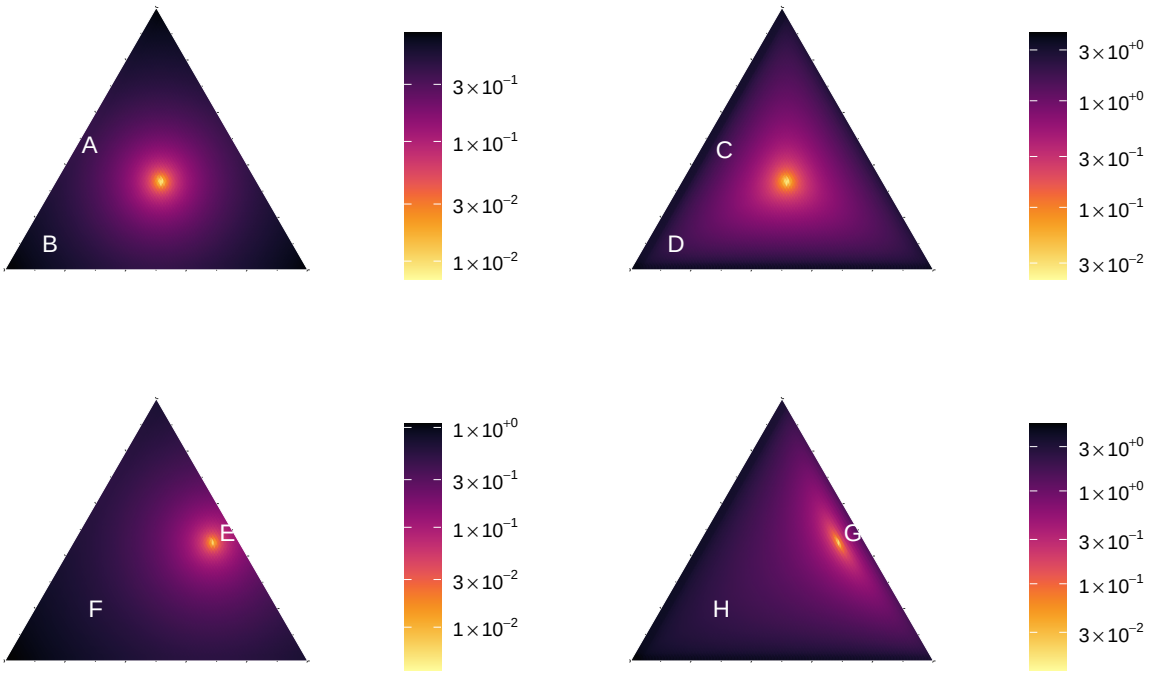

Figure 3.1: Visualisation of the Euclidean distance (left) and Aitchison distance (right) between points on $\mathbb{S}^{d-1}_{+(1)}$ and the reference points $(1/3, 1/3, 1/3)$ (top) and $(0.1, 0.45, 0.45)$ (bottom). The labelled points A-H are referred to in the body text.

3.3 Compositional PCA for extremes

3.3.1 Motivation

Compositional principal component analysis (CoDA-PCA) aims at finding low-dimensional representations of compositional data via projections onto linear subspaces of the Aitchison simplex (**aitchisonPrincipalComponentAnalysis1983**). It is inadvisable to apply classical PCA to compositions for three reasons. First, spurious correlations arising from the compositional constraint mean it is unreliable as an exploratory tool for analysing the correlation structure between variables (**aitchisonStatisticalAnalysisCompositional1982**). Second, compositional data often exhibit curvature that cannot be captured by traditional linear methods, leading to a loss of efficiency (**aitchisonPrincipalComponentAnalysis1983**). Third, interpretability is hindered by the fact that low-dimensional projections of the data are not guaranteed to lie in the simplex. These problems are solved by performing PCA in the Aitchison Hilbert space.

3.3.2 Methodology

Suppose $\boldsymbol{X} \in \mathcal{RV}_+^d(\alpha)$ and let H be the angular measure with respect to the L_1-norm $\|\cdot\|_1$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random vectors with the same distribution as $\boldsymbol{X}$. Following the spirit of **dreesPrincipalComponentAnalysis2021**, we seek to minimise the limiting mean-squared reconstruction error associated with the extremal angles $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(k)}$ upon projecting onto low-dimensional linear subspace. For any Aitchison subspace $\mathcal{S} \subset \mathbb{S}_{+(1)}^{d-1}$, let $\Pi_{\mathcal{S}}$ denote the orthogonal projection (matrix) onto $\mathcal{S}$ and define the true asymptotic risk

$$R(\mathcal{S}) := \mathbb{E}_H[\|\boldsymbol{\Theta} \ominus \Pi_{\mathcal{S}}\boldsymbol{\Theta}\|_a^2].$$

For some $1 \leq p \leq d-1$, let $\mathcal{V}_p$ be the class of linear subspaces of dimension p in $\mathbb{S}_{+(1)}^{d-1}$. As usual, the minimisers of $R(\mathcal{S})$ over $\mathcal{V}_p$ corresponds to the principal eigenspace of a covariance matrix. The following result explains how the CoDA-PCA procedure outlined by **aitchisonStatisticalAnalysisCompositional1982** applies to our setting.

Proposition 3.1. *Let $\tilde{\boldsymbol{\Theta}} \sim d^{-1}H$ be a random composition. Assume $\tilde{\boldsymbol{\Theta}}$ is centred (without loss of generality) and has finite total variance. Let $\Gamma = \mathrm{Cov}(\mathrm{clr}(\tilde{\boldsymbol{\Theta}}))$ be the CLR-covariance matrix of $\tilde{\boldsymbol{\Theta}}$ and $\Gamma = \Omega D \Omega^T$ its eigendecomposition, where D is a diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{d-1} \geq \lambda_d = 0$ and Ω is an orthonormal $d \times d$ matrix whose columns are the eigenvectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{d-1} \in \mathcal{T}^{d-1}$ and $\boldsymbol{\omega}_d \propto \mathbf{1}_d$. The back-transformed eigenvectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d-1}\} := \{\mathrm{clr}^{-1}(\boldsymbol{\omega}_1), \ldots, \mathrm{clr}^{-1}(\boldsymbol{\omega}_{d-1})\}$ form an orthonormal basis of $\mathbb{S}_{+(1)}^{d-1}$ and for any $1 \leq p \leq d-1$,*

$$\mathcal{S}_p^\star := \mathrm{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\} = \left\{ \bigoplus_{j=1}^{p} (\tau_j \odot \boldsymbol{u}_j) : \tau_1, \ldots, \tau_p \in \mathbb{R} \right\}$$

minimises R over $\mathcal{V}_p$.

Proof. Let $\tilde{\boldsymbol{Y}} := \mathrm{clr}(\tilde{\boldsymbol{\Theta}})$. By assumption, $\tilde{\boldsymbol{Y}}$ has zero mean and finite second order moments, since

$$\mathbb{E}[\tilde{\boldsymbol{Y}}] = \arg\min_{\boldsymbol{\mu}} \mathbb{E}[d_e^2(\tilde{\boldsymbol{Y}}, \boldsymbol{\mu})] = \arg\min_{\boldsymbol{\mu}} \mathbb{E}[d_a^2(\tilde{\boldsymbol{\Theta}}, \mathrm{clr}^{-1}(\boldsymbol{\mu}))] = \mathrm{clr}(\boldsymbol{e}_a) = \mathbf{0}$$

and

$$\mathrm{Var}(\tilde{Y}_j) = \mathrm{Var}([\mathrm{clr}(\tilde{\boldsymbol{\Theta}})]_j)) < \infty.$$

Using well-known results from classical PCA – e.g. Theorem 5.3 in **seberMultivariateObservations1984** – we have that

$$\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p\} = \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[\|\tilde{\boldsymbol{Y}} - \Pi_{\operatorname{span}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\}}\tilde{\boldsymbol{Y}}\|_2^2\right].$$

Observe that for any scalar $a \in \mathbb{R}$ and compositional vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}_{+(1)}$,

$$\operatorname{clr}((a \odot \boldsymbol{x}) \oplus \boldsymbol{y}) = a \operatorname{clr}(\boldsymbol{x}) + \operatorname{clr}(\boldsymbol{y})$$

and therefore

$$\begin{aligned}
\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p\} &= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[\|\tilde{\boldsymbol{Y}} - \Pi_{\operatorname{span}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\}}\tilde{\boldsymbol{Y}}\|_2^2\right] \\
&= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}[d_e^2(\tilde{\boldsymbol{Y}}, \Pi_{\operatorname{span}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\}}\tilde{\boldsymbol{Y}})] \\
&= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}[d_a^2(\tilde{\boldsymbol{\Theta}}, \operatorname{clr}^{-1}(\Pi_{\operatorname{span}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\}}\tilde{\boldsymbol{Y}}))] \\
&= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[d_a^2\left(\tilde{\boldsymbol{\Theta}}, \operatorname{clr}^{-1}\left(\sum_{j=1}^{p}\langle\tilde{\boldsymbol{Y}}, \boldsymbol{v}_j\rangle\boldsymbol{v}_j\right)\right)\right] \\
&= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[d_a^2\left(\tilde{\boldsymbol{\Theta}}, \bigoplus_{j=1}^{p}\langle\tilde{\boldsymbol{\Theta}}, \operatorname{clr}^{-1}(\boldsymbol{v}_j)\rangle_a \odot \operatorname{clr}^{-1}(\boldsymbol{v}_j)\right)\right].
\end{aligned}$$

This optimisation depends on each vector $\boldsymbol{v}_j$ only through $\operatorname{clr}^{-1}(\boldsymbol{v}_j)$, so can map the problem to the simplex, yielding

$$\begin{aligned}
\{\operatorname{clr}^{-1}(\boldsymbol{\omega}_1), \ldots, \operatorname{clr}^{-1}(\boldsymbol{\omega}_p)\} &= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{S}^{d-1}_{+(1)}}{\arg\min}\ \mathbb{E}\left[d_a^2\left(\tilde{\boldsymbol{\Theta}}, \bigoplus_{j=1}^{p}\langle\tilde{\boldsymbol{\Theta}}, \boldsymbol{v}_j\rangle_a \odot \boldsymbol{v}_j\right)\right] \\
&= \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p \in \mathbb{S}^{d-1}_{+(1)}}{\arg\min}\ R(\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\})
\end{aligned}$$

Since $\mathcal{S}_p^\star = \operatorname{span}\{\operatorname{clr}^{-1}(\boldsymbol{\omega}_1), \ldots, \operatorname{clr}^{-1}(\boldsymbol{\omega}_p)\}$, this completes the proof.

□

The finite total variance condition prohibits the angular measure from placing mass on the simplex boundary and restricts how much mass may be concentrated near the boundary. A sufficient condition is $\operatorname{Var}_H(\log(\Theta_i/\Theta_j)) < \infty$ for all $i \neq j$. No such condition is

required by **dreesPrincipalComponentAnalysis2021** because finite variances are guaranteed by boundedness of the simplex in $\mathbb{R}^d$. The assumption that $\tilde{\boldsymbol{\Theta}}$ is centred at the simplex barycentre is not equivalent to the moment constraint (1.12) on H. Informally, the moment constraints dictate the arithmetic mean of H, whereas the compositional centre represents the geometric mean. Finally, we point out that $\operatorname{rank}(\Gamma) = d - 1$ in accordance with the dimensions of $\mathbb{S}^{d-1}_{+(1)}$ and $\mathcal{T}^{d-1}$. For compositional PCA we need only consider reconstructions up to rank $d-1$, whereas **dreesPrincipalComponentAnalysis2021** and **cooleyDecompositionsDependenceHighdimensional2019** require all d components to be retained to guarantee perfect reconstruction (in general).

In practice H is unknown, so we must resort to minimising an empirical risk based on the k most extreme data points. The empirical risk is defined by

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

By replacing H with $\hat{H}$ in Proposition 3.1, one can show that the minimiser of $\hat{R}$ in $\mathcal{V}_p$ is simply the principal subspace

$$\hat{S}_p = \operatorname{span}\{\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_p\},$$

where $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_{d-1} \in \mathbb{S}^{d-1}_{+(1)}$ are the back-transformed principal eigenvectors of the empirical CLR-covariance matrix (**egozcueModellingCompositionalData2018**)

$$\hat{\Gamma} = \frac{1}{k-1}\sum_{i=1}^{k} \operatorname{clr}(\boldsymbol{\Theta}_{(i)} \ominus \bar{\boldsymbol{\Theta}})\operatorname{clr}(\boldsymbol{\Theta}_{(i)} \ominus \bar{\boldsymbol{\Theta}})^T, \qquad \bar{\boldsymbol{\Theta}} = \frac{1}{k}\bigoplus_{i=1}^{k} \boldsymbol{\Theta}_{(i)}.$$

Our setup is identical to **dreesPrincipalComponentAnalysis2021**, except that all algebraic-geometric concepts (linear, projection, orthogonal, etc.) are interpreted in the Aitchison sense. This subtle change fundamentally alters the problem in two different ways. While we still optimise over the class of linear subspaces, we refer to a different notion of linearity to **dreesPrincipalComponentAnalysis2021**. Linear trends are now represented by compositional straight lines $\{\boldsymbol{a} \oplus (\tau \odot \boldsymbol{b}) : \tau \in \mathbb{R}\} \subset \mathbb{S}^{d-1}_{+(1)}$ for some $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{S}^{d-1}_{+(1)}$. From a Euclidean perspective these 'lines' are curved, so compositional PCA is able to capture features of the data that appear non-linear – see Figure 1b in

aitchisonPrincipalComponentAnalysis1983 for an illustration of this phenomenon. The second crucial difference is that our objective criterion perceives reconstruction errors in the Aitchison rather than Euclidean metric. Section XX showed that these two measures can produce wildly different assessments of how close two points are; in fact, for any $\epsilon > 0$ and $\varepsilon > 0$, there exists $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}_{+(1)}$ such that $d_e(\boldsymbol{x}, \boldsymbol{y}) < \epsilon$ and $d_a(\boldsymbol{x}, \boldsymbol{y}) > \epsilon$. The Aitchison loss is especially sensitive to reconstruction errors near the boundary, steering the optimiser to target these regions more closely.

3.3.3 Simulation experiments

We now compare the performance of our proposed procedure against **dreesPrincipalComponentAnalys** Herein referred to as CoDA-PCA and DS-PCA, respectively. Let $\boldsymbol{X} \in \mathcal{RV}^d_+(1)$ be on 1-Fréchet margins with angular measure H. We generate independent realisations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ and, for a fixed level k, compute the optimal linear subspaces $\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_{p-1}$ using CoDA-PCA and DS-PCA.

3.3.3.1 Performance metrics

We assess performance in two ways: mean-squared reconstruction error (MSRE) and tail event probability estimation. To quantify reconstruction error we use the asymptotic risk of $\mathcal{S}$ under the Aitchison and Euclidean metrics, i.e.

$$\mathcal{L}_a(\mathcal{S}) = \mathbb{E}_H[\|\boldsymbol{\Theta} \ominus \Pi_{\mathcal{S}}\boldsymbol{\Theta}\|_a^2 \mathbf{1}\{\Pi_{\mathcal{S}}\boldsymbol{\Theta} \in (0,\infty)^d\}],$$
$$\mathcal{L}_e(\mathcal{S}) = \mathbb{E}_H[\|\boldsymbol{\Theta} - \Pi_{\mathcal{S}}\boldsymbol{\Theta}\|_2^2\}].$$

These quantities can be computed to arbitrary precision using Monte Carlo estimation. To sample from the true angular measure we employ the `rmevspec` function in the `mev` package. The inclusion of $\mathbf{1}\{\Pi_{\mathcal{S}}\boldsymbol{\Theta} \in (0,\infty)^d\}$ ensures that $\mathcal{L}_a$ is well-defined in cases where DS-PCA yields reconstructed vectors that lie outside of the positive orthant. In such cases, the Aitchison distance is undefined. Naively, one might consider projecting invalid reconstructions on to the positive orthant, but this does not really resolve the core issue. Naturally such points should be sent somewhere near the coordinate axes, but by adjusting *how close* you place the point the Aitchison loss can be made arbitrarily large.

Instead, we simply ignore such points. Arguably this choice favours DS-PCA, since we do not penalise its failure to produce valid reconstructions.

Measuring reconstruction error gives a high-level summary of how closely the low-dimensional approximation of H resembles the true measure over its entire support. The second class of performance metrics focusses on assessing performance in specific regions of the simplex that are relevant for estimating tail event probabilities. Consider the estimands

$$p_{\min} := \lim_{u\to\infty} \mathbb{P}(\min \boldsymbol{X} > u \mid \|\boldsymbol{X}\|_1 > u) = \mathbb{E}_H \left[\min_{j=1,\dots,d} \Theta_j \right], \tag{3.2}$$

$$p_{\max} := \lim_{u\to\infty} \mathbb{P}(\max \boldsymbol{X} > u \mid \|\boldsymbol{X}\|_1 > u) = \mathbb{E}_H \left[\max_{j=1,\dots,d} \Theta_j \right]. \tag{3.3}$$

These probabilities specify the minimal and maximal contributions of variable to the norm. The true probabilities may be computed via Monte Carlo simulation or analytically. To assess a principal subspace $\hat{\mathcal{S}}_p$, we compute estimates of $p_{\min}$ and $p_{\max}$ using the empirical angular measure *after* projecting the data onto $\hat{\mathcal{S}}_p$. Specifically, for any subspace $\mathcal{S}$ we define

$$\hat{H}_{\mathcal{S}}(\cdot) := \frac{d}{\sum_{i=1}^k \mathbf{1}\{\Pi_{\mathcal{S}} \boldsymbol{\Theta}_{(i)} \in (0,\infty)^d\}} \sum_{i=1}^k \delta_{\Pi_{\mathcal{S}} \boldsymbol{\Theta}_{(i)}}(\cdot) \mathbf{1}\{\Pi_{\mathcal{S}} \boldsymbol{\Theta}_{(i)} \in (0,\infty)^d\}$$

and compute

$$\hat{p}_{\min}(\mathcal{S}) := \mathbb{E}_{\hat{H}_{\mathcal{S}}} \left[\min_{j=1,\dots,d} \Theta_j \right], \qquad \hat{p}_{\max}(\mathcal{S}) := \mathbb{E}_{\hat{H}_{\mathcal{S}}} \left[\max_{j=1,\dots,d} \Theta_j \right].$$

The performance of each dimension reduction method is assessed by examining the sequence of estimates $\hat{p}_{\min}(\hat{\mathcal{S}}_1), \dots, \hat{p}_{\min}(\hat{\mathcal{S}}_{p-1})$ and comparing against the true probability $p_{\min}$ and the (uncompressed) empirical estimate $\hat{p}_{\min}(\mathbb{S}^{d-1}_{+(1)})$.

3.3.3.2 Max-linear model with compositionally colinear factors

Our first experiment involves max-linear random vectors. The parameter matrix A will be constructed in a particular way so as to nicely illustrates the benefits of our proposed approach. We stock the factor matrix $A \in \mathbb{R}_+^{d\times q}$ with power-perturbation combinations of linearly independent compositions $\boldsymbol{v}_1, \dots, \boldsymbol{v}_s \in \mathbb{S}^{d-1}_{+(1)}$, that is $A = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_q)$ and for each

$j = 1, \ldots, q$,

$$\mathcal{C}\boldsymbol{a}_j = \bigoplus_{l=1}^{s} (\beta_{jl} \odot \boldsymbol{v}_l)$$

for some scalars $\beta_{j1}, \ldots, \beta_{js} \in \mathbb{R}$. Then the angular measure of the associated max-linear random vector $\boldsymbol{X} \in \mathcal{RV}_+^d(1)$ is

$$H(\cdot) = \sum_{j=1}^{q} \|\boldsymbol{a}_j\|_1 \delta_{\bigoplus_{l=1}^{s}(\beta_{jl} \odot \boldsymbol{v}_l)}(\cdot)$$

The support of H is spanned (in the Aitchison geometry) by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s$. If $s < d-1$, then the angular measure may be represented by a low-dimensional object with no information loss incurred.

We start with a low-dimensional example to facilitate visualisation. Let $d = 3$, $s = 1$, $q = 50$ and $\boldsymbol{v}_1 = (0.12, 0.58, 0.3)$. The values $\beta_{11}, \ldots, \beta_{q1}$ are sampled uniformly between -4 and 4. The resulting matrix is shown in Figure 3.3 (left). Since the angular measure is discrete, the true probabilities (3.2) and (3.3) can be computed analytically as $p_{\min} \approx 0.636$ and $p_{\max} \approx 0.068$. Realisations of $\boldsymbol{X}$ are generated using the max-stable (MS) construction $\boldsymbol{X} = A \times_{\max} \boldsymbol{Z}$ and the transformed-linear (TL) construction $\boldsymbol{X} = A \otimes \boldsymbol{Z}$ defined in (1.21) and (1.22). For both processes, we generate 50 datasets of size $n = 5 \times 10^3$ and run the PCA methods with $k/n = 0.01$. *The results for other values of k and n are very similar – see Appendix XX.*

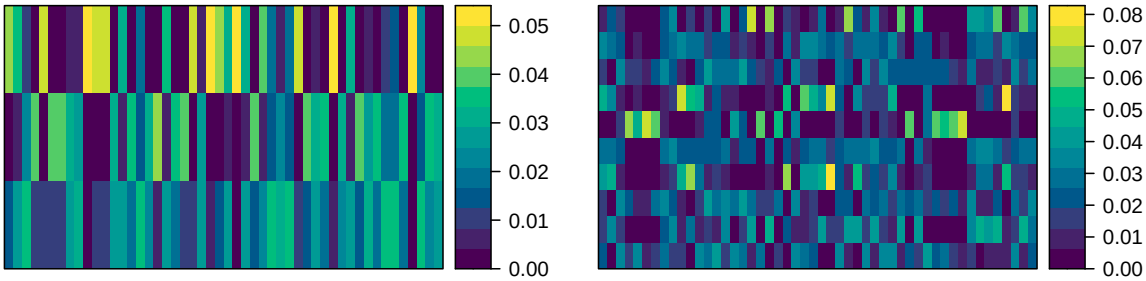

Figure 3.2: Blah.

The exploratory plots in Figure 3.3 provides some insight into the data structure and the mechanics of the two procedures. In each plot, the support of the angular measure (i.e. the points $\mathcal{C}\boldsymbol{a}_1, \ldots, \mathcal{C}\boldsymbol{a}_{50}$) is represented by the green diamonds. By construction, these points exhibit a one-dimensional structure and lie on the compositional straight line $\{\tau \odot (0.12, 0.58, 0.3) : \tau \in \mathbb{R}\}$. The black points are the extremal angles $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)}$

taken from one set of realisations from the TL max-linear random vector (left) and the MS max-linear random vector (middle and left). The TL data are noisier in the sense that the extremal angles do not tend to lie exactly on the green points, whereas the MS angles converge to the limiting distribution very quickly. The left and middle plots show the results of CoDA-PCA. The red and blue dashed lines represents the first and second (sample) principal axes, respectively. The method correctly identifies the true low-dimensional structure, though there is some estimation error visible in the TL case (left). The red crosses represent the rank-one reconstructions obtained by projecting the black points onto the red line. The rank-two reconstructions always have zero error since $\mathbb{S}^2_{+(1)}$ is a two-dimensional space. The right-hand plot depicts the results of DS-PCA for the same MS dataset. Now we add blue crosses showing the rank-two reconstructions. Clearly, there are several issues to discuss. The one-dimensional projections lie along a straight line (the process of visualising the line on a ternary plot distorts it slightly) and do a poor job of describing the data. Some points lie outside of the triangle; this means the projected vector had a negative component. The rank-two reconstructions do a much better job, but are still imperfect because DS-PCA treats the data as points in $\mathbb{R}^3$.

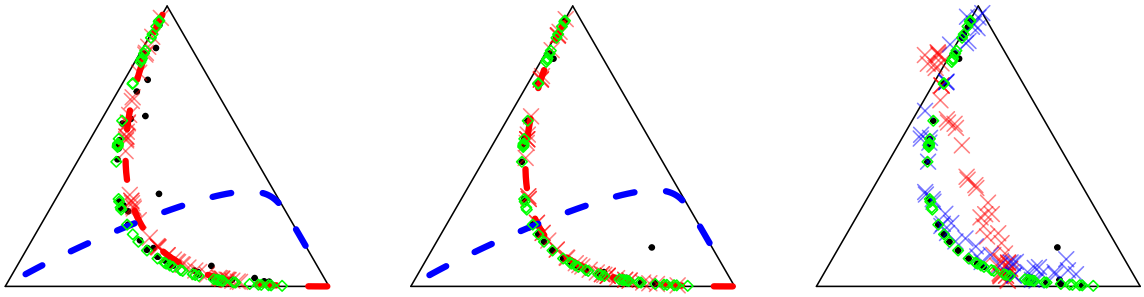

Figure 3.3: Examples of CoDA-PCA (left and middle) and DS-PCA (right) applied to max-linear data from the TL construction (left) and MS construction (middle and right). The green diamonds represent the true angular measure. The black points are the angular components associated with the $k = 50$ largest 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.

The full results are shown in Figure 3.4. The four sub-plots show the distributions of the four performance measures: the Aitchison MSRE $\mathcal{L}_a$ (top left), the Euclidean MSRE (top right), estimates $\hat{p}_{\min}$ (bottom left), and estimates $\hat{p}_{\max}$ (bottom right). Within each sub-plot, the two sub-panels are based on the two data generating processes. The horizontal

axis indicates which PCA method is being used and the number of retained principal components, p. As expected from Figure 3.3, CoDA-PCA produces excellent reconstructions when $p = 1$ under *both* the Aitchison and Euclidean loss. The picture for DS-PCA is more mixed. According to $\mathcal{L}_e$, adding a second component improves the reconstructions considerably. The Aitchison loss $\mathcal{L}_a$ disagrees, suggesting that the two-dimensional projections are poor approximations of points near the simplex boundary. This can be observed in Figure 3.3 (right) near the upper and right-hand vertices. In terms of reconstruction error, there is not any noticeable difference between the MS and TL results. Now consider the tail probability estimates in the bottom row. The boxplots are the estimates $\hat{p}_\bullet(\hat{\mathcal{S}}_p)$ and the blue dashed line marks the true probability $p_\bullet$. CoDA-PCA with $p = 1$ yields very good estimates in the sense of (i) being close to the true value and (ii) being as good as the $p = 2$ estimate resulting from the uncompressed empirical angular measure. The estimates are slightly biased for the TL data, suggesting that the extremal angles converge to their limiting distribution in such a way that biases the principal eigenvector of the empirical CLR-covariance matrix. For DS-PCA, $\hat{p}_{\min}(\hat{\mathcal{S}}_1)$ and $\hat{p}_{\max}(\hat{\mathcal{S}}_1)$ show a large positive and negative bias, respectively. Adding a second component reduces the bias considerably but the uncertainty is quite large. Euclidean loss treats all errors equally irrespective of where they occur in the simplex, so the PCA solution finds a compromise between good fit near the centre and near the boundary. In contrast, the Aitchison loss prioritises good reconstruction of points near the boundary, resulting in more stable PCA solutions. Analogous plots based on samples of size $n = 50,000$ are shown in Appendix XX. The results are virtually identical, confirming that errors are due to methodological deficiencies, not a lack of data. Even with infinite samples DS-PCA will fit a straight line to curved data!

The three-dimensional example is instructive but limited in terms of comparing dimension reducing capabilities. Our next experiment is based on the same construction with $d = 10$, $q = 50$ and $s = 2$. The vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{S}^9_{+(1)}$ and scalars $\{\beta_{jl} : j = 1, \ldots, 50,\, l = 1, 2\}$ are generated randomly producing the matrix in Figure 3.3 (right). The results in Figure 3.5 are based on 25 repeated simulations from the MS construction with $n = 10^4$ and $k/n = 0.02$. The left-hand plot shows the average cumulative proportion of the total variance as a function of the number of retained components. These are computed as $\sum_{j \le p} \hat{\lambda}_j / \sum_{j=1}^d \hat{\lambda}_j$, where $\{\hat{\lambda}_j : j = 1, \ldots, 10\}$ are the sample eigenvalues of the TPDM or CLR-covariance matrix. Note that for CoDA-PCA it is the total variance $\text{totVar}_a(\Pi_{\hat{S}_p} \boldsymbol{\Theta})$ being measured.

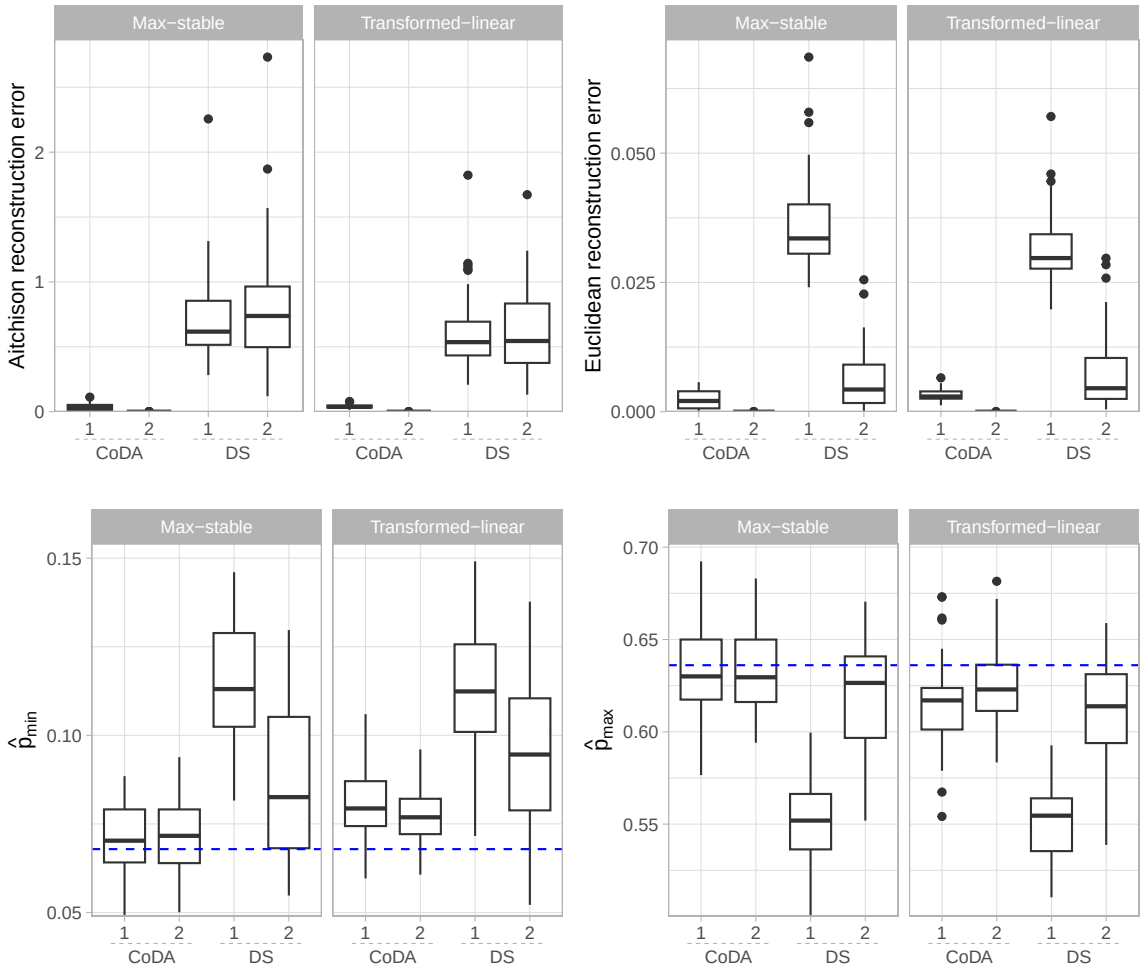

Figure 3.4: Blah.

The dashed line represents the 95% threshold. In PCA a common rule-of-thumb is to retain as many components as necessary to explain 95% of the variance. Based on this criterion, CoDA-PCA correctly identifies that two components are sufficient, whereas DS-PCA requires one additional dimension. However, this only means that three components are adequate for capturing *Euclidean* variability in the angles. Hopefully the reader is by now persuaded that this is not sufficient for a good description of extremal dependence. The right-hand plot confirms our hypothesis. Here we see estimates $\hat{p}_{\max}(\hat{\mathcal{S}}_p)$ against p for each PCA algorithm. CoDA-PCA estimates $p_{\max}$ very accurately with $p = 2$. DS-PCA requires at least six or seven components. This stark contrast emphasises our point that diagnosing PCA performance using the Euclidean notions of reconstruction error or proportion of variance can be rather misleading.

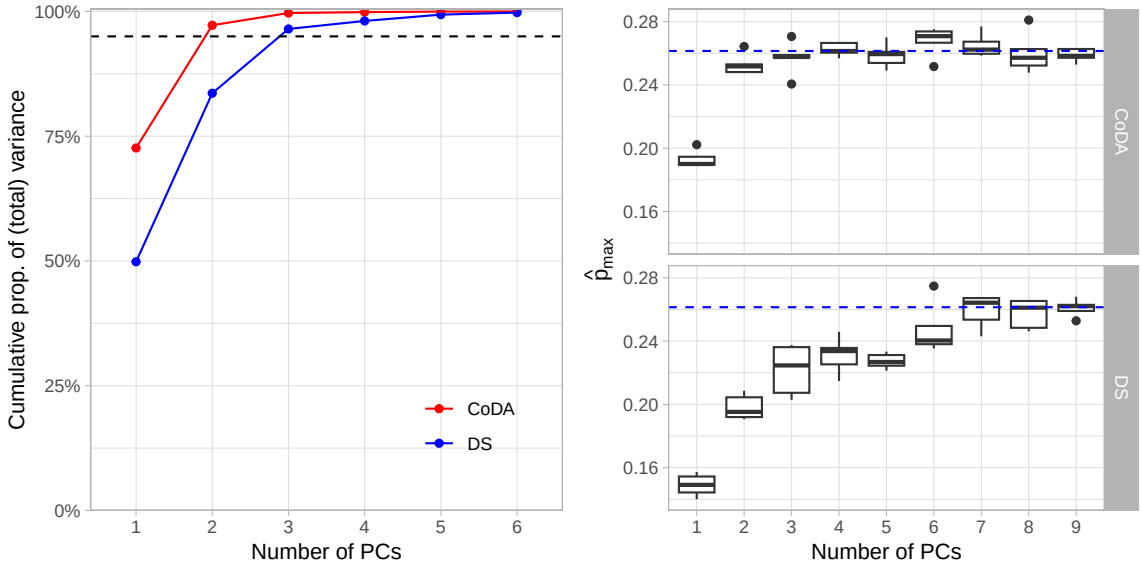

Figure 3.5: Blah.

3.3.3.3 Hüsler-Reiss

Now we consider more realistic examples where the data are generated from four different Hüsler-Reiss models. For $i = 1, \ldots, 4$, suppose $\boldsymbol{X}_i \in \mathcal{RV}_+^d(1)$ follows a Hüsler-Reiss distribution parametrised by $\Lambda_i \in \mathbb{R}_+^{d \times d}$. For the first three examples we set $d = 3$ and

$$\Lambda_1 = \begin{pmatrix} 0 & 0.025 & 0.311 \\ 0.025 & 0 & 0.168 \\ 0.311 & 0.168 & 0 \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} 0 & 0.034 & 0.009 \\ 0.034 & 0 & 0.024 \\ 0.009 & 0.024 & 0 \end{pmatrix}, \quad \Lambda_3 = \begin{pmatrix} 0 & 0.001 & 0.322 \\ 0.001 & 0 & 0.309 \\ 0.322 & 0.309 & 0 \end{pmatrix}$$

The fourth model is 10-dimensional example with $\Lambda_4 \in \mathbb{R}_+^{10 \times 10}$ constructed in such a way that there are three clusters of asymptotically dependent variables – see the corresponding pairwise tail dependence coefficients χ_{ij} in Figure 3.6 (left). The entries of $\Lambda_1, \ldots, \Lambda_4$ were randomly generated following the steps in Appendix B1 in **fomichovSphericalClusteringDetection2023**.

The exploratory plots in Figure 3.7 show that $\Lambda_1, \Lambda_2, \Lambda_3$ induce qualitatively different dependence structures. The plots' interpretation is the same as in Figure 3.3. Each ternary plot displays $k = 250$ extremal angles extracted from $n = 10^4$ samples of $\boldsymbol{X}$ under the three models. For Λ_1 (left) and Λ_3 (right) the data variability is approximately one-dimensional with differing degrees of curvature. Under Λ_2 (middle) the data cloud is two-dimensional.

The dependence structure corresponding pairwise tail dependence coefficients χ_{ij} is shown in Figure 3.6 (left). There are three clusters of asymptotically dependent variables. The pairwise dependencies are strong in clusters 1 and 3 and more variable in cluster 2. This structure can also be seen via the CLR-covariance matrix eigenvectors (Figure 3.6, right). The leading eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ separate out the three clusters. The localised behaviour of cluster 2 is captured by eigenvectors $\boldsymbol{u}_3$, $\boldsymbol{u}_4$ and $\boldsymbol{u}_5$. The remaining four (non-zero) eigenvectors relate to fine-scale behaviour in clusters 1 and 3.

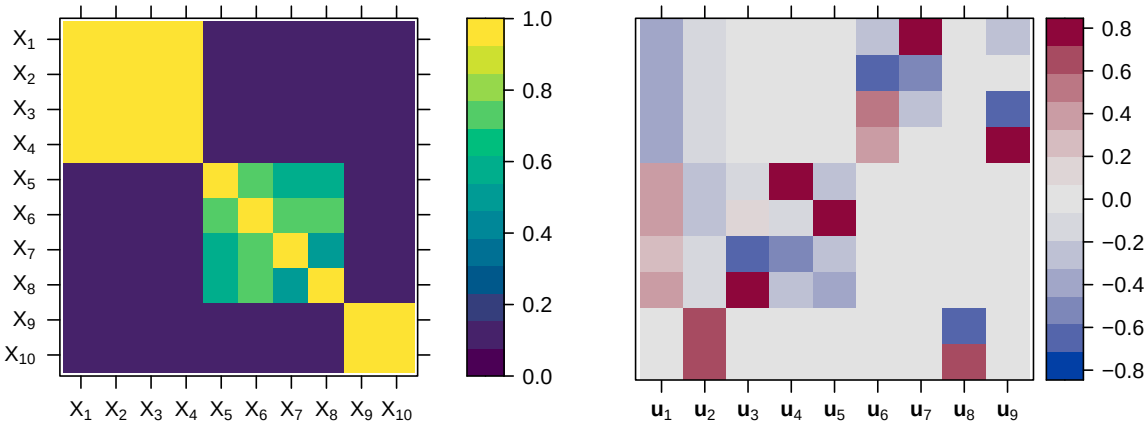

Figure 3.6: Matrix of tail dependence coefficients $\chi_{ij} = 2\bar{\Phi}(\Lambda)$.

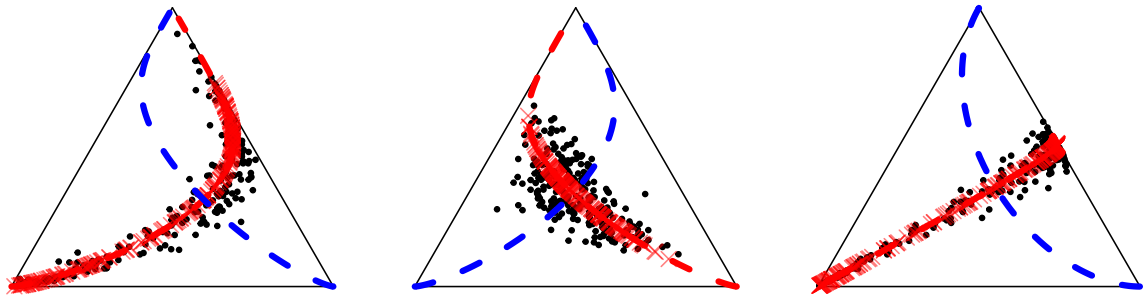

Figure 3.7: Example data from the three trivariate Hüsler-Reiss models. Based on $n = 10^4$ and $k = 250$.

Similar to before, we repeatedly generate n samples of each random vector $\boldsymbol{X}_i$, $i = 1, \ldots, 4$ and perform PCA based on the k largest observations in norm. We choose $n = 5,000$ and $k = 50$ when $d = 3$ and $n = 10^4$ and $k = 200$ when $d = 10$. The true probabilities $p_{\min}$ and $p_{\max}$ are computed empirically from $n = 10^6$ samples based on $u = 100$. To three decimal places, the true values of $p_{\min}$ (resp. $p_{\max}$) under the four sub-models are found to be 0.089, 0.228, 0.081 and 0.000 (resp. 0.603, 0.456, 0.586 and 0.385).

Figure 3.8 illustrates the Aitchison loss $\mathcal{L}_a$ for the first three models. As expected, models

1 and 3 permit an accurate one-dimensional representation, whereas model 2 requires both CoDA components. Under model 2, extremal dependence is strong and the angular measure is concentrated near the barycentre of the simplex, where the Euclidean and Aitchison geometries are most similar. Therefore both methods produce comparable outcomes. On the other hand, the extremal angles of $\boldsymbol{X}_1$ and $\boldsymbol{X}_3$ lie near the simplex boundary, so we expect the methods to diverge. It is obvious from Figure 3.7 (left) that DS-PCA with $p = 1$ will fail for model 1; this is borne out in the results. Interestingly, it also performs poorly on model 3, despite the apparently linear structure of the data. This is because lines that are visually straight in a ternary diagram do not generally correspond to Euclidean straight lines. (Line segments that are visually straight and near the simplex centre may be well-approximated by a Euclidean line segment.)

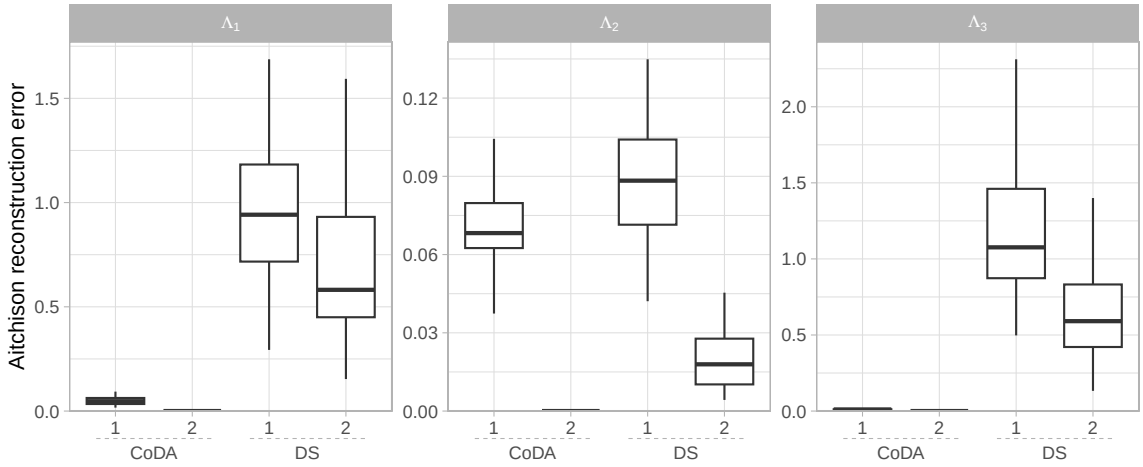

Figure 3.8: Blah.

Finally, we analyse the quality of the dimension reduction in terms of the tail event probabilities. The boxplots in Figure 3.9 represents the empirical distributions of $\hat{p}_{\min}$ (top two rows) and $\hat{p}_{\max}$ (top two rows) across the four test cases. The true probabilities are indicated by the blue dashed lines. For models 1-3, the results follow the same pattern as for the Aitchison loss: for models 1 and 3 one component is sufficient for CoDA but not for DS; for model 2 the PCA methods perform similarly. Recall that $p_{\min} \approx 0$ under model 4, owing to the fact that there are three clusters which act (almost) independently. DS-PCA is founded on the empirical TPDM, which tends to overestimate weak dependence – recall the bias issue described in Section XX – inducing a positive bias in the estimates of $p_{\min}$. The CoDA-based estimates do not suffer the same issue, suggesting the

CLR-covariance matrix is more robust in such settings. Angles corresponding to events occurring in different clusters are mapped to very distant points in CLR-space. This creates large separation between clusters, enabling accurate detection of weakly dependent variables. Both methods perform similar in terms of $p_{\max}$, requiring three or four principal components to attain good estimates. However, our CoDA estimates have lower variance and are therefore superior in terms of mean squared error.

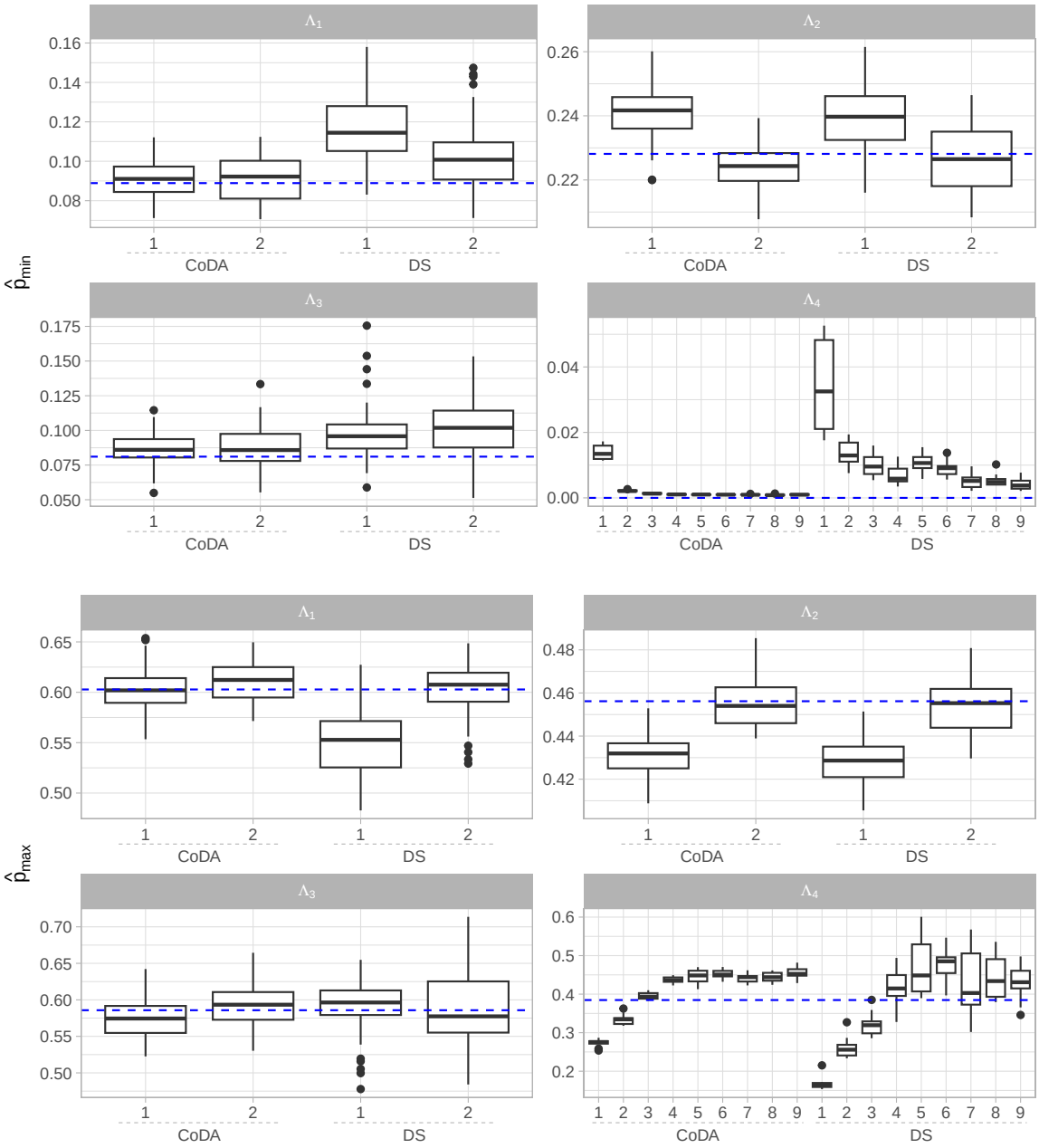

Figure 3.9: Blah.

3.4 Compositional classification for extremes

3.4.1 Motivation and framework

To lay out the framework for this section, we formulate the problem of binary classification in extreme regions as set out in **jalalzaiBinaryClassificationExtreme2018**. 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 $\|\cdot\|_1$. Suppose $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ is a labelled training set comprising independent copies of $(\boldsymbol{X}, Y)$. In general classification settings, the goal is to learn a classifier $g : \mathbb{R}^d \to \{-1, +1\}$ that minimises the expected classification error rate

$$\mathcal{L}(g) := \mathbb{P}(Y \neq g(\boldsymbol{X})).$$

By the law of total probability, this decomposes as

$$\mathcal{L}(g) = \mathbb{P}(Y \neq g(\boldsymbol{X}) \mid \|\boldsymbol{X}\| \leq t)\mathbb{P}(\|\boldsymbol{X}\| \leq t) + \mathbb{P}(Y \neq g(\boldsymbol{X}) \mid \|\boldsymbol{X}\| > t)\mathbb{P}(\|\boldsymbol{X}\| > t)$$

for some large $t > 0$. **jalalzaiBinaryClassificationExtreme2018** point out that the optimal classifier need not perform well in extreme regions of the predictor space $\{\|\boldsymbol{X}\| > t\}$, since such regions exert a negligible influence over the global prediction error. They propose 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 and a certain class $\mathcal{G}$, the optimal tail classifier $g^\star := \arg\min_{g \in \mathcal{G}} \mathcal{L}_\infty(g)$ is of the form $g^\star(\boldsymbol{x}) = g^\star(\boldsymbol{x}/\|\boldsymbol{x}\|)$ (**jalalzaiBinaryClassificationExtreme2018**). That is, the optimal classifier depends on $\boldsymbol{x}$ only through its angular component. In practice, they suggest estimating $g^\star$

by minimising the empirical risk

$$\hat{\mathcal{L}}_t(g) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\{g(\boldsymbol{\Theta}_{(i)}) \neq Y_{(i)}\}, \tag{3.4}$$

where $Y_{(i)}$ is the class label associated with $\boldsymbol{X}_{(i)}$. The remainder of their paper is devoted to providing theoretical guarantees for this learning principle. They leave aside the practical issue of designing algorithms for solving (3.4), instead resorting to general-purpose classifiers such as k-NN and random forests.

Being familiar with CoDA, we may recognise (3.4) as simply a compositional binary classification problem. The CoDA community has developed bespoke algorithms for classifying compositional data in a way that accounts for their unique geometry (**jooBinaryClassificationCompositional2021**; **martin-fernandezCriticalApproachNonParamet**
tsagrisImprovedClassificationCompositional2016). The contribution of this section is to showcase the use of off-the-shelf CoDA methods for extremal classification and demonstrate their superiority over traditional classifiers intended for unconstrained data.

3.4.2 Compositional classifiers and the α-metric

Classifiers are intrinsically linked to the geometry of the sample space. Suppose we observe training samples $(\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)$ of $(\boldsymbol{X}, Y)$ and wish to classify a new, unseen data point $\boldsymbol{x}^\star$. Consider three popular classifiers: k-nearest neighbours (k-NN), support vector machines (SVM) and random forest (RF). We assume basic familiarity with these algorithms and refer the reader to **hastieElementsStatisticalLearning2009** for more details. The k-NN algorithm allocates $\boldsymbol{x}^\star$ to the majority class among its k nearest neighbours. The notion of neighbours implicitly assumes an underlying distance metric. An SVM finds the optimal hyperplane that maximally separates data into different classes. The definition of a hyperplane is inherently dependent on the ambient space and speaking of maximal separation assumes a distance metric. Random forests are less dependent on the geometry, but it still influences the model in terms of how points are distributed over the feature space. If $\boldsymbol{X}$ is a compositional random vector, then the CoDA philosophy dictates that each algorithm should be

tailored to the geometry of the simplex. The Aitchison Hilbert space is an obvious choice. Under this approach, the data undergo a CLR-transformation before being classified in the usual way. However, **greenacreChiPowerTransformationValid2024** argues that for supervised problems where an objective performance criterion (e.g. out-of-sample classification error rate) is available, we should not be wedded to the Aitchison geometry and might consider alternatives such as the α-metric proposed by **tsagrisImprovedClassificationCompositional2016**. The α-metric is similar to the Aitchison matrix, except the CLR-transformation is substituted with a Box-Cox type transformation.

Definition 3.1. For $\alpha \geq 0$, the α-transformation is defined by

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

where H is any $(d-1) \times d$ real matrix with orthonormal rows and the $\alpha = 0$ case is interpreted as $\boldsymbol{z}_0(\boldsymbol{x}) := \lim_{\alpha \downarrow 0} \boldsymbol{z}_\alpha(\boldsymbol{x})$.

Typically H is chosen as the Helmert matrix with its first row removed, but the classification algorithms we consider are invariant to this choice. Note that α is completely unrelated to the tail index of $\boldsymbol{X}$, also denoted by α, which we have fixed equal to 1 throughout this section.

Definition 3.2. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}_{+(1)}$ 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 $\alpha = 1$ are equivalent to the Aitchison and Euclidean distances, respectively. For any Euclidean classifier $g_e : \mathbb{R}^d \to \{-1, 1\}$ and $\alpha \geq 0$, the α-transformed compositional classifier is simply

$$g_\alpha : \mathbb{S}^{d-1}_{+(1)} \to \{-1, 1\}, \qquad \boldsymbol{\theta} \mapsto g_e(\boldsymbol{z}_\alpha(\boldsymbol{\theta})).$$

Adapting k-NN, SVM and RF in this way, **tsagrisImprovedClassificationCompositional2016** define three families of compositional classification algorithms, which we denote by k-

NN(α), SVM(α) and RF(α). Each family encompasses a Euclidean- and Aitchison-based classifier corresponding to g_0 and g_1, respectively. The hyperparameter α may be selected by standard tuning procedures such as cross-validation and provides a way of comparing prediction error across different geometries. Crucially, if the classification error rate is lower for $\alpha = 0$ than for $\alpha = 1$, then it confirms that switching to a simplicial geometry is beneficial for performance.

3.4.3 Simulation experiments

For our simulation experiments, we generate realisations of $\boldsymbol{X} \mid Y = y$ in $d = 3$ dimensions on 1-Fréchet margins from one of three MEV models: symmetric logistic (SL), negative symmetric logistic (NL), and bilogistic (BL). The classes are balanced globally and asymptotically, meaning

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

The negative ($y = -1$) and positive ($y = +1$) class instances are generated using different (scalar) dependence parameters, denoted ϑ_{-1} and ϑ_1, respectively. For the SL and NL models, ϑ_σ corresponds to the parameter $1/\gamma$ and $1/\gamma$ in Definition 1.1 and Definition 1.3. The BL model sets all of its parameters equal to $\vartheta_\sigma \in [0, 1]$ – see the documentation of the `mev` package for more details.

From each model, we simulate $n = 5 \times 10^3$ training samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ and use $\{(\boldsymbol{\theta}_{(i)}, y_{(i)}) : i = 1, \ldots, k\}$ with $k = 500$ to train each tail classifier. The set of classifiers we consider is k-NN(α) and RF(α) for $\alpha \in \{0, 0.1, \ldots, 0.9, 1\}$ and SVM(α) for $\alpha \in \{0, 0.2, \ldots, 0.8, 1\}$. Implementations of these algorithms are readily available in the `Compositional` and `CompositionalML` R packages. The hyperparameters (e.g. the number of neighbours in k-NN(α)) are chosen by minimising the estimated out-of-sample classification error through 10-fold cross-validation. The use of cross-validation for estimating asymptotic classification risk is studied in more detail in **aghbalouCrossvalidationExtremeRegions2024**. The cross-validated empirical risk is used as a measure of the training error of the optimal classifier. To assess its out-of-sample, asymptotic performance, we compute an approximation of the asymptotic

risk $\mathcal{L}_\infty$ by measuring the error rate across 10^5 realisations from the true limit model, i.e. angles sampled from H_- and H_+ generated via the `rmevspec` function from the `mev` package. All reported results are based on 100 repeated simulations.

Figure 3.10 illustrates one realisation of the extremal angles in the training set for each model and different combinations of $\vartheta_{-1}, \vartheta_1$. In the top row, $\vartheta_1/\vartheta_{-1} = 1.5$ and the classes' dependence structures are very similar, making prediction very challenging. For the bottom row, the ratio becomes $\vartheta_1/\vartheta_{-1} = 3$ and the class separation becomes much greater. In general classification settings the classes may concentrate in different sub-regions of the feature space. Our setting is complicated by the fact the angular measures H_+ and H_- are subject to the constraints (1.12). It is impossible to encounter a scenario where, say, the red points are all near the bottom edge and the blue points are all near the opposing vertex. Instead, any class separation that occurs tends to relate to whether the points are close to the centre or near the boundary, as shown in the bottom middle and bottom right plots. For the other sub-models, the plots suggest that predictive performance will vary across the features space. For example, under the BL model with $\vartheta_1/\vartheta_{-1} = 1.5$ (top left), points near the right-most edge can be classified as $Y = 1$ with a reasonable degree of confidence, but points near the centre are near-impossible to classify.

Figure 3.11 plots the asymptotic risk as a function of α for each algorithm and generative process. The coloured lines are the median risk across the 100 simulations while the shaded regions demarcate the interquartile range. As expected, the ratio between 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 classifiers perform worst when $\alpha = 1$, i.e. in the Euclidean geometry. This supports our main argument that traditional multivariate techniques are ill-suited to tasks of this nature. For the BL and SL data, the asymptotic risk is minimised by taking $\alpha = 0$. For the NL data, the optimum occurs at some intermediate value, say $0.2 \leq \alpha \leq 0.4$. Thus the optimal classifiers fall somewhere under the CoDA umbrella and are significantly different from the naive Euclidean classifiers. 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. A full comparison of the relative merits of each classifier is beyond the scope of our investigation. However, it should be pointed out that for each $\vartheta_1/\vartheta_0 = 3$

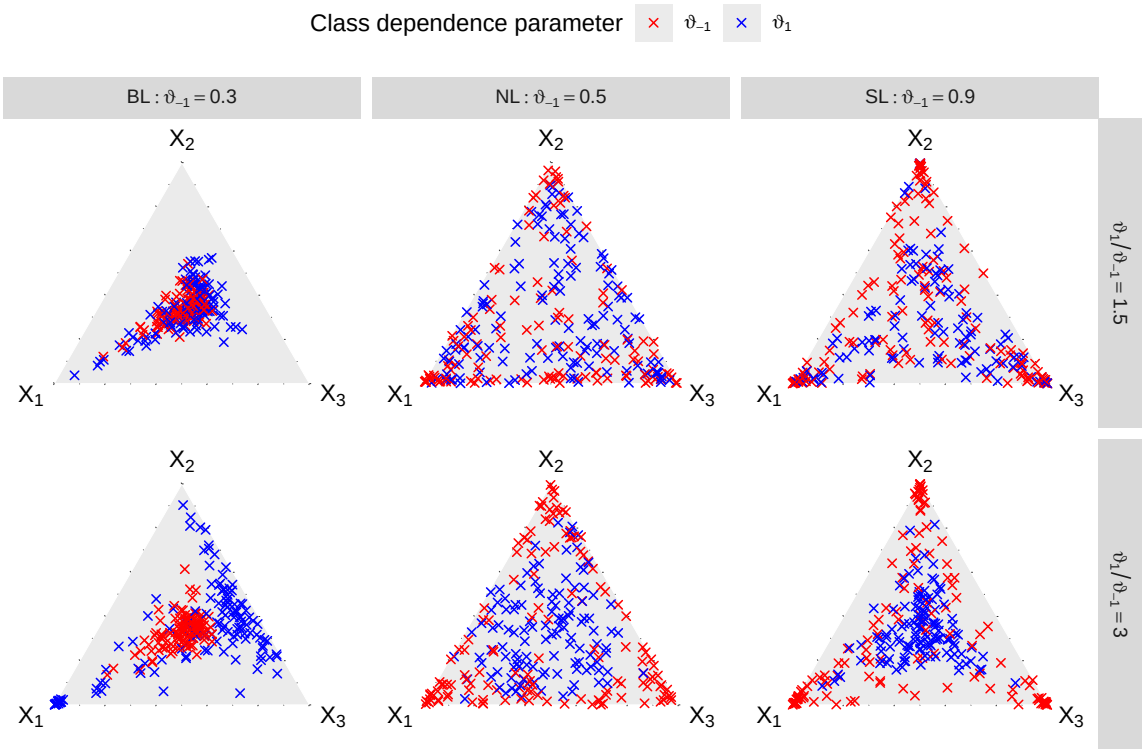

Figure 3.10: Blah.

sub-case (bottom row), the worst-performing Aitchison classifier is as good as the best-performing Euclidean classifier. Switching to a more appropriate geometry reaps similar performance benefits to utilising more sophisticated classifiers.

In practice, we cannot choose α by examining the asymptotic risk. Instead, α becomes an additional hyperparameter to be selected by minimising the cross-validation training error along with the other tuning parameters. Let $\hat{\alpha}$ denote the selected value of α (i.e. the α that attains the lowest training error) and $\alpha^\star$ the true optimum with respect to the asymptotic risk. Figure 3.12 depicts kernel density estimates of the distribution of $\hat{\alpha}$ across the 100 data sets. In accordance with Figure 3.11, the mode of $\hat{\alpha}$ is close to $\alpha^\star = 0$ for SL and BL and $\alpha^\star \approx 0.3$ for the NL model. However, there is significant variation and even $\hat{\alpha} = 1$ is selected in some instances. The profile of $\hat{\alpha}$ indicates the degree to which the difficulty of the classification task varies across geometries. For example, consider the BL model with $\vartheta_1/\vartheta_0 = 3$ (top left). Here, the large variation in $\hat{\alpha}$ can be attributed to the minimal separation between the classes (see Figure 3.10, top left). Adjusting the geometry does little to improve this predicament, so the error rate will be quite insensitive to the choice of α. Noisy estimates of $\alpha^\star$ are the result. Another example can be seen in the SL

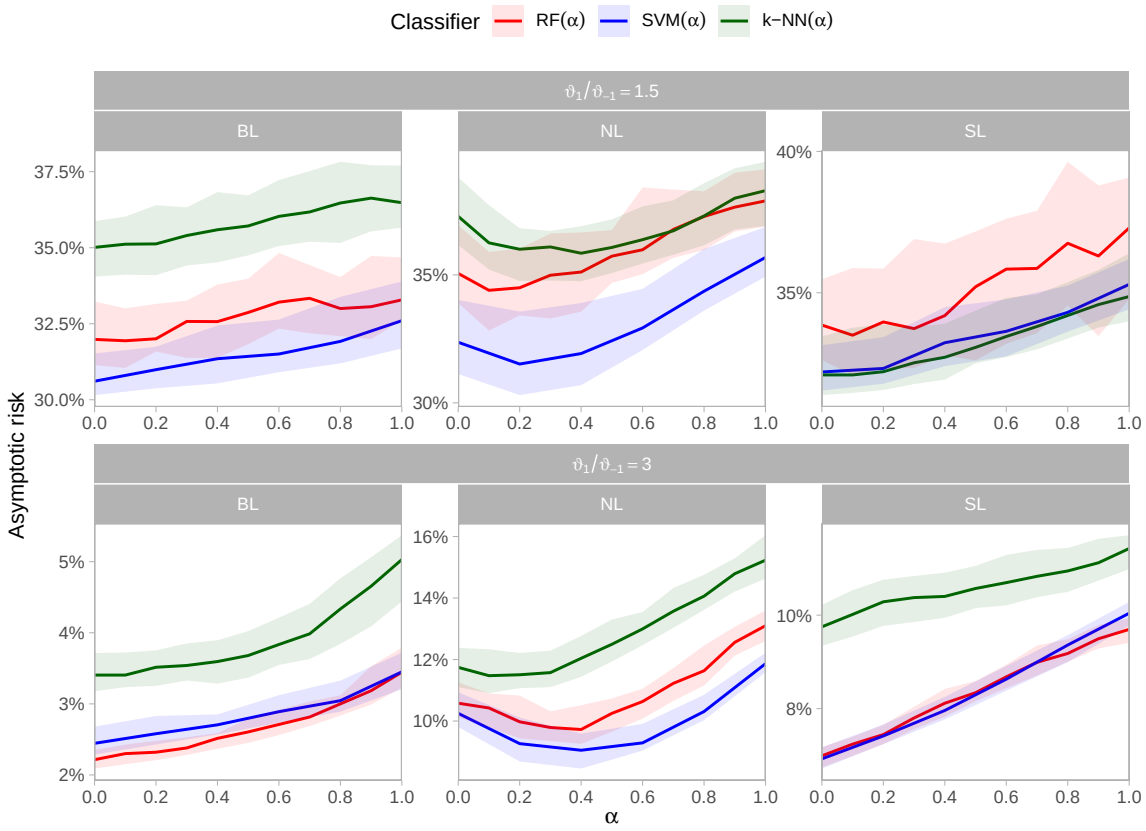

Figure 3.11: Blah.

data with $\vartheta_1/\vartheta_0 = 3$. In this case, classification is relatively easy (Figure 3.10, bottom right). However, the classes are so well separated that Euclidean nearest-neighbours are the same as Aitchison nearest-neighbours, so the performance of k-NN(α) does not change dramatically between $\alpha = 0$ and $\alpha = 1$. For SVM(α) and RF(α), adjusting the geometry affects the way they partition the feature space, so classification performance is sensitive to this choice and *hat*α concentrates near $\alpha^\star$.

3.4.4 Discussion and outlook

Discuss conclusions of CoDA PCA classification stuff here.

TO DO: - Write captions and discussion section for Compositional. - Write future work section for ChangingExtDep. - Check I have addressed all feedback points for Compositional and ChangingExtDep. - Quick proofread through Chapters 2-4 and address anything that comes up.

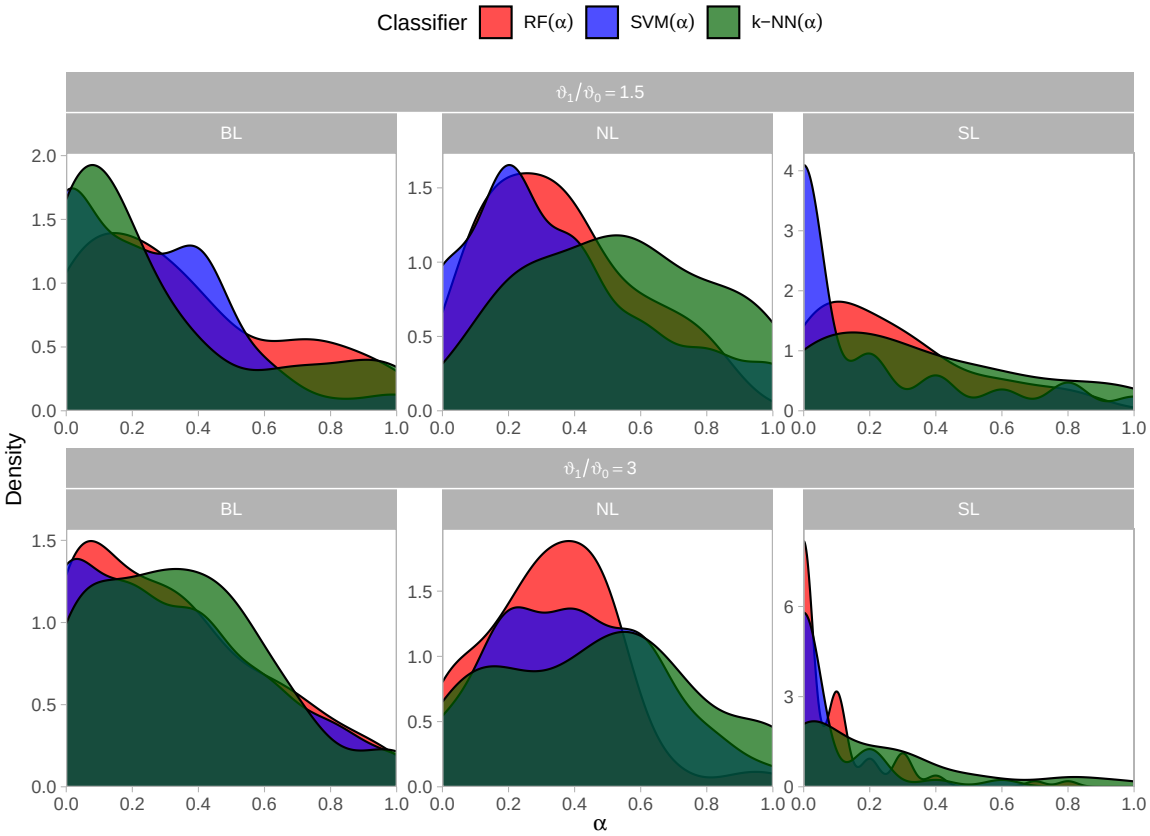

Figure 3.12: Blah.

4 EVA (2023) Data Challenge: Extreme value statistics for analysing simulated environmental extremes

4.1 Preamble

Write introductory page explaining context of the Data Challenge.

Appendix B: Statement of Authorship

(For help with copyright and permissions contact openaccess@bath.ac.uk).

This declaration concerns the article entitled:			
Publication status (tick one) **Draft manuscript** ☐ **Submitted** ☐ **In review** ☐ **Accepted** ☐ **Published** ☐			
Publication details (reference)			
Copyright status (tick the appropriate statement) The material has been published with a CC-BY license ☐ The publisher has granted permission to replicate the material included here ☐			
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the… Formulation of ideas: Design of methodology: Experimental work: Presentation of data in journal format:		
Statement from Candidate	This paper reports on original research I conducted during the period of my Doctoral Degree candidature.		
Signed (typed signature)		**Date**	

4.2 Abstract

We present the methods employed by team 'Uniofbathtopia' as part of a competition organised for the 13th International Conference on Extreme Value Analysis (EVA2023), including our winning entry for the third sub-challenge. Our approaches unite ideas from extreme value theory, which provides a statistical framework for the estimation of probabilities/return levels associated with rare events, with techniques from unsupervised statistical learning, such as clustering and support identification. The methods are demonstrated on the data provided for the EVA (2023) Conference Data Challenge – environmental data sampled from the fantasy country of 'Utopia' – but the underlying assumptions and frameworks should apply in more general settings and applications.

4.3 Introduction

In recent decades, the field of environmental sciences has experienced significant advancements, particularly through the utilisation of sophisticated modelling techniques to better understand extreme events. Extreme value analysis is the branch of statistical modelling that focuses on quantifying the frequency and severity of very rare events. Notably, the Peaks over Threshold (PoT) approach [**DavSmith1990**, **pickands75**}, has played a pivotal role in enhancing our comprehension of extreme environmental phenomena. Within climate science, significant strides have been made in the modelling of a broad spectrum of variables, including temperature (**clarkson23**), precipitation (**katz99**), wind speeds [**Kunz2010**,**FawWalsh06**}, as well as other broader environmental topics including hydrology [**Towler10**,**KATZ2002**} and air pollution (**GOULD2022**).

In this paper, we outline the techniques utilised by the team 'Uniofbathtopia' for the data challenge organised for the 13$^{\text{th}}$ International Conference on Extreme Value Analysis (EVA2023). A full description of the tasks can be found in the editorial (**Rohr23**). We outline our methodologies for each of the four sub-challenges, in which we complement traditional methods from extreme value statistics with other statistical modelling techniques according to the requirements of each task. The challenges involve the estimation of extreme marginal quantiles, marginal exceedance probabilities and joint tail probabilities

within the context of an environmental application, designed on the fictitious country of 'Utopia'. The organisers of the competition simulated the data using known parameters, so that teams' models could be validated and compared, and in such a way as to mimic the rich, complex behaviour exhibited by real-world processes. Therefore, we expect that the performance of our proposed methods should extend to general settings and applications.

In the univariate tasks we utilise the generalised Pareto distribution (GPD) and use model-based clustering methods including hierarchical models (**hastibfri2009**) and mixture models (**fraley2002**), as well as Markov chain Monte Carlo (MCMC) for parameter estimation [**coles1996@**. We also use bootstrapping methods for confidence interval estimation (**gill2020**). For the multivariate problems our approaches are heavily based on the parametric family of max-linear combinations of regularly varying random variables (**fougeres2013**). We introduce novel methods for performing inference for these models, advancing existing approaches (**cooley2019**) using modern statistical learning techniques including sparsity-inducing projections and clustering. The novel aspects of our work are: exploring MCMC parameter estimation bias for systems with large uncertainty in tail behaviour, and proposing a new estimator for the noise coefficient matrix of a max-linear model based on sparse projections onto the simplex.

The format of the paper is as follows: Section 4.4 describes our solutions for the univariate challenges with each challenge split into methodology and results sections. Section 4.5 introduces the requisite background theory from multivariate extremes before outlining the methodological frameworks and the results attained for Challenges 3 and 4. We conclude with some final discussion of our performance in Section 4.6.

4.4 Univariate Challenges

The first two challenges both involve estimating univariate extreme marginal quantiles, so we initially describe some of the theory that will be used across both tasks. Suppose that a generalised Pareto distribution with scale and shape parameters, σ and ξ respectively, is

a suitable model for exceedances of a high threshold u by a variable Y. Then, for $y > u$,

$$\mathbb{P}(Y > y \mid Y > u) = \begin{cases} \left(1 + \xi\left(\frac{y-u}{\sigma}\right)\right)_+^{-1/\xi}, & \xi \neq 0, \\ \exp\left(-\frac{y-u}{\sigma}\right), & \xi = 0, \end{cases}$$

where $\xi, u \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$ and $(\cdot)_+ = \max(\cdot, 0)$. Given the probability $\zeta_u = \mathbb{P}(Y > u)$, the probability of $Y > y$ can be expressed as

$$\mathbb{P}(Y > y) = \zeta_u \mathbb{P}(Y > y \mid Y > u). \tag{4.1}$$

We can find the pth percentile of the distribution of Y by rearranging the distribution function to form the quantile function,

$$q(p) = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{\zeta_u}{1-p}\right)^{\xi} - 1\right), & \xi \neq 0, \\ u + \sigma \log\left(\frac{\zeta_u}{1-p}\right), & \xi = 0, \end{cases}$$

where $p > 1 - \zeta_u$ to ensure that $q(p) > u$. However, we can also write down this expression for a return level, that is, the level y_T that is exceeded on average once every T years,

$$y_T = \begin{cases} u + \frac{\sigma}{\xi}((T n_{\text{yr}} \zeta_u)^{\xi} - 1), & \xi \neq 0, \\ u + \sigma \log(T n_{\text{yr}} \zeta_u), & \xi = 0. \end{cases} \tag{4.2}$$

where n_{yr} is the number of observations per year. For the purposes of the second challenge, $n_{\text{yr}} = 300$ as we are given that a year in Utopia consists of 12 months and 25 days per month, and there is one observation per day.

These expressions imply that, having chosen an appropriate threshold u, we can estimate quantiles and return levels once we obtain estimates of σ and ξ. The method of obtaining these values is different for different tasks and will be described, for the univariate challenges, in Section 4.4.2 and Section 4.4.2. We also require an estimate of ζ_u, the probability of an individual observation exceeding the threshold u. We can achieve this by using the empirical probability of Y exceeding u,

$$\hat{\zeta}_u = \sum_{i=1}^{n} \mathbf{1}\{y_i > u\}/n, \tag{4.3}$$

that is, the proportion of the total number of observations that exceed the threshold for observations $y_1, \dots y_n$, where for the univariate challenges $n = 21,000$, and where $\mathbf{1}\{y_i > u\}$ is the indicator function,

$$\mathbf{1}\{y_i > u\} = \begin{cases} 1 & \text{if } y_i > u, \\ 0 & \text{otherwise.} \end{cases}$$

4.4.1 Data

We are interested in the extreme values of the environmental variable Y which, for the two univariate tasks, we denote Y_i for day i. For each day, we also have a vector of covariates $\boldsymbol{X}_i = (V_{1,i}, \dots, V_{8,i})$ with variables $(V_1, \dots, V_4)$ representing unnamed covariates and (V_5, V_6, V_7, V_8) representing season, wind speed, wind direction and atmosphere respectively. Season is a factor variable, where $V_5 \in \{1, 2\}$, and each season covers half of the year. We perform our methodology for this challenge on each season individually. Given the covariates $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, $Y_1, \dots, Y_n$ are independent (**Rohr23**). For the first challenge, the data is divided into a training set comprising 70 years' worth of daily environmental data and a test set featuring 100 days of environmental data, showcasing diverse combinations of the covariates. We denote the test data points by $\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_{100}$. A comprehensive description of the dataset for this challenge is available in the competition editorial (**Rohr23**).

4.4.2 Challenge 1

4.4.2.1 Methodology

In the first task, we are required to provide point estimates and central 50% confidence intervals for extreme quantiles. As such, we need a model for the distribution of $Y \mid \boldsymbol{X}$ in order to estimate the conditional quantiles, $q_1, \dots, q_{100}$, where

$$\mathbb{P}(Y \leq q_k) \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k) = 0.9999.$$

We approach this problem with a two step strategy. We first partition the covariate space into clusters and then analyse the extremes of each cluster individually. In each resulting cluster, it is assumed that the extreme values can be adequately modelled using a GPD,

characterised by common scale and shape parameters, along with a common threshold and exceedance probability. To implement this, we propose partitioning the observations into groups by fitting a Gaussian mixture model to the environmental covariates. The dataset contained some missing values. We assumed that the missing observations where missing completely at random and removed them from the analysis, which was a valid assumption based on the editorial **Rohr23**.

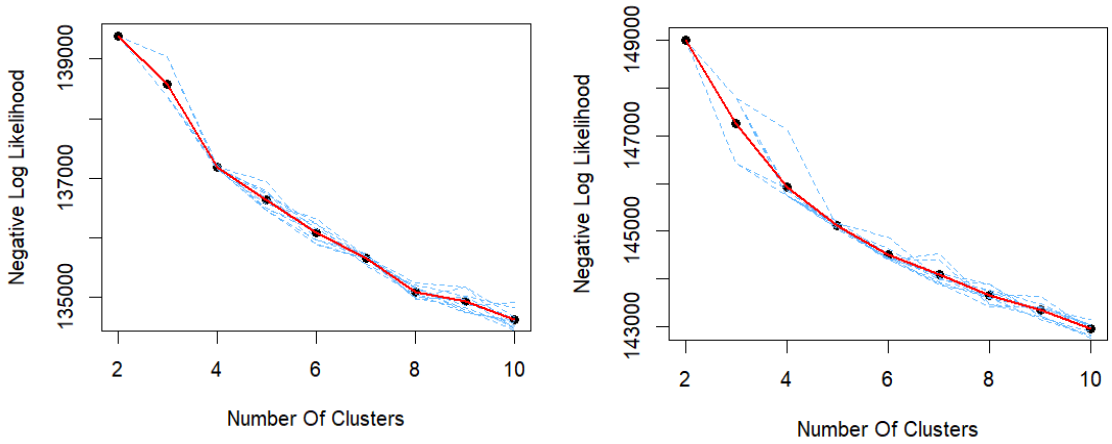

Figure 4.1: An elbow plot to determine an optimum number of clusters for season 1 (top) and season 2 (bottom). The figure displays the negative log-likelihood for different simulations (blue) as well as the average (red).

We perform the clustering of the covariates for each season individually by using the method in **fraley2003**. Let $\boldsymbol{x}_1^{(s)}, \ldots, \boldsymbol{x}_N^{(s)}$, where $N = n/2$, denote the observations for season $s \in \{1, 2\}$. We then assume that the covariates are sampled from a mixture of multivariate normal distributions. More formally, we want to estimate a d-dimensional Gaussian mixture model (where $d = 7$ in this challenge) with $J^{(s)}$ components, mean vectors $\boldsymbol{\mu}_1^{(s)}, \ldots, \boldsymbol{\mu}_{J^{(s)}}^{(s)} \in \mathbb{R}^d$, mixture probabilities $\boldsymbol{\alpha}^{(s)} = \{\alpha_1^{(s)}, \ldots, \alpha_{J^{(s)}}^{(s)}\}$, where $\alpha_j^{(s)} > 0$ and $\sum_{j=1}^{J^{(s)}} \alpha_j^{(s)} = 1$, and covariance matrices $\boldsymbol{\Sigma}_1^{(s)}, \ldots, \boldsymbol{\Sigma}_{J^{(s)}}^{(s)} \in \mathbb{R}^{d \times d}$, which are symmetric and positive definite. In the remainder of the subsection, we drop the season index s for brevity. The probability density function of the observation $\boldsymbol{x}_i$ can then be written as

$$p(\boldsymbol{x}_i \mid \boldsymbol{\Psi}) = \sum_{j=1}^{J} \alpha_j \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{4.4}$$

where $\boldsymbol{\Psi} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ are the parameters of the mixture model and we use $\phi(\cdot\,; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ to denote the probability density function of the Gaussian distribution with mean vector

$\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Assuming that an appropriate fixed J has been found for each season, the mixture model parameters $\boldsymbol{\Psi}$ are unknown and to be found. In order to achieve this, we use the expectation maximisation (EM) algorithm to perform maximum likelihood estimation (**mclachlan2000**), which is executed using the `mclust` package in `R` (**fraley2003**). The optimum number of clusters is then found by using the elbow method; we derive the negative log-likelihood for each choice of J, and identify the point at which this value begins to plateau.

In addition to the cluster parameters, the EM algorithm gives a cluster allocation for the observations. We use this allocation to split the observations of Y, and we denote by $\mathcal{Y}^{(j)}$ the data points in the jth cluster ($j = 1, \ldots, J$). We are then able to perform extreme value analysis on each set of points $\mathcal{Y}^{(j)}$ separately, by fitting a GPD, with parameters $\hat{\sigma}_j$ and $\hat{\xi}_j$, to the extremal data, using maximum likelihood estimation, as outlined at the start of Section 4.4. We make the assumption that the effects of the covariates on Y are entirely captured by the cluster assignments. The threshold, $\hat{u}_j$, is chosen by interpreting mean residual life plots and parameter stability plots (**Coles2001**).

The cluster estimates are then used to estimate q_k. We introduce a latent variable Z_k, where $Z_k = j$ refers to $\tilde{\boldsymbol{x}}_k$ being allocated to the jth cluster. Using the law of total probability, we then write $\mathbb{P}(Y > q_k \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k)$ as

$$\mathbb{P}(Y > q_k \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k) = \sum_{j=1}^{J} \mathbb{P}(Z_k = j \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k)\mathbb{P}(Y > q_k \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k, Z_k = j). \tag{4.5}$$

We can write the first probability in this expression using (4.4),

$$\mathbb{P}(Z_k = j \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k) = \frac{\alpha_j \phi(\tilde{\boldsymbol{x}}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{p(\tilde{\boldsymbol{x}}_k \mid \boldsymbol{\Psi})},$$

and using (4.1), it is possible to express the second probability as

$$\mathbb{P}(Y > q_k \mid Z_k = j, \boldsymbol{X} = \tilde{\boldsymbol{x}}_k) = \mathbb{P}(Y > q_k \mid Y > \hat{u}_j, Z_k = j)\mathbb{P}(Y > \hat{u}_j | Z_k = j).$$

The exceedance probability, $\mathbb{P}(Y > \hat{u}_j | Z_k = j)$ is calculated empirically using (4.3), and the other probability in the expression is found by fitting a GPD to the clusters. We then have all the components in (4.5) to find q_k that satisfies $\mathbb{P}(Y > q_k \mid \boldsymbol{X} = \tilde{\boldsymbol{x}}_k) = 0.0001$.

The challenge also requires central 50% confidence intervals for the estimates of these extreme conditional quantiles. For each set of points in a cluster, $\mathcal{Y}^{(j)}$, we are able to find an estimate of the quantile by performing a jackknife resampling (**efron1981**) of the extremal data and use this to fit the GPD. In the jackknife resampling method, sets of samples are created by leaving out one observation at a time and calculating the quantile point estimate based on the remaining observations. For each observation i in the extremal dataset $\mathcal{Y}^{(j)} \mid \mathcal{Y}^{(j)} > \hat{u}_j$, we create a new dataset $\mathcal{Y}^{(j)}_{(-i)} \mid \mathcal{Y}^{(j)} > \hat{u}_j$ by excluding the ith observation. We then fit a new GPD to this data and, using these parameter estimates, calculate the quantile point estimate in (**??**). Once we have undertaken this process for every observation in the dataset, we calculate the empirical 25th and 75th percentiles for the jackknife samples.

4.4.2.2 Results

As mentioned above, we perform clustering separately for each of the two seasons - Season 1 and Season 2. Elbow plots are used to determine suitable number of clusters; these results are displayed in Figure 4.1. It is clear from these figures that the negative log-likelihood does not obviously level off for up to 10 clusters. After balancing the trade-off between maximising the number of covariate clusters, and maximising the number of data points within each cluster, the number of covariate clusters was determined to be $J = 5$ for each season. By choosing $J = 5$, the average number of observations in each cluster was 2,580 (with a standard deviation of 535).

The performance of this method was disappointing (see editorial **Rohr23**), and there are a few reasons why this could have been the case. Firstly, there were no clearly defined clusters. This points to the fact that the fitted Gaussian mixture distribution may have been unable to fully capture the non-Gaussian distribution of the covariate values. In Figure 4.1, the point at which the negative log-likelihood plateaus is not clear, illustrating this argument. The second reason could be that covariates were not used when fitting the GPDs to the extreme data in each cluster, only for cluster assignment. A GPD with covariate models for each cluster may have resulted in better quantile estimates. A further development could be to explore generalised Pareto regression trees, as they are able to perform clustering using trends in the covariates (**Fark21**).

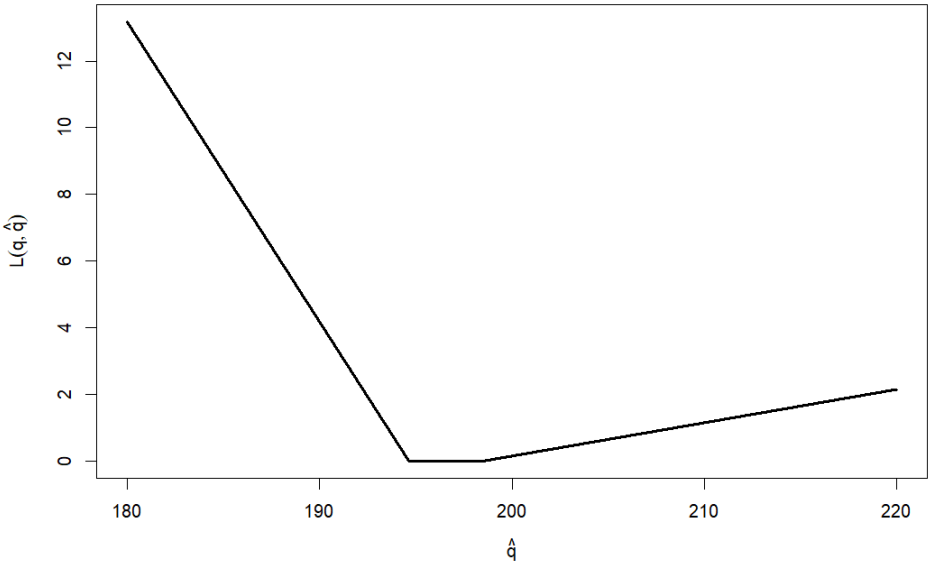

Figure 4.2: A diagram of the loss function using the true quantile value $q = 196.6$.

4.4.3 Challenge 2

In the second challenge, we are required to estimate the marginal quantile $q \in \mathbb{R}$ such that

$$\mathbb{P}(Y > q) = \frac{1}{300T},$$

where $T = 200$, that is, an estimate of the 200 year return level for Amaurot. This challenge was assessed using a loss function that penalises under-estimating the quantile more than over-estimating it. Formally, for a given estimate $\hat{q}$ and the true marginal quantile q, the loss is calculated as,

$$L(q, \hat{q}) = \begin{cases} 0.9(0.99q - \hat{q}) & \text{if } 0.99q > \hat{q}, \\ 0, & \text{if } |q - \hat{q}| \leq 0.01q, \\ 0.1(\hat{q} - 1.01q), & \text{if } 1.01q < \hat{q}. \end{cases} \tag{4.6}$$

This replicates real-world scenarios. For example, in a hydrology context, over-estimating flood defences leads to increased costs, while under-estimating them results in severe consequences, such as fatalities. After completing the challenge, we were given the correct

quantile $q = 196.6$ and so we can substitute this value into the loss function, as displayed in Figure 4.2.

For this analysis, we make the underlying assumption that the extreme values of Y are independent and identically distributed (IID), that is, we ignore any dependence on $\boldsymbol{X}$. This assumption is made to simplify the modelling process. We then assume that the exceedances of Y above a high threshold u follow a GPD,

$$Y - u \mid Y > u \sim \text{GPD}(\sigma, \xi).$$

Using (**??**), the 200-year return level is then

$$y_{200} = \hat{u} + \frac{\hat{\sigma}}{\hat{\xi}}((200 \times 300\hat{\zeta}_u)^{\hat{\xi}} - 1),$$

where we have substituted $T = 200$ and $n_y = 300$ as we have 300 daily observations a year.

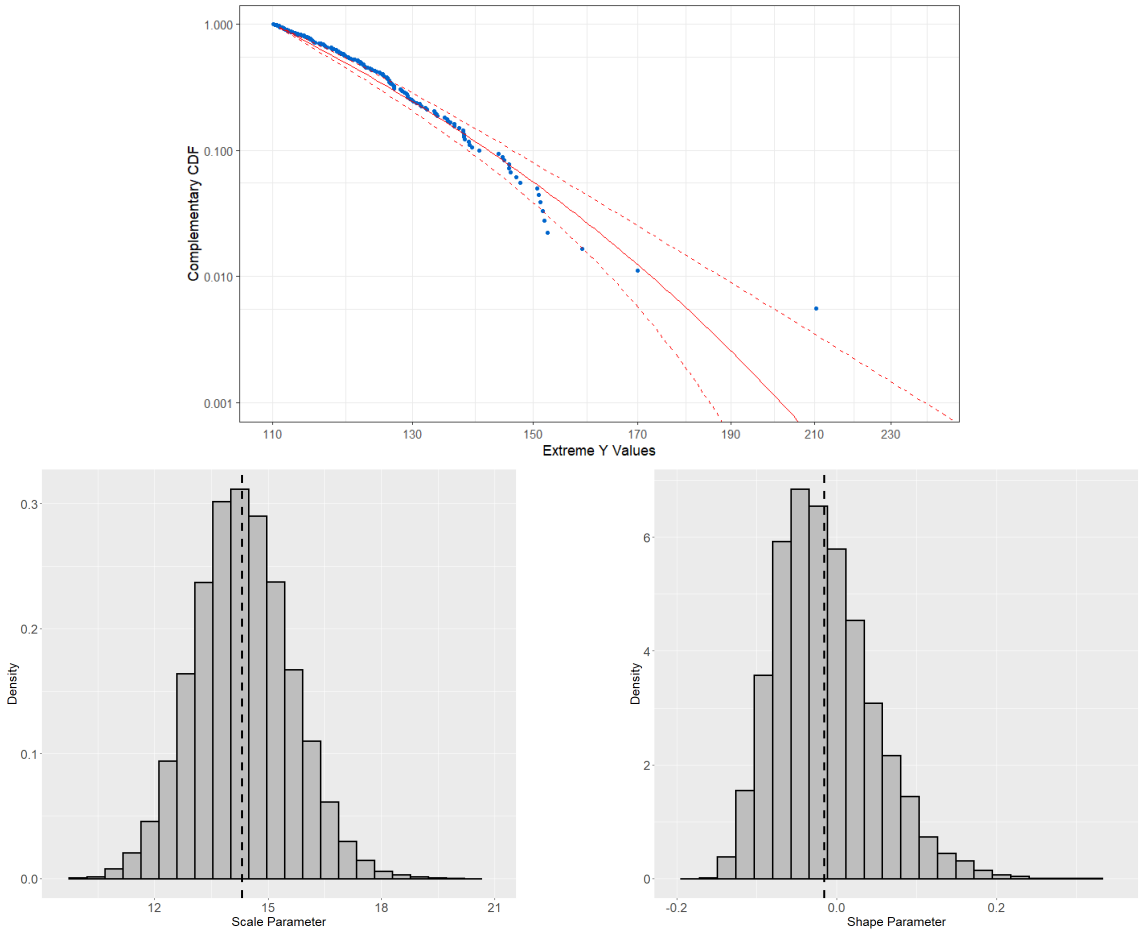

Figure 4.3: A GPD fitted to extreme Y values above threshold 110, with a survival curve (top) fitted using the mean of the posterior parameter values, with a 90% confidence interval as dashed lines, and the respective posterior distributions of sampled scale (bottom left) and shape (bottom right) parameters, with the mean values as dashed lines.

To define a Bayesian framework, we choose prior distributions for σ and ξ, $\sigma \sim \text{Gamma}(4, 1)$, $\xi \sim N(0, 1)$. These prior distributions are selected to reflect minimal prior knowledge about the exact parameter values (whilst ensuring that σ remain positive). Subsequent analysis demonstrated that the parameter estimates were robust and largely unaffected by the specific choice of priors, which indicated that the data provided sufficient information to primarily determine the posterior distributions.

We use Markov chain Monte Carlo methods (**coles1996**) to sample from the resulting posterior distribution of (σ, ξ). This method was preferred over maximum likelihood estimation as we wanted a full distribution of the parameters, allowing for more insight into

the uncertainty of the estimates.

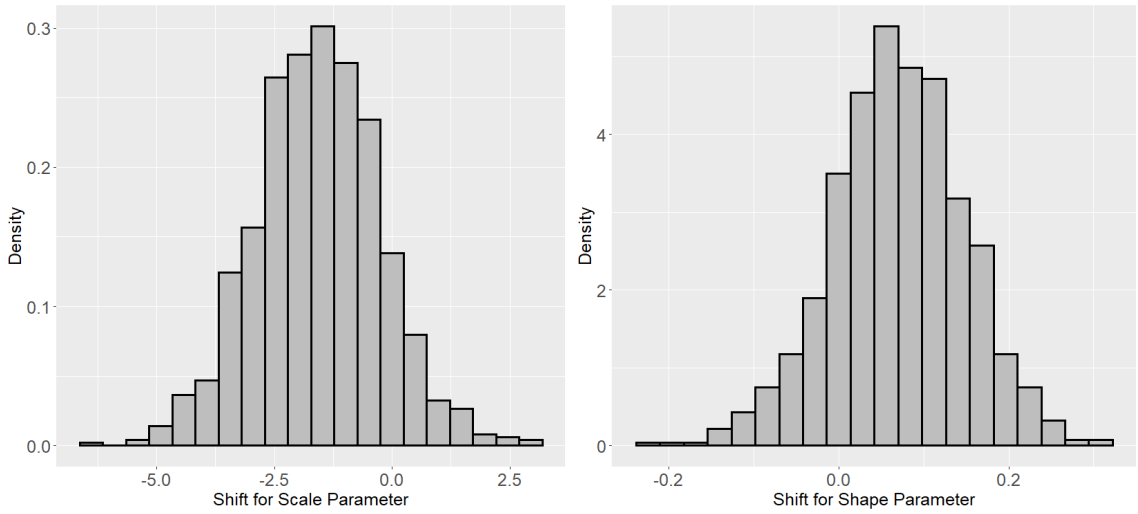

Figure 4.4: A histogram of residuals of predicted and simulated scale (left) and shape (right) parameters. These residuals are then used as a bias correction for the return level estimate.

To assess how well our approach recovers the true parameter values, we generate sets of simulated data from GPDs with parameters that were representative of the ones sampled via our MCMC algorithm. Formally, for each simulated dataset, we sample values $\tilde{\sigma}$ and $\tilde{\xi}$, and then generate data points $\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_m$ from a GPD$(\tilde{\sigma}, \tilde{\xi})$, where m is equal to the number of exceedances of Y above u. We considered two approaches to sample $\tilde{\sigma}$ and $\tilde{\xi}$, and we generated 1000 datasets for each. The first approach is to sample these parameters from the posterior distribution. The other approach is to sample the parameters from uniform distributions, $\tilde{\sigma} \sim U(11, 18)$ and $\tilde{\xi} \sim \text{Unif}(-0.15, 0.20)$; these limits were selected based on the posterior samples. For each generated dataset, we again run our MCMC algorithm and store the posterior mean estimate for $\tilde{\sigma}$ and $\tilde{\xi}$. Finally, we apply the mean corrections to the initial GPD parameter estimates to calculate an adjusted return level.

4.4.3.1 Results

We determine a threshold of $\hat{u} = 110$ by using mean excess plots, resulting in $m = 180$. We can then use this to calculate the empirical estimate for the exceedance probability $\hat{\zeta}_u = 180/21000 \approx 0.0086$. Figure 4.3 shows the posterior samples of the shape and scale distributions for the initial GPD fit. The range of these distributions are then used to

inform the sampling boundaries for the simulated datasets as described in Section **??**. We also plot a survival function to assess the model fit (Figure 4.3), where the dashed red lines reflect a 90% credible interval for the fit. We find that the posterior distribution of the shape parameter spanned both negative and positive values, indicating uncertainty in the tail behaviour of the underlying distribution. It was also clear from inspecting smaller values of the extremes in the survival curve, that the GPD fit could be improved.

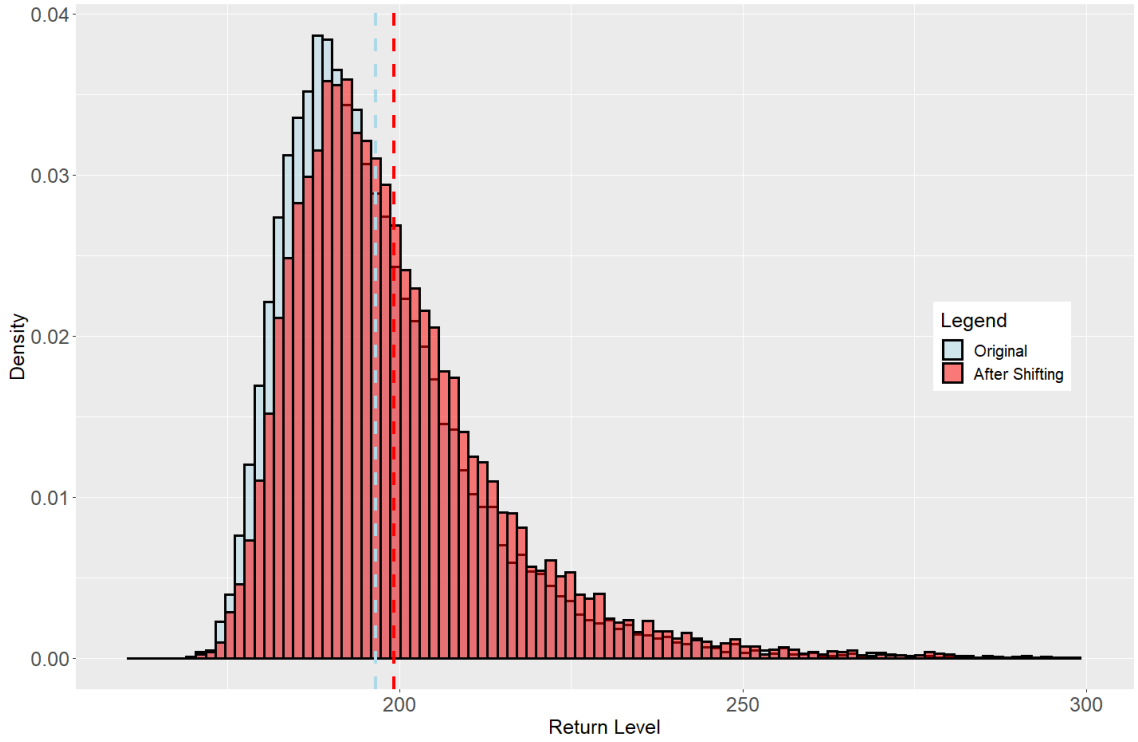

Figure 4.5: The return level samples before shifting (blue) and after shifting (red). The dashed lines denote the mean of the samples.

We opted for our second approach to create the simulated datasets, that is, we sampled the parameters from uniform distributions to make the corrections for our final results. We choose this approach as, by employing uniform coverage of the parameter space, we return a more risk-averse return level with respect to the challenge loss function. The distributions of the residuals of the mean predicted parameters and the true (simulated) parameters are displayed in Figure 4.4. There are significant negative residuals for the scale parameter and significant positive residuals for the shape parameter. We wanted to use these residual distributions to make a correction to the initial parameter estimates. We apply the mean corrections to the initial posterior parameter values and compared the

mean return levels. These are shown in Figure 4.5. This non-standard approach was used for simplicity, however, some further work could be to more carefully incorporate these results into the initial posterior parameter distributions. This gives an overall increase in the average return level from 196.4 to 199.4.

The correct quantile value was $q = 196.6$. We were interested to find that the initial GPD fit to the data gave a return level of 196.4, which led to a loss of $L(196.6, 196.4) = 0$. It is possible that small effects of the covariates are compensated for by other factors, giving a correct return level under the misspecified model assumption. Although our final answer, with the correction, increased our return level, it also increased the loss from $L(196.6, 196.4) = 0$ to $L(196.6, 199.4) = 0.0834$ (using (4.6)). Before the correct quantile value was announced, we were encouraged that the method resulted in an increase in the return level, as an over-estimate was preferable to an under-estimate in the context of the competition. It is worth noting that if we had opted to sample the parameters of the simulated data from the posterior distributions, we would have got an average return level of 198.7, which would have given a smaller loss, $L(196.6, 198.7) = 0.0134$. The simplicity of the model, given the underlying assumption about the data, creates ambiguity regarding whether the large residuals in the parameter estimates are caused by bias or by model misspecification. For future work, it would be appropriate to incorporate a more sophisticated model of the covariates to reduce model misspecification and to investigate whether any residuals in parameter estimates are due to bias.

4.5 Multivariate Challenges

In Challenges 3 and 4 we are presented with a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d) \sim F_{\boldsymbol{X}}$, representing the value of an environmental variable at d sites in Utopia, where $F_{\boldsymbol{X}}$ is an unknown joint distribution function. Our goal is to estimate the probability that $\boldsymbol{X}$ lies in a given extreme 'failure region' based on a sample of independent observations of $\boldsymbol{X}$. The failure regions of interest are such that certain components of $\boldsymbol{X}$ are simultaneously large while all remaining components (if any) are of lower order. The inherent difficulty of the task stems from the fact that the events' return periods are of similar order or even significantly longer than the observation period over

which the data are collected; empirical methods based solely on the relative frequency of event occurrences are fruitless. Instead, we use the observed data to infer an estimate for the probabilistic structure of the joint tail of $\boldsymbol{X}$ and subsequently compute tail event probabilities under this model. This encapsulates the philosophy of multivariate extreme value statistics.

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

4.5.1 Background

4.5.1.1 Multivariate regular variation and the angular measure

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

4.5.1.2 Multivariate regular variation and the angular measure

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

Definition 4.1. We say that $\boldsymbol{X}$ is *multivariate regularly varying* with index $\alpha > 0$, denoted $\boldsymbol{X} \in \mathrm{RV}_+^d(\alpha)$, if it satisfies the following (equivalent) definitions (**resnickHeavytailPhenomenaProbabilistic2007**):

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

$$n\mathbb{P}(b_n^{-1}\boldsymbol{X} \in \cdot) \xrightarrow{\text{v}} \nu_{\boldsymbol{X}}(\cdot), \qquad (n \to \infty), \tag{4.7}$$

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

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

$$n\mathbb{P}((b_n^{-1}R, \boldsymbol{\Theta}) \in \cdot) \xrightarrow{\text{v}} \nu_\alpha \times H_{\boldsymbol{X}}(\cdot), \qquad (n \to \infty), \tag{4.8}$$

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

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

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

The probabilistic tail of $\boldsymbol{X}$ decomposes into a univariate α-regularly varying radial component (**resnickHeavytailPhenomenaProbabilistic2007**), that is asymptotically independent of the angular component. The angular measure represents the limiting distribution of the angular component and encodes all information about the tail dependence structure.

The MRV property implies that the margins of $\boldsymbol{X}$ are heavy-tailed with a common tail index. Henceforth, assume that the components of $\boldsymbol{X} \in \mathrm{RV}_+^d(\alpha)$ are Fréchet distributed with shape parameter α, that is $\mathbb{P}(X_i < x) = \Psi_\alpha(x) := \exp(-x^{-\alpha})$ for $x > 0$ and $i = 1, \ldots, d$. (The data for Challenges 3 and 4 are on Gumbel margins but can be accommodated into our framework following a suitable transformation.) Moreover, we choose $\|\cdot\| = \|\cdot\|_\alpha$, the L_α-norm on $\mathbb{R}^d$, and specify that the normalising sequence in (4.8) is $b_n = n^{1/\alpha}$. With these particular choices the marginal variables have unit scale

(**kluppelbergEstimatingExtremeBayesian2021**) and $H_{\boldsymbol{X}}(\mathbb{S}_+^{d-1}) = d$.

The problem of modelling the angular measure has attracted considerable attention in recent years – a survey of the related literature can be found in **engelkeSparseStructuresMultivariate202** One research avenue concerns learning which sets of variables may be concurrently extreme; this can be posed as a support detection problem (**goixSparseRepresentationMultivariate2017**; **simpsonDeterminingDependenceStructure2020**). Consider the index set $\mathbb{V}(d) := \{1, \ldots, d\}$ and denote by $\mathcal{P}_d^\star = \mathcal{P}(\mathbb{V}(d)) \setminus \emptyset$ its power set excluding the empty set. A set $\beta \in \mathcal{P}_d^\star$ is termed an extremal direction of $\boldsymbol{X}$ if $H_{\boldsymbol{X}}$ places mass on the subspace

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

Another branch of research aims at developing dimension reduction techniques for analysing the angular measure in high dimensions. To this end, one often considers a summary of the full dependence structure encoded in a matrix of pairwise extremal dependence metrics. One such matrix, originally proposed in **larssonExtremalDependenceMeasure2012** and later popularised by **cooleyDecompositionsDependenceHighdimensional2019**, is the tail pairwise dependence matrix (TPDM). The TPDM of $\boldsymbol{X} \in \mathrm{RV}_+^d(2)$ is the $d \times d$ matrix $\Sigma = (\sigma_{ij})$ with entries

$$\sigma_{ij} = \int_{\mathbb{S}_+^{d-1}} \theta_i \theta_j \, \mathrm{d}H_{\boldsymbol{X}}(\boldsymbol{\theta}), \qquad (i, j = 1, \ldots, d). \tag{4.9}$$

The diagonal entries are the squared marginal scales, i.e. $\sigma_{ii} = 1$ for $i = 1, \ldots, d$. The off-diagonal entries measure extremal dependence between the associated pairs of variables. In particular, $\sigma_{ij} = 0$ if and only if X_i and X_j are asymptotically independent. For asymptotically dependent variables the magnitude of σ_{ij} represents the dependence strength. An important property of the TPDM is its complete positivity (**cooleyDecompositionsDependenceHighdimensional2019**).

Definition 4.2. A matrix S is *completely positive* if there exists a matrix A with non-negative entries such that $S = AA^T$. We call A (resp. AA^T) a *CP-factor* (resp. *CP-decomposition*) of S.

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

4.5.1.3 The max-linear model for multivariate extremes

Our proposed methods for Challenges 3 and 4 employ the max-linear model, a parametric model based on the class of random vectors constructed by max-linear combinations of independent Fréchet random variables (**fougeresDenseClassesMultivariate2013**). This model is appealing for several reasons. First, it is flexible, in the sense that any regularly varying random vector can be arbitrarily well-approximated by a max-linear model with sufficiently many parameters (**fougeresDenseClassesMultivariate2013**). Since neither Challenge 3 nor Challenge 4 provide any prior information about the underlying data-generating processes, it is preferable to avoid imposing overly restrictive assumptions on the tail dependence structure. Secondly, although the number of parameters grows rapidly – at least $\mathcal{O}(d)$ but often even $\mathcal{O}(d^2)$ – efficient inference procedures are available even in high dimensions. Scalability is critical for Challenge 4. Finally, extremal directions and failure probabilities can be straightforwardly identified and computed directly from the model parameters (**kirilioukEstimatingProbabilitiesMultivariate2022**).

Definition 4.3. For some $q \geq 1$ and $\alpha > 0$, let $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$ be a random vector with independent components $Z_1, \ldots, Z_q \sim \Psi_\alpha$ and let $A = (a_{ij}) \in \mathbb{R}_+^{d \times q}$ be a deterministic matrix. If

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

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

cooleyDecompositionsDependenceHighdimensional2019 show that $\boldsymbol{X} = A \circ \boldsymbol{Z} \in \mathrm{RV}_+^d(\alpha)$ and

$$H_{\boldsymbol{X}}(\cdot) = \sum_{j=1}^{q} \|\boldsymbol{a}_j\|_\alpha^\alpha \delta_{\boldsymbol{a}_j / \|\boldsymbol{a}_j\|_\alpha}(\cdot), \tag{4.10}$$

where $\delta_{\boldsymbol{x}}(A) := \mathbf{1}\{\boldsymbol{x} \in A\}$ is the Dirac mass function. The angles along which extremes can occur are (in the limit) precisely the q self-normalised columns of A. Therefore $\beta \in \mathcal{P}_d^\star$ is an extremal direction of $\boldsymbol{X}$ if and only if there exists $j \in \{1, \ldots, q\}$ such that $\boldsymbol{a}_j / \|\boldsymbol{a}_j\|_\alpha \in C_\beta$. When it comes to model fitting, the testing procedure of **kirilioukHypothesisTestingTail2020** can provide guidance for choosing q; for our purposes, it either represents a tuning parameter (Challenge 3) or takes a fixed value owing

to computational/algorithmic restrictions (Challenge 4). Substituting (4.10) into (4.9), we observe that $\boldsymbol{X} \sim \text{MaxLinear}(A, 2)$ has TPDM $\Sigma_{\boldsymbol{X}} = AA^T$. In other words, the noise coefficient matrix is a CP-factor of the model TPDM. Conversely, given an arbitrary random vector $\boldsymbol{X} \in \text{RV}_+^d(2)$ with TPDM Σ, any CP-factor A of Σ parametrises a max-linear model with identical pairwise tail dependence metrics to $\boldsymbol{X}$.

kiriliouk2022 give examples of classes of tail events $\mathcal{R} \subset \mathbb{E}_0$ for which $\mathbb{P}(\boldsymbol{X} \in \mathcal{R})$ can be well-approximated by a function of the parameter matrix A. With a view to Challenges 3 and 4, we focus on tail events where $\boldsymbol{X}$ is large in the $s \leq d$ components indexed by $\beta = \{\beta_1, \ldots, \beta_s\} \in \mathcal{P}_d^\star$, while all $d - s$ remaining components are of lower order. Formally, for $\boldsymbol{u} = (u_1, \ldots, u_s) \in \mathbb{R}_+^s$ a vector of high thresholds and $\boldsymbol{l} \in \mathbb{R}_+^{d-s}$ a vector of comparatively low thresholds, we consider

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

When $\beta = \mathbb{V}(d)$ the threshold vector $\boldsymbol{l}$ is superfluous and is omitted. **kiriliouk2022** specify an approximate formula for $\mathbb{P}(A \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}})$ in terms of A. We derive an analogous formula for $\mathbb{P}(A \circ \boldsymbol{Z} \in \mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}})$ for general $\beta \in \mathcal{P}_d^\star$ stated as follows: if $\boldsymbol{X} \sim \text{MaxLinear}(A; \alpha)$, then

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}}) \approx \hat{\mathbb{P}}(\boldsymbol{X} \in \mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}}) := \sum_{j:\, \frac{\boldsymbol{a}_j}{\|\boldsymbol{a}_j\|_\alpha} \in C_\beta} \min_{i=1,\ldots,s} \left(\frac{a_{\beta_i,j}}{u_i}\right)^\alpha. \tag{4.11}$$

The approximation should be understood as being valid as $u_1, \ldots, u_s \to \infty$ while $\boldsymbol{l}$ remains fixed. Note that the entries of $\boldsymbol{l}$ do not appear in the right-hand side of (4.11). This reflects the fact that, asymptotically, the magnitude of the probability of the event $\{\boldsymbol{X}_\beta > \boldsymbol{u}\} \cap \{\boldsymbol{X}_{-\beta} < \boldsymbol{l}\}$ is predominantly determined by the threshold exceedance event $\{\boldsymbol{X}_\beta > \boldsymbol{u}\}$ while the relative contribution of the threshold non-exceedance event $\{\boldsymbol{X}_{-\beta} < \boldsymbol{l}\}$ vanishes. Henceforth, the threshold $\boldsymbol{l}$ may at times be suppressed for notational convenience. From (4.7) we have that, provided $u_1, \ldots, u_s$ are sufficiently large,

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

where the last step exploits homogeneity of the exponent measure. Transforming to pseudo-

polar coordinates this becomes

$$\nu_{\boldsymbol{X}}(\mathcal{C}_{\beta,\boldsymbol{u},\boldsymbol{l}}) = \int_{\{(r,\boldsymbol{w}):\, r\boldsymbol{w}_\beta > \boldsymbol{u},\, r\boldsymbol{w}_{-\beta} < \boldsymbol{l}\}} \alpha r^{-\alpha-1}\,\mathrm{d}r\,\mathrm{d}H_{\boldsymbol{X}}(\boldsymbol{w}).$$

Based on (4.10), there are only q possible angles along which extremes can occur, but only those with $\boldsymbol{a}_j/\|\boldsymbol{a}_j\|_\alpha \in C_\beta$ will contribute to the integral. To see this, consider $\boldsymbol{w} = \boldsymbol{a}/\|\boldsymbol{a}\|_\alpha \in C_\gamma$, where $\boldsymbol{a}$ denotes an arbitrary column of A and $\gamma \in \mathcal{P}_d^\star$. First, suppose $\gamma \subset \beta$ (strictly) so that there exists $i \in \beta$ such that $i \notin \gamma$. Then, for all $r > 0$, we have $r\boldsymbol{w}_\beta \not> \boldsymbol{u}$, since $rw_i = 0$. Similarly, suppose $\gamma \supset \beta$, in which case there exists $i \in \gamma$ with $i \notin \beta$. Let ℓ denote the entry of $\boldsymbol{l}$ corresponding to component i. An extreme event along $\boldsymbol{w}$ requires $rw_i < \ell$ and $r\boldsymbol{w}_\beta > \boldsymbol{u}$, but for $\boldsymbol{u}$ sufficiently large these inequalities have no solution $r > 0$. On the other hand, if $\beta = \gamma$, then $\boldsymbol{w}_{-\beta} = \boldsymbol{0} \in \mathbb{R}^{d-s}$ and so

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

This yields the final result

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

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

4.5.1.4 Existing approaches to inference for max-linear models

Suppose $\boldsymbol{X} \sim \text{MaxLinear}(A, \alpha)$, where α is known and A must be estimated from a sample $\{\boldsymbol{x}_t = (r_t, \boldsymbol{\theta}_t) : t = 1,\ldots,n\}$ of $\boldsymbol{X} = (R, \boldsymbol{\Theta})$. For $j \in \{1,\ldots,n\}$, let $r_{(j)}$ denote the jth upper order statistic of $\{r_1,\ldots,r_n\}$ and let $\boldsymbol{x}_{(j)}, \boldsymbol{\theta}_{(j)}$ denote the corresponding observation and angular component, respectively.

One estimate of A is motivated by (4.10), which says its normalised columns represent the set of realisable extremal angles (**cooleyDecompositionsDependenceHighdimensional2019**)

Definition 4.4. The empirical estimate of A based on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is $\hat{A} = (\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_k) \in \mathbb{R}_+^{d\times k}$, where $1 \leq k \leq n$ and $\hat{\boldsymbol{a}}_j = (d/k)^{1/\alpha}\boldsymbol{\theta}_{(j)}$ for $j = 1, \ldots, k$.

The quantity k is the customary tuning parameter that represents the number of 'extreme' observations with norm not less than the implied radial threshold $r_{(k)}$. The associated angular measure (for any α) and TPDM (for $\alpha = 2$) are given by

$$\hat{H}_{\boldsymbol{X}}(\cdot) := H_{\hat{A}\times_{\max}\boldsymbol{Z}}(\cdot) = \frac{d}{k}\sum_{t=1}^{n}\mathbf{1}\{\boldsymbol{\theta}_t \in \cdot, r_t \geq r_{(k)}\} \tag{4.12}$$

$$\hat{\sigma}_{\boldsymbol{X}_{ij}} := \sigma_{\hat{A}\times_{\max}\boldsymbol{Z},ij} = \frac{d}{k}\sum_{t=1}^{n}\theta_{ti}\theta_{tj}\mathbf{1}\{r_t \geq r_{(k)}\} = [\hat{A}\hat{A}^T]_{ij}. \tag{4.13}$$

These are the empirical angular measure (**einmahlMaximumEmpiricalLikelihood2009**) and empirical TPDM (**cooleyDecompositionsDependenceHighdimensional2019**), respectively. In view of (4.13), alternative estimates of A can be obtained by CP decomposition of the empirical TPDM.

Definition 4.5. CP-factors of $\hat{\Sigma}_{\boldsymbol{X}} = (\hat{\sigma}_{\boldsymbol{X}_{ij}})$ are called CP-estimates of A, denoted $\tilde{A}$.

An estimate $\tilde{A}$ induces a different angular measure $\tilde{H}_{\boldsymbol{X}} := H_{\tilde{A}\times_{\max}\boldsymbol{Z}}$ but, by construction, the empirical and CP models share identical tail pairwise dependencies, since $\tilde{\Sigma}_{\boldsymbol{X}} := \Sigma_{\tilde{A}\times_{\max}\boldsymbol{Z}} = \tilde{A}\tilde{A}^T = \hat{\Sigma}_{\boldsymbol{X}}$. Note that $\tilde{A}$ implicitly depends on the same tuning parameter k as $\hat{A}$ via the empirical TPDM. All CP-estimates in this paper are obtained using the decomposition algorithm of **kirilioukEstimatingProbabilitiesMultivariate2022**, which efficiently factorises moderate- to high-dimensional TPDMs. Their algorithm takes as input a strictly positive TPDM and a permutation $(i_1, \ldots, i_d)$ of $\mathbb{V}_d$ and (for some permutations) returns a square CP-factor $\tilde{A} \in \mathbb{R}_+^{d\times d}$ whose columns satisfy $\tilde{\boldsymbol{a}}_j/\|\tilde{\boldsymbol{a}}_j\|_2 \in C_{\mathbb{V}_d\setminus\{i_l : l<j\}}$ for $j = 1, \ldots, d$.

4.5.1.5 Inference for max-linear models based on sparse projections

A limitation of $\hat{A}$ and $\tilde{A}$ is that they fail to capture the true extremal directions of $\boldsymbol{X}$. In the case of $\hat{A}$ this is because the angles $\boldsymbol{\Theta} = \boldsymbol{X}/\|\boldsymbol{X}\|$ lie in the simplex interior almost surely,

meaning $\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_k \in \mathbb{V}_d$ and $\hat{\mathbb{P}}(\hat{A} \times_{\max} \boldsymbol{Z} \in \mathcal{C}_{\beta,\boldsymbol{u}}) = 0$ for any $\beta \neq \mathbb{V}_d$. The d extremal directions of $\tilde{A} \times_{\max} \boldsymbol{Z}$ are fully determined by the input path $(i_1, \ldots, i_d)$ and need not bear any resemblance to the extremal directions suggested by the data. To address this gap, we propose augmenting the empirical estimate with an alternative notion of angle based on Euclidean projections onto the L_1-simplex (**meyer2023**).

Definition 4.6. The Euclidean projection onto the L_1-simplex is defined by

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

This projection is useful because $\pi(\boldsymbol{v})$ may lie on a boundary of the simplex even if $\boldsymbol{v}/\|\boldsymbol{v}\|_1$ lies in its interior. Assume now that $\alpha = 1$.

Definition 4.7. The *sparse empirical estimate* of A based on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is $\hat{A}^\star = (\hat{\boldsymbol{a}}_1^\star, \ldots, \hat{\boldsymbol{a}}_k^\star) \in \mathbb{R}_+^{d \times k}$, where $1 \leq k < n$ and $\hat{\boldsymbol{a}}_j^\star = (d/k)\pi(\boldsymbol{x}_{(j)}/r_{(k+1)})$ for $j = 1, \ldots, k$.

The corresponding angular measure

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

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

4.5.2 Challenge 3

4.5.2.1 Data

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

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

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

4.5.2.2 Methodology

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

$$p_1 = \mathbb{P}(X_1 > e^6, X_2 > e^6, X_3 > e^6), \quad p_2 = \mathbb{P}\left(X_1 > e^7, X_2 > e^7, X_3 < \frac{1}{\log 2}\right). \tag{4.14}$$

The thresholds e^6, e^7, and $1/\log(2)$ correspond approximately to the 99.8%, 99.9% and 50% quantiles of Ψ_1, respectively. Our solution models $\boldsymbol{X}$ as max-linear, infers A using the sparse empirical estimator (for some hyperparameter k) and computes estimates for p_1 and p_2 via (4.11), that is

$$\hat{p}_1 = \hat{\mathbb{P}}(\hat{A}^\star \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(3), e^6}), \qquad \hat{p}_2 = \hat{\mathbb{P}}(\hat{A}^\star \circ \boldsymbol{Z} \in \mathcal{C}_{\{1,2\}, e^7, 1/\log(2)}). \tag{4.15}$$

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

4.5.2.3 Results

The results presented are based on $k = 500 \approx 2.5\% \times n$. This value was used for our competition submission and the results' sensitivity to this choice will be examined later.

The ternary plots in Figure 4.6 depict the angular components associated with the k largest observations in norm, i.e. exceeding the radial threshold $r_{(k+1)} \approx 138.77$. The left-hand plot represents the self-normalised vectors $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)}$. We find points lying near the centre of the triangle and in the neighbourhood of all its edges and vertices. This suggests that the angular measure spreads mass across all seven faces of $\mathbb{S}_+^2$. However, we reiterate that no points lie *exactly* on the boundary. In contrast, the sparse angles $\{\pi(\boldsymbol{x}_t / r_{(k+1)}) : r_t > r_{(k+1)}\}$ in the right-hand plot lie in the interior (black, 40 points), along the edges (red, 139 points), and on the vertices (blue, 321 points) of the closed

simplex. Only the 40 vectors in $C_{\mathbb{V}(3)}$ and 23 vectors in $C_{\{1,2\}}$ will enter into the estimates of (4.14).

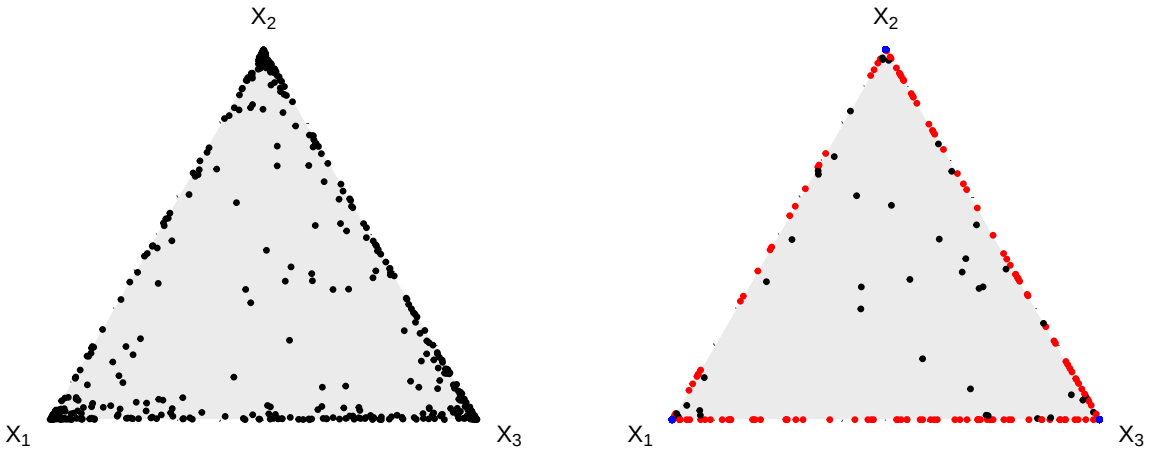

Figure 4.6: The angles $\boldsymbol{x}_{(j)}/\|\boldsymbol{x}_{(j)}\|_1 \in \mathbb{S}_+^2$ (left) and Euclidean projections $\pi(\boldsymbol{x}_{(j)}/r_{(k+1)}) \in \mathbb{S}_+^2$ (right) for the 500 largest observations in Challenge 3. Points are coloured according to whether they lie in the interior (black), on an edge (red) or on a vertex (blue).

The projected vectors are collated to form the 3×500 matrix $\hat{A}^\star$. The first 100 columns of the matrices $\hat{A}$ and $\hat{A}^\star$ are represented visually in Figure 4.7. As expected, $\hat{A}$ is dense, albeit populated with small values, while $\hat{A}^\star$ exhibits a high degree of sparsity, in the sense that most of its columns satisfy $\|\hat{\boldsymbol{a}}_j^\star\|_0 < \|\hat{\boldsymbol{a}}_j\|_0 = 3$. Duplicated columns in $\hat{A}^\star$ could be merged and re-weighted to produce a compressed estimate $\hat{A}^\star_{\text{comp}}$ with $q_{\text{comp}} = 40 + 139 + 3 = 182$ unique columns, but ultimately they parametrise identical models since $H_{\hat{A}^\star \circ \boldsymbol{Z}} = H_{\hat{A}^\star_{\text{comp}} \circ \boldsymbol{Z}}$.

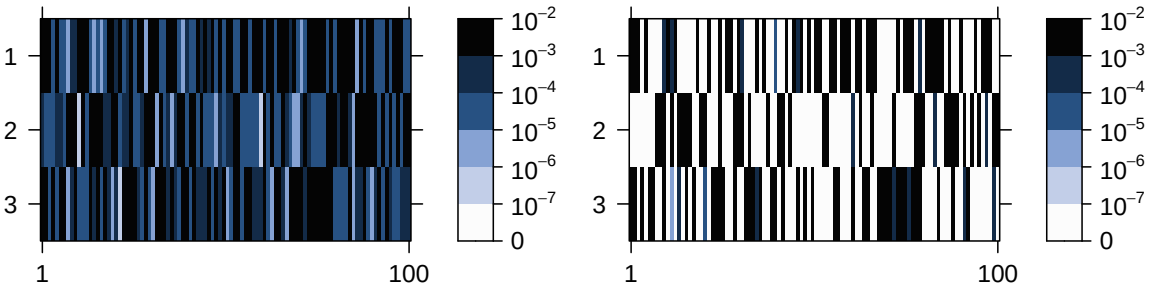

Figure 4.7: The first 100 columns of $\hat{A}$ (left) and $\hat{A}^\star$ (right) for Challenge 3. The colour intensity of each cell represents the magnitude of the corresponding matrix entry.

Substituting $\hat{A}^\star$ into (**??**) yields the final point estimates $\hat{p}_1 = 3.36 \times 10^{-5}$ and $\hat{p}_2 = 2.76 \times 10^{-5}$, to three significant figures. We are pleased to find that our estimates are very

close to the true values given in **Rohr23**. Furthermore, Figure 4.8 shows that our estimates of p_2 are fairly stable across a range of 'reasonable' choices of k, say, $1\% \leq k/n \leq 5\%$. When k is smaller than this, the estimates become highly variable due to the limited effective sample size. If k is too large, we risk introducing bias by including observations that do not reflect the true tail dependence structure. The estimates of p_1 exhibit greater sensitivity to the choice of threshold. The development of theoretically justified diagnostics/procedures for selecting the threshold represents an avenue for future work.

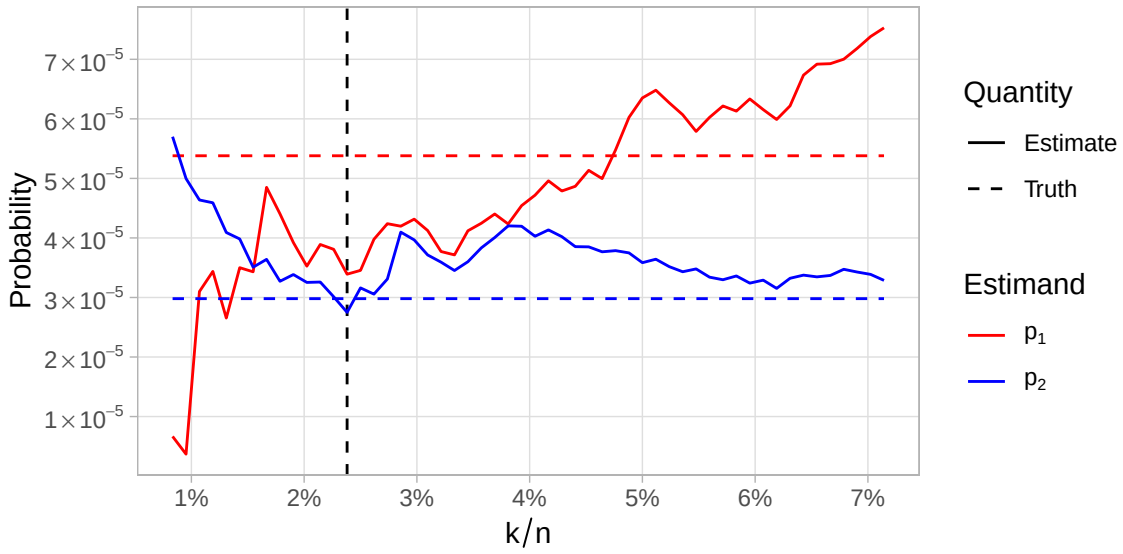

Figure 4.8: Challenge 3 results for different choices for the tuning parameter, k. The horizontal dashed lines indicate the estimands' true values. Our submitted solution was based on $k = 500$ (vertical dashed line).

4.5.3 Challenge 4

4.5.3.1 Data

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

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

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

4.5.3.2 Methodology

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

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

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

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

At a high level, our method proceeds in a similar vein to Challenge 3: we will model $\boldsymbol{X}$ as max-linear and compute estimates for p_1 and p_2 based on (4.11). However, our exploratory analysis revealed that not all components of $\boldsymbol{X}$ are asymptotically dependent, implying that $H_{\boldsymbol{X}}(\mathbb{V}(d)) = 0$. This has important ramifications for how we construct our solution. On the one hand, empirical or CP-estimates of A will assert that $C_{\mathbb{V}(d)}$ is an extremal direction, resulting in a misspecified model. On the other hand, a model that correctly identifies $C_{\mathbb{V}(d)}$ as a $H_{\boldsymbol{X}}$-null set will yield $\hat{p}_1 = \hat{p}_2 = 0$, despite the true values being non-zero (**Rohr23**). How do we resolve this difficulty? The key is to identify clusters of asymptotically dependent variables prior to model fitting and estimate the probability of concurrent joint exceedance events in all clusters. Our working hypothesis – that the marginal variables can be partitioned such that asymptotic independence is present between clusters but not within them – is formalised below.

Assumption 1. There exists $2 \le K \le d$ and a partition $\beta_1,\dots,\beta_K$ of $\mathbb{V}(d)$ such that the angular measure is supported on the closed subspaces $\bar{C}_{\beta_1},\dots,\bar{C}_{\beta_K} \subset \mathbb{S}_+^{d-1}$, where

$\bar{C}_\beta := \{C_{\beta'} : \beta' \subseteq \beta\}$ for any $\beta \in \mathcal{P}_d^\star$. That is, $H_{\boldsymbol{X}}(\cup_{l=1}^K \bar{C}_{\beta_l}) = H_{\boldsymbol{X}}(\mathbb{S}_+^{d-1})$.

This scenario has been considered before, cf. Assumption 1 in **fomichovSphericalClusteringDetection2**
If $\boldsymbol{X}$ is max-linear with parameter $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q) \in \mathbb{R}_+^{d\times q}$, we may restate the assumption as follows.

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

Under Assumption 2, $\boldsymbol{X}$ divides into random sub-vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$, where $\boldsymbol{X}^{(l)} := (X_j : j \in \beta_l) \sim \text{MaxLinear}(A_\phi^{(l)}; \alpha)$ for $l = 1, \ldots, K$. (Henceforth, assume for notational convenience that the columns of A are already appropriately ordered, so that $A = A_\phi$.) While the clustering assumption is simplistic and cannot be expected to hold in general applications, it reflects what we suspect to be the true dependence structure for Challenge 4. Moreover, it is dimension reducing because the original d-dimensional problem is transformed to a set of K independent problems with dimensions $d_1, \ldots, d_K < d$. This ameliorates the curse of dimensionality to some extent.

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

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

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

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}}) \approx \prod_{l=1}^K \hat{\mathbb{P}}(A^{(l)} \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}). \tag{4.16}$$

Table 4.1: Summary statistics for the Challenge 4 clusters and their empirical TPDMs.

(a) Challenge 4: cluster summary statistics.

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

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

4.5.3.3 Results

We now present our results for Challenge 4. First, the variables $X_1, \ldots, X_d$ are partitioned into K groups based on asymptotic (in)dependence using the clustering algorithm of **bernardClusteringMaximaSpatial2013**. This entails constructing a distance matrix $\mathcal{D} = (\hat{d}_{ij})$, where $\hat{d}_{ij}$ denotes a non-parametric estimate of the F-madogram distance between variables X_i and X_j. The distance metric is connected to the strength of extremal dependence between X_i and X_j, with $\hat{d}_{ij} \approx 0$ implying strong asymptotic dependence and $\hat{d}_{ij} = 1/6$ in the case of asymptotic independence. The partition around medoids (PAM) clustering algorithm (**kaufmanFindingGroupsData1990**) returns a partition $\beta_1, \ldots, \beta_K$ of $\mathbb{V}(d)$ based on $\mathcal{D}$. The number of clusters K is a pre-specified tuning parameter; we identify $K = 5$ clusters whose sizes are given in Table 4.1. By consulting **Rohr23**, we have verified that our clustering was correct. Defining cluster membership variables $\mathcal{M}_1, \ldots, \mathcal{M}_d \in \{1, \ldots, K\}$ by $\mathcal{M}_i = l \iff i \in \beta_l$ for $i = 1, \ldots, d$, we find that $\max\{\hat{d}_{ij} : \mathcal{M}_i = \mathcal{M}_j\} = 0.113 < 1/6$ and $\min\{\hat{d}_{ij} : \mathcal{M}_i \neq \mathcal{M}_j\} = 0.164 \approx 1/6$. These summary statistics are consistent with Assumption 1.

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

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

We set $k_1 = \ldots = k_K =: k = 250$, corresponding to a sampling fraction of $k/n = 2.5\%$ for each cluster. The empirical TPDMs for the first two clusters are displayed in Figure 4.9; summary statistics for all clusters' TPDMs are listed in Table 4.1. Asymptotic dependence is strongest in clusters 2 and 3 and weakest in clusters 1 and 4.

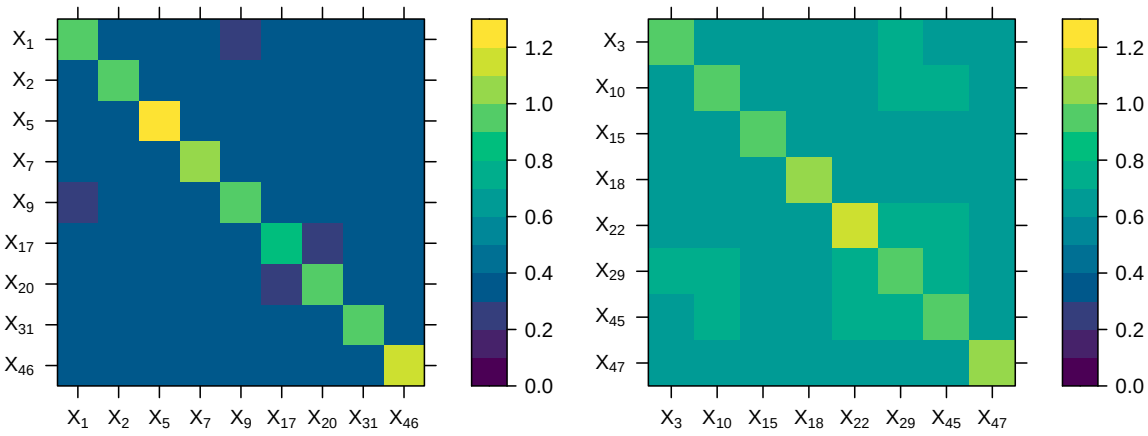

Figure 4.9: The estimated TPDMs for the first two clusters, $\hat{\Sigma}_{\boldsymbol{X}^{(1)}}$ (left) and $\hat{\Sigma}_{\boldsymbol{X}^{(2)}}$ (right), based on the $k = 250$ most extreme observations in norm.

By repeated application of the CP-decomposition algorithm of **kiriliouk**EstimatingProbabilitiesMulti with randomly chosen inputs, we obtain $N_{\text{cp}} = 50$ CP-estimates of each matrix $A^{(l)}$. The resulting CP-factors are denoted by $\tilde{A}_1^{(l)}, \ldots, \tilde{A}_{N_{\text{cp}}}^{(l)}$. Note that among $\tilde{A}_1^{(l)}, \ldots, \tilde{A}_{N_{\text{cp}}}^{(l)}$ there are at most d_l distinct leading columns, because there are only d_l unique ways to initialise the (deterministic) algorithm. But $\hat{\mathbb{P}}(\tilde{A} \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}})$ is fully determined by $\tilde{\boldsymbol{a}}_1$, so in fact $\{\hat{\mathbb{P}}(\tilde{A}_r^{(l)} \circ \boldsymbol{Z} \in \mathcal{C}_{\mathbb{V}(d_l),\boldsymbol{u}^{(l)}}) : r = 1, \ldots, N_{\text{cp}}\}$ contains at most d_l distinct values. These values are represented by black points in the left-hand plot in Figure 4.10. Clusters 2 and 3 are deemed most likely to experience a joint extreme event, because they contain a small number ($d_2 = d_3 = 8$) of highly dependent variables. The effect of changing the threshold

from $\boldsymbol{u}_1$ to $\boldsymbol{u}_2$ is most pronounced in cluster 5, since it is primarily composed of sites in area U2.

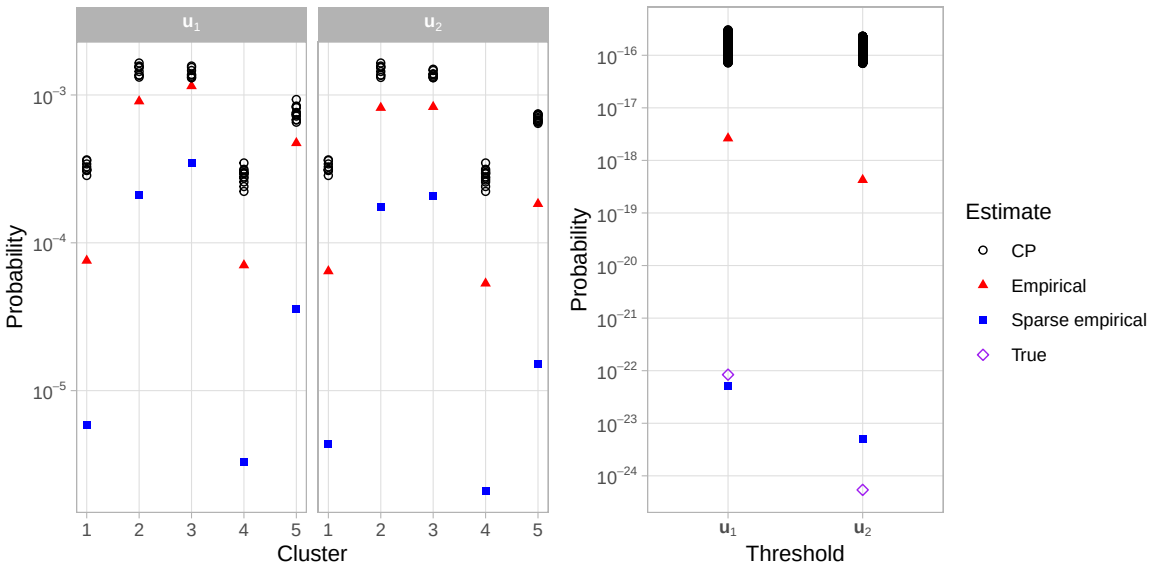

Figure 4.10: Joint exceedance probability estimates for each cluster sub-vector (left) and the full vector (right) based on different estimators for the parameters $A^{(1)}, \ldots, A^{(5)}$.

Substituting each CP-estimate into (4.16) and enumerating over all possible combinations, we produce sets of estimates of p_i, for $i = 1, 2$, given by

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

Each set has size $N_{\text{cp}}^K \approx 3 \times 10^8$ (including repeated values) and contains $\prod_{l=1}^K d_l = 89,856$ distinct values. The distributions of the estimates are represented by the black points in the right-hand panel of Figure 4.10. Our final point estimates are taken as the median values $\tilde{p}_1 := \text{median}(\tilde{P}_1) = 1.4 \times 10^{-16}$ and $\tilde{p}_2 := \text{median}(\tilde{P}_2) = 1.3 \times 10^{-16}$, respectively, to two significant figures.

4.5.3.4 Improving performance using sparse empirical estimates

Unfortunately, it transpires that our method dramatically over-estimated the true probabilities. In hindsight, this could have been anticipated in view of the simulation study in Section 5.1 in **kiriliouk EstimatingProbabilitiesMultivariate2022**, where the authors remark that failure regions of the the type $\mathcal{C}_{\mathbb{V}(d),\boldsymbol{u}}$ are poorly summarised by the TPDM.

This prompts us to investigate whether using empirical or sparse empirical estimates instead of a CP-based approach would have improved our performance. For the former, this simply involves replacing $A^{(l)}$ with the precomputed matrix $\hat{A}^{(l)}$ in (4.16). The approach based on sparse empirical estimates proceeds analogously except that we must revert to the $\alpha = 1$ setting and transform the data and thresholds accordingly. The resulting estimates for the joint exceedance probabilities are represented by the red and blue points in Figure 4.10. The values are smaller (better) in all clusters and consequently the final estimates are closer to the true event probabilities (purple). In fact, using sparse empirical estimates yields $\hat{p}_1^\star = 5.1 \times 10^{-23}$ and $\hat{p}_2^\star = 5.0 \times 10^{-24}$, which are remarkably close to the correct solutions $p_1 = 8.4 \times 10^{-23}$ and $p_2 = 5.4 \times 10^{-25}$. Submitting these values would have significantly improved our ranking for this sub-challenge.

4.6 Conclusion

We propose four different methods in order to solve the EVA (2023) Conference Data Challenge. These methods allow for the estimation of extreme quantiles and probabilities. In the univariate cases, we estimate the conditional quantile by attempting to cluster the covariates to categorise the extremes. Once clustered, bootstrapping was used in order to estimate central 50% confidence intervals of the extreme quantiles. In practice however, we found no clearly defined clusters and that the Gaussian mixture distribution was not able to capture the non-Gaussian distribution of the covariates. This could also be extended by using GPDs with covariate models for each cluster, to try to improve quantile estimates. In the second challenge, we were able to assess the GPD fit of a misspecified model and apply a correction to an initial fit where there was large uncertainty in the tail behaviour. This resulted in a change in the return level from an under-estimate to a potentially less-severe over-estimate. By using a more sophisticated model for the covariates, further work can be done to articulate the extent to which the correction was due to a bias in the model fit.

Our performance in the multivariate challenges demonstrates that the max-linear model provides a good framework for estimating tail event probabilities. By connecting this model with sparse simplex projections, one can achieve exceptional performance on

both Challenges 3 and 4. Given these results, further research on the theoretical properties of the sparse empirical estimator $\hat{A}^\star$ is warranted. An obvious shortcoming of our Challenge 3 methodology is that it ignores the available covariate information. Incorporating covariates into the max-linear/TPDM framework remains an unexplored area. The upwards bias in our Challenge 4 probability estimates is partly caused by over-estimating the dependence strength between certain groups of variables. The empirical TPDM has a known tendency to over-estimate weak/moderate dependence (**fixSimultaneousAutoregressiveModels2021**), as is the case in clusters 1, 3, and 5. A better TPDM estimator that takes this issue into account would likely improve our final results. In Challenges 3 and 4 we switch between $\alpha = 1$ and $\alpha = 2$ in a way that is somewhat unsatisfactory (mathematically and presentationally). In hindsight, employing a generalised version of the TPDM (**kiriliioukEstimatingProbabilitiesMultivariate2022**) would have allowed us to fix $\alpha = 1$ throughout. Finally, uncertainty quantification is a crucial aspect of risk assessment in practice, especially when dealing with the exceedingly rare events encountered in Challenge 4. **fixSimultaneousAutoregressiveModels2021** utilise a bootstrapping procedure in the context of TPDM-based inference for extremes and their approach would transfer to our methods very naturally.

5 Bias-corrected estimation of the TPDM

5.1 Introduction and motivation

Section XX demonstrated a deficiency of the empirical TPDM estimator, which we termed the bias issue. Specifically, the empirical TPDM $\hat{\Sigma}$ has a tendency to overestimate weak pairwise dependence – see (1.65). The bias issue has arisen at several subsequent points throughout the thesis: our test for time-varying dependence becomes less powerful when dependence is weak (Figure 2.13); projections onto principal eigenspaces of $\hat{\Sigma}$ result in poor reconstructions near the simplex boundary (Figure 3.9, second row and right column); tail probability estimates based on completely positive factorisations of $\hat{\Sigma}$ are biased (Figure XX).

This chapter addresses the problem of reducing bias in the empirical TPDM. Bias reduction requires grappling with two interconnected problems. First, what *form* should a bias-corrected estimator take? The procedure proposed in **fixSimultaneousAutoregressiveModels2021** (see Section XX) shrinks all entries of $\hat{\Sigma}$ equally. We contend that this is not ideal, given that the bias is greatest for pairs of variables where dependence is weak. Inspired by **rothmanGeneralizedThresholdingLarge2009**, a flexible extension based on the adaptive lasso is proposed. This permits control over whether shrinkage is applied equally to all entries or targeted towards smaller entries, at the cost of introducing an additional parameter. The second problem concerns *how much* bias correction should be applied. **fixSimultaneousAutoregressiveModels2021** solve this by assuming that dependence strength is a decreasing function of spatial distance and is approximately zero for the most distant pair(s) of sites in the study region. However, these assumptions are only meaningful (let alone satisfied) in spatial settings, precluding its use in general applications. While we were unable to find a general solution to this research problem,

we made progress in a special case where bias-reduction is achieved via linear shrinkage techniques (**ledoitImprovedEstimationCovariance2003**). For this class, we are able to devise a theoretically justified, data-driven procedure for selecting the magnitude of the shrinkage.

5.2 Regularised TPDM estimators: thresholding and shrinkage

In high-dimensional settings, where the number of variables is comparable to the number of observations, the empirical covariance matrix may be noisy and poorly conditioned. Regularisation may be applied to obtain more stable and accurate estimates. In particular, thresholding is used to enforce sparsity by shrinking small values towards zero, enhancing stability and interpretability in cases where many variables are weakly correlated (**rothmanGeneralizedThresholdingLarge2009**). On the other hand, shrinkage methods move the empirical estimate towards a biased but highly structured pre-specified target matrix (**ledoitImprovedEstimationCovariance2003**). In both cases, more accurate estimates are obtained by reducing the entries of the empirical estimator in some way. Thus, we may consider applying these techniques to the empirical TPDM to achieve bias-reduction. This may have the additional benefit of improving stability in cases where d is comparable to the number of extremes k, but this aspect will not be explored.

5.2.1 Thresholded TPDM estimators

The core idea of thresholded estimation is to apply a thresholding operator to the raw, empirical estimate (in our case $\hat{\Sigma}$). The thresholding operator controls the nature of the regularisation and must satisfy certain properties.

Definition 5.1. Let $\lambda \geq 0$ be a fixed threshold. A function $s_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is called a thresholding operator if, for all $x \in \mathbb{R}_+$,

$$s_\lambda(x) \leq x, \tag{5.1}$$

$$x \leq \lambda \implies s_\lambda(x) = 0, \tag{5.2}$$

$$|s_\lambda(x) - x| \leq \lambda. \tag{5.3}$$

The tuning parameter λ controls the strength of the regularisation. In the two extreme cases, we have $s_0(x) = x$ (no regularisation) and $s_\infty(x) = 0$ for all $x \in \mathbb{R}_+$ (full shrinkage). Condition (5.1) states that s_λ should push x towards zero. Condition (5.2) enforces a threshold below which values are mapped to zero, potentially inducing sparsity in the output. Condition (5.3) limits the amount of shrinkage. For simplicity, we assume the threshold in (5.2) and the maximal shrinkage in (5.3) are identical. This could be relaxed by introducing separate parameters λ_1 and λ_2 if desired (**antoniadisRegularizationWaveletApproximations2001**). Thresholding operators may be equivalently formulated as solutions to regularised optimisation problems of the form

$$s_\lambda(x) = \arg\min_{y \in \mathbb{R}_+} \left[\frac{1}{2}(y - x)^2 + p_\lambda(y; x) \right] \tag{5.4}$$

for some penalty function p_λ. While we do not explicitly make use of this formulation, the penalty function is useful for providing insight into an operator's behaviour.

We will consider three popular thresholding operators: hard-thresholding, soft-thresholding, and adaptive lasso. For reference, the corresponding functions $x \mapsto s_{0.1}(x)$ are shown in Figure 5.1. The hard-thresholding rule

$$s_\lambda^H(x) = x\mathbf{1}\{x > \lambda\} \tag{5.5}$$

applies no shrinkage to values above λ. The jump discontinuity in $s_\lambda^H(x)$ at $\lambda = x$ can be observed in Figure 5.1. In the space of thresholding operators, it is diametrically opposed to the soft-thresholding operator

$$s_\lambda^S(x) = (x - \lambda)_+, \tag{5.6}$$

which induces the maximal amount of shrinkage permitted by (5.3). Soft-thresholding corresponds to solving (5.4) with lasso penalty function $p_\lambda(y; x) = \lambda y$. Note that $p_\lambda(y; x)$ does not depend on x, reflecting the fact that equal shrinkage is applied to large and small values (exceeding λ) alike. Other thresholding operators offer some kind of compromise between these two cases. The adaptive lasso penalty function $p_\lambda(y; x) = \lambda^{\eta+1} y x^{-\eta}$ down-

weights larger values, leading to the thresholding rule

$$s^{AL}_{\lambda,\eta}(x) = (x - \lambda^{\eta+1}x^{-\eta})_+, \qquad (\eta \geq 0). \tag{5.7}$$

The degree of down-weighting is controlled by the tuning parameter $\eta \geq 0$. Setting $\eta = 0$ results in no down-weighting, yielding the soft-thresholding operator $s^{AL}_{\lambda,0}(x) = s^S_\lambda(x)$. On the other hand, $s^{AL}_{\lambda,\eta}(x) \to s^H_\lambda(x)$ as $\eta \to \infty$. Thus, soft- and hard-thresholding are special (limiting) cases of adaptive lasso. The clipped L_1-penalty and the smoothly clipped absolute deviations (SCAD) penalty behave in a similar fashion (**antoniadisRegularizationWaveletApproximations2001**).

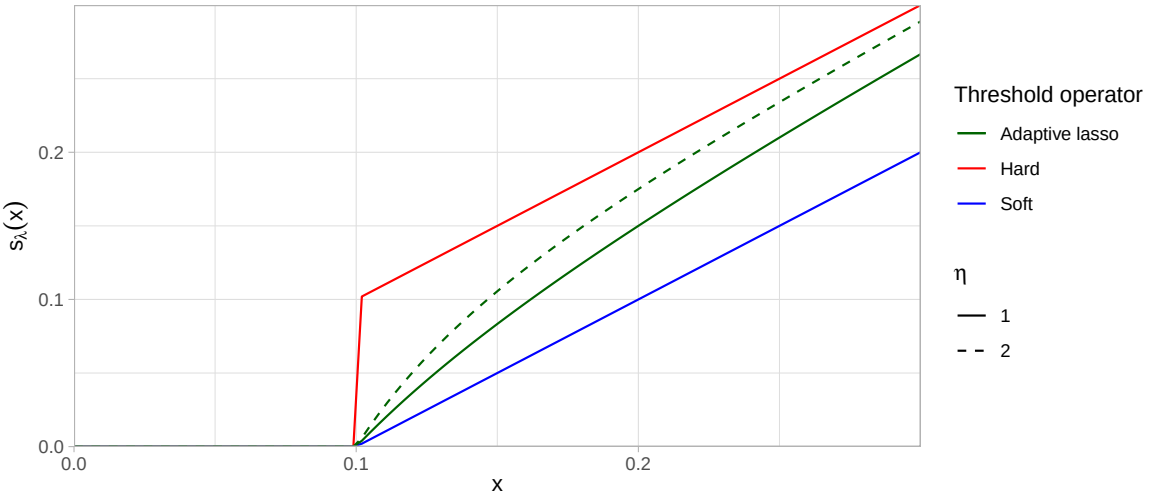

Figure 5.1: Illustration of popular thresholding operators $s_\lambda(x)$ with $\lambda = 0.1$.

We now proceed to apply thresholding to the empirical TPDM to define our first class of bias-reduced estimators. Let $A = (a_{ij}) \in \mathbb{R}^{m\times n}_+$ be a non-negative matrix and $s_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ a thresholding operator. The thresholded matrix $s_\lambda(A) := (s_\lambda(a_{ij}))$ is defined by element-wise application of s_λ to A. Clearly this is very simple and fast to compute.

Definition 5.2. Let s_λ be a thresholding operator. The thresholded TPDM estimator is the $d \times d$ matrix

$$\tilde{\Sigma}(\lambda) = (\tilde{\sigma}_{ij}(\lambda)), \qquad \tilde{\sigma}_{ij} = \begin{cases} \hat{\sigma}_{ij}, & i = j, \\ s_\lambda(\hat{\sigma}_{ij}), & i \neq j. \end{cases} \tag{5.8}$$

Note that we do not shrink the diagonal entries, but in any case the performance of estimators will be evaluated using the off-diagonal entries only. The soft-thresholded TPDM

$s^S_\lambda(\hat{\Sigma})$ coincides with estimator proposed by **fixSimultaneousAutoregressiveModels2021**. However, Definition 5.2 includes a much broader class of estimators with more flexible behaviours. In particular, the adaptive lasso TPDM $s^{AL}_{\lambda,\eta}(\hat{\Sigma})$ allows us to concentrate shrinkage towards small entries in $\hat{\Sigma}$, where the bias is presumed to be greater. Figure 5.2 provides a visual comparison between the three thresholding methods using the Red Sea surface temperature data from Section XX. Each matrix represents a regularised TPDM $\tilde{\Sigma}$ of the random vector $\boldsymbol{X}$, where X_i corresponds to the ith site in Figure 2.9 (ordered by longitude and then by latitude) for $i = 1, \ldots, 200$. The figure shows how the hard-thresholded, soft-thresholded and adaptive lasso ($\eta = 2$) estimates vary with the regularisation parameter $\lambda \in \{0, 0.1, \ldots, 0.5\}$. The empirical TPDM is estimated from the proportion $k/n = 0.15$ of largest observations; this matrix is shown in the sub-panels where $\lambda = 0$. Since the components of $\boldsymbol{X}$ are ordered according to the spatial locations, $\hat{\Sigma}$ generally exhibits stronger dependencies among variable pairs closer to the diagonal and weaker dependencies for those farther away. Hard-thresholding (top left) has no effect when $\lambda < \min\{\hat{\sigma}_{ij} : i \neq j\} \approx 0.249$. Increasing the threshold introduces abrupt shifts, resulting in a non-smooth sequence of solutions. In contrast, soft-thresholding yields smoother solutions because all off-diagonal entries receive some shrinkage when $\lambda > 0$. However, the regularisation acts indiscriminately by applying an equal penalty to all entries. Consequently, as λ increases the dependence strength is close to zero for all pairs of sites, including neighbouring ones. The adaptive lasso (bottom left) combines the advantages of hard- and soft-thresholding, yielding smooth solution paths but prioritising shrinkage of entries that are close to zero.

A drawback of these estimators is that thresholding fails to preserve certain matrix properties, including positive semi-definiteness and complete positivity. Evidence of this is provided in Figure 5.3, which depicts the eigenvalues of the hard-thresholded TPDMs $\tilde{\Sigma}(\lambda)$ from Figure 5.2. Specifically, the plot shows the absolute value of the eigenvalues, with positive eigenvalues in black and negative eigenvalues in red. As λ increases and the thresholding takes effect, the eigenvalues of $\tilde{\Sigma}(\lambda)$ get shrunk, occasionally causing them to become negative. Thus, $\tilde{\Sigma}$ may need to undergo pre-processing before it can be incorporated into existing TPDM-based modelling tools such as PCA. This motivates us to propose an alternative approach to bias-reduction based on linear shrinkage methods. While this class is less flexible than thresholded TPDMs, its simplicity ensures estimates

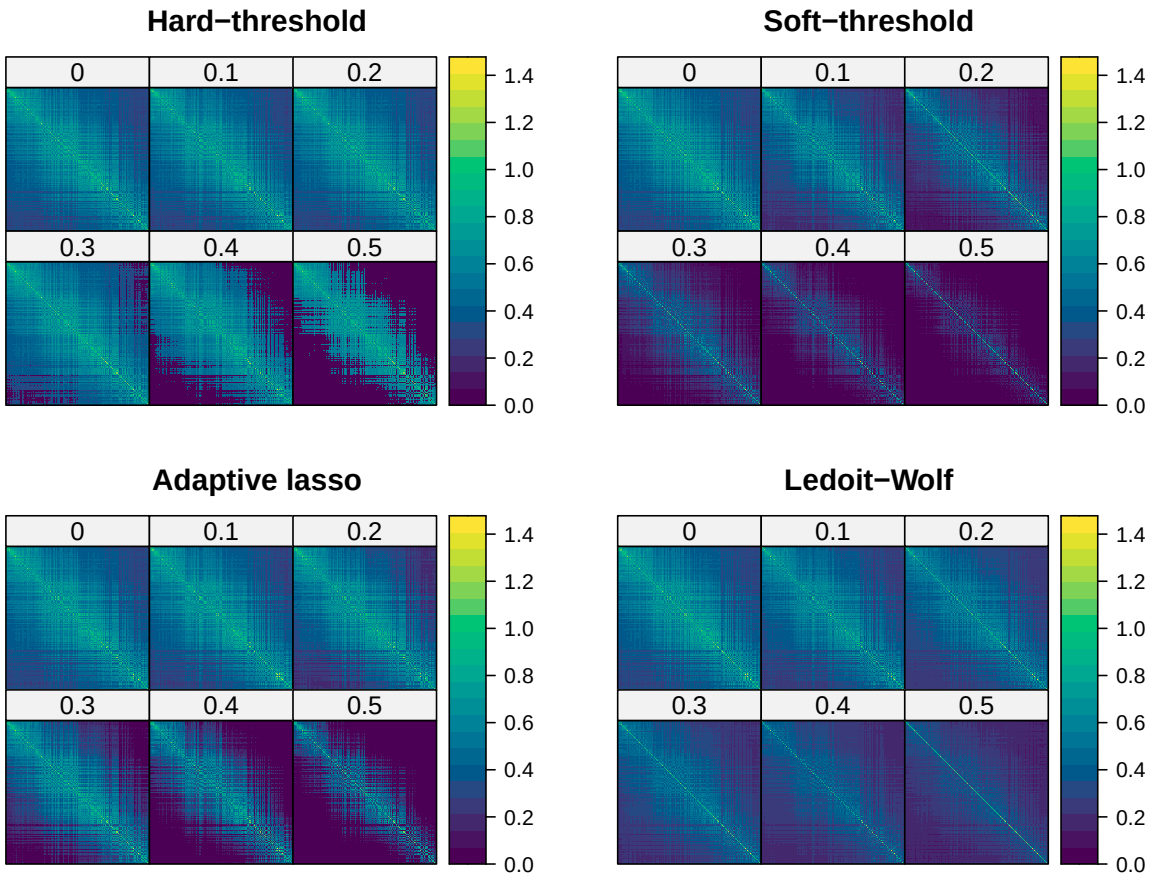

Figure 5.2: Regularised TPDM estimates $\tilde{\Sigma}(\lambda)$ based on the Red Sea surface temperature data from Section XX. The $d = 200$ sites in Figure 2.9 are ordered by longitude and then latitude and $\hat{\Sigma}$ is estimated with $k/n = 0.15$. The values at the top of each sub-panel represent the regularisation parameter $\lambda \in \{0, 0.1, \ldots, 0.5\}$. For the adaptive lasso estimates we take $\eta = 2$.

possess nice properties and are more amenable to mathematical analysis.

5.2.2 The Ledoit-Wolf TPDM estimator

Shrinkage is a regularisation technique dating back to **steinInadmissibilityUsualEstimator1956**, whereby a sample estimate is shrunk towards some pre-specified target. The simplest shrinkage method is Ledoit-Wolf linear shrinkage (**ledoitImprovedEstimationCovariance2003**). Their approach to regularising the sample covariance matrix transfers easily to the TPDM.

Definition 5.3. Let T be a $d \times d$ target matrix. For $\lambda \in [0, 1]$, the Ledoit-Wolf TPDM is

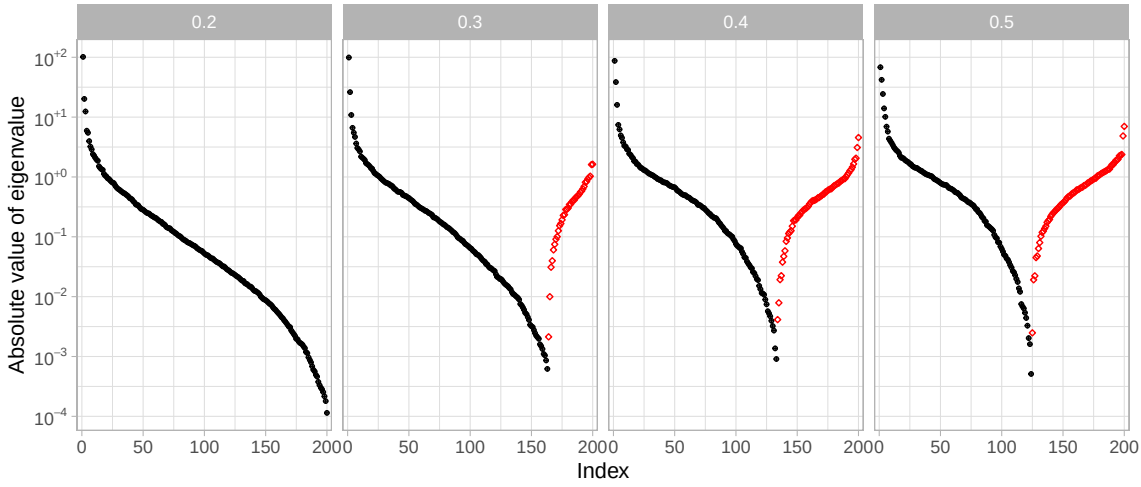

Figure 5.3: Absolute values of the eigenvalues of the hard-thresholded TPDM estimates $\tilde{\Sigma}(\lambda)$ from Figure 5.2 (top left) for $\lambda \in \{0.2, 0.3, 0.4, 0.5\}$. Positive and negative eigenvalues are coloured black and red, respectively.

the $d \times d$ matrix

$$\tilde{\Sigma}(\lambda) = \lambda T + (1 - \lambda)\hat{\Sigma}. \tag{5.9}$$

The key idea is to bias the empirical TPDM $\hat{\Sigma}$ towards a highly structure target T by forming a convex linear combination between the two. The shrinkage intensity $\lambda \in [0, 1]$ determines the weight allocated to each matrix. No shrinkage occurs when $\lambda = 0$. If $\lambda = 1$ the estimator returns T, which is biased but perfectly stable (it has zero variance). The choice of target depends on the context at hand. We will always choose $T = I_d$, the $d \times d$ identity matrix. This means the off-diagonal entries are shrunk towards zero and the diagonal entries are left unaltered. Similar to soft-thresholding, the Ledoit-Wolf TPDM counters the bias by applying equal shrinkage to all entries. A slight difference is that the Ledoit-Wolf shrinkage intensity is applied multiplicatively, i.e. $\tilde{\sigma}_{ij} = (1 - \lambda)\hat{\sigma}_{ij}$, whereas thresholding works additively. Figure 5.2 (bottom right) shows the sequence of Ledoit-Wolf TPDMs for the Red Sea data, as described earlier. We observe that linear shrinkage behaves quite differently to any of the thresholding estimators; the general structure of the matrix does not change all that much as λ increases. While this behaviour is not necessarily desirable from a modelling perspective, it explains why the Ledoit-Wolf TPDM retains many of the mathematical properties of the empirical TPDM.

Proposition 5.1. *The Ledoit-Wolf TPDM $\tilde{\Sigma}(\lambda)$ is symmetric, positive semi-definite, and*

completely positive.

Proof. For brevity, we suppress dependence on λ. Symmetry follows immediately because for any i, j,

$$\tilde{\sigma}_{ij} = \lambda \mathbf{1}\{i = j\} + (1 - \lambda)\hat{\sigma}_{ij} = \lambda \mathbf{1}\{j = i\} + (1 - \lambda)\hat{\sigma}_{ji} = \tilde{\sigma}_{ji}.$$

For positive semi-definiteness, recall that $\hat{\Sigma}$ is positive semi-definite and so for any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$,

$$\boldsymbol{y}^T \tilde{\Sigma} \boldsymbol{y} = \lambda \boldsymbol{y}^T \boldsymbol{y} + (1 - \lambda)\boldsymbol{y}^T \hat{\Sigma} \boldsymbol{y} = \lambda \|\boldsymbol{y}\|_2^2 + (1 - \lambda)\boldsymbol{y}^T \hat{\Sigma} \boldsymbol{y} \geq 0.$$

The identity matrix is trivially completely positive since $I_d = I_d I_d^T$ and $\hat{\Sigma}$ is completely positive by Proposition 1.3. The class of completely positive matrices is a convex cone (**bermanCompletelyPositiveMatrices2003**), so complete positivity is preserved under convex combinations.

□

Positive semi-definiteness means the Ledoit-Wolf estimator may replace $\hat{\Sigma}$ in PCA techniques. Complete positivity implies that there exists a max-linear random vector $\boldsymbol{X}_\lambda$ whose TPDM is $\tilde{\Sigma}(\lambda)$. Therefore the Ledoit-Wolf TPDM is more straightforward to use in practice compared to the thresholded estimators.

5.3 Selecting the regularisation parameter

5.3.1 Frobenius risk minimisation

Statistically, the main challenge is to select the regularisation parameter λ. A natural criterion is to choose λ so as to minimise the discrepancy between the true TPDM Σ and the regularised estimate $\tilde{\Sigma}(\lambda)$. We are generally only interested in the error in the off-diagonal entries, so we will tend to refer to the upper-half vectorised quantities

$$\boldsymbol{\sigma} = \mathrm{vecu}(\Sigma), \quad \hat{\boldsymbol{\sigma}} = \mathrm{vecu}(\hat{\Sigma}), \quad \tilde{\boldsymbol{\sigma}}(\lambda) = \mathrm{vecu}(\tilde{\Sigma}(\lambda)),$$

where the vecu operator was defined in (1.61). For a given class $\{\tilde{\Sigma}(\lambda) : \lambda \geq 0\}$, the risk associated with λ is defined as

$$\mathcal{R}(\lambda) := \mathbb{E}[\|\tilde{\boldsymbol{\sigma}}(\lambda) - \boldsymbol{\sigma}\|_2^2] = \mathbb{E}\left[\sum_{i<j} (\tilde{\sigma}_{ij}(\lambda) - \sigma_{ij})^2\right]. \tag{5.10}$$

We refer to $\mathcal{R}(\lambda)$ as the Frobenius risk. The minimiser

$$\lambda^\star := \arg\min_{\lambda} \mathcal{R}(\lambda)$$

corresponds to the regularisation parameter which gives the TDPM estimate closest to the true TPDM in terms of the mean-square error. Of course, the true TPDM is always unknown in practice, so the risk $\mathcal{R}(\lambda)$ cannot be evaluated or minimised directly. We call $\lambda^\star$ the oracle parameter to signify that it is the value that would be chosen by an oracle with knowledge of the underlying true TPDM. Our goal is to devise a statistical procedure for estimating $\lambda^\star$.

The task of estimating $\lambda^\star$ becomes much simpler if we are privy to additional information about the data-generating process. For example, **fixSimultaneousAutoregressiveModels2021** infer a reasonable soft-thresholding parameter $\lambda = \beta_2$ in (1.66) by exploiting prior information about the spatial configuration and the spatial extent of the modelled phenomenon. Similarly, in supervised learning settings where an objective loss function and labelled training data are available, one may resort to standard hyperparameter tuning procedures such as cross validation. We study the problem in its most general, challenging form, where no additional structure/information is available.

For covariance matrices, a standard strategy for selecting the regularisation parameter is to minimise an empirical estimate of $\mathcal{R}$, i.e. an approximation of the risk function constructed from the data (**yiSUREtunedTaperingEstimation2013**). Unfortunately, this strategy does not transfer easily to the TPDM context. Upon noting that

$$(\tilde{\sigma}_{ij}(\lambda) - \sigma_{ij})^2 = (\tilde{\sigma}_{ij}(\lambda) - \hat{\sigma}_{ij})^2 + 2(\tilde{\sigma}_{ij}(\lambda) - \sigma_{ij})(\hat{\sigma}_{ij}(\lambda) - \sigma_{ij}) - (\hat{\sigma}_{ij} - \sigma_{ij})^2$$

the risk function (5.10) can be equivalently expressed as

$$\mathcal{R}(\lambda) = \mathbb{E}[\|\tilde{\boldsymbol{\sigma}}(\lambda) - \hat{\boldsymbol{\sigma}}\|_2^2] + 2\sum_{i<j} \mathbb{E}[(\tilde{\sigma}_{ij}(\lambda) - \sigma_{ij})(\hat{\sigma}_{ij} - \sigma_{ij})] - \mathbb{E}[\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2^2] \tag{5.11}$$

These three terms are called apparent error, optimism, and constant (**yiSUREtunedTaperingEstimatio**
The apparent error acts as a penalty against departing too far away from the empirical TPDM. This term can be estimated using the unbiased estimator $\|\tilde{\boldsymbol{\sigma}}(\lambda) - \hat{\boldsymbol{\sigma}}\|_2^2$. The constant term is irrelevant and can be ignored for the risk minimization as it does not depend on λ. The fundamental issue stems from the optimism term. For covariance matrices, the empirical estimator is unbiased and the summand may be replaced with $\mathrm{Cov}(\tilde{\sigma}_{ij}(\lambda), \hat{\sigma}_{ij})$, which may be estimated via bootstrapping (**fangTuningparameterSelectionRegularized2016**). This approach cannot be replicated for TPDMs, because the empirical estimator is biased. This impediment prevented us from finding a solution to the general problem of tuning arbitrary regularised TPDMs. However, we were able to devise a procedure for optimising Ledoit-Wolf estimators by taking a completely different approach to the one sketched above.

5.3.2 The optimal Ledoit-Wolf shrinkage intensity

ledoitImprovedEstimationCovariance2003 provide a formula for the asymptotically optimal Ledoit-Wolf shrinkage intensity for covariance matrices. Inspired by their approach, we are able to obtain an analogous result for TPDMs. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ is a sequence of independent copies of $\boldsymbol{X}$ and let $\tilde{\Sigma}(\lambda)$ denote a Ledoit-Wolf estimator of the TPDM. For any $\lambda \in [0, 1]$, the associated Frobenius risk is

$$\begin{aligned}
\mathcal{R}(\lambda) &= \mathbb{E}[\|(1-\lambda)\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2^2] \\
&= \sum_{i<j} \mathbb{E}\left[((1-\lambda)\hat{\sigma}_{ij} - \sigma_{ij})^2\right] \\
&= \sum_{i<j} \left\{(1-\lambda)^2 \mathbb{E}[\hat{\sigma}_{ij}^2] - 2(1-\lambda)\sigma_{ij}\mathbb{E}[\hat{\sigma}_{ij}] + \sigma_{ij}^2\right\}.
\end{aligned}$$

The first-order optimality condition gives

$$\mathcal{R}'(\lambda) = 0 \implies \sum_{i<j} 2(\lambda - 1)\mathbb{E}[\hat{\sigma}_{ij}^2] + 2\sigma_{ij}\mathbb{E}[\hat{\sigma}_{ij}] = 0 \implies \lambda^\star := \frac{\sum_{i<j}\left\{\mathbb{E}[\hat{\sigma}_{ij}^2] - \sigma_{ij}\mathbb{E}[\hat{\sigma}_{ij}]\right\}}{\mathbb{E}[\|\hat{\boldsymbol{\sigma}}\|_2^2]}.$$

It is simple to check that $\mathcal{R}''(\lambda) = \sum_{i<j} 2\mathbb{E}[\hat{\sigma}_{ij}^2] > 0$, confirming that $\lambda^\star$ is indeed the risk minimiser. The formula for $\lambda^\star$ still depends on the true TPDM, so this is not helpful by itself. However, we may recognise the numerator of $\lambda^\star$ as being related to the variance of $\hat{\sigma}_{ij}$ and recall from Proposition 1.5 that the asymptotic variance of $\sqrt{k}\hat{\sigma}_{ij}$ is ν_{ij}^2. For clarity, we now include dependence on the sample size n explicitly. Under the conditions of Proposition 1.5, we have that

$$\begin{aligned}
&\lim_{n\to\infty} k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]\lambda^\star(n) \\
&\quad = \lim_{n\to\infty} \sum_{i<j} k(n)\left\{\mathbb{E}[\hat{\sigma}_{ij}(n)^2] - \sigma_{ij}\mathbb{E}[\hat{\sigma}_{ij}(n)]\right\} \\
&\quad = \lim_{n\to\infty} \sum_{i<j} k(n)\left\{\mathbb{E}[\hat{\sigma}_{ij}(n)^2] - \mathbb{E}[\hat{\sigma}_{ij}(n)]^2\right\} \\
&\quad = \lim_{n\to\infty} \sum_{i<j} k(n)\mathrm{Var}(\hat{\sigma}_{ij}(n) \\
&\quad = \sum_{i<j} \nu_{ij}^2.
\end{aligned}$$

What's the best way of doing Line 2 to line 3? Am just using the fact that $\mathbb{E}[\hat{\sigma}_{ij}] \to \sigma_{ij}$**.** Since $\sum_{i<j}\nu_{ij}^2 > 0$ and $\|\hat{\boldsymbol{\sigma}}(n)\|_2^2 > 0$ almost surely, it follows that $\lambda^\star(n) > 0$ for sufficiently large n. The above result gives that, asymptotically, the optimal shrinkage behaves as a constant divided by $k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]$. As the asymptotic constant is unknown, we might naively try estimating the asymptotic variances ν_{ij}^2. Indeed, this is the approach taken in the context of covariance matrices and a consistent estimator $\hat{\nu}_{ij}^2$ of ν_{ij}^2 is given in (1.60) and used in Chapter XX. However, for a finite sample size $\hat{\nu}_{ij}^2$ is biased in a way that mirrors the bias in the empirical TPDM, so this approach is not viable. Instead, we are forced to circumvent estimation of $\sum_{i<j}\nu_{ij}^2$. To eliminate the asymptotic constant, we take ratios at different values of n. Given a fixed proportion $\kappa \in (0, 1)$, it is easy to see that

$$\lim_{n\to\infty} k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]\lambda^\star(n) = \sum_{i<j}\nu_{ij}^2 = \lim_{n\to\infty} k(\lfloor\kappa n\rfloor)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(\lfloor\kappa n\rfloor)\|_2^2]\lambda^\star(\lfloor\kappa n\rfloor).$$

and the algebra of limits then yields

$$\lim_{n\to\infty} \frac{k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]\lambda^\star(n)}{k(\lfloor\kappa n\rfloor)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(\lfloor\kappa n\rfloor)\|_2^2]\lambda^\star(\lfloor\kappa n\rfloor)} = 1.$$

For sufficiently large n all quantities in the denominator are non-zero, so the limit is well-defined. Moreover it holds that

$$\frac{\lambda^\star(\lfloor\kappa n\rfloor)}{\lambda^\star(n)} \approx \frac{k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]}{k(\lfloor\kappa n\rfloor)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(\lfloor\kappa n\rfloor)\|_2^2]} =: \zeta(n;\kappa).$$

In other words, the optimal shrinkage intensities at sample sizes n and $\lfloor\kappa n\rfloor$ are $\lambda^\star(n)$ and $\zeta(n;\kappa)\lambda^\star(n)$, respectively. The ratio $\zeta(n;\kappa)$ describes the rate of decay of the optimal shrinkage as more data becomes available. It is easy to see that $\zeta(n,\kappa) \to 1$ when $\kappa \to 1$. Additionally $\zeta(n;\kappa) \to k(n)/k(\lfloor\kappa n\rfloor)$ as $n \to \infty$ due to consistency of $\hat{\boldsymbol{\sigma}}$. This limit will usually be some function of κ, e.g. choosing $k(n) \propto \sqrt{n}$ gives $\zeta(n;\kappa) \to \kappa^{-1/2}$ as $n \to \infty$. Crucially, $\zeta(n;\kappa)$ is defined in terms of quantities that are known or may be estimated without bias. Let $\hat{\zeta} = \hat{\zeta}(n;\kappa)$ be an estimate of $\zeta = \zeta(n;\kappa)$ to be defined later.

Knowing its asymptotic rate of decay is not sufficient to deduce $\lambda^\star(n)$. The missing piece of information is its magnitude. Consider the empirical TPDMs $\hat{\Sigma}(n)$ and $\hat{\Sigma}(\lfloor\kappa n\rfloor)$ based on samples of sizes n and $\lfloor\kappa n\rfloor$, respectively. If these matrices are very similar (resp. different), then intuitively the absolute difference between $\lambda^\star(n)$ and $\lambda^\star(\lfloor\kappa n\rfloor)$ should be small (resp. large). Now, our large sample theory tells us that our best estimates of $\boldsymbol{\sigma}$ are of the form $(1-\lambda^\star(n))\hat{\boldsymbol{\sigma}}(n)$ and $(1-\hat{\zeta}\lambda^\star(n))\hat{\boldsymbol{\sigma}}(\lfloor\kappa n\rfloor)$. This hints at estimating $\lambda^\star(n)$ by minimising the discrepancy between these two estimates:

$$\hat{\lambda}(n) := \operatorname*{arg\,min}_{\lambda\in[0,\min(1,\zeta^{-1})]} \left\|(1-\lambda)\hat{\boldsymbol{\sigma}}(n) - (1-\hat{\zeta}\lambda)\hat{\boldsymbol{\sigma}}(\lfloor\kappa n\rfloor)\right\|_2^2. \tag{5.12}$$

5.3.3 Practical and statistical considerations

Having outlined the general approach, we fill in some missing details regarding inference. First we address estimation of ζ. To estimate $\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]$ and $\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(\lfloor\kappa n\rfloor)\|_2^2]$, we produce a collection of bootstrapped samples $\mathcal{X}_n^{(b)} = \{\boldsymbol{X}_1^{(b)},\dots,\boldsymbol{X}_n^{(b)}\}$ and $\mathcal{X}_{\lfloor\kappa n\rfloor}^{(b)} = \{\boldsymbol{X}_1^{(b)},\dots,\boldsymbol{X}_{\lfloor\kappa n\rfloor}^{(b)}\}$ from $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$, for $b = 1,\dots,B$. Next, we compute empirical TPDMs $\hat{\Sigma}^{(1)}(n),\dots,\hat{\Sigma}^{(B)}(n)$ and $\hat{\Sigma}^{(1)}(\lfloor\kappa n\rfloor),\dots,\hat{\Sigma}^{(B)}(\lfloor\kappa n\rfloor)$ using the $k(n)$ and $k(\lfloor\kappa n\rfloor)$ largest ob-

servations from $\mathcal{X}_n^{(1)}, \dots, \mathcal{X}_n^{(B)}$ and $\mathcal{X}_{\lfloor \kappa n \rfloor}^{(1)}, \dots, \mathcal{X}_{\lfloor \kappa n \rfloor}^{(B)}$, respectively. An estimate of ζ is then given by

$$\hat{\zeta}(n;\kappa) := \frac{k(n) \sum_{b=1}^{B} \|\hat{\boldsymbol{\sigma}}^{(b)}(n)\|_2^2}{k(\lfloor \kappa n \rfloor) \sum_{b=1}^{B} \|\hat{\boldsymbol{\sigma}}^{(b)}(\lfloor \kappa n \rfloor)\|_2^2}. \tag{5.13}$$

The bootstrapped empirical TPDMs can be reused in the final step, so that instead of (5.12), we actually set

$$\hat{\lambda}(n) := \operatorname*{arg\,min}_{\lambda \in [0, \zeta^{-1}]} \left\| \frac{1}{B} \sum_{b=1}^{B} \left[(1-\lambda)\hat{\boldsymbol{\sigma}}^{(b)}(n) - (1-\hat{\zeta}\lambda)\hat{\boldsymbol{\sigma}}^{(b)}(\lfloor \kappa n \rfloor) \right] \right\|_2^2. \tag{5.14}$$

Bootstrapping is intended to reduce noise in the estimation of ζ and $\lambda^\star$, but it is not fundamental to our procedure. The only place where sub-sampling is mandatory is to obtain an empirical TPDM based on $\lfloor \kappa n \rfloor$ observations.

For fixed n, selection of κ is subject to the familiar bias-variance trade-off. If κ is small, then $\lfloor \kappa n \rfloor$ may be too small for the underlying asymptotic approximation to be valid. If κ is very close to one, then the estimate $\hat{\lambda}$ will be quite sensitive to small changes in $\hat{\zeta}$. This is borne out in our simulation experiments. A formal analysis of how to select κ is beyond the scope of our investigation.

5.4 Simulation experiments

5.4.1 Symmetric logistic

We start by testing our regularised TPDM estimators in the simplest setting, where the data are generated from the symmetric logistic model. Suppose $\boldsymbol{X} = (X_1, \dots, X_d) \in \mathrm{RV}_+^d(2)$ follows a symmetric logistic model with dependence parameter $\gamma \in (0, 1]$. Then the true TPDM is of the form

$$\sigma_{ij} = \begin{cases} 1, & i = j, \\ \sigma, & i \neq j, \end{cases}$$

where $\sigma = \sigma(\gamma)$ is some constant – see Example 1.5. Recall from Figure 1.5 that the empirical TPDM overestimates σ when $\gamma \approx 1$. Since all off-diagonal entries of the TPDM are equal, the symmetric logistic model is one of the rare cases where penalising all entries equally is actually desirable. Assume for simplicity that $\hat{\boldsymbol{\sigma}}(n) = \hat{\sigma}(n)\mathbf{1}_{\binom{d}{2}}$ for some scalar

$\hat{\sigma}(n) > \sigma$. It is straightforward to show that the optimal Ledoit-Wolf shrinkage intensity is $\lambda^\star_{\mathrm{LW}(n)} = 1 - \sigma/\hat{\sigma}(n)$, since then

$$\tilde{\boldsymbol{\sigma}}(\lambda^\star_{\mathrm{LW}}(n)) = [1 - (1 - \lambda^\star_{\mathrm{LW}}(n))]\hat{\boldsymbol{\sigma}} = \frac{\sigma}{\hat{\sigma}(n)}\hat{\sigma}(n)\mathbf{1}_{\binom{d}{2}} = \boldsymbol{\sigma}.$$

Similarly, the optimal thresholds under soft-thresholding and adaptive lasso are

$$\begin{aligned}\lambda^\star_{\mathrm{S}}(n) &= \hat{\sigma}(n) - \sigma = \hat{\sigma}(n)\lambda^\star_{\mathrm{LW}},\\ \lambda^\star_{\mathrm{AL}}(n) &= [\hat{\sigma}(n)^\eta(\hat{\sigma}(n) - \sigma)]^{1/(\eta+1)} = \hat{\sigma}(n)[\lambda^\star_{\mathrm{LW}}(n)]^{1/(\eta+1)},\end{aligned}$$

respectively. Note that $\lambda^\star_{\mathrm{S}}$ and $\lambda^\star_{\mathrm{AL}}$ are expressed in terms of $\lambda^\star_{\mathrm{LW}}$. Thus, finding a good estimate for the Ledoit-Wolf shrinkage is sufficient to obtain reasonable estimates for the other regularisation parameters. On this basis, we allow ourselves to focus exclusively on testing the Ledoit-Wolf estimator.

The setup for our experiment is as follows. We generate n independent observations of $\boldsymbol{X} \in \mathrm{RV}^3_+(2)$ from a symmetric logistic model with dependence parameter γ. The intermediate sequence is set as $k(n) = \lfloor 4\sqrt{n} \rfloor$. This satisfies the rate conditions (1.43) and gives a reasonable number of threshold exceedances in practice. The ratio $\zeta(n;\kappa)$ is estimated using (5.13) with $B = 10$ and $\hat{\lambda}$ computed via (5.14) using the `optim()` function in R. Finally, we also compute the optimal shrinkage that minimises the Frobenius loss between the bootstrapped empirical TPDMs and the true TPDM; when averaged over many simulations these provide a good approximation to $\lambda^\star$. This experimental procedure is repeated for a sequence of sample sizes $5 \times 10^3 \leq n \leq 2 \times 10^5$, proportions $\kappa \in \{0.5, 0.75, 0.9\}$, and true dependence parameters $\gamma \in \{0.5, 0.8, 0.9\}$. The corresponding dependence strengths are $\sigma(0.5) = 0.85$ (strong dependence), $\sigma(0.8) = 0.48$ (moderately weak dependence) and $\sigma(0.9) = 0.27$ (weak dependence). Results are based on 100 repetitions for each combination of parameters.

The left plot in Figure 5.4 shows the average Frobenius loss $\mathcal{R}(0) = \|\boldsymbol{\sigma} - \hat{\boldsymbol{\sigma}}(n)\|_2^2$ as a function of n for each value of γ. This summarises the magnitude of the bias of the empirical TPDM; an error $\mathcal{R}(0) = \epsilon$ is equivalent to overestimating each entry of $\boldsymbol{\sigma}$ by $\epsilon/\sqrt{3}$. As expected, the magnitude of the bias is decreasing in n and increasing in γ. However, we note that even at $n = 200,000$ there is non-negligible bias present, showing that bias-

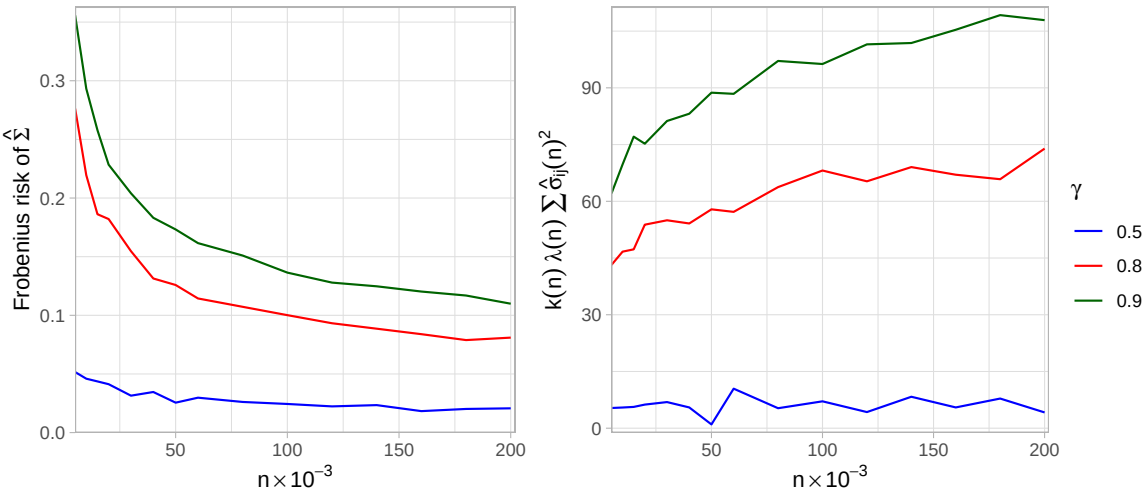

Figure 5.4: Exploratory plots for understanding the performance of the Ledoit-Wolf TPDM in the symmetric logistic experiment. Left: the Frobenius risk $\mathcal{R}(0)$ associated with the empirical TPDM $\hat{\Sigma}$ as a function of n. Right: $k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]\lambda^\star(n)$ as a function of n.

corrected estimation has a role to play even when n is large. This is an important point, considering that our procedure relies on large sample approximations. The right-hand plot shows the mean of $k(n)\mathbb{E}[\|\hat{\boldsymbol{\sigma}}(n)\|_2^2]\lambda^\star(n)$ against n. Our tuning procedure requires that this quantity is approximately constant beyond $\lfloor \kappa n \rfloor$. For $\gamma = 0.5$, the convergence is rapid, so we expect our procedure will perform well even when n is small. When $\gamma = 0.8$, the quantity approximately levels off at approximately $n = 100,000$, while for $\gamma = 0.9$ it is still increasing at $n = 200,000$.

Figure 5.5 compares the mean estimates $\hat{\lambda}(n)$ for different values of κ (dashed/dotted lines) against the oracle value $\lambda^\star(n)$ (solid line). The solid lines represent the true values of the optimal shrinkage parameter $\lambda(n)$. For $\gamma = 0.5$, the optimal shrinkage is close to zero and the estimates $\hat{\lambda}(n)$ reflect this, even when n is small. As predicted, the selection procedure performs well when $\gamma = 0.8$ provided the sample size is sufficiently large. In particular, we obtain good estimates when $n \geq 150,000$ (i.e. $\lfloor \kappa n \rfloor \geq 75,000$), which aligns reasonably closely with our earlier comments concerning the convergence in Figure 5.4. Our tuning procedure does not work particularly well in the weakest dependence scenario ($\gamma = 0.9$). While the intensities are sub-optimal, conservative estimates that err towards the empirical TPDM are probably preferable to over-shrinking. Our theoretical analysis provides assurance that these estimates would improve eventually as n increases. The plot also shows the influence of the tuning parameter κ. When κ is close to one the

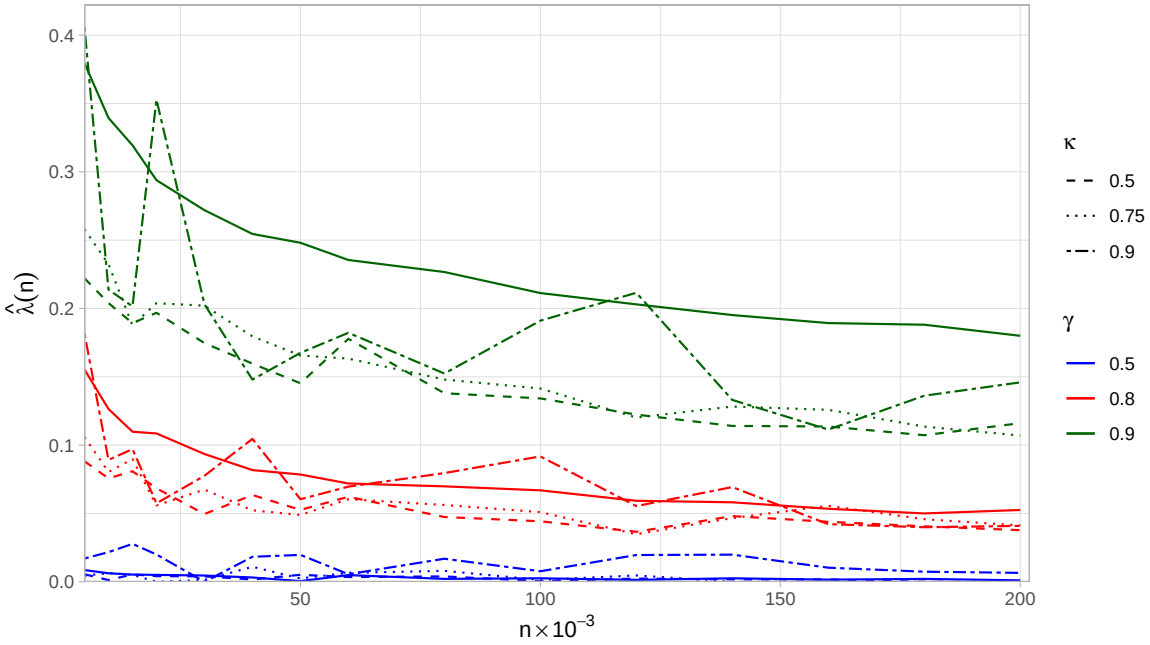

Figure 5.5: The optimal Ledoit-Wolf shrinkage intensity $\lambda^\star(n)$ (solid lines) and the mean estimates $\hat{\lambda}(n)$ obtained from our tuning procedure (dashed/dotted lines) as a function of n for various values of γ and κ.

estimates $\hat{\lambda}(n)$ have high uncertainty, while conservative choices of κ yield more stable but (negatively) biased estimates.

5.4.2 EVA (2023) Data Challenge 4

Recall from Chapter XX that our submission to Challenge 4 in the EVA (2023) Data Challenge utilised completely positive (CP) decompositions of the empirical TPDMs corresponding to five separate clusters. It transpired that our approach overestimated the true event probabilities of interest. Subsequent investigation revealed that we overestimated dependence. This was especially true in the first, fourth and fifth clusters, where dependence is weak (**rohrbeckEditorialEVA20232023**), leading us to suspect that the bias issue might be the root cause. In our post-hoc analysis (Section XX), we proposed a remedy using sparse simplex projections. The intention and effect of this was to spread the mass of the empirical angular measure towards/onto the simplex boundary, thereby artificially 'weakening' the dependence. This approach showed great promise, but required wholesale changes to our overall methodology. Armed with the Ledoit-Wolf TPDM estimator, we are now in a position to propose yet another strategy. The new approach

follows our original methodology exactly, except we replace the empirical TPDM with the Ledoit-Wolf counterpart. The substitution is seamless due to complete positivity of $\tilde{\Sigma}$ (Proposition 5.1).

Visual representations of the five clusters' empirical TPDMs are provided in Figure 5.6 (left). From Table XX in Chapter XX we know that the median values of $\{\hat{\sigma}_{ij} : i \neq j\}$ in each cluster are 0.33, 0.68, 0.67, 0.33, and 0.50. Crucially, within each cluster the range of $\{\hat{\sigma}_{ij} : i \neq j\}$ is at most 0.15, implying that the within-cluster pairwise dependence strengths are relatively equal. In such cases, the Ledoit-Wolf estimator is adequate. If this were not the case, more flexible regularisers such as adaptive lasso would be required. Let $\hat{\Sigma}^{(1)}, \ldots, \hat{\Sigma}^{(5)}$ denote the clusters' empirical TPDMs based on the k largest observations among the $n = 10^4$ samples provided. We estimate shrinkage parameters $\hat{\lambda}^{(1)}, \ldots, \hat{\lambda}^{(5)}$ for each cluster using our selection procedure with $k(n) = \lfloor k\sqrt{n}/100 \rfloor$, $\kappa = 0.75$ and $B = 10$. The remaining steps follow Section XX: we apply the CP-factorisation algorithm of **kiriliouk EstimatingProbabilitiesMultivariate2022** to each Ledoit-Wolf TPDM and compute probability estimates using the formula XX.

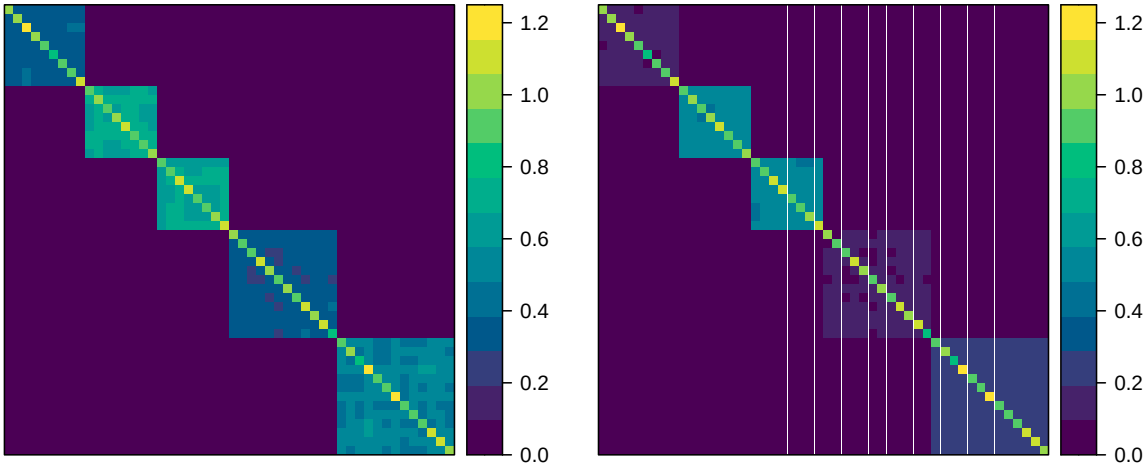

Figure 5.6: The empirical TPDM $\hat{\Sigma} = \text{diag}(\hat{\Sigma}^{(1)}, \ldots, \hat{\Sigma}^{(5)})$ (left) and the Ledoit-Wolf TPDM $\tilde{\Sigma} = \text{diag}(\tilde{\Sigma}^{(1)}(\hat{\lambda}_1), \ldots, \tilde{\Sigma}^{(5)}(\hat{\lambda}_5))$ for the EVA (2023) Data Challenge based on $k = 250$.

Figure 5.7 reveals the amount of each shrinkage for each cluster for $0.02 \leq k/n \leq 0.07$. As expected, clusters 1 and 4 are penalised the most and clusters 2 and 3 the least. The shrinkage intensities are sensitive to the choice of k. Counterintuitively the parameters $\hat{\lambda}^{(1)}, \ldots, \hat{\lambda}^{(5)}$ seem to increase as k decreases, despite the fact that the bias in the TPDM is usually smaller at higher thresholds (**fixSimultaneousAutoregressiveModels2021**).

This warrants further investigation in future. For the present study, we focus on the case where $k = 250$, since this is the number of threshold exceedances used in our competition entry. The corresponding Ledoit-Wolf TPDM estimate is shown in Figure 5.6 (right). Comparing against the empirical TPDM on the left, there is a dramatic reduction in the entries in clusters 1 and 4. Consequently, this matrix is presumably closer to the (unknown) true TPDM than the raw estimate. The final probability estimates in Figure 5.8 support this conclusion. In each panel, the horizontal dashed line shows the true value of the probabilities p_1 (left) and p_2 (right). The grey line corresponds to our original method based on CP-decompositions of the empirical TPDM. It drastically overestimates the event probabilities. The green and blue lines result from fitting a max-linear model using the k largest angles defined via self-normalisation and Euclidean projections, respectively; these methods were described in detail in Section XX. The red line is derived from our new procedure based on CP-decompositions of Ledoit-Wolf TPDMs. At $k = 250$, it gives a near-perfect estimate of p_1 and a very good estimate of p_2. When k is large the probability estimates are too large because the shrinkage intensities are too small, especially in the first, fourth and fifth clusters (see Figure 5.7).

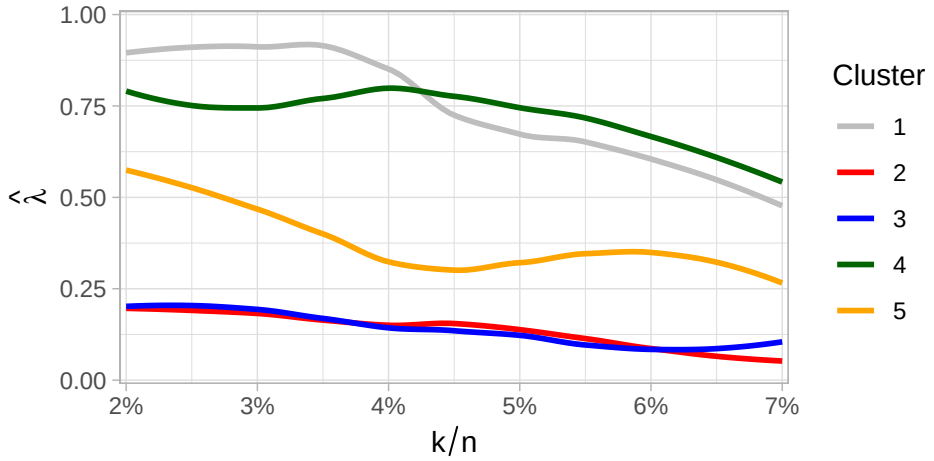

Figure 5.7: Estimated shrinkage intensities $\hat{\lambda}^{(1)}, \ldots, \hat{\lambda}^{(5)}$ associated with the clusters' empirical TPDMs $\hat{\Sigma}^{(1)}, \ldots, \hat{\Sigma}^{(5)}$ for the EVA (2023) Data Challenge.

5.4.3 Extremal SAR model

Our final experiment emulates the simulation study in **fixSimultaneousAutoregressiveModels2021**. The goal is to estimate the dependence parameter ρ of the extremal SAR model (Section

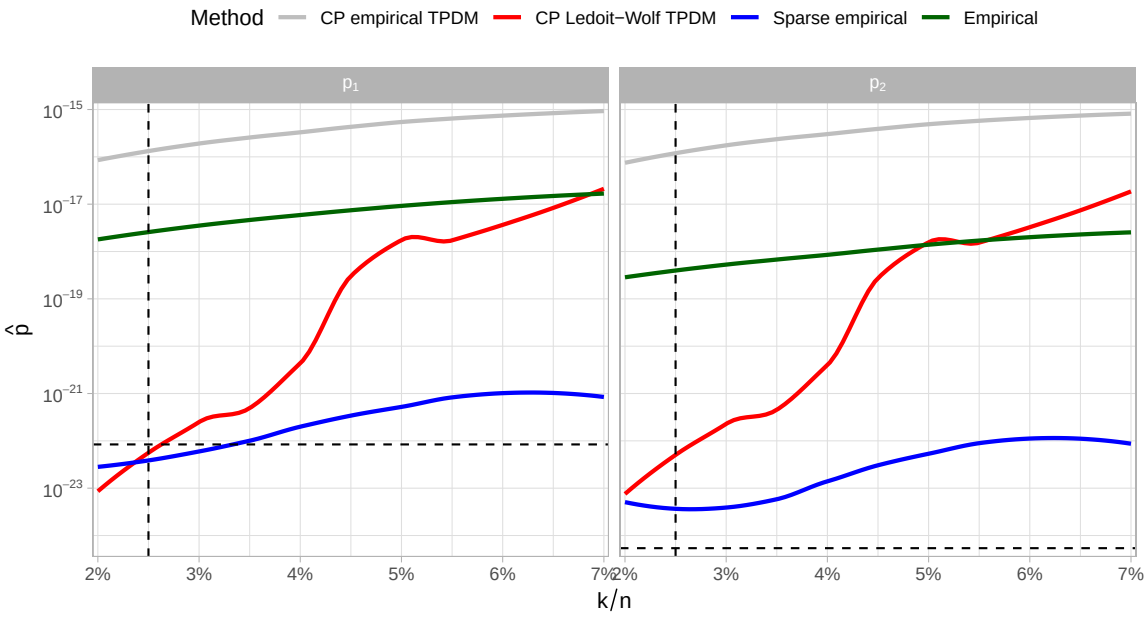

Figure 5.8: Probability estimates for the EVA (2023) Data Challenge from all four proposed methods. The true probabilities p_1 (left) and p_2 (right) are indicated by the horizontal dashed lines. The methodologies associated with the grey, blue and green lines are described in Chapter XX. The red line shows the results based on CP decompositions of the Ledoit-Wolf estimates. The vertical dashed line is at $k = 250$, the number of threshold exceedances used in our competition entry.

XX), a sub-case of the max-linear model used for spatial extremes. This experiment will reinforce the effectiveness of our tuning procedure, while also highlighting the limitations of Ledoit-Wolf shrinkage compared to the more flexible thresholding estimators. Suppose $\boldsymbol{X}$ is as described in Section XX, with (known) neighbourhood matrix W and (unknown) dependence parameter ρ. Following **fixSimultaneousAutoregressiveModels2021**, ρ is to be estimated by minimising the Frobenius difference between the model TPDM (viewed as a function of ρ) and a bias-corrected estimate of the TPDM. This can be posed as a hierarchical/bilevel optimisation problem, whereby the upper-level objective function (estimating ρ) depends on the solution to a lower-level problem (estimating Σ). Specifically, we define

$$\hat{\rho} := \arg\min_{\varrho} \hat{\mathcal{L}}(\varrho), \qquad \hat{\mathcal{L}}(\varrho) := \|\boldsymbol{\sigma}(\varrho) - \tilde{\boldsymbol{\sigma}}(\hat{\lambda})\|_2^2. \tag{5.15}$$

The loss function $\mathcal{L}(\varrho)$ depends on the estimate $\tilde{\boldsymbol{\sigma}}(\hat{\lambda})$ of $\boldsymbol{\sigma}$, so accurate estimation of the TPDM is critical. If $\tilde{\boldsymbol{\sigma}}(\hat{\lambda})$ is not representative of $\boldsymbol{\sigma}$, then the loss may not be informative for

ρ. In **fixSimultaneousAutoregressiveModels2021**, $\tilde{\boldsymbol{\sigma}}$ represents the soft-thresholded estimator and $\hat{\lambda}$ is obtained by fitting a non-linear model relating the pairwise dependence strengths and inter-site distances. The soft-thresholded estimator achieves a reasonable fit – see Figure 3 (right) in **fixSimultaneousAutoregressiveModels2021**.

Our simulation setup follows Section 4.2.1 in **fixSimultaneousAutoregressiveModels2021** with some minor adjustments. We consider $d = 36$ sites arranged in an 6×6 grid and generate n independent realisations of a 6×6 spatial field from an extremal SAR model with dependence parameter $\rho = 0.15$. With this parameter choice, dependence vanishes at a distance of approximately 4 units. The empirical TPDM $\hat{\Sigma}$ is computed using the $k(n) = \lfloor 4\sqrt{n} \rfloor$ largest observations. Next, we use the true TPDM (Example 1.7) to compute $\lambda^\star$ for the soft-threshold, Ledoit-Wolf, and adaptive lasso ($\eta = 2$) TPDMs. From each oracle TPDM $\tilde{\boldsymbol{\sigma}}(\lambda^\star)$, we compute an estimate of ρ by solving (5.15). These values represent the estimates of ρ that would be found by an oracle who is constrained to a particular type of regularisation. Finally, we repeat the same steps using regularisation parameters selected using our data-driven procedure (for Ledoit-Wolf with $\kappa = 0.75$ and $B = 5$) or the spatial method of **fixSimultaneousAutoregressiveModels2021** (for soft-threshold). Currently there are no available methods for tuning λ and/or η for the adaptive lasso. The procedure is repeated 100 times at a sequence of sample sizes varying between $10^4 \leq n \leq 2 \times 10^5$.

The left-hand plot in Figure 5.9 shows how the Frobenius loss (on a log scale) varies with n for each oracle/estimated TPDM. The colours correspond to the regularisation method used; the line type indicates whether the regularisation parameter is estimated or an oracle based on knowledge of the true TPDM. By definition, the risk attained by the oracle losses provides a lower bound on the risk that can be attained for a given estimator class. If the solid line lies close to the oracle loss, then λ has been well selected. In this respect, the blue lines show that the soft-threshold tuning procedure of **fixSimultaneousAutoregressiveModels2021** performs very well, with the usual caveat that it makes use of extra information to do so. Moreover, the soft-thresholded estimates are very close to the true TPDM in Frobenius norm. For Ledoit-Wolf (red) our data-driven procedure gives reasonably good results provided n is large. The Ledoit-Wolf estimates are worse than the soft-threshold or adaptive lasso, but are still far su-

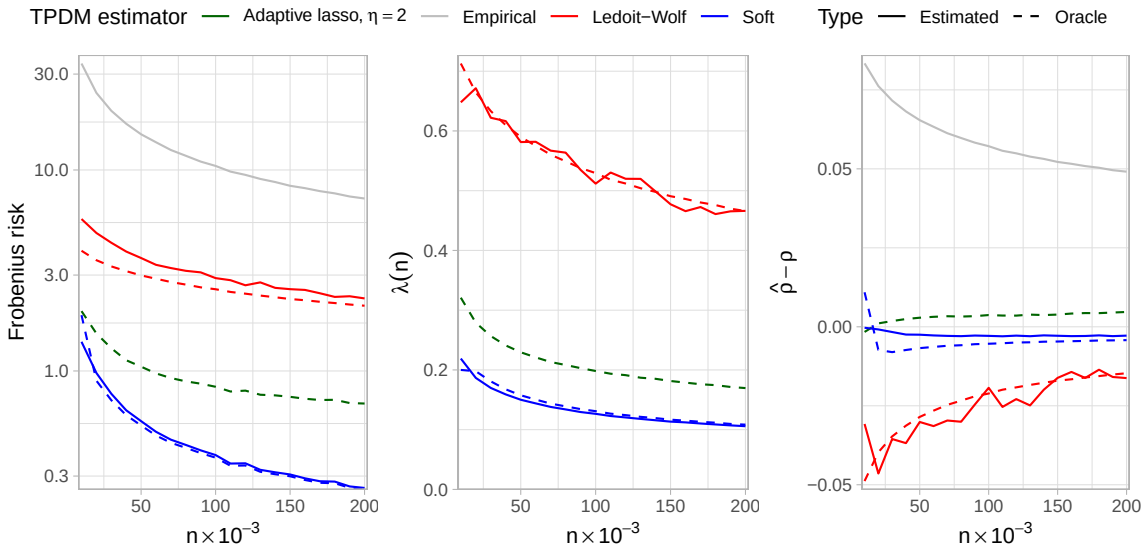

Figure 5.9: Results for our experiments based on the extremal SAR model. Left: the mean Frobenius error of the oracle and estimated TPDMs. Middle: the values of the regularisation parameter selected by the oracle or tuning procedure. Right: the empirical bias of the estimates of ρ computed by solving (5.15).

perior to the empirical estimate. The middle plot provides more insight into the selected values of λ. The soft-threshold parameter is very close to the oracle value. Our shrinkage parameters also track the oracle quite closely, even for small n, though these small errors seem to translate into non-negligible errors in the overall Frobenius norm. The right-hand plot examines the empirical bias $\hat{\rho} - \rho$ in the final estimates of the spatial dependence parameter. The empirical TPDM (grey) overestimates dependence and wildly overestimates ρ. The soft-threshold estimate is interesting, with the estimator of **fixSimultaneousAutoregressiveModels2021** actually outperforming the oracle. This is not a contradiction: a smaller Frobenius loss does not guarantee a better estimate of ρ. **fixSimultaneousAutoregressiveModels2021** found that their estimator exhibited a slight negative bias for ρ. Our results confirm this and provide additional insight into why this occurs. Since the oracle suffers the same issue, we can infer that the bias is due to misspecification (i.e. being restricted to the soft-thresholded TPDMs) rather than a deficiency of their tuning procedure. As predicted, the Ledoit-Wolf estimator performs comparatively poorly and generally underestimates ρ, especially when n is small. However, we reiterate that this is a shortcoming of using a restrictive shrinkage method and not due to our our tuning procedure, since the oracle estimates (dashed line) are no better than our estimated (solid line). Adaptive lasso achieves comparable results to soft-thresholding,

except the sign of the bias is reversed. There is scope for further improvement by varying the down-weighting hyperparameter η.

5.5 Conclusions and outlook

The main contributions of this chapter are to counteract the bias issue of the empirical TPDM by proposing two classes of bias-corrected estimators. The first class, based on thresholding, generalises the estimator proposed in **fixSimultaneousAutoregressiveModels2021** and includes more flexible regularisers, such as adaptive lasso. The second class employs linear shrinkage to reduce the bias. Admittedly, the restrictive form of this estimator limits its applicability. However, a key advantage is that it possesses nice mathematical properties that enable it to be used immediately in existing TPDM-based modelling frameworks. Moreover, we construct a data-driven statistical procedure for tuning these estimators, the first of its kind. The procedure is underpinned by asymptotic approximations, but our simulation experiments demonstrate reasonable performance with finite sample sizes.

Move (some of?) the comments below to final chapter?

There is ample opportunity to develop these methods further. In terms of our approach, more investigation into how to choose the proportion κ and the number of bootstrap samples B is needed. Ways to improve the flexibility of the Ledoit-Wolf estimator would be a welcome enhancement. As an initial suggestion, is it possible to apply different levels of shrinkage to groups of similarly sized entries in $\hat{\Sigma}$? Establishing a general solution to the problem of data-driven tuning for regularised estimators would be a significant breakthrough. This would permit the use of the soft-thresholded TPDM in non-spatial applications, for example. It would be interesting to meditate further on the results in Section XX, pertaining to the EVA (2023) Data Challenge. There we provided two very different approaches to mitigate the problem of 'overestimating weak dependence' that yield almost identical results. The first approach utilises sparse simplex projections to directly modify the data (pseudo angles); the second approach treats bias-correction as a post-processing step. Can these two disparate approaches be connected or unified? Finally, one might take a broader perspective and consider alternative loss functions instead of the

Frobenius risk. For example, if one is interested in applying the TPDM for PCA purposes, a more appropriate criterion might be some measure of agreement between the principal eigenspaces of Σ and $\tilde{\Sigma}(\lambda)$, such as the K_p coefficient (CITE Krzanowski (1979))

$$K_p(\lambda) := \sum_{i=1}^{p}\sum_{j=1}^{p} \langle \tilde{\boldsymbol{u}}_i(\lambda), \boldsymbol{u}_j \rangle^2 , \qquad (1 \leq p \leq d),$$

where $\tilde{\boldsymbol{u}}_i(\lambda)$ denotes the ith eigenvectors of $\tilde{\Sigma}(\lambda)$. Alternatively, some applications (e.g. clustering) might call for identification of the sparsity structure of Σ, i.e. the positions of its zero/non-zero entries. For this, one might adopt the performance criteria

$$\mathrm{TPR}(\lambda) := \frac{|\{(i,j) : i < j,\, \tilde{\sigma}_{ij}(\lambda) \neq 0,\, \sigma_{ij} \neq 0\}|}{|\{(i,j) : i < j,\, \sigma_{ij} \neq 0\}|}, \tag{5.16}$$

$$\mathrm{FPR}(\lambda) := \frac{|\{(i,j) : i < j,\, \tilde{\sigma}_{ij}(\lambda) \neq 0,\, \sigma_{ij} = 0\}|}{|\{(i,j) : i < j,\, \sigma_{ij} = 0\}|}, \tag{5.17}$$

Note that this only makes sense for threshold-based estimates, since Ledoit-Wolf shrinkage does not produce zeroes unless $\lambda = 1$. In such settings, the threshold λ would typically be selected by inspecting a receiver operating characteristic (ROC) curve and/or maximising the area under the curve (AUC). The obvious impediment to this is that the true TPDM is unknown, so the TPR and FPR cannot be evaluated. To circumvent this, we suggest replacing Σ in (5.16) and (5.17) with a modified version of the empirical TPDM based on Euclidean simplex projections. *I'll come back to this when I write up the future work section in Chapter 7, I think I'll just write down all my thoughts about the sparse TPDM there.* By conducting a series of appropriate experiments, one can ascertain with this pseudo-ROC curve is close to the true ROC curve, thereby providing a method for selecting the optimum threshold.

References

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 **fixSimultaneousAutoregressiveModels2021**. 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 (1.30) with $\alpha = 1$

and $H_1(\mathrm{d}\boldsymbol{\theta}) = h_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$,

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

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

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

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

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

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

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

□

In the trivial case $\alpha = 1$ the formula reduces to $\tilde{h}_1(\theta) = \tilde{h}_1(\theta)$, as one would hope. Setting $\alpha = 2$ yields $\tilde{h}_2(\theta) = 2\theta\tilde{h}_1(\theta^2)$, which matches the formula gives in **fixSimultaneousAutoregressiveModels2021**. 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 1.13 is $H_\alpha = m^{-1}H$, where m is the mass of H. Therefore, (1.52) 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 1.6

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 1.5. 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

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

and therefore

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

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

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

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 $\mathbb{S}^{d-1}_{+(1)}$. However, in each case, the Hilbert space in question will be isometric to the usual Euclidean space $(\mathbb{R}^d, \langle \cdot, \cdot \rangle)$. That is, there exists an isomorphism $h : \mathcal{H} \to \mathbb{R}^d$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$,

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

We present PCA for random vectors in $\mathbb{R}^d$, with the understanding that the data may have undergone an isometric transformation in pre-processing and outputs may need to be back-transformed to lie in the original space. This transform/back-transform approach is equivalent to conducting the analysis in the original space with appropriately generalised notions of mean, variance, etc. (**pawlowsky-glahnGeometricApproachStatistical2001**).

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 $\mathcal{S} \subseteq \mathbb{R}^d$

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

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 $\mathbb{E}[V_i V_j] = \lambda_i \mathbf{1}\{i = j\}$. For $1 \leq p < d$, the truncated expansion

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

produces the optimal p-dimensional projection of $\boldsymbol{Y}$. In other words, the subspace $\mathcal{S}_p = \mathrm{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p\}$ minimises the criterion (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 $j = 1, \ldots, d$. For $p = 1, \ldots, d$ and $i = 1, \ldots, n$, the rank-p reconstruction of $\boldsymbol{Y}_i$ is given by

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

where $\hat{V}_{ij} := \langle \boldsymbol{Y}_i, \boldsymbol{u}_j \rangle$. The subspace $\hat{\mathcal{S}}_p = \text{span}\{\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_p\}$ minimises (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 **cadimaRelationshipsUncentredColumnCentred2009**. 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 **cooleyDecompositionsDependenceHighdimensional2019** has been applied for exploratory purposes in the context of climatology (**jiangPrincipalComponentAnalysi** **szemkusSpatialPatternsIndices2024**), finance (**cooleyDecompositionsDependenceHighdimensi** and sport (**russellAnalyzingDependenceMatrices2018**).

jiangPrincipalComponentAnalysis2020 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 (1.56) 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.

szemkusSpatialPatternsIndices2024 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 **russellAnalyzingDependenceMatrices2018**. 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. **russellAnalyzingDependenceMatrices2018** 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.

rohrbeckSimulatingFloodEvent2023 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. **rohrbeckSimulatingFloodEvent2023** show how to generate approximate samples from H, even in high-dimensions, by leveraging the PCA method of **cooleyDecompositionsDependenceHighdimensional2019**. 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 $(V_1, \dots, V_p)$ for some appropriately chosen $p < d$. To achieve this, they use a spherical kernel density estimate to flexibly model the dependence between $V_1, \dots, V_p$ and additionally between $(V_1, \dots, V_p)$ and $(V_{p+1}, \dots, V_d)$. The dependence structure of $(V_{p+1}, \dots, V_d)$ 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 $\tilde{\boldsymbol{x}}_1^{[p]}, \dots, \tilde{\boldsymbol{x}}_N^{[p]}$ for a range of values p, and then assessing whether any of the generated samples resemble the discarded event using

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

where $\varrho(\cdot,\cdot)$ is an angular dissimilarity measure. After repeating for all extreme events $i = 1 \dots, k$, 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 **chautruDimensionReductionMultivariate2015**, **janssenKmeansClusteringExtremes2020** and **medinaSpectralLearningMultivariate2021** 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. **goixSparseRepresentationMultivariate2017** and

simpsonDeterminingDependenceStructure2020 identify clusters according to whether observations fall within appropriately sized rectangular/conic neighbourhoods of the corresponding axis in $\mathbb{R}^d_+$. **meyerDetectionExtremalDirections2020** take a different approach, whereby the angular component is defined with respect to the Euclidean projection (**liuEfficientEuclideanProjections2009**) 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. **bernardClusteringMaximaSpatial2013** propose grouping components using the k-medoids algorithm (**kaufmanFindingGroupsData1990**) with a dissimilarity matrix populated with pairwise measures of tail dependence, similar to χ_{ij} and σ_{ij}. The approaches of **fomichovSphericalClusteringDetection2023** and **richardsModernExtremeValue2024** involve the TPDM; these are reviewed in greater detail below.

fomichovSphericalClusteringDetection2023 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 $K = 1$:

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

Note that $\boldsymbol{u}_1 \in \mathbb{S}^{d-1}_{+(2)}$ is assumed to be suitably normalised with all entries being non-negative; the Perron-Frobenius theorem guarantees this is possible. This result informs an iterative clustering procedure called spherical k-principal-components. Consider a set of extremal angles $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(k)} \in \mathbb{S}^{d-1}_{+(2)}$ and current centroids $\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_K \in \mathbb{S}^{d-1}_{+(2)}$. A single iteration of their procedure yields new centroids $\hat{\boldsymbol{w}}^\star_1, \ldots, \hat{\boldsymbol{w}}^\star_K \in \mathbb{S}^{d-1}_{+(2)}$ given by the respective principal eigenvectors of

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

The matrix $\hat{\Sigma}^{[i]}$ represents the empirical TPDM (up to some multiplicative constant) based on the nearest neighbours of the ith centroid. **fomichovSphericalClusteringDetection2023** prove that, under certain conditions, the limiting centroids lie in a neighbourhood of the faces of interest $C_{\beta_1}, \ldots, C_{\beta_K}$. Thresholding the centroid vectors yields the final partition $\beta_1, \ldots, \beta_K$.

richardsModernExtremeValue2024 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 **richardsModernExtremeValue2024** and the accompanying code made available at https://github.com/matheusguerrero/yalla. Define the dissimilarity between X_i and X_j as $\varrho_{ij} = 1 - \sigma_{ij}$. This satisfies the properties of a dissimilarity measure (CITE: A MATHEMATICAL THEORY FOR CLUSTERING IN METRIC SPACES):

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

The $d \times d$ dissimilarity matrix $\mathcal{D} = 1 - \Sigma = (\varrho_{ij})$ can be fed into standard hierarchical clustering algorithms. Agglomerative hierarchical clustering initially assigns each variable belongs to its own cluster, i.e. $\beta_i = \{i\}$ for $i = 1, \ldots, d$. The algorithm proceeds iteratively,

repeatedly joining together the two closest clusters until some stopping criterion is satisfied. Under complete-linkage clustering, the distance between clusters $\beta \neq \beta'$ is given by $\max\{\varrho_{ij} : i \in \beta, j \in \beta'\}$. The merging process may be stopped when there is a sufficiently small number of clusters or when the clusters are sufficiently separated.

E Summary of dreesStatisticalInferenceChanging2023

dreesStatisticalInferenceChanging2023 tests (2.2) against (2.3) 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

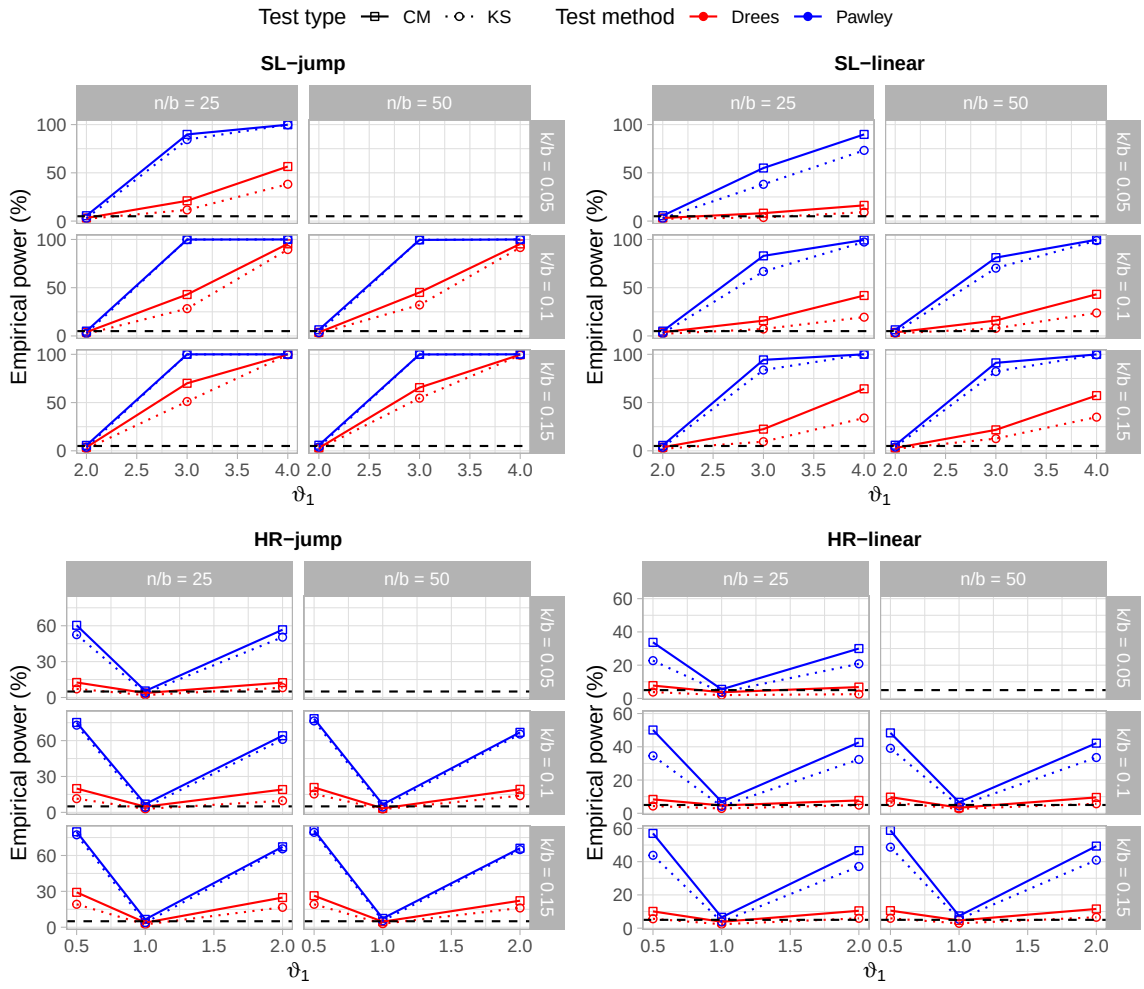

Figure F.1: Empirical power (%) as a function of the dependence parameter ϑ_1 for different combinations of tuning parameters b and k. Based on 1000 simulations with $n = 2,500$ and $d = 2$. For the SL and HR models, $\vartheta_0 = 2$ and $\vartheta_0 = 1$, respectively. Tests are conducted at the 5%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

G Max-linear example with other combinations of n and k

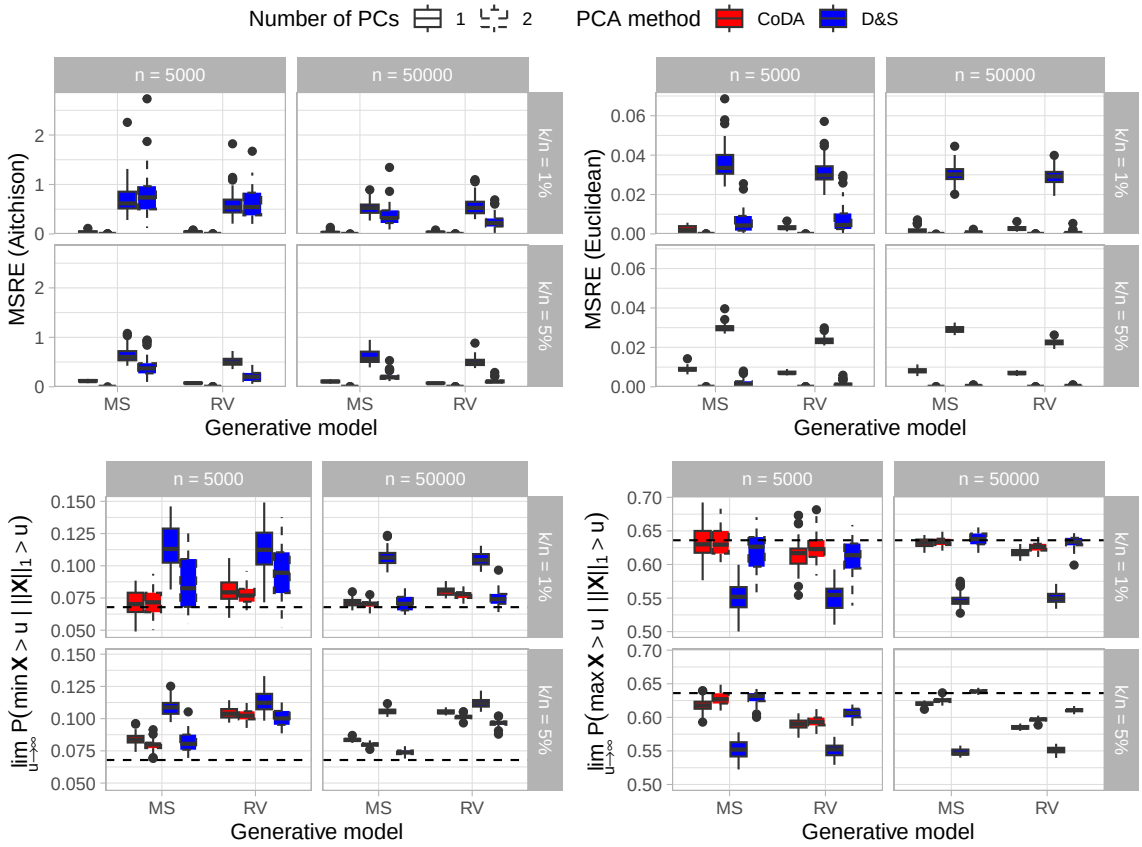

Figure G.1: PCA performance metrics based on trivariate max-linear data.

G.0.1 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.

G.0.2 Training risk and test risk for classification

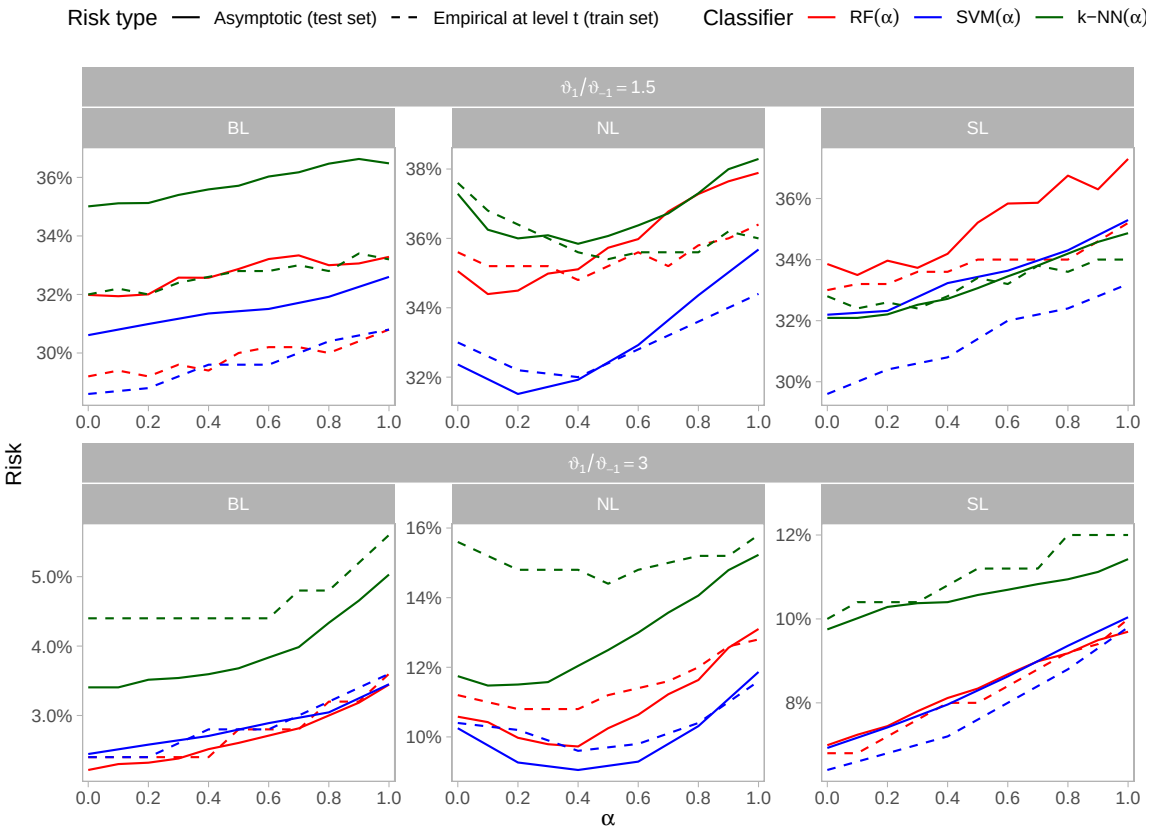

Figure G.2: Blah.