



**HAL**  
open science

# Variable clustering of multivariate time series according to the dependence of their extremes

Alexis Boulin

► **To cite this version:**

Alexis Boulin. Variable clustering of multivariate time series according to the dependence of their extremes. Statistics [math.ST]. Université Côte d'Azur, 2024. English. NNT: 2024COAZ5039 . tel-04767333

**HAL Id: tel-04767333**

**<https://theses.hal.science/tel-04767333v1>**

Submitted on 5 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# THÈSE DE DOCTORAT

Partitionnement des variables de séries temporelles  
multivariées selon la dépendance de leurs extrêmes

**Alexis Boulin**

Laboratoire J.A. Dieudonné

**Présentée en vue de l'obtention  
du grade de docteur en  
mathématiques**

d'Université Côte d'Azur

**Dirigée par** : Thomas Laloë / Gwladys  
Toulemonde

**Co-encadrée par** : Elena Di  
Bernardino

**Soutenue le** : 30/09/2024

**Devant le jury, composé de :**

Stéphan Cléménçon, Professeur,  
Télécom Paristech

Elena Di Bernardino, Professeur,  
Université Côte d'Azur

Carlo Gaetan, Professeur, Università  
Ca' Foscari - Venezia

Armelle Guillou, Professeur, Université  
de Strasbourg

Philippe Naveau, Directeur de  
Recherche, LSCE

Johan Segers, Professeur, Université  
Catholique de Louvain

Vincent Vandewalle, Professeur,  
Université Côte d'Azur



Je dédie cette thèse à mon papa.



# Partitionnement des variables de séries temporelles multivariées selon la dépendance de leurs extrêmes

## **Jury:**

### **Rapporteurs :**

Stéphan Cléménçon, Professeur, Télécom Paris  
Johan Segers, Professeur, Université Catholique de Louvain

### **Examineurs :**

Carlo Gaetan, Professeur, Università Ca' Foscari - Venezia  
Armelle Guillou, Professeur, Université de Strasbourg  
Philippe Naveau, Directeur de Recherche, Laboratoire des Sciences du Climat et de l'Environnement  
Vincent Vandewalle, Professeur, Université Côte d'Azur

### **Direction :**

Elena Di Bernardino, Professeure, Université Côte d'Azur  
Thomas Laloë, Maître de Conférences, Université Côté d'Azur (Invité)  
Gwladys Toulemonde, Professeure, Université de Montpellier



# Résumé et mots clés

Dans un grand éventail d'applications allant des sciences du climat à la finance, des événements extrêmes avec une probabilité loin d'être négligeable peuvent se produire, entraînant des conséquences désastreuses. Les extrêmes d'événements climatiques tels que le vent, la température et les précipitations peuvent profondément affecter les êtres humains et les écosystèmes, entraînant des événements tels que des inondations, des glissements de terrain ou des vagues de chaleur. Lorsque l'emphase est mise sur l'étude de variables mesurées dans le temps sur un grand nombre de stations ayant une localisation spécifique, comme les variables mentionnées précédemment, le partitionnement de variables devient essentiel pour résumer et visualiser des tendances spatiales, ce qui est crucial dans l'étude des événements extrêmes. Cette thèse explore plusieurs modèles et méthodes pour partitionner les variables d'un processus stationnaire multivarié, en se concentrant sur les dépendances extrémales.

[Le chapitre 1](#) présente les concepts de modélisation de la dépendance via les copules, fondamentales pour la dépendance extrême. La notion de variation régulière est introduite, essentielle pour l'étude des extrêmes, et les processus faiblement dépendants sont abordés. Le partitionnement est discuté à travers les paradigmes de séparation-proximité et de partitionnement basé sur un modèle. Nous abordons aussi l'analyse non-asymptotique pour évaluer nos méthodes dans des dimensions fixes.

[Le chapitre 2](#) est à propos de la dépendance entre valeurs maximales est cruciale pour l'analyse des risques. Utilisant la fonction de copule de valeur extrême et le madogramme, ce chapitre se concentre sur l'estimation non paramétrique avec des données manquantes. Un théorème central limite fonctionnel est établi, démontrant la convergence du madogramme vers un processus Gaussien tendu. Des formules pour la variance asymptotique sont présentées, illustrées par une étude numérique.

[Le chapitre 3](#) propose les modèles asymptotiquement indépendants par blocs (AI-blocs) pour le partitionnement de variables, définissant des clusters basés sur l'indépendance des maxima. Un algorithme est introduit pour récupérer les clusters sans spécifier leur nombre à l'avance. L'efficacité théorique de l'algorithme est démontrée, et une méthode de sélection de paramètre basée sur les données est proposée. La méthode est appliquée à des données de neurosciences et environnementales, démontrant son potentiel.

[Le chapitre 4](#) adapte des techniques de partitionnement pour analyser des événements extrêmes composites sur des données climatiques européennes. Les sous-régions présentant une



---

dépendance des extrêmes de précipitations et de vitesse du vent sont identifiées en utilisant des données ERA5 de 1979 à 2022. Les clusters obtenus sont spatialement concentrés, offrant une compréhension approfondie de la distribution régionale des extrêmes. Les méthodes proposées réduisent efficacement la taille des données tout en extrayant des informations cruciales sur les événements extrêmes.

**Le chapitre 5** propose une nouvelle méthode d'estimation pour les matrices dans un modèle linéaire à facteurs latents, où chaque composante d'un vecteur aléatoire est exprimée par une équation linéaire avec des facteurs et du bruit. Contrairement aux approches classiques basées sur la normalité conjointe, nous supposons que les facteurs sont distribués selon des distributions de Fréchet standards, ce qui permet une meilleure description de la dépendance extrême. Une méthode d'estimation est proposée garantissant une solution unique sous certaines conditions. Une borne supérieure adaptative pour l'estimateur est fournie, adaptable à la dimension et au nombre de facteurs.

**Mots clés :** Analyse non-asymptotique, Analyse probabiliste d'algorithmes, Modélisation de la dépendance extrême, Partitionnement de variables, Processus faiblement dépendant, Processus stationnaire multivarié, Théorie des valeurs extrêmes, Variation régulière

# Abstract and keywords

In a wide range of applications, from climate science to finance, extreme events with a non-negligible probability can occur, leading to disastrous consequences. Extremes in climatic events such as wind, temperature, and precipitation can profoundly impact humans and ecosystems, resulting in events like floods, landslides, or heatwaves. When the focus is on studying variables measured over time at numerous specific locations, such as the previously mentioned variables, partitioning these variables becomes essential to summarize and visualize spatial trends, which is crucial in the study of extreme events. This thesis explores several models and methods for partitioning the variables of a multivariate stationary process, focusing on extreme dependencies.

**Chapter 1** introduces the concepts of modeling dependence through copulas, which are fundamental for extreme dependence. The notion of regular variation, essential for studying extremes, is introduced, and weakly dependent processes are discussed. Partitioning is examined through the paradigms of separation-proximity and model-based clustering. Non-asymptotic analysis is also addressed to evaluate our methods in fixed dimensions.

**Chapter 2** study the dependence between maximum values is crucial for risk analysis. Using the extreme value copula function and the madogram, this chapter focuses on non-parametric estimation with missing data. A functional central limit theorem is established, demonstrating the convergence of the madogram to a tight Gaussian process. Formulas for asymptotic variance are presented, illustrated by a numerical study.

**Chapter 3** proposes asymptotically independent block (AI-block) models for partitioning variables, defining clusters based on the independence of maxima. An algorithm is introduced to recover clusters without specifying their number in advance. Theoretical efficiency of the algorithm is demonstrated, and a data-driven parameter selection method is proposed. The method is applied to neuroscience and environmental data, showcasing its potential.

**Chapter 4** adapts partitioning techniques to analyze composite extreme events in European climate data. Sub-regions with dependencies in extreme precipitation and wind speed are identified using ERA5 data from 1979 to 2022. The obtained clusters are spatially concentrated, offering a deep understanding of the regional distribution of extremes. The proposed methods efficiently reduce data size while extracting critical information on extreme events.

**Chapter 5** proposes a new estimation method for matrices in a latent factor linear model, where each component of a random vector is expressed by a linear equation with factors and

---

noise. Unlike classical approaches based on joint normality, we assume factors are distributed according to standard Fréchet distributions, allowing a better description of extreme dependence. An estimation method is proposed, ensuring a unique solution under certain conditions. An adaptive upper bound for the estimator is provided, adaptable to dimension and the number of factors.

**Keywords:** Extremal dependence modeling, Extreme value theory, Multivariate stationary process, Non-asymptotic analysis, Regular variation, Variables clustering, Weakly dependent process

## Remerciements

Mes premiers remerciements vont tout d'abord à mes encadrants de thèse, Elena Di Bernardino, Thomas Laloë et Gwladys Toulemonde. Durant ces trois années, vous m'avez offert une grande liberté pour explorer les mathématiques qui m'intéressaient, tout en me suggérant des pistes particulièrement pertinentes auxquelles je n'avais pas pensé. Je repense avec une certaine nostalgie au moment où Elena m'a proposé d'étudier théoriquement mon tout premier algorithme dans le cadre des processus de mélange. Sans cette demande spécifique, je n'aurais probablement pas eu le courage de m'y atteler, et je ne regrette pas de m'y être consacré. Comment pourrais-je oublier les efforts considérables de Thomas et Gwladys, qui m'ont encouragé à appliquer mes méthodes à des problématiques concrètes? Leur soutien a été fondamental pour me motiver dans cette démarche et apporter une dimension cruciale à cette thèse. Je tiens également à remercier à nouveau Gwladys pour son accueil chaleureux à Montpellier et pour les nombreuses discussions au tableau à refaire les mathématiques!

Je voudrais aussi remercier grandement Stéphan Cléménçon et Johan Segers, qui ont accepté de rapporter cette thèse et m'ont donné des retours très positifs sur ce manuscrit. Cela m'honore et m'engage pour la suite. Je souhaite aussi exprimer ma gratitude envers Carlo Gaetan, Armelle Guillou, Philippe Naveau et Vincent Vandevale qui ont gentiment accepté d'être dans mon jury de thèse.

Je souhaite tout particulièrement remercier l'ensemble des doctorants et post-doctorants qui m'ont accompagné tout au long de ma thèse. Merci aux soirées repas, aux soirées jeux de société, aux soirées cinéma, les retraites des doctorants à Fréjus et dans les Cévennes, les journées des jeunes statisticiens à Porquerolles, les journées des statistiques. Ensemble, nous avons écrit un chapitre qui a été accepté pour publication dans l'ouvrage collectif de ma vie. Merci tout particulièrement à Alba, Alex, Antoine, Bruno, Cambyse, Chloé, Domenico, Enrico, Gustave, Jérémie, Lamine, Lorenzo, Mariem, Matthieu, Meriem, Ryan, Sebastian, Sophie, Thomas, Titouan Tommaso, Victor, Zakaria. Merci pour les rires et les repas qu'on a partagé. Je tiens également à exprimer mes remerciements envers la MED et la CAF, gardiennes des enfers, dont ADOOM est la porte d'entrée. Je tiens également à remercier tous mes amis qui, au détour d'une conversation discord ou d'une partie de jeu vidéo, ont su me permettre de relâcher la pression.

Ces dernières lignes sont consacrées à ma famille qui me soutiennent inconditionnellement. Merci à mes parents, mon grand frère et ma petite soeur.



# Table of contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	A survival guide for high-dimensional extremal dependence modeling . . . . .	1
1.1.1	Dependence modeling . . . . .	1
1.1.2	Extremal dependence modeling . . . . .	8
1.1.3	Multivariate regularly varying random vectors . . . . .	15
1.1.4	Weakly dependent random processes . . . . .	21
1.1.5	Clustering . . . . .	28
1.1.6	Mathematics of high dimension . . . . .	31
1.2	Outline and contributions . . . . .	42
<b>2</b>	<b>Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework</b>	<b>49</b>
2.1	Introduction . . . . .	49
2.2	Non parametric estimation of the Madogram with missing data . . . . .	52
2.3	Numerical results . . . . .	58
2.3.1	Presentation of the models . . . . .	58
2.3.2	Description of numerical experiments . . . . .	60
2.3.3	Results of experiments . . . . .	61
2.4	Extremal dependence rainfall analysis via hybrid madogram . . . . .	64
2.5	Conclusions . . . . .	66
	<b>Appendix A Proofs of Chapter 2</b>	<b>69</b>
A.1	Proofs . . . . .	69
A.1.1	Proofs of main results . . . . .	69
A.1.2	Proofs of auxiliary results . . . . .	78
<b>3</b>	<b>High-dimensional variable clustering based on maxima of a weakly dependent random process</b>	<b>81</b>
3.1	Introduction . . . . .	81
3.2	A model for variable clustering . . . . .	83
3.2.1	Background setting . . . . .	83
3.2.2	Proposed AI-block models . . . . .	85
3.2.3	Extremal dependence structure for AI-block models . . . . .	86
3.3	Consistent estimation of minimally separated clusters . . . . .	88
3.3.1	Multivariate tail coefficient . . . . .	88

## Table of contents

---

3.3.2	Inference in AI-block models . . . . .	90
3.3.3	Estimation in growing dimensions . . . . .	91
3.3.4	Data-driven selection of the threshold parameter . . . . .	93
3.4	Hypotheses discussion for a multivariate random persistent process . . . . .	94
3.5	Numerical examples . . . . .	96
3.5.1	Numerical results . . . . .	96
3.5.2	Comparison with competitors . . . . .	98
3.6	Real-data applications . . . . .	103
3.6.1	Clustering brain extreme from EEG channel data . . . . .	103
3.6.2	Extremes on river network . . . . .	104
3.7	Conclusions . . . . .	105
<b>Appendix B Proofs of Chapter 3</b>		<b>107</b>
B.1	Proofs of main results . . . . .	107
B.1.1	Proofs of Section 3.2 . . . . .	107
B.1.2	Proofs of Section 3.3 . . . . .	108
B.1.3	Proofs of Section 3.4 . . . . .	116
B.2	Additional results . . . . .	117
B.2.1	Additional results of Section 3.2 . . . . .	117
B.2.2	Additional results of Section 3.3 . . . . .	123
B.3	Further results . . . . .	125
B.3.1	A usefull Glivenko-Cantelli result for the copula with known margins in a weakly dependent setting . . . . .	125
B.3.2	Weak convergence of an estimator of $\mathcal{A}^{(O)} - \mathcal{A}$ . . . . .	126
<b>4 Identifying regions of concomitant compound precipitation and wind speed extremes over Europe</b>		<b>131</b>
4.1	Introduction . . . . .	131
4.2	A clustering algorithm for compound extreme events . . . . .	135
4.2.1	A measure for evaluating dependence between compound extremes . . . . .	135
4.2.2	Clustering for compound extremes . . . . .	137
4.3	Detecting concomitant extremes of compound precipitation and wind . . . . .	139
4.3.1	Non-serially independent . . . . .	139
4.3.2	Exploratory analysis . . . . .	140
4.3.3	Clustering with constrained AI block models . . . . .	141
4.3.4	Results . . . . .	143
4.3.5	Alternative clustering method using SECO . . . . .	145
4.4	Conclusion and perspectives . . . . .	147
<b>Appendix C Supplementary materials of Chapter 4</b>		<b>149</b>
C.1	Axioms for a valid dependence measure . . . . .	149
C.2	A coherent measure for extreme value random vectors . . . . .	151
C.3	Incompleteness tail dependence structure estimation in high dimension . . . . .	153
C.4	Consistent estimation of SECO . . . . .	155
C.5	Definition of the Adjusted Rand Index (ARI) . . . . .	160
C.6	Supplementary Figures . . . . .	160

<b>5</b>	<b>Estimating Regularly Varying Random Vectors with Discrete Spectral Measure using Model-Based Clustering</b>	<b>165</b>
5.1	Introduction . . . . .	165
5.2	Identifiability . . . . .	169
5.3	Estimation . . . . .	173
5.3.1	Estimation of $I$ and $\mathcal{I}$ . . . . .	174
5.3.2	Estimation of the allocation matrix $A$ and soft clusters. . . . .	175
5.4	Statistical guarantees . . . . .	177
5.4.1	Statistical guarantees for $\hat{K}$ , $\hat{I}$ and $\hat{\mathcal{I}}$ . . . . .	179
5.4.2	Statistical guarantees for $\hat{A}$ . . . . .	180
5.5	Numerical results . . . . .	181
5.5.1	A data-driven selection method for the tuning parameter . . . . .	181
5.5.2	Performance of the proposed methodology in finite sample setting . . . . .	181
5.5.3	Numerical comparisons . . . . .	184
5.6	Applications . . . . .	186
5.6.1	Extreme precipitations in France . . . . .	186
5.6.2	Wildfires in French Mediterranean . . . . .	188
5.7	Discussion . . . . .	189
<b>Appendix D Supplementary materials of Chapter 5</b>		<b>193</b>
D.1	Investigation into the computation time of clique algorithm . . . . .	193
D.2	Algorithm . . . . .	193
D.3	Proofs of Section 5.2 . . . . .	195
D.4	Proof of Section 5.3 . . . . .	198
D.5	Proof of Section 5.4 . . . . .	200
D.5.1	Proof of Section 5.4.1 . . . . .	200
D.5.2	Proof of Section 5.4.2 . . . . .	202
D.6	Proofs of Section 5.5 . . . . .	206
D.7	Supplementary Lemmata . . . . .	209
<b>6</b>	<b>Conclusion and perspectives</b>	<b>213</b>
6.1	Strong Consistency of madogram-based K-means under mixing conditions . . . . .	215
6.2	Estimating Sparse Linear Regression with Randomly Varying Design . . . . .	216
6.3	Changepoint Detection for High-Dimensional Extremal Dependence . . . . .	218
<b>References</b>		<b>221</b>
<b>Appendix E A Python Package for Sampling from Copulae</b>		<b>235</b>
E.1	Introduction . . . . .	235
E.2	Classes . . . . .	236
E.2.1	The Archimedean class . . . . .	237
E.2.2	The Extreme class . . . . .	238
E.3	Random vector generator . . . . .	239
E.3.1	The bivariate case . . . . .	239
E.3.2	The multivariate case . . . . .	242
E.4	Case study: modeling pairwise dependence between spatial maximas with missing data . . . . .	243



## Table of contents

---

E.4.1	Background . . . . .	244
E.4.2	Numerical results . . . . .	245
E.5	Discussion . . . . .	248
E.5.1	Comparison of <code>clayton</code> with R packages . . . . .	248
E.5.2	Conclusion . . . . .	249
E.6	Bivariate Archimedean models . . . . .	251
E.7	Implemented bivariate extreme models . . . . .	253
E.8	Multivariate Archimedean copulae . . . . .	254
E.9	Multivariate extreme models . . . . .	255
E.10	Multivariate elliptical dependencies . . . . .	257

# CHAPTER 1

## INTRODUCTION

### 1.1 A survival guide for high-dimensional extremal dependence modeling

#### Notation

We introduce here notation that will be used throughout this and later chapters. Vector order is taken componentwise, i.e., for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  we write  $\mathbf{x} \leq \mathbf{y}$  if and only if  $x^{(j)} \leq y^{(j)}$ ,  $1 \leq j \leq d$ . Multivariate intervals are defined as follows: for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  such that  $\mathbf{x} \leq \mathbf{y}$ ,

$$[\mathbf{x}, \mathbf{y}] = \{\mathbf{x} \in \mathbb{R}^d : x^{(j)} \leq y^{(j)}, 1 \leq j \leq d\}.$$

Open and semi-open intervals are defined similarly. For  $\mathbf{x} \in \mathbb{R}^{\mathbb{Z}}$ , we recall the notation

$$\bigvee_{i \in \mathbb{Z}} x_i = \max_{i \in \mathbb{Z}} x_i, \quad \bigwedge_{i \in \mathbb{Z}} x_i = \min_{i \in \mathbb{Z}} x_i.$$

For  $d$ -dimensional vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n$ ,  $\bigvee_{i=1}^d \mathbf{x}_i$  and  $\bigwedge_{i=1}^d \mathbf{x}_i$  denote componentwise maxima and minima, respectively, i.e.,

$$\bigvee_{i=1}^n \mathbf{x}_i = \left( \bigvee_{i=1}^n \mathbf{x}_i^{(1)}, \dots, \bigvee_{i=1}^n \mathbf{x}_i^{(d)} \right), \quad \bigwedge_{i=1}^n \mathbf{x}_i = \left( \bigwedge_{i=1}^n \mathbf{x}_i^{(1)}, \dots, \bigwedge_{i=1}^n \mathbf{x}_i^{(d)} \right).$$

For a set  $A \subset \mathbb{R}^d$  and  $y > 0$ , we define the dilated set  $yA$  by

$$yA = \{ya : a \in A\}.$$

Given an arbitrary norm on  $\mathbb{R}^d$  denoted  $\|\cdot\|$ , we define the open and closed balls and sphere with center  $\mathbf{x} \in \mathbb{R}^d$  and radius  $r > 0$  by

$$\begin{aligned} B(\mathbf{x}, r) &= \{\mathbf{y} \in \mathbb{R}^d, \|\mathbf{x} - \mathbf{y}\| < r\}, & \bar{B}(\mathbf{x}, r) &= \{\mathbf{y} \in \mathbb{R}^d, \|\mathbf{x} - \mathbf{y}\| \leq r\} \\ S(\mathbf{x}, r) &= \{\mathbf{y} \in \mathbb{R}^d, \|\mathbf{x} - \mathbf{y}\| = r\}, & \mathbb{S}_{d-1} &= S(0, 1). \end{aligned}$$

A sequence  $\mathbf{u} \in \mathbb{R}^{\mathbb{Z}}$  will be named as scaling if it is non negative, i.e.,  $u_n > 0$  for  $n \in \mathbb{Z}$  and non-decreasing, meaning that  $u_n \leq u_{n+1}$  for  $n \in \mathbb{Z}$ .

#### 1.1.1 Dependence modeling

In this introductory chapter, I have chosen to present to the reader methods of reasoning by fully writing out certain proofs of results. This choice allows the reader to more easily grasp

mathematical concepts by directly manipulating them. From my experience, understanding of mathematics truly deepens when definitions are laid out and mathematical objects are rigorously manipulated through the demonstration of theorems. Although this approach may result in an increase in the length of the text, I align with one of N. Bourbaki's principles - avoiding paper savings - particularly when addressing beginners in a field. By analogy, learning Latin does not involve studying fragments of tablets discovered during ancient excavations in Rome, but rather through the reading of well-written texts whose meaning is clear. Similarly, I have chosen to present proofs that highlight ideas rather than calculations. These proofs may be exercises taken from existing works or proofs taken from these works, to which I have possibly added elements to facilitate understanding.

In our discussion, we frequently have to consider random variables  $X^{(1)}, \dots, X^{(d)}$  defined on a shared probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . A  $d$ -dimensional random vector, denoted as  $\mathbf{X}$ , serves as a measurable mapping from  $\Omega$  into  $\mathbb{R}^d$ . The term "measurable" signifies that the inverse image

$$\mathbf{X}^{-1}(B) = \{\omega \in \Omega, \mathbf{X}(\omega) \in B\},$$

of every Borel set  $B$  in  $\mathcal{B}(\mathbb{R}^d)$  belongs to  $\mathcal{F}$ . Considering  $\mathbf{X}$  a random vector on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , a probability measure  $\mathbb{P}_{\mathbf{X}}$  can be defined on the measurable space  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$  by

$$\forall B \in \mathcal{B}(\mathbb{R}^d), \quad \mathbb{P}_{\mathbf{X}}(B) := \mathbb{P} \left\{ \mathbf{X}^{-1}(B) \right\}.$$

The probability measure  $\mathbb{P}_{\mathbf{X}}$  is referred to as the law or distribution of  $\mathbf{X}$ . Independence, a fundamental concept in probability theory, statistics, and numerous related fields, plays a crucial role. The assumption of independence is a cornerstone in numerous statistical models. As a simple illustration, consider the linear model  $Y = \mathbf{X}\beta + \epsilon$  under random design, which often assumes, independence between  $\mathbf{X}$  and  $\epsilon$ . Describing this concept in terms of random variables,  $X^{(1)}, \dots, X^{(d)}$  defined on  $(\Omega, \mathcal{F}, \mathbb{P})$  are deemed independent if, for every choice of  $d$  Borel sets  $B^{(1)}, \dots, B^{(d)}$ , one has

$$\mathbb{P} \left\{ X^{(1)} \in B^{(1)}, \dots, X^{(d)} \in B^{(d)} \right\} = \mathbb{P} \left\{ \bigcap_{j=1}^d (X^{(j)} \in B^{(j)}) \right\} = \prod_{j=1}^d \mathbb{P} \left\{ X^{(j)} \in B^{(j)} \right\}.$$

As commonly known, understanding the law of  $\mathbb{P}_{\mathbf{X}}$  of a random vector  $\mathbf{X}$  is facilitated by knowledge of its distribution function. Let us commence by considering the case  $d = 1$ , specifically when a single random variable  $X$  is in focus.

**Definition 1.1.1.** The distribution function  $F_X$  of a random variable  $X$  on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is the function  $F_X : \mathbb{R} \rightarrow [0, 1]$  defined by:

$$F_X(x) = \mathbb{P} \{ X \leq x \},$$

such that  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  and  $\lim_{x \rightarrow \infty} F_X(x) = 1$ .

The distribution function of a random variable can be characterised in terms of its analytical properties.

**Theorem 1.1.1.** Let  $F_X : \mathbb{R} \rightarrow [0, 1]$ . The following statements are equivalent:

- (A) there is a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and a random variable  $X$  on it such that  $F_X$  is the distribution function of  $X$ ;

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

**(B)**  $F$  satisfies the following properties:

- (a)**  $F_X$  is continuous at the right at every point of  $\mathbb{R}$ , i.e., for every  $x \in \mathbb{R}$ ,  $\ell^+(F_X(x)) = F_X(x)$ ;
- (b)**  $F_X$  is increasing, i.e.,  $F_X(x) \leq F_X(x')$  whenever  $x \leq x'$ ;
- (c)**  $F_X$  satisfies the following limits

$$\lim_{x \rightarrow -\infty} F_X(x) = 0, \quad \lim_{x \rightarrow \infty} F_X(x) = 1.$$

Under appropriate conditions, it becomes feasible to convert any random variable  $X$  into another random variable  $U$  uniformly distributed over the interval  $[0, 1]$ , denoted  $U \sim \mathcal{U}([0, 1])$ . This transformation requires a suitable inverse of a distribution function, which is elucidated below.

**Definition 1.1.2.** For a distribution function  $F : \mathbb{R} \mapsto [0, 1]$ , the generalised inverse of  $F$  is the function  $F^{\leftarrow} : [0, 1] \rightarrow \mathbb{R}$  given, for every  $x \in ]0, 1]$ , by

$$F^{\leftarrow}(x) = \inf \{t \in \mathbb{R}, F(t) \geq x\},$$

with  $F^{\leftarrow}(0) := \inf \{t \in \mathbb{R}, F(t) > 0\}$ .

The generalised inverse of a distribution function  $F_X$  aligns with the standard inverse when  $F_X$  is both continuous and strictly increasing. For latter uses, we here define the quantile function defined on  $[0, 1]$  of  $X$ ,

$$Q_X(t) = F_X^{\leftarrow}(1 - t). \tag{1.1}$$

Here, we gather a set of widely recognised properties associated with the generalised inverse of a distribution function.

**Theorem 1.1.2.** Let  $F_X$  be a distribution function and let  $F_X^{\leftarrow}$  be its generalised inverse. Then

- (a)**  $F_X^{\leftarrow}$  is increasing. In particular, if  $F_X$  is continuous, then  $F_X^{\leftarrow}$  is strictly increasing;
- (b)**  $F_X^{\leftarrow}$  is left continuous on  $[0, 1]$ ;
- (c)** If  $x \in \text{Ran}(F_X)$ ,  $F_X(F_X^{\leftarrow}(x)) = x$ . In particular, if  $F_X$  is continuous,  $F_X(F_X^{\leftarrow}(x)) = x$  for every  $x \in [0, 1]$ ;
- (d)**  $F_X^{\leftarrow}(F_X(x)) \leq x$  for every  $x \in \mathbb{R}$ . In particular, if  $F_X$  is strictly increasing, then  $F_X^{\leftarrow}(F_X(x)) = x$  for every  $x \in \mathbb{R}$ ;
- (e)** For every  $x \in \mathbb{R}$  and  $t \in [0, 1]$ ,  $F_X(x) \geq t$  if, and only if,  $x \geq F_X^{\leftarrow}(t)$ .

Thanks to these properties, the following classical result follows.

**Theorem 1.1.3.** Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, \mathbb{P})$  whose distribution function is given by  $F_X$ .

- (a)** If  $F_X$  is continuous then  $F_X \circ X$  is uniformly distributed on  $[0, 1]$ .
- (b)** If  $U$  is a random variable that is uniformly distributed on  $[0, 1]$ , then  $F_X^{\leftarrow} \circ U$  has distribution function equal to  $F_X$ .

Nevertheless, the statement mentioned above can be adapted to this general construction, as presented in a exercise from [Rio \(1999\)](#).

## Introduction

---

**Theorem 1.1.4.** Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, \mathbb{P})$  whose distribution function is given by  $F_X$ . Let  $\delta$  be a random variable with uniform distribution over  $[0, 1]$ , independent of  $X$ . Set

$$V = \ell^-(F_X(X)) + \delta(F_X(X) - \ell^-(F_X(X))).$$

Then  $V$  has the uniform distribution over  $[0, 1]$  and, almost surely  $F^{\leftarrow}(V) = X$ .

**Proof** Let  $v(x, t) = \ell^-(F_X(x)) + t(F_X(X) - \ell^-(F_X(X)))$ . The defined mapping  $v$  is measurable with respect to  $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}([0, 1])$ . Hence  $V = v(X, \delta)$  is a real-valued random variable. Let  $a$  be any real in  $[0, 1]$ . Let us consider

$$b = \ell^+(F_X^{\leftarrow}(x)) = \sup\{x \in \mathbb{R}, F_X(x) \leq a\}.$$

If  $F_X$  is continuous at point  $b$ , then  $F_X(b) = a$ . In that case  $(v(x, t) \leq a)$  if and only if  $(x \leq b)$ , which ensures that  $\mathbb{P}\{v \leq a\} = F_X(b) = a$ . If  $F_X$  is not continuous at point  $b$ , then  $a \in [\ell^-(F_X(b)), F_X(b)]$ , which implies that

$$a = v(b, u) \text{ for some } u \in [0, 1].$$

In that case,  $(v(x, t) \leq a)$  if and only if either  $(x \leq b)$  or  $(x = b \text{ and } t \leq u)$ . Then

$$\mathbb{P}\{V \leq a\} = \ell^+(F_X(b)) + u(F_X(b) - \ell^+(F_X(b))) = a.$$

Consequently,  $V$  has the uniform distribution over  $[0, 1]$ . Now, since  $F_X(x) \geq v(x, t)$  for any  $t \in [0, 1]$ , we have:

$$x \geq F_X^{\leftarrow}(v(x, t)) \text{ for any } t \in [0, 1].$$

It follows that  $x \geq F_X^{\leftarrow}(V)$  almost surely. Let  $\Phi$  be the distribution function the standard normal law. Since  $\{F_X^{\leftarrow}(V) > x\}$  if and only if  $\{V > F_X(x)\}$ , we have:

$$\begin{aligned} \mathbb{E}[\Phi(F_X^{\leftarrow}(V))] &= \int_{\mathbb{R}} \mathbb{P}\{F_X^{\leftarrow}(V) > x\} \Phi'(x) dx = \int_{\mathbb{R}} \mathbb{P}\{V > F_X(x)\} \Phi'(x) dx \\ &= \int_{\mathbb{R}} (1 - F_X(x)) \Phi'(x) dx = \mathbb{E}[\Phi(X)]. \end{aligned}$$

It follows that  $\mathbb{E}[\Phi(X)] = \mathbb{E}[\Phi(F_X^{\leftarrow}(V))]$ . Since  $\Phi(X) \geq \Phi(F_X^{\leftarrow}(V))$  almost surely, it implies that  $\Phi(X) = \Phi(F_X^{\leftarrow}(V))$  almost surely. Hence  $X = F_X^{\leftarrow}(V)$  almost surely, which completes the proof of the theorem.  $\square$

The concept of a distribution function can be similarly defined in higher dimensions.

**Definition 1.1.3.** The distribution function of a random vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is defined by

$$F_{\mathbf{X}}(x^{(1)}, \dots, x^{(d)}) = \mathbb{P}\{X^{(1)} \leq x^{(1)}, \dots, X^{(d)} \leq x^{(d)}\}$$

for all  $x^{(1)}, \dots, x^{(d)}$  in  $\mathbb{R}$ .

We are now in a position to define the functions - copulae - that serve as the primary tools in dependence modeling. The fundamental idea behind modeling stochastic dependence using copulae is as follows: let  $\mathbf{X} = (X^{(1)}, X^{(2)})$  be a random vector with continuous marginal

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

distributions. The probability integral transform (see Theorem 1.1.4) applied to  $X^{(1)}$  and  $X^{(2)}$  defines two random variables  $U^{(1)} = F_{X^{(1)}}(X^{(1)}) := F^{(1)}(X^{(1)})$  and  $U^{(2)} = F_{X^{(2)}}(X^{(2)}) := F^{(2)}(X^{(2)})$ , and since the transforms are invertible, specifying the dependence between  $X^{(1)}$  and  $X^{(2)}$  is the same as specifying the dependence between  $U^{(1)}$  and  $U^{(2)}$ . Consequently, the task of investigating stochastic dependence has been reduced to the problem of investigating stochastic dependence with uniform marginals, which is the copula. Below, we provide the formal definition of this mathematical object, building upon the previously mentioned intuition.

**Definition 1.1.4.** For every  $d \geq 2$ , a  $d$ -dimensional copula is a  $d$ -dimensional distribution function concentrated on  $[0, 1]^d$  whose univariate marginals are uniformly distributed on  $[0, 1]$ .

**Remark 1.1.1.** The term copula (plural copulae or copulas) is derived from the Latin word for a link or tie that connects two different things. In linguistics, a copula refers to a word or phrase that links the subject of a sentence to a predicate, as seen in examples like “the food smells food” where *smells* is the copula. This linguistic application of the term “copula” was the primary inspiration for Sklar to designate a function linking a multidimensional distribution to its one-dimensional margins. The same concept can be alternatively labeled as “uniform representation” (Kimeldorf and Sampson (1989)) or “dependence function” (see, for instance, Deheuvels (1979); Galambos (1977); Hsing (1989)).

Every copula corresponds to a random vector  $\mathbf{U}$  defined on an appropriate probability space, where the joint distribution of  $\mathbf{U}$  is represented by  $C$ . This probabilistic characterisation enables the introduction of those fundamental examples of copulae.

**Example 1.1.1 (The comonotonicity copula  $M_d$ ).** Let  $U$  be a random variable defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Suppose that  $U$  is uniformly distributed on  $[0, 1]$ . Consider the random vector  $\mathbf{U} = (U, \dots, U)$ . Then for every  $\mathbf{u} \in [0, 1]^d$

$$\mathbb{P}\{\mathbf{U} \leq \mathbf{u}\} = \mathbb{P}\left\{\mathbf{U} \leq \min(u^{(1)}, \dots, u^{(d)})\right\} = \min(u^{(1)}, \dots, u^{(d)}).$$

Thus the distribution function given, for every  $\mathbf{u} \in [0, 1]^d$  by

$$M_d(u^{(1)}, \dots, u^{(d)}) := \min(u^{(1)}, \dots, u^{(d)}),$$

is a copula, which will be called the comonotonicity copula.

**Example 1.1.2 (The independence copula  $\Pi_d$ ).** Let  $U^{(1)}, \dots, U^{(d)}$  be independent random variables defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Suppose that each  $U^{(j)}$  is uniformly distributed on  $[0, 1]$ . Consider the random variable  $\mathbf{U} = (U^{(1)}, \dots, U^{(d)})$ . Then, for every  $\mathbf{u} \in [0, 1]^d$ ,

$$\mathbb{P}\{\mathbf{U} \leq \mathbf{u}\} = \prod_{j=1}^d \mathbb{P}\{U^{(j)} \leq u^{(j)}\} = \prod_{j=1}^d u^{(j)},$$

is a copula, which will be called the independence copula.

**Example 1.1.3 (The countercomonotonicity copula  $W_2$ ).** Let  $U$  be a random variable on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Suppose that  $U$  is uniformly distributed on  $[0, 1]$ . Consider the random vector  $\mathbf{U} = (U, 1 - U)$ . Then for every  $\mathbf{u} \in [0, 1]^2$ ,

$$\mathbb{P}\{\mathbf{U} \leq \mathbf{u}\} = \mathbb{P}\left\{U \leq u^{(1)}, 1 - U \leq u^{(2)}\right\} = \max(0, u^{(1)} + u^{(2)} - 1).$$

## Introduction

---

Thus the distribution function given, for every  $\mathbf{u} \in [0, 1]^2$ , by

$$W_2(u^{(1)}, u^{(2)}) := \max(0, u^{(1)} + u^{(2)} - 1),$$

is a copula, which will be called the countercomonotonicity copula.

The motivation leading to the definition of copula is summarised in the well-known Sklar's theorem (Sklar (1959)), which forms the basis for most applications of copulae, it turns out that relaxing the assumption of continuity of the marginals leads to non-uniqueness of the associated copula.

**Theorem 1.1.5.** *Let a random vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  be given on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , let  $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}\{X^{(1)} \leq x^{(1)}, \dots, X^{(d)} \leq x^{(d)}\}$  be the joint distribution function of  $\mathbf{X}$  and let  $F^{(j)}(x^{(j)}) := F_{X^{(j)}}(x^{(j)})$ ,  $j = 1, \dots, d$ , be its marginals. Then there exists a copula  $C$  such that for every point  $\mathbf{x} = (x^{(1)}, \dots, x^{(d)}) \in \mathbb{R}^d$ ,*

$$F_{\mathbf{X}}(x^{(1)}, \dots, x^{(d)}) = C(F^{(1)}(x^{(1)}), \dots, F^{(d)}(x^{(d)})).$$

*If the marginals  $F^{(1)}, \dots, F^{(d)}$  are continuous, then the copula  $C$  is uniquely defined.*

**Proof** To gain a nuanced understanding of the proof, we initially assume that all marginals  $F^{(j)}$ ,  $j = 1, \dots, d$ , are continuous. This condition will be leveraged later on. In view of the univariate probability integral transformation (see Theorem 1.1.3),  $F^{(j)} \circ X^{(j)}$  is distributed uniformly over  $[0, 1]$  for each  $j = 1, \dots, d$ . Thus the distribution function of  $(F^{(1)} \circ X^{(1)}, \dots, F^{(d)} \circ X^{(d)})$  has uniform univariate marginals and, hence, it is a copula. Moreover, for every point  $\mathbf{x} \in \mathbb{R}^d$ , one has

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}\{X^{(1)} \leq x^{(1)}, \dots, X^{(d)} \leq x^{(d)}\} \\ &= \mathbb{P}\{F^{(1)}(X^{(1)}) \leq F^{(1)}(x^{(1)}), \dots, F^{(d)}(X^{(d)}) \leq F^{(d)}(x^{(d)})\} \\ &= C(F^{(1)}(x^{(1)}), \dots, F^{(d)}(x^{(d)})), \end{aligned}$$

which is the assertion.

Take  $F^{(j)}$  as a not necessarily continuous function for  $j = 1, \dots, d$ , let  $\delta$  be independent of  $\mathbf{X}$  and uniformly distributed on  $[0, 1]$ . For  $j = 1, \dots, d$ , consider the transformation  $U^{(j)} = \ell^-(F^{(j)}(X^{(j)})) + \delta \left( F^{(j)}(X^{(j)}) - \ell^-(F^{(j)}(X^{(j)})) \right)$ . According to Theorem 1.1.4,  $U^{(j)}$  is uniformly distributed over  $[0, 1]$  and  $X^{(j)} = (F^{(j)})^{\leftarrow}(U^{(j)})$  almost surely,  $j = 1, \dots, d$ . Thus, defining  $C$  to be the distribution function of  $\mathbf{U} = (U^{(1)}, \dots, U^{(d)})$  one has

$$\begin{aligned} F_{\mathbf{X}} &= \mathbb{P}\left\{\bigcap_{j=1}^d (X^{(j)} \leq x^{(j)})\right\} = \mathbb{P}\left\{\bigcap_{j=1}^d \left((F^{(j)})^{\leftarrow}(U^{(j)}) \leq x^{(j)}\right)\right\} \\ &= \mathbb{P}\left\{\bigcap_{j=1}^d \left(U^{(j)} \leq F^{(j)}(x^{(j)})\right)\right\} \\ &= C(F^{(1)}(x^{(1)}), \dots, F^{(d)}(x^{(d)})). \end{aligned}$$

□

This theorem first appeared in Sklar (1959), and its significance for the concept of stochastic dependence cannot be overstated. Every joint distribution function, which inherently contains

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

all information about the dependence of a random vector can be decomposed into a copula and the marginal distributions. Since the marginals are incapable of explaining any dependence, the dependence has to be entirely characterised by the copula. This is emphasised by the fact that standard measures of dependence like Kendall's  $\tau$  or Spearman's  $\rho$  are functions of copulae. In the following proposition, we compile some fundamental properties of copulae. To this end, we consider the function called the lower Fréchet-Hoeffding bound  $W_d : [0, 1]^d \rightarrow [0, 1]$  defined by:

$$W_d = \max \left( 0, \sum_{j=1}^d u^{(j)} - (d-1) \right).$$

**Proposition 1.1.1.** *Let a random vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  and  $F_{\mathbf{X}}$  be the joint distribution function of  $\mathbf{X}$  and let  $F^{(j)}$ ,  $j = 1, \dots, d$  be its marginals and  $C$  its copula. Then the following results hold:*

- (a) **Fréchet-Hoeffding bound** For every copula  $C$  and for every point  $\mathbf{u} = (u^{(1)}, \dots, u^{(d)}) \in [0, 1]^d$  one has

$$W_d(\mathbf{u}) \leq C(\mathbf{u}) \leq M_d(\mathbf{u}).$$

*Irrespective of the dimension, the upper bound  $M_d$  corresponds to perfect positive dependence in the sense that  $U^{(j)}$  is a strictly increasing function of  $U^{(1)}, \dots, U^{(j-1)}, U^{(j+1)}, \dots, U^{(d)}$ . In dimension 2,  $W_2$  corresponds to perfect negative dependence and is a copula only for  $d = 2$ .*

- (b) **Independence** If  $F^{(1)}, \dots, F^{(d)}$  are continuous, then  $X^{(1)}, \dots, X^{(d)}$  are independent if and only  $C(\mathbf{u}) = \prod_{j=1}^d u^{(j)} = \Pi_d(\mathbf{u})$ .
- (c) **Lipschitz continuity**  $C$  is Lipschitz-continuous with respect to the 1-norm on  $[0, 1]^d$

$$|C(\mathbf{u}) - C(\mathbf{v})| \leq \|\mathbf{u} - \mathbf{v}\|_1, \quad \forall \mathbf{u}, \mathbf{v} \in [0, 1]^d.$$

- (d) **Differentiability** For all  $u^{(1)}, \dots, u^{(j-1)}, u^{(j+1)}, \dots, u^{(d)} \in [0, 1]$ , it holds that the partial derivatives  $\partial_j C(\mathbf{u})$  exists for  $\lambda$ -almost  $u^{(j)}$ . Furthermore  $0 \leq \partial_j C(\mathbf{u}) \leq 1$ .
- (e) **Invariance under increasing transformations** If  $F^{(1)}, \dots, F^{(d)}$  are continuous and  $\alpha^{(j)}$ ,  $j = 1, \dots, d$  are strictly increasing, then the copula of  $(\alpha^{(1)} \circ X^{(1)}, \dots, \alpha^{(d)} \circ X^{(d)})$  is  $C$ .
- (f) **Spearman's  $\rho$ , Kendall's  $\tau$**  If  $F^{(1)}$  and  $F^{(2)}$  are continuous, then the Spearman's  $\rho$  of  $X^{(1)}$  and  $X^{(2)}$  is given by

$$\rho(X^{(1)}, X^{(2)}) = \rho(C) = 12 \int_{[0,1]^2} C(u, v) dudv - 3.$$

The Kendall's  $\tau$  of  $X^{(1)}$  and  $X^{(2)}$  is given by

$$\tau(X^{(1)}, X^{(2)}) = \tau(C) = 4 \int_{[0,1]^2} C(u, v) dudv - 1.$$

For detailed proofs of these results and a comprehensive understanding of copula theory, we recommend consulting the following monographs: [Durante and Sempi \(2015\)](#) for mathematical principles, [Joe \(2014\)](#) for insights on dependence modeling with copulae, and [Nelsen \(2006\)](#) for an introduction to the theory and applications of copulae.



### 1.1.2 Extremal dependence modeling

One approach to evaluate dependence is through sample (cross)-correlations. In extreme modeling, there is no guarantee that theoretical moments like correlations exist, but samples versions will always be available. However, correlation is a somewhat basic measure of dependence, really informative only between jointly Gaussian variables. It is simple but not subtle, as it does not distinguish between large values and small values. In this section, we will formally introduce the concept of tail dependence and embed it into the framework of copulae. Loosely speaking, bivariate tail dependence is the amount of upper quadrant-tail of a bivariate distribution. The concept is deeply related to multivariate extreme value theory, i.e., the limiting distributions of componentwise maxima of i.i.d.  $d$ -variate random vectors. We will establish the de Haan-Resnick representation theorem for so-called max-infinitely divisible random vectors. The subsequent representation introduces common dependence function for tail dependence, namely the stable tail dependence function and the Pickands dependence function.

Commencing with the work of [Geffroy \(1958, 1959\)](#) and [Sibuya et al. \(1960\)](#), consider a two-dimensional vector  $\mathbf{X} = (X^{(1)}, X^{(2)})$  with a joint distribution function  $F_{\mathbf{X}}$ , marginal distributions  $F^{(1)}, F^{(2)}$ , and copula  $C$ . The subsequent definition introduces the common scalar measure of tail dependence, namely the extremal correlation.

**Definition 1.1.5.**  $X^{(1)}$  and  $X^{(2)}$  are said to be tail dependent, extremally dependent or asymptotically dependent if the tail dependence parameter

$$\chi(1, 2) = \lim_{q \rightarrow 1} \mathbb{P} \left\{ X^{(1)} > (F^{(1)})^{\leftarrow}(q) \mid X^{(2)} > (F^{(2)})^{\leftarrow}(q) \right\}$$

exists and is strictly positive.  $X^{(1)}$  and  $X^{(2)}$  are said to be tail independent, extremally independent or asymptotically independent if  $\chi(1, 2) = 0$ .

As one might anticipate, for continuous marginal distributions, tail dependence is a property of the copula of  $\mathbf{X}$ . If  $\bar{C}(\mathbf{u}) = u^{(1)} + u^{(2)} - 1 + C(1 - u^{(1)}, 1 - u^{(2)})$  denotes the survival copula of  $\mathbf{X}$ , a straightforward calculation based on the definition of conditional probabilities justifies the following proposition.

**Proposition 1.1.2.** *Let  $\mathbf{X} = (X^{(1)}, X^{(2)})$  be continuous bivariate random vector, then*

$$\chi(1, 2) = \lim_{q \rightarrow 1} \frac{1 - 2q + C(q, q)}{q} = \lim_{q \rightarrow 1} \frac{\bar{C}(q, q)}{q}.$$

In the following, we will explore the concept of extreme value within the framework of multivariate extreme value theory. It will be revealed that a fundamental condition of extreme value theory, i.e., the cumulative distribution of the  $d$ -variate maxima of  $\mathbf{X}$  converges to a max-stable distribution equivalently stated as the max-domain of attraction assumption, is sufficient for the extremal correlation to exist. It suffices that the copula of  $F_{\mathbf{X}}$  is in some domain of attraction and this attractor will be called the extreme value copula. Furthermore, the extreme value copula characterises the extremal dependence structure of  $\mathbf{X}$ . We will further showcase its connections to fundamental objects characterising the extremal dependence.

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

Let  $\mathbf{X}_i = (X_i^{(1)}, \dots, X_i^{(d)})$ ,  $i = 1, \dots, n$  represent i.i.d. random vectors with a common distribution function  $F_{\mathbf{X}}$ . The  $d$ -variate maximum is defined by

$$\bigvee_{i=1}^n \mathbf{X}_i = \left( \bigvee_{i=1}^n X_i^{(1)}, \dots, \bigvee_{i=1}^n X_i^{(d)} \right),$$

it immediately falls down that

$$\mathbb{P} \left\{ \bigvee_{i=1}^n \frac{\mathbf{X}_i - b_n}{a_n} \leq \mathbf{x} \right\} = F_{\mathbf{X}}^n(a_n \mathbf{x} + b_n), \quad (1.2)$$

for a certain vector  $a_n > 0$  and  $b_n$ .

The extreme value distributions correspond to all possible nondegenerate limits in Equation (1.2). The only conceivable limit laws as  $n \rightarrow \infty$  are max-stable distribution functions, i.e. distributions where:

$$H_{\mathbf{X}}^n(b_n + a_n \mathbf{x}) = H_{\mathbf{X}}(\mathbf{x}), \quad \forall n \in \mathbb{N}.$$

If  $H_{\mathbf{X}}$  is max-stable, then the marginals also exhibit this characteristic. This implies that a multivariate max-stable distribution is continuous (according to (Falk et al., 2010, Lemma 2.2.6)). In the subsequent discussion, our primary focus is on max-stable distribution function with standard Fréchet margins, denoted as  $H_1^{(j)} = e^{-x^{-1}}$ ,  $x > 0$ ,  $j = 1, \dots, d$ .

This standardisation in the univariate margins can always be accomplished through a straightforward transformation. If  $H_{\mathbf{X}}$  is max-stable with margins  $H^{(j)}$ ,  $j = 1, \dots, d$  then:

$$H_{\mathbf{X}} \left( (H_1^{(1)})^{\leftarrow} (H^{(1)}(x^{(1)})), \dots, (H_1^{(d)})^{\leftarrow} (H^{(d)}(x^{(d)})) \right),$$

defines a max-stable distribution function with margins  $H_1^{(j)}$  for any  $j = 1, \dots, d$ . Because of its importance, we are going to give a first max-stable distribution function.

**Lemma 1.1.1.** *For every  $K \in \mathbb{N}$ , let  $Z^{(1)}, \dots, Z^{(K)}$  be i.i.d. random variables with standard Fréchet distribution functions. Let  $A_{ja} > 0$  for  $a = 1, \dots, K$  and  $j = 1, \dots, d$ . Then*

$$\mathbb{P} \left\{ \bigvee_{a=1}^K A_{ja} Z^{(a)} \leq x^{(j)}, j = 1, \dots, d \right\} = e^{-\left\{ \sum_{a=1}^K \bigvee_{j=1}^d \frac{A_{ja}}{x^{(j)}} \right\}}, \quad \mathbf{x} \geq 0. \quad (1.3)$$

*Thus obtaining a  $d$ -variate max-stable distribution function with Fréchet margins. If, in addition,*

$$\sum_{a=1}^K A_{ja} = 1, \quad j = 1, \dots, d$$

*then the univariate margins are standard Fréchet.*

**Proof** Notice that  $\{\bigvee_{a=1}^K A_{ja} Z^{(a)} \leq x^{(j)}\}$ ,  $j = 1, \dots, d$ , if and only if  $\{Z^{(a)} \leq \bigwedge_{j=1}^d \frac{x^{(j)}}{A_{ja}}\}$ ,  $a = 1, \dots, K$ . Thus (1.3) follows from the independence of  $Z^{(1)}, \dots, Z^{(K)}$  and using  $(\bigwedge_{j=1}^d \frac{x^{(j)}}{A_{ja}})^{-1} =$

## Introduction

---

$\bigvee_{j=1}^d \frac{A_{ja}}{x^{(j)}}$ . The max-stability is obvious. We see that the  $j$ th marginal  $H^{(j)}$  is given by

$$H^{(j)}(x) = e^{-\left\{\sum_{a=1}^K \frac{A_{ja}}{x^{(j)}}\right\}}, \quad x > 0,$$

and hence, the assertion concerning the univariate margins holds.  $\square$

Now, we broaden our framework from max-stable to max-infinitely divisible distribution functions. A random vector  $\mathbf{X}$  is max-infinitely divisible if, for every  $n \in \mathbb{N}$ , it has the stochastic representation:

$$\mathbf{X} \stackrel{d}{=} \bigvee_{i=1}^n \mathbf{X}_i^{(n)},$$

for some i.i.d. vectors  $\mathbf{X}_1^{(n)}, \dots, \mathbf{X}_n^{(n)}$  with distribution function  $F_n$ . The above equation can be equivalently stated in terms of the distribution function with

$$F_{\mathbf{X}} = F_n^n, \quad \forall n \in \mathbb{N}.$$

It is evident that  $F_{\mathbf{X}}$  is max-infinitely divisible if it is max-stable.

Consider a Borel measure  $\Lambda$  defined on the punctured  $d$ -dimensional Euclidean space  $\mathcal{E} = \mathbb{R}^d \setminus \{0\}$ , where  $\Lambda$  is finite for all Borel sets that are away from the origin. Measures of  $\Lambda$  in this context are fundamental as they provide a characterisation of max-infinitely divisible random vectors. To simplify the discussion, let us focus on a  $d$ -dimensional max-infinitely divisible random vector with Fréchet margins  $H_{\alpha}^{(j)} = e^{-x^{-\alpha}}$ ,  $x > 0$ ,  $j = 1, \dots, d$ ,  $\alpha > 0$ . The following theorem articulates this result:

**Theorem 1.1.6.** *Let  $\mathbf{X}$  a random vector with joint distribution function  $H_{\mathbf{X}}$  and  $d \in \mathbb{N}$ . Then there is a 1-1 correspondence between*

- (A) *Max-infinitely divisible random vector  $\mathbf{X}$  on  $\mathbb{R}^d$  with Fréchet margins  $H_{\alpha}^{(j)}(x) = e^{-x^{-\alpha}}$ ,  $x > 0$ ,  $\alpha > 0$ ,  $j = 1, \dots, d$ .*
- (B) *Borel measure  $\Lambda_{\mathbf{X}}$  on  $\mathcal{E}_+ = [0, \infty)^d \setminus \{0\}$  satisfying*
- (a)  *$\Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus [0, \mathbf{x}]) < \infty$  for every  $\mathbf{x} > 0$ .*

*This correspondence is given by*

$$H_{\mathbf{X}}(\mathbf{x}) = \begin{cases} e^{-\Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus [0, \mathbf{x}])}, & \mathbf{x} > 0 \\ 0 & \mathbf{x} \leq 0. \end{cases} \quad (1.4)$$

A proof of a more general statement, but similar to the one mentioned above is provided by (Resnick, 2008, Proposition 5.8). Now, suppose that  $\mathbf{X}$  is max-stable with Fréchet marginals  $H_{\alpha}^{(j)}(x) = e^{-x^{-\alpha}}$ , then there exist  $a_n > 0$  and  $b_n \in \mathbb{R}^d$  such that

$$H_{\mathbf{X}}(\mathbf{x}) = H_{\mathbf{X}}^n(a_n \mathbf{x} + b_n).$$

Since marginals are Fréchet, by taking  $\mathbf{x} = (x^{(1)}, \infty, \dots, \infty)$

$$H_{\alpha}^{(1)}(x^{(1)}) = (H_{\alpha}^{(1)})^n(a_n x^{(1)} + b_n),$$

## 1.1 A survival guide for high-dimensional extremal dependence modeling

which implies that  $a_n^{(1)} = n^{1/\alpha}$  and  $b_n^{(1)} = 0$ . Then we deduce that  $a_n^{(j)} = n^{1/\alpha}$  and  $b_n^{(j)} = 0$  for any  $j \in \{1, \dots, d\}$ . Thus, for any  $n \in \mathbb{N}$  and  $\mathbf{x} \in \mathbb{R}^d$ , the following holds:

$$H_{\mathbf{X}}^n(n^{1/\alpha}\mathbf{x}) = H_{\mathbf{X}}(\mathbf{x}).$$

This yields

$$H_{\mathbf{X}}^{n/m}((n/m)^{1/\alpha}\mathbf{x}) = H_{\mathbf{X}}(\mathbf{x}), \quad n, m \in \mathbb{N}, \mathbf{x} > 0.$$

Choose now  $n, m$  such that  $n/m \rightarrow t^\alpha$ . Then the continuity of  $H$  implies

$$H_{\mathbf{X}}^{t^\alpha}(t\mathbf{x}) = H_{\mathbf{X}}(\mathbf{x}), \quad t > 0, \mathbf{x} \in \mathbb{R}^d.$$

From (1.4), we obtain for any  $\mathbf{x} > 0$  and any  $t > 0$ ,

$$\Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus [0, \mathbf{x}]) = t^\alpha \Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus [0, t\mathbf{x}]) = t^\alpha \Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus t[0, \mathbf{x}]).$$

This equation can be readily be extended to hold for all rectangles contained in  $\mathcal{E}_+$ . The equality

$$\Lambda_{\mathbf{X}}(B) = t^\alpha \Lambda_{\mathbf{X}}(tB), \tag{1.5}$$

thus holds on a generalising class closed under intersections and is therefore true for any Borel subset of  $E$  using Dynkin's theorem. Denote by  $\|\cdot\|$  an arbitrary norm of  $\mathbf{x} \in \mathbb{R}^d$ . From (1.5), we obtain for any  $t > 0$  and any Borel subset  $A$  of the unit sphere  $\mathbb{S}_{\mathcal{E}_+} := \{\mathbf{z} \in \mathcal{E}_+ : \|\mathbf{z}\| = 1\}$  in  $\mathcal{E}_+$ .

$$\begin{aligned} \Lambda_{\mathbf{X}}\left(\left\{\mathbf{x} \in \mathcal{E}_+ : \|\mathbf{x}\| \geq t, \frac{\mathbf{x}}{\|\mathbf{x}\|} \in A\right\}\right) &= t^{-\alpha} \Lambda_{\mathbf{X}}\left(\left\{\frac{\mathbf{x}}{t} \in \mathcal{E}_+ : \|\mathbf{x}\| \geq t, \frac{\mathbf{x}}{\|\mathbf{x}\|} \in A\right\}\right) \\ &= t^{-\alpha} \Lambda_{\mathbf{X}}\left(\left\{\mathbf{y} \in \mathcal{E}_+ : \|\mathbf{y}\| > 1, \frac{\mathbf{y}}{\|\mathbf{y}\|} \in A\right\}\right) \\ &= t^{-\alpha} \Phi(A), \end{aligned}$$

where  $\Phi$  is the spectral measure. Define the bijective function  $T$  as the transformation of a vector onto its polar coordinates with the norm  $\|\cdot\|$ . From the above equation, we deduce that the measure  $(T\Lambda_{\mathbf{X}})(B) = \Lambda_{\mathbf{X}}(T^{-1}(B))$  induced by  $\Lambda_{\mathbf{X}}$  and  $T$ , satisfies

$$\begin{aligned} (T\Lambda_{\mathbf{X}})([t, \infty) \times A) &= \Lambda_{\mathbf{X}}\left(\left\{\mathbf{x} \in \mathcal{E}_+ : \|\mathbf{x}\| \geq t, \frac{\mathbf{x}}{\|\mathbf{x}\|} \in A\right\}\right) = t^{-\alpha} \Phi(A) \\ &= \int_A \int_{[t, \infty)} \alpha^{-1} r^{-(\alpha+1)} dr \Phi(d\mathbf{a}) = \int_{[t, \infty) \times A} \alpha^{-1} r^{-(\alpha+1)} dr \Phi(d\mathbf{a}). \end{aligned}$$

We have for an arbitrary vector  $\mathbf{z} \in \mathbb{R}^d$

$$\begin{aligned} T(\mathcal{E}_+ \setminus [0, \mathbf{x}]) &= T\left(\left\{\mathbf{y} \in \mathcal{E}_+ : y^{(j)} > x^{(j)}, \text{ for some } j = 1, \dots, d\right\}\right) \\ &= \left\{(r, \mathbf{a}) \in [0, \infty) \times \mathbb{S}_{\mathcal{E}_+} : (ra^{(j)}) > x^{(j)}, \text{ for some } j = 1, \dots, d\right\} \\ &= \left\{(r, \mathbf{a}) \in [0, \infty) \times \mathbb{S}_{\mathcal{E}_+} : r > \bigwedge_{j=1}^d \frac{x^{(j)}}{a^{(j)}}\right\}. \end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus [0, \mathbf{x}]) &= (T\Lambda_{\mathbf{X}})T(\mathcal{E}_+ \setminus [0, \mathbf{x}]) \\
&= (T\Lambda_{\mathbf{X}}) \left\{ (r, \mathbf{a}) \in [0, \infty) \times \mathbb{S}_{\mathcal{E}_+} : r > \bigwedge_{j=1}^d \frac{x^{(j)}}{a^{(j)}} \right\} \\
&= \int_{\mathbb{S}_{\mathcal{E}_+}} \int_{\bigwedge_{j=1}^d \frac{x^{(j)}}{a^{(j)}}, \infty[} \alpha^{-1} r^{-(\alpha+1)} dr \Phi(d\mathbf{a}) \\
&= \int_{\mathbb{S}_{\mathcal{E}_+}} \left( \frac{1}{\bigwedge_{j=1}^d \frac{x^{(j)}}{a^{(j)}}} \right)^{-\alpha} \Phi(d\mathbf{a}) = \int_{\mathbb{S}_{\mathcal{E}_+}} \left( \bigvee_{j=1}^d \frac{a^{(j)}}{x^{(j)}} \right)^{\alpha} \Phi(d\mathbf{a}).
\end{aligned}$$

We have established, therefore, the de Haan-Resnick representation theorem that we state below.

**Theorem 1.1.7.** *Let  $\mathbf{X}$  a  $d$ -dimensional random vector with joint distribution function  $H_{\mathbf{X}}$ . The following statements are equivalent:*

- (a)  $H_{\mathbf{X}}$  is a multivariate extreme value distribution with Fréchet marginals  $H_{\alpha}^{(j)}(x) = e^{-x^{-\alpha}}$ ,  $x > 0$ ,  $\alpha > 0$ ,  $j = 1, \dots, d$ .
- (b) There exists a finite measure  $\Phi$  on

$$\mathbb{S}_{\mathcal{E}_+} = \{\mathbf{y} \in \mathcal{E}_+ : \|\mathbf{y}\| = 1\}$$

satisfying

$$\int_{\mathbb{S}_{\mathcal{E}_+}} (a^{(j)})^{\alpha} \Phi(d\mathbf{a}) = 1, \quad 1 \leq j \leq d$$

such that for  $\mathbf{x} \in \mathbb{R}_+^d$

$$H_{\mathbf{X}}(\mathbf{x}) = \exp \left\{ - \int_{\mathbb{S}_{\mathcal{E}_+}} \left( \bigvee_{j=1}^d \frac{a^{(j)}}{x^{(j)}} \right)^{\alpha} \Phi(d\mathbf{a}) \right\}.$$

The theorem mentioned above extends the assertion of Lemma 1.1.1 by incorporating in Theorem 1.1.7 the measure  $\Phi = \sum_{a=1}^K \|A_{\cdot a}\| \delta_{\{A_{\cdot a}/\|A_{\cdot a}\|\}}$  with  $\alpha = 1$ . When normalising all marginals to standard Fréchet distribution, the stable tail dependence function is defined as follows:

$$L(v^{(1)}, \dots, v^{(d)}) = \Lambda_{\mathbf{X}}(\mathcal{E}_+ \setminus [0, 1/\mathbf{v}]) = \int_{\mathbb{S}_{\mathcal{E}_+}} \bigvee_{j=1}^d a^{(j)} v^{(j)} \Phi(d\mathbf{a}).$$

Expressed in terms of the original max-stable distribution function  $H_{\mathbf{X}}$ , it is given by:

$$L(v^{(1)}, \dots, v^{(d)}) = -\ln H_{\mathbf{X}} \left\{ (H^{(1)})^{\leftarrow}(e^{-v^{(1)}}), \dots, (H^{(d)})^{\leftarrow}(e^{-v^{(d)}}) \right\}, \quad \mathbf{v} \in [0, \infty)^d.$$

Conversely, we can reconstruct a max-stable distribution function  $H_{\mathbf{X}}$ , from its margins  $H^{(j)}$  and its stable tail dependence function  $L$  through

$$-\ln(H_{\mathbf{X}}(\mathbf{x})) = L \left( -\ln \left( H^{(1)}(x^{(1)}) \right), \dots, -\ln \left( H^{(d)}(x^{(d)}) \right) \right), \quad \mathbf{x} \in \mathbb{R}^d.$$

## 1.1 A survival guide for high-dimensional extremal dependence modeling

A stable tail dependence function  $L$  has the following properties.

**Proposition 1.1.3.** *Let  $\mathbf{X}$  a multivariate extreme value distribution with Fréchet marginals  $H_\alpha^{(j)}(x) = e^{-x^{-\alpha}}$ ,  $x > 0$ ,  $\alpha > 0$ ,  $j = 1, \dots, d$ , with stable tail dependence function  $L$ . Then  $L$  has the following properties.*

- (a) **Homogeneity**  $L(s \cdot) = sL(\cdot)$  for  $0 < s < \infty$ ;
- (b) **Groundedness**  $L(\mathbf{e}^{(j)}) = 1$  for  $j = 1, \dots, d$ , where  $\mathbf{e}^{(j)}$  is the unit vector in  $\mathbb{R}^d$ .
- (c) **Fréchet-Hoeffding bounds**  $\bigvee_{j=1}^d v^{(j)} \leq L(\mathbf{v}) \leq \sum_{j=1}^d v^{(j)}$  for  $\mathbf{v} \in [0, \infty)^d$ . The upper and lower Fréchet-Hoeffding bounds are itself valid stable tail dependence functions, the lower bound corresponds to complete dependence  $H_{\mathbf{X}}(\mathbf{x}) = \bigwedge_{j=1}^d H^{(j)}(x^{(j)})$ , whereas the upper bound corresponds to independence  $H_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^d H^{(j)}(x^{(j)})$ .
- (d) **Convexity**  $L$  is convex, that is  $L(\lambda \mathbf{v} + (1 - \lambda)\mathbf{w}) \leq \lambda L(\mathbf{v}) + (1 - \lambda)L(\mathbf{w})$  for  $\lambda \in (0, 1)$ .

Note that, except for the bivariate case, properties (a)-(d) do not characterise the class of stable tail dependence functions. That is, a function  $L$  satisfying properties (a)-(d) is not necessarily a stable tail dependence function. To illustrate this with a counterexample in the trivariate case, consider  $L(v^{(1)}, v^{(2)}, v^{(3)}) = (v^{(1)} + v^{(2)}) \vee (v^{(2)} + v^{(3)}) \vee (v^{(1)} + v^{(3)})$ . It is evident that properties (a)-(d) are fulfilled. However,  $L$  cannot be a stable tail dependence function because  $L(1, 1, 0) = L(1, 0, 1) = L(0, 1, 1) = 2$ , which would imply pairwise independence and, as we will show below, full independence. This is contradiction with  $L(1, 1, 1) = 2 \neq 3$ .

**Proposition 1.1.4.** *Suppose  $\mathbf{X}$  has max-stable distribution. The following are equivalent:*

- (A) *The components of  $\mathbf{X}$ , namely  $X^{(1)}, \dots, X^{(d)}$ , are independent random variables, i.e.,*

$$H_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^d H^{(j)}(x^{(j)}), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

- (B) *There exists  $\mathbf{y} \in \mathbb{R}^d$  with  $0 < H^{(j)}(y^{(j)}) < 1$  for all  $j = 1, \dots, d$  such that*

$$H_{\mathbf{X}}(\mathbf{y}) = \prod_{j=1}^d H^{(j)}(y^{(j)}).$$

- (C) *The components of  $\mathbf{X}$  are pairwise independent : for every  $1 \leq i < j \leq d$*

$$X^{(i)} \text{ and } X^{(j)},$$

*are independent random variables, i.e.,  $\forall \mathbf{x} \in \mathbb{R}^d$*

$$H^{(i,j)}(\mathbf{x}^{(i,j)}) = H^{(i)}(x^{(i)})H^{(j)}(x^{(j)}).$$

**Proof** We only show (A)  $\iff$  (B). For a proof of (A)  $\iff$  (C), see (Resnick, 2008, Corollary 5.25). The direct sense (A)  $\implies$  (B) is direct and holds more generally for a random vector  $\mathbf{X}$ , not necessarily max-stable. The direction (B)  $\implies$  (A) may be proved as follows. Denoting  $v^{(j)} = -\ln H^{(j)}(y^{(j)})$ , we must have

$$\int_{\mathbb{S}_{\mathcal{E}_+}} \left\{ \sum_{j=1}^d a^{(j)} v^{(j)} - \bigvee_{j=1}^d a^{(j)} v^{(j)} \right\} \Phi(d\mathbf{a}) = 0.$$

Since the integrand is non-negative, the  $\Phi$ -measure of the set where it is positive must be zero. But then, since  $0 < v^{(j)} < \infty$  for all  $j = 1, \dots, d$ , the set  $\{\mathbf{a} \in \mathbb{S}_{\mathcal{E}_+}, \exists 1 \leq i < j \leq d, a^{(i)} >$

$0, a^{(j)} > 0\}$  must have  $\Phi$ -measure zero. Consequently,  $\Phi$  is concentrated on the complement of the set above which is equal to  $\{\mathbf{e}^{(1)}, \dots, \mathbf{e}^{(d)}\}$ . We must have  $\Phi(\mathbf{e}^{(j)}) = 1$  for  $j = 1, \dots, d$ , which implies independence (see (Resnick, 2008, Corollary 5.25)).  $\square$

From Theorem 1.1.7, we can deduce that a  $d$ -variate max-infinitely divisible distribution function  $H_{\mathbf{X}}$  with standard Fréchet marginals  $H_1^{(j)}$ ,  $j = 1, \dots, d$  can be expressed in terms of the Pickands dependence function  $\mathcal{A} : \Delta_{d-1} \rightarrow [0, \infty)$ , where the domain  $\Delta_{d-1}$  is defined on the  $(d - 1)$ -dimensional unit simplex:

$$\Delta_{d-1} := \left\{ (w^{(1)}, \dots, w^{(d-1)}) \in [0, 1]^{d-1} : \sum_{j=1}^{d-1} w^{(j)} \leq 1 \right\}.$$

For  $\mathbf{x} = (x^{(1)}, \dots, x^{(d)}) \in [0, \infty)^d$ ,  $\mathbf{x} \neq 0$ , we have

$$\begin{aligned} H_{\mathbf{X}}(\mathbf{x}) &= \exp \left\{ - \int_{\mathbb{S}_{\mathcal{E}_+}} \bigvee_{j=1}^d a^{(j)} v^{(j)} \Phi(d\mathbf{a}) \right\} \\ &= \exp \left\{ - (v^{(1)} + \dots + v^{(d)}) \int_{\mathbb{S}_{\mathcal{E}_+}} \bigvee_{j=1}^d a^{(j)} \frac{v^{(j)}}{v^{(1)} + \dots + v^{(d)}} \Phi(d\mathbf{a}) \right\} \\ &= \exp \left\{ - (v^{(1)} + \dots + v^{(d)}) \mathcal{A} \left( \frac{v^{(1)}}{v^{(1)} + \dots + v^{(d)}}, \dots, \frac{v^{(d)}}{v^{(1)} + \dots + v^{(d)}} \right) \right\}, \end{aligned}$$

where  $\Phi$  is the spectral measure defined in Theorem 1.1.7 and

$$\mathcal{A}(w^{(1)}, \dots, w^{(d-1)}) = \int_{\mathbb{S}_{\mathcal{E}_+}} \max \left( a^{(1)} w^{(1)}, \dots, a^{(d-1)} w^{(d-1)}, a^{(d)} \left( 1 - \sum_{j=1}^{d-1} w^{(j)} \right) \right) \Phi(d\mathbf{a}),$$

is the Pickands dependence function.

If the random vector  $\mathbf{X}$  follows the max-stable distribution function  $H_{\mathbf{X}}$ , then the scenarios where  $\mathcal{A}(\mathbf{w}) = 1$  and  $\mathcal{A}(\mathbf{w}) = \max\{w^{(1)}, \dots, w^{(d-1)}, 1 - \sum_{j=1}^{d-1} w^{(j)}\}$  characterises the cases of independence and complete dependence of the random variables  $X^{(1)}, \dots, X^{(d)}$ . Below, we outline some important properties of the Pickands dependence function.

**Proposition 1.1.5.** *Let  $\mathbf{X}$  a max-stable random vector with a Pickands dependence function  $\mathcal{A}$ , then the latter function has the following properties.*

- (a) **Continuity** *The Pickands dependence function is continuous.*
- (b) **Groundedness**  $\mathcal{A}(\mathbf{e}^{(j)}) = 1$  and the  $\mathbf{e}^{(j)}$ ,  $1 \leq j \leq d - 1$  are the extremal points of the convex set  $\Delta_{d-1}$ .
- (c) **Fréchet-Hoeffding bounds**  $\frac{1}{d} \leq \max\{w^{(1)}, \dots, w^{(d-1)}, 1 - \sum_{j=1}^{d-1} w^{(j)}\} \leq \mathcal{A}(\mathbf{w}) \leq 1$  for any  $\mathbf{w} \in \Delta_{d-1}$ .
- (d) **Convexity** *The function  $\mathcal{A}$  is convex, that is, for  $\mathbf{w}_1, \mathbf{w}_2 \in \Delta_{d-1}$  and  $\lambda \in [0, 1]$*

$$\mathcal{A}(\lambda \mathbf{w}_1 + (1 - \lambda) \mathbf{w}_2) \leq \lambda \mathcal{A}(\mathbf{w}_1) + (1 - \lambda) \mathcal{A}(\mathbf{w}_2)$$

- (e) **Extreme value copula** - The copula function  $C_\infty$  of the extreme value distribution  $H_{\mathbf{X}}$ , with Pickands dependence function, the so-called extreme value copula, is

$$\begin{aligned} C_\infty(\mathbf{u}) &= H_{\mathbf{X}}\left(\ln(u^{(1)}), \dots, \ln(u^{(d)})\right) \\ &= \left(\prod_{j=1}^d u^{(j)}\right)^{\mathcal{A}\left(\frac{\ln(u^{(1)})}{\sum_{j=1}^d \ln u^{(j)}}, \dots, \frac{\ln(u^{(d-1)})}{\sum_{j=1}^d \ln u^{(j)}}\right)}, \quad \mathbf{u} \in (0, 1]^d. \end{aligned}$$

We refer the reader to (Falk et al., 2010, Section 4.3) for a proof of these results.

The extreme value copula obviously satisfies for any  $t > 0$

$$C_\infty(\mathbf{u}^t) = C_\infty(\mathbf{u})^t,$$

which is a characterising property of extreme value copula (see Gudendorf and Segers (2010) for an extensive overview). Take  $d = 2$ , we can notice that  $\chi(1, 2) = 2(1 - \mathcal{A}(1/2))$  and thus, the convexity of  $\mathcal{A}$  implies that  $\chi(1, 2) = 0$  is equivalent to the condition  $\mathcal{A}(z) = 1$ ,  $z \in (0, 1)$ .

### 1.1.3 Multivariate regularly varying random vectors

Up to this point, we have characterised the extremes of the multivariate random vector  $\mathbf{X}$  by examining the scale-normalised componentwise maxima of independent copies. Another approach involves studying the distribution of the scale-normalized exceedances

$$u^{-1}\mathbf{X} \mid \bigvee_{j=1}^d X^{(j)} > u,$$

of the random vector  $\mathbf{X}$ , conditioned on the event that at least one component  $X^{(j)}$  exceeds a large threshold. We only mention this second approach without providing specific details (for a more technical exposition, we refer to Rootzén and Tajvidi (2006)). The only conceivable limits of these peak-over-threshold as  $u \rightarrow \infty$  are multivariate Pareto distributions. The probability laws of these distributions are induced by a homogeneous measure  $\Lambda_{\mathbf{X}}$  on the (non-rectangular) set  $\mathcal{L} = \mathcal{E}_+ \setminus [0, 1]^d$  and take the form  $\mathbb{P}_{\mathcal{L}}(d\mathbf{y}) = \Lambda_{\mathbf{X}}(d\mathbf{y})/\Lambda_{\mathbf{X}}(\mathcal{L})$ . An apparent connection between these two approaches is the exponent measure  $\Lambda_{\mathbf{X}}$ , which characterises the distribution function of both multivariate max-stable distribution and multivariate Pareto distribution. Indeed, this connection is established through a fundamental limiting result that links the two approaches via regular variation.

In this section, we introduce the concept of regular variation for finite-dimensional random vector. To do so, we employ the concept of vague convergence on  $\mathcal{E}$  and utilise the exponent measure, which characterises the extremal behaviour of the vector in a given sector of  $\mathcal{E}$ . An important characterisation of multivariate regularly varying random vectors is the dichotomy between extremal dependence and extremal independence. We delve into the problem of obtaining the limiting conditional distribution of the random vector given an extrem event. When such limits exist, they can be used to define various extremal dependence measures, which be of interest throughout this thesis.

Let  $\mathbb{R}^d$  be endowed with its usual topology (defined by an arbitrary norm  $\|\cdot\|$ ) which makes it Polish, and the related Borel  $\sigma$ -field. Consider  $\mathcal{C}_K(\mathbb{R}^d)$  the set of functions defined on  $\mathbb{R}^d$



## Introduction

---

with compact support and  $\mathcal{C}_b(\mathbb{R}^d)$  the set of functions defined on  $\mathbb{R}^d$  which are continuous and bounded. There are numerous concepts of convergence of measures on a Polish space, each characterised by a particular class of test functions. If we test convergence on bounded continuous functions, then the sequence of measures  $(\mu_n, n \in \mathbb{N})$  converge weakly to  $\mu$  if

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} f(\mathbf{x}) \mu_n(d\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}) \mu(d\mathbf{x}), \quad \forall f \in \mathcal{C}_b(\mathbb{R}^d).$$

If we choose continuous function with compact support, the classical notion of vague convergence is recovered, i.e.,

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} f(\mathbf{x}) \mu_n(d\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}) \mu(d\mathbf{x}), \quad \forall f \in \mathcal{C}_K(\mathbb{R}^d).$$

These two notions of convergence of measures are equivalent in  $\mathbb{R}^d$  provided that  $\lim_{n \rightarrow \infty} \mu_n(\mathbb{R}^d) = \mu(\mathbb{R}^d)$ . It appears that what really matters for vague convergence is the notion of bounded sets, which is not a topological notion, but can be defined intrinsically without reference to a metric.

**Definition 1.1.6.** Let  $E$  be a set. A boundedness  $\mathcal{B}$  on  $E$  is a collection of subsets of  $E$ , called bounded sets, with the following properties,

- (a) a finite union of bounded set is bounded;
- (b) a subset of a bounded set is a bounded set.

Using this notion of boundedness, we may define the new notion of vague convergence due to [Hult and Lindskog \(2006\)](#) and further developed by [Lindskog et al. \(2014\)](#); [Segers et al. \(2017\)](#).

**Definition 1.1.7.**

- A set  $A$  is said to be separated from 0 if there exists an open set  $U$  such that  $0 \in U$  and  $A \subset U^c$ .
- The collection of all sets separated from 0 is a boundedness and denoted  $\mathcal{B}_0$ .
- A measure  $\mu$  on  $\mathbb{R}^d \setminus \{0\}$  is said to be  $\mathcal{B}_0$ -boundedly finite if  $\mu(A) < \infty$  for all Borel measurable set  $A$  separated from 0.
- A sequence of  $\mathcal{B}_0$ -boundedly finite measures  $\{\mu_t, t \in \mathcal{T}\}$  (with  $\mathcal{T} = \mathbb{N}$  or  $\mathcal{T} = \mathbb{R}_+$ ) on  $\mathcal{E}$  converges  $\mathcal{B}_0$ -vaguely<sup>#</sup> to a measure  $\mu$  on  $\mathcal{E}$ , written  $\mu_t \xrightarrow{v^\#} \mu$ , if

$$\lim_{t \rightarrow \infty} \int_{\mathbb{R}^d} f(\mathbf{x}) d\mu_t(\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}) d\mu(\mathbf{x}),$$

for all bounded continuous functions  $f$  with  $\mathcal{B}_0$ -bounded support.

Many useful applications of weak convergence rely on the Portmanteau theorem. Next, we derive the corresponding of this theorem translated for vague<sup>#</sup> convergence.

**Theorem 1.1.8.** *Let  $\{\mu_n, n \in \mathbb{N}\}$  be a sequence of  $\mathcal{B}_0$ -boundedly finite measure. The following statement are equivalent:*

- (A)  $\mu_n \xrightarrow{v^\#} \mu$ ;
- (B)  $\lim_{n \rightarrow \infty} \mu_n(A) = \mu(A)$  for all bounded Borel sets such that  $\mu(\partial A) = 0$ ;

(C) for all bounded Borel sets,

$$\mu(A^\circ) \leq \underline{\lim} \mu_n(A) \leq \overline{\lim} \mu_n(A) \leq \mu(\bar{A}).$$

We can now define the regular variation of a random vector.

**Definition 1.1.8.** A  $d$ -dimensional positive random vector  $\mathbf{X}$  is regularly varying if there exists a non-zero  $\mathcal{B}_0$ -boundedly finite measure  $\Lambda_{\mathbf{X}}$  on  $\mathcal{E}_+$ , called an exponent measure of  $\mathbf{X}$ , and a scaling sequence  $\{c_n\}$  such that the sequence of measure  $n\mathbb{P}\{c_n\mathbf{X} \in \cdot\}$  converges  $\mathcal{B}_0$ -vaguely<sup>#</sup> on  $\mathbb{R}_+^d \setminus \{0\}$  to the measure  $\Lambda_{\mathbf{X}}$ .

**Theorem 1.1.9.** The following statements are equivalent:

- (A) The vector  $\mathbf{X}$  is regularly varying in the sense of Definition 1.1.8.
- (B) There exists  $\alpha > 0$ , a function  $g : (0, \infty) \rightarrow (0, \infty)$  which is regularly varying at infinity with index  $\alpha$ , i.e., for  $t > 0$

$$\lim_{x \rightarrow \infty} \frac{g(tx)}{g(x)} = t^\alpha,$$

and a  $\mathcal{B}_0$ -boundedly finite non-zero measure  $\Lambda_{\mathbf{X}}$  such that  $g(t)\mathbb{P}\{t^{-1}\mathbf{X} \in \cdot\}$  converges  $\mathcal{B}_0$ -vaguely<sup>#</sup> to  $\Lambda_{\mathbf{X}}$  as  $t \rightarrow \infty$ .

If (A) and (B) hold, the sequence  $\{c_n\}$  is regularly varying at infinity with index  $1/\alpha$  and the measure  $\Lambda_{\mathbf{X}}$  is  $(-\alpha)$ -homogeneous, i.e., for  $y > 0$  and  $A$  separated from 0,

$$\Lambda_{\mathbf{X}}(yA) = y^{-\alpha} \Lambda_{\mathbf{X}}(A).$$

We proceed with two elementary examples of regularly varying random vectors and provide their exponent measures.

**Example 1.1.4 (Independence).** Assume that  $\mathbf{X}$  represents a positive  $d$ -dimensional random vector with independently and identically distributed regularly varying components. In this context, the exponent measure  $\Lambda_{\mathbf{X}}$  concentrates on the axes. Opting for the selection  $c_n = Q_{X^{(1)}}(1/n)$ , where  $Q_{X^{(1)}}$  is defined in (1.1), produces the specific outcome:

$$\Lambda_{\mathbf{X}}(du^{(1)}, \dots, du^{(d)}) = \sum_{j=1}^d \delta_0(du^{(1)}) \otimes \dots \otimes \Lambda_\alpha(du^{(j)}) \otimes \dots \otimes \delta_0(du^{(d)}),$$

where  $\Lambda_\alpha(du) = \alpha u^{-\alpha-1} du$ . In particular, for  $y > 0$  and  $j = 1, \dots, d$ ,  $\Lambda_{\mathbf{X}}(\{x : x^{(j)} > y\}) = y^{-\alpha}$ . This means that only one of the  $d$  components  $X^{(1)}, \dots, X^{(d)}$  can be extremely large at a time.

**Example 1.1.5 (Total dependence).** Assume we have a random vector  $\mathbf{X}$  defined as  $\mathbf{X} = (X^{(1)}, \dots, X^{(1)})$ , where  $X^{(1)}$  exhibits regular variation with an index of  $\alpha$ . Additionally, consider positive constants  $v^{(j)}$  for each components with  $j = 1, \dots, d$ . Now, with the specific choice of

## Introduction

---

$c_n = Q_{X^{(1)}}(1/n)$ , where the quantile function  $Q_{X^{(1)}}$  is defined in (1.1), we have

$$\begin{aligned} \Lambda_{\mathbf{X}}([0, \mathbf{v}]^c) &= \lim_{n \rightarrow \infty} n\mathbb{P} \left\{ X^{(1)} > c_n v^{(j)}, \text{ for some } j = 1, \dots, d \right\} \\ &= \left( \bigwedge_{j=1}^d v^{(j)} \right)^{-\alpha} = \left( \bigvee_{j=1}^d \frac{1}{v^{(j)}} \right)^{\alpha}. \end{aligned}$$

Thus we conclude that the exponent measure is concentrated on the line  $v^{(1)} = \dots = v^{(d)}$ .

It is essential to highlight the connection between the exponent measure and the convergence of scaled componentwise maxima. For sequences comprising i.i.d. random vectors with non-negative components, this convergence aligns with the concept of multivariate regular variation. This correlation elucidates the use of the term ‘‘exponent measure’’ in this context.

**Theorem 1.1.10.** *Let  $\mathbf{X}_i$ ,  $i \geq 1$  be independent copies of the regularly varying vector  $\mathbf{X}$  with non-negative components and exponent measure  $\Lambda_{\mathbf{X}}$  associated to the scaling sequence  $\{c_n\}$ . Then for  $\mathbf{u} \in \mathcal{E}_+$ ,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \bigvee_{i=1}^n \mathbf{X}_i \leq c_n \mathbf{u} \right\} = e^{-\Lambda_{\mathbf{X}}([0, \mathbf{u}]^c)}.$$

**Proof** If  $\mathbf{X}_i$  are i.i.d., then

$$\mathbb{P} \left\{ \bigvee_{i=1}^n \mathbf{X}_i \leq c_n \mathbf{u} \right\} = \left\{ 1 - n^{-1} n\mathbb{P} \left\{ \mathbf{X} \in c_n [0, \mathbf{u}]^c \right\} \right\}^n.$$

The set  $[0, \mathbf{u}]^c$  is separated from 0 in  $\mathbb{R}_+^d \setminus \{0\}$ , therefore

$$\lim_{n \rightarrow \infty} n\mathbb{P} \left\{ \mathbf{X} \in c_n [0, \mathbf{u}]^c \right\} = \Lambda_{\mathbf{X}}([0, \mathbf{u}]^c).$$

□

The limiting distribution which appears in Theorem 1.1.10 is indeed a max-stable distribution since the componentwise maxima of a finite number of i.i.d. random vectors with this distribution will again have the same distribution after scaling. This is an immediate consequence of the fact that it is a limiting distribution of maxima. The marginal distributions are Fréchet distribution with the distribution function

$$H_{\alpha, c}^{(j)} = e^{-c^\alpha x^{-\alpha}}, \quad x \geq 0,$$

where  $\alpha > 0$  is the tail index and the constant  $c > 0$  will be referred to as its scale parameter. Since if  $X^{(j)}$  has distribution  $H_{\alpha, c}^{(j)}$ , then  $c^{-1}X^{(j)}$  has distribution  $H_{\alpha, 1}^{(j)} := H_{\alpha}^{(j)}$  as detailed in Section 1.1.2.

Now, we provide conditions for a transformation of a regularly varying vector  $g(\mathbf{X})$  to also be regularly varying vector. Before delving into these conditions, it is pertinent to recall a useful lemma available in Lindskog et al. (2014).

**Lemma 1.1.2.** *Let  $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$  be a measurable mapping and consider the following statements:*

- (a) *The mapping  $g$  is continuous at  $0_{\mathbb{R}^d}$  and  $g(0_{\mathbb{R}^d}) = 0_{\mathbb{R}^k}$ ;*
- (b) *For every Borel set  $A$  with  $0_{\mathbb{R}^k} \notin A$  it holds that  $0_{\mathbb{R}^d} \notin \overline{g^{-1}(A)}$  in  $\mathbb{R}^d$ ;*

## 1.1 A survival guide for high-dimensional extremal dependence modeling

(c) For every  $\epsilon > 0$  there exists  $\delta > 0$  such that  $B(0_{\mathbb{R}^d}, \delta) \subset g^{-1}(B(0_{\mathbb{R}^k}, \epsilon))$ .

**Proof** To prove (b)  $\iff$  (c) notice that  $0_{\mathbb{R}^k} \notin \overline{A}$  if and only if there exists  $\epsilon > 0$  such that  $A \subset \mathbb{R}^k \setminus B(0_{\mathbb{R}^k}, \epsilon)$  and that  $0_{\mathbb{R}^d} \notin \overline{g^{-1}(A)}$  if and only if there exists  $\delta > 0$  such that  $g^{-1}(A) \subset \mathbb{R}^d \setminus B(0_{\mathbb{R}^d}, \delta)$ . Hence (b) holds if and only if for every  $\epsilon > 0$ , there exists  $\delta > 0$  such that  $g^{-1}(\mathbb{R}^k \setminus B(0_{\mathbb{R}^k}, \epsilon)) \subset \mathbb{R}^d \setminus B(0_{\mathbb{R}^d}, \delta)$  if and only if  $B(0, \delta) \subset g^{-1}(B(0_{\mathbb{R}^k}, \epsilon))$ . Taking complements shows that  $g^{-1}(\mathbb{R}^k \setminus B(0_{\mathbb{R}^k}, \epsilon)) \subset \mathbb{R}^d \setminus B(0_{\mathbb{R}^d}, \delta)$  if and only if  $B(0_{\mathbb{R}^d}, \delta) \subset g^{-1}(B(0_{\mathbb{R}^k}, \epsilon))$ .

To obtain (a)  $\iff$  (c), note that (a) implies that for every  $\epsilon > 0$  there exists  $\delta > 0$  such that  $g(B(0_{\mathbb{R}^d}, \delta)) \subset B(0_{\mathbb{R}^k}, \epsilon)$ , which implies that  $g^{-1}(g(B(0_{\mathbb{R}^d}, \delta))) \subset g^{-1}(B(0_{\mathbb{R}^k}, \epsilon))$ . Since  $g(g^{-1}(B(0_{\mathbb{R}^d}, \delta))) \subset B(0_{\mathbb{R}^d}, \delta)$  holds for any  $g$ , it follows that (c) holds.

Now, (c) implies that for every  $\epsilon > 0$  there exists  $\delta > 0$  such that  $g(B(0_{\mathbb{R}^d}, \delta)) \subset B(0_{\mathbb{R}^k}, \delta)$  which implies  $g^{-1}(g(B(0_{\mathbb{R}^d}, \delta))) \subset g^{-1}(B(0_{\mathbb{R}^k}, \delta))$ . Since  $B(0_{\mathbb{R}^k}, \delta) \subset g^{-1}(g(B(0_{\mathbb{R}^k}, \delta)))$  holds for any  $g$  it follows that (a) holds.  $\square$

**Proposition 1.1.6.** Let  $\mathbf{X}$  be a regularly varying random vector with tail index  $\alpha$ , exponent measure  $\Lambda_{\mathbf{X}}$  on  $\mathbb{R}_+^d \setminus \{0\}$  associated to the scaling sequence  $\{c_n\}$ . Let  $g$  be a continuous map from  $\mathbb{R}_+^d$  to  $\mathbb{R}_+^k$  such that  $g(t\mathbf{x}) = t^\gamma g(\mathbf{x})$  for some  $\gamma > 0$ . If the measure is not identically zero on  $\mathbb{R}_+^k \setminus \{0\}$ , then  $g(\mathbf{X})$  is regularly varying with tail index  $\alpha/\gamma$  and exponent measure  $\Lambda_{\mathbf{X}} \circ g^{-1}$  associated to the scaling sequence  $\{c_n^\gamma\}$

**Proof** Let  $\Lambda_n = n\mathbb{P}\{c_n\mathbf{X}^{-1} \in \cdot\}$ . By hypothesis,  $\Lambda_n$  converges vaguely<sup>#</sup> to  $\Lambda_{\mathbf{X}}$  in  $\mathbb{R}_+^d \setminus \{0\}$ . The continuity and homogeneity of  $g$  implies that  $g(0_{\mathbb{R}^d}) = 0_{\mathbb{R}^k}$ . Let us consider a Borel set  $A$  with  $0_{\mathbb{R}^k} \notin \overline{A}$  and  $\Lambda_{\mathbf{X}}(g^{-1}(\partial A)) = 0$ . Since  $\partial g^{-1}(A) \subset g^{-1}(\partial A) \cup D_g$ , where  $D_g \subset \mathbb{R}^d$  be the set of discontinuity of  $g$ , we have  $\Lambda_{\mathbf{X}}(\partial g^{-1}(A)) \leq \Lambda_{\mathbf{X}}(g^{-1}(\partial A)) + \Lambda_{\mathbf{X}}(D_g) = 0$ . Since  $\Lambda_n \xrightarrow{v^\#} \Lambda_{\mathbf{X}}$ ,  $\Lambda_{\mathbf{X}}(\partial g^{-1}(A)) = 0$ , and, by Lemma 1.1.2,  $0_{\mathbb{R}^d} \notin \overline{g^{-1}(A)}$ , it follows by Theorem 1.1.8 that  $\Lambda_n(g^{-1}(A)) \xrightarrow{v^\#} \Lambda_{\mathbf{X}}(g^{-1}(A))$ . Hence  $\Lambda_n \circ g^{-1} \xrightarrow{v^\#} \Lambda_{\mathbf{X}} \circ g^{-1}$ . It suffices to identify the scaling sequence. If  $\Lambda_{\mathbf{X}} \circ g^{-1}(\partial B) = 0$ , the inclusion  $\partial(g^{-1}(B)) \subset g^{-1}(\partial B)$  implies

$$\lim_{n \rightarrow \infty} n\mathbb{P}\{c_n^{-\gamma} g(\mathbf{X}) \in B\} = \lim_{n \rightarrow \infty} n\mathbb{P}\{c_n^{-1} \mathbf{X} \in g^{-1}(B)\} = \Lambda_{\mathbf{X}} \circ g^{-1}(B).$$

Therefore,  $\{c_n^\gamma\}$  is a scaling sequence for  $\Lambda_n \circ g^{-1}$ .  $\square$

**Corollary 1.1.1.** Let  $\mathbf{Z} = (Z^{(1)}, \dots, Z^{(K)})$  be a regularly varying random vector with tail index  $\alpha$ , exponent measure  $\Lambda_{\mathbf{Z}}$  associated to the scaling sequence  $\{c_n\}$ . Let  $A$  be a  $d \times K$  matrix with positive entries. Then the vector

$$\left( \bigvee_{a=1}^K A_{1a} Z^{(a)}, \dots, \bigvee_{a=1}^K A_{da} Z^{(a)} \right),$$

is regularly varying with tail index  $\alpha$ .

Let us consider a set  $A$  that remains separated from zero. This separation implies the existence of  $\epsilon > 0$ , such that the complement of  $A$  encompasses a ball centered at zero with radius of  $\epsilon$ . This  $\epsilon$  may vary based on the chosen norm but always exists. In this context, an exceedance relative to  $A$  can be defined as the event  $X \in xA$  indicating that  $\mathbf{X} > x\epsilon$ . This signifies not only that  $\mathbf{X}$  is large in the conventional sense but also concerning this specific choice of the event  $A$ . There are various possible choices for  $A$ , let us mention for instance the following

ones:

$$\left\{ \bigvee_{j=1}^d x^{(j)} > 1 \right\}, \quad \left\{ \bigwedge_{j=1}^d x^{(j)} > 1 \right\}, \quad \left\{ \sum_{j=1}^d x^{(j)} > 1 \right\}, \quad \left\{ \prod_{j=1}^d x^{(j)} > 1 \right\},$$

and contributions (unions or intersections) of these events are also valid. In leveraging the property of regular variation to analyse the exceedance  $\mathbf{X} \in xA$  as  $x \rightarrow \infty$ , it becomes necessary to ensure that  $\Lambda_{\mathbf{X}}(A) > 0$  in addition to  $\Lambda_{\mathbf{X}}(\partial A) = 0$ . In this scenario, the regular variation implies that as  $x \rightarrow \infty$ ,

$$\mathbb{P} \left\{ \frac{\mathbf{X}}{x} \in \cdot \mid \mathbf{X} \in xA \right\} \xrightarrow[n \rightarrow \infty]{v\#} \frac{\Lambda_{\mathbf{X}}(A \cap \cdot)}{\Lambda_{\mathbf{X}}(A)}. \quad (1.6)$$

By employing the spectral decomposition of the exponent measure, see, e.g., Theorem 1.1.7, this limit can be expressed in terms of the spectral measure. For all measurable subset  $B$  of  $\mathcal{E}_+$ , we have:

$$\frac{\Lambda_{\mathbf{X}}(A \cap B)}{\Lambda_{\mathbf{X}}(A)} = \frac{\int_0^\infty \alpha r^{-(\alpha+1)} dr \int_{\mathbb{S}_{\mathcal{E}_+}} \mathbf{1}_{A \cap B}(r\mathbf{a}) \Phi(d\mathbf{a})}{\int_0^\infty \alpha r^{-(\alpha+1)} dr \int_{\mathbb{S}_{\mathcal{E}_+}} \mathbf{1}_A(r\mathbf{a}) \Phi(d\mathbf{a})}.$$

Instead of dividing by  $x$  in (1.6), an alternative is to opt for a norm in  $\mathbb{R}^d$  and divide by  $\|\mathbf{X}\|$ . This leads to the following expression: for all measurable subsets of  $B$  of  $\mathbb{S}_{\mathcal{E}_+}$  such that  $\Lambda_{\mathbf{X}}(\partial B) = 0$ ,

$$\lim_{x \rightarrow \infty} \mathbb{P} \left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in B \mid \mathbf{X} \in xA \right\} = \frac{\Lambda_{\mathbf{X}}(A \cap B^*)}{\Lambda_{\mathbf{X}}(A)}, \quad (1.7)$$

where  $B^*$  is the cone with base  $B$ , that is  $B^* = \{\mathbf{x} \in \mathbb{R}_+^d : \mathbf{X}/\|\mathbf{X}\| \in B\}$ . Choosing now  $A = \{\mathbf{x} : \|\mathbf{x}\| > 1\}$ , we simply obtain

$$\lim_{x \rightarrow \infty} \mathbb{P} \left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in B \mid \|\mathbf{X}\| > x \right\} = \Phi(B). \quad (1.8)$$

All these expressions are mathematically equivalent. However, from a statistical standpoint, considering the specific problem and available data, one expression may prove to be more practically relevant than the others.

A noteworthy scenario occurs when  $d \geq 2$ , and the vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  with the exponent measure  $\Lambda_{\mathbf{X}}$  is divided into two subvectors:  $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ , where  $\mathbf{X}^{(1)} = (X^{(1)}, \dots, X^{(h)})$  and  $\mathbf{X}^{(2)} = (X^{(h+1)}, \dots, X^{(d)})$  with  $h \in \{1, \dots, d-1\}$ . Let  $C$  be a set separated from 0 in  $\mathbb{R}_+^d$  such that  $\Lambda_{\mathbf{X}}(C \times \mathbb{R}_+^{d-h}) > 0$ . According to Proposition 1.1.6,  $\mathbf{X}^{(1)}$  is regularly varying with the exponent measure  $\Lambda_{\mathbf{X}}(\cdot) = \Lambda_{\mathbf{X}}(\cdot \times \mathbb{R}_+^{d-h})$ . For such a set  $C$  we can explore the possibility of conditioning on the event  $\{\mathbf{X}^{(1)} \in xC\}$ . Subsequently by taking a measurable set  $D$  in  $\mathbb{R}_+^{d-h}$  such that  $\Lambda_{\mathbf{X}}(\partial(C \times D)) = 0$ ,

$$\lim_{x \rightarrow \infty} \mathbb{P} \left\{ \frac{\mathbf{X}^{(2)}}{x} \in D \mid \mathbf{X}^{(1)} \in xC \right\} = \frac{\Lambda_{\mathbf{X}}(C \times D)}{\Lambda_{\mathbf{X}^{(1)}}(C)}.$$

The limits of this conditional probability for different choices of the set  $C$ , have been employed as measure of extremal dependence. Heuristically, they quantify the tendency of some or all components of the vector to be jointly extremely large. These measures are instrumental in defining certain indices, particularly in applications such as climate sciences on risk managements. In the following we introduce some of these quantities.

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

**Extremal coefficient** - Let  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  be a regularly varying random vector and with distribution  $H_{\mathbf{X}}$  and  $\mathbf{X}^{(1)} = (X^{(1)}, \dots, X^{(h)})$  a subvector of  $\mathbf{X}$  and  $h \in \{1, \dots, d-1\}$ . The coefficients

$$\begin{aligned} \theta_h &= \Lambda_{\mathbf{X}} \left( (\mathcal{E}_+ \setminus [0, 1]^h) \times \mathbb{R}_+^{d-h} \right) = \lim_{x \rightarrow \infty} \mathbb{P} \left\{ x^{-1} \mathbf{X} \in (\mathcal{E}_+ \setminus [0, 1]^h) \times \mathbb{R}_+^{d-h} \right\} \\ &= \int_{\mathbb{S}_{d-1}} \bigvee_{j=1}^h a^{(j)} \Phi(d\mathbf{a}). \end{aligned}$$

In particular, stronger extremal dependence corresponds to smaller extremal coefficients  $\theta_h$ . Clearly  $\theta_0 = 0$  and  $\theta_1 = 1$ , so that only relevant coefficient  $\theta_h$  are those for which  $h \geq 2$ . We have  $1 \leq \theta_h \leq h$ , the upper and lower bounds correspond to independence and complete dependence, respectively.

**Extremal dependence measure** - Let  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  be a regularly varying random vector. Since  $f(\mathbf{x}) = \prod_{j=1}^d x^{(j)}$  is a continuous and bounded function on  $\mathbb{S}_{\mathcal{E}_+}$ , applying weak convergence in (1.8) yields

$$\lim_{x \rightarrow \infty} \mathbb{E} \left[ f \left( \frac{\mathbf{X}}{\|\mathbf{X}\|} \right) \mid \|\mathbf{X}\| > x \right] = \int_{\mathbb{S}_{\mathcal{E}_+}} \prod_{j=1}^d a^{(j)} \Phi(d\mathbf{a}).$$

In particular, for a bivariate regularly varying random vector  $(X^{(1)}, X^{(2)})$ , the expression on the RHS is called the extremal dependence measure (EDM):

$$EDM(X^{(1)}, X^{(2)}) = \int_{\mathbb{S}_{\mathcal{E}_+}} a^{(1)} a^{(2)} \Phi(d\mathbf{a}).$$

We can interpret EDM as a covariance-like quantity, computed with respect to the spectral measure. The EDM vanished whenever the spectral measure  $\Phi$  is concentrated on the axes, that is, in the case of extremal independence.

**Extremal correlation** - Let  $\mathbf{X} = (X^{(1)}, X^{(2)})$  be a bivariate regularly varying random vector in  $\mathbb{R}_+^2$  with exponent measure  $\Lambda_{\mathbf{X}}$ . Choosing in particular  $C = D = (1, \infty)$  in (1.7), then we obtain the extremal correlation

$$\chi(2, 1) = \lim_{x \rightarrow \infty} \mathbb{P} \left\{ X^{(2)} > x \mid X^{(1)} > x \right\} = \Lambda_{\mathbf{X}}((1, \infty) \times (1, \infty)) / \Lambda_{X^{(1)}}((1, \infty)).$$

### 1.1.4 Weakly dependent random processes

To understand how the limiting distribution deviates, establishing upper bounds for algebraic moments or exponential inequalities of a partial sum of real-valued random variables is of prime interest. One of the key steps in this process is to analyse the variance of this sum. While the variance of the sum equals the sum of individual variances for independent random variables, this statement does not hold true for dependent random variables, except for martingale difference sequences. Let  $(X_k, k \in \mathbb{Z})$  be a sequence of random variables on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  that is strictly stationary. Drawing inspiration from martingales, we can gain

an intuition of weak dependence for the sequence  $(X_k, k \in \mathbb{Z})$  if for any  $k \in \mathbb{Z}$ ,

$$\mathbb{E} [|\mathbb{E}[X_{k+n}|X_1, \dots, X_n]|] \xrightarrow{n \rightarrow \infty} 0. \quad (1.9)$$

Indeed, martingale differences satisfy condition (1.9), as do mixingales differences sequences. Gaussian sequences with pairwise correlation coefficients tending to zero, lacunary random variables, and many others satisfy (1.9), either directly or in slightly modified form. This definition serves reasonably well if one is interested in sums of the random variables  $X$ . However, for statistics, this definition falls short. To understand why, consider  $(\xi_k, k \in \mathbb{Z})$  to be a stationary sequence of real-valued random variables with common distribution function  $F_\xi$ , the fundamental components for the empirical process are the discontinuous random functions:

$$X_k := X_k(x, \omega) = \mathbb{1}_{\{\xi_k(\omega) \leq x\}} - F_\xi(x), \quad x \in \mathbb{R}.$$

So, even if the sequence  $(\xi_k, k \in \mathbb{Z})$  satisfies relation (1.9), the function  $(X_k, k \in \mathbb{Z})$  may not. In this section, we will outline the primary mixing conditions examined in this thesis and their implications for covariance inequalities and coupling lemmas. Unlike condition (1.9), most of these conditions do not assume the existence of a finite expectation. Let  $\mathcal{F}_a^b$  denote the  $\sigma$ -field generated by the random variables  $X_a, \dots, X_b$  when  $-\infty \leq a < b \leq \infty$ .

**Definition 1.1.9.**

- The sequence  $(X_k, k \in \mathbb{Z})$  is called strongly mixing if

$$\alpha(n) := \sup \left\{ |\mathbb{P}\{A \cap B\} - \mathbb{P}\{A\}\mathbb{P}\{B\}| : A \in \mathcal{F}_{-\infty}^k, B \in \mathcal{F}_{k+n}^\infty, k \geq 1 \right\} \xrightarrow{n \rightarrow \infty} 0.$$

- The sequence is called absolutely regular if

$$\beta(n) := \sup \left\{ \left| \mathbb{P}_{\mathcal{F}_{-\infty}^k \otimes \mathcal{F}_{k+n}^\infty} (C) - \mathbb{P}_{\mathcal{F}_{-\infty}^k} \otimes \mathbb{P}_{\mathcal{F}_{k+n}^\infty} (C) \right| : C \in \mathcal{F}_{-\infty}^k \otimes \mathcal{F}_{k+n}^\infty, k \geq 1 \right\} \xrightarrow{n \rightarrow \infty} 0,$$

where  $\mathbb{P}_{\mathcal{F}_{-\infty}^k \otimes \mathcal{F}_{k+n}^\infty}$  to be defined on  $(\Omega \times \Omega, \mathcal{F}_{-\infty}^k \otimes \mathcal{F}_{k+n}^\infty)$  is the image of  $\mathbb{P}$  under the canonical injection  $i$  from  $(\Omega, \mathcal{F}, \mathbb{P})$  into  $(\Omega \times \Omega, \mathcal{F}_{-\infty}^k \otimes \mathcal{F}_{k+n}^\infty, \mathbb{P})$  denoted by  $i(\omega) = (\omega, \omega)$ . Also,  $\mathbb{P}_{\mathcal{F}_{-\infty}^k}$  (resp.  $\mathbb{P}_{\mathcal{F}_{k+n}^\infty}$ ) is the restriction of  $\mathbb{P}$  to  $\mathcal{F}_{-\infty}^k$  (resp.  $\mathcal{F}_{k+n}^\infty$ ).

- The sequence is called  $\varphi$ -mixing or uniformly mixing if

$$\varphi(n) := \sup \left\{ \left| \frac{\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} \right| ; A \in \mathcal{F}_{-\infty}^k, \mathbb{P}(A) > 0, B \in \mathcal{F}_{k+n}^\infty, k \geq 1 \right\} \xrightarrow{n \rightarrow \infty} 0.$$

It is worth emphasising that all mixing coefficients, when regarded as a function of  $n$ , decrease or remain constant. The strong mixing condition was introduced by [Rosenblatt \(1956b\)](#), the  $\beta$ -mixing coefficient by [Volkonskii and Rozanov \(1959\)](#), and the  $\varphi$ -mixing coefficient by [Ibragimov \(1962\)](#). The following relations between these coefficients are valid

$$2\alpha(n) \leq \beta(n) \leq \varphi(n) \leq 1,$$

hence a  $\varphi$ -mixing sequence is absolutely regular which is consequently strongly mixing. Here, we provide few standard examples of weakly dependent processes.

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

**Example 1.1.6.** A sequence  $(X_k, k \in \mathbb{Z})$  is called  $m$ -dependent if the  $\sigma$ -fields  $\mathcal{F}_{-\infty}^k$  and  $\mathcal{F}_{k+n}^\infty$  are independent for each  $k \geq 1$ . Obviously,  $m$ -dependent sequences are  $\alpha$ -mixing with  $\alpha(k) = 0$ ,  $\forall k \geq n$  and  $\alpha(k) > 0$ ,  $\forall k < n$ .

**Example 1.1.7.** A sequence  $(X_k, k \in \mathbb{Z})$  be a strongly stationary Markov chains in an countable space. If  $(X_k, k \in \mathbb{Z})$  is irreducible and aperiodic, then  $\beta(n) \rightarrow 0$ , as  $n \rightarrow \infty$ .

For two  $\sigma$ -fields  $\mathcal{A}, \mathcal{B}$ , we define

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup \{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{A}, B \in \mathcal{B} \}. \quad (1.10)$$

Notice that  $\alpha(\mathcal{A}, \mathcal{B})$  measures the degree of dependence between the  $\sigma$ -fields  $\mathcal{A}$  and  $\mathcal{B}$ . Indeed, for independent cases,  $\alpha(\mathcal{A}, \mathcal{B}) = 0$ . Also, note that the mixing coefficients  $\alpha(n)$  can be expressed as  $\alpha(n) = \sup_k \alpha(\mathcal{F}_{-\infty}^k, \mathcal{F}_{k+n}^\infty)$ . For random variables  $X$  and  $Y$ , let  $\sigma(X)$  and  $\sigma(Y)$  denote the  $\sigma$ -fields generated by them.

**Lemma 1.1.3.** *Let  $X$  and  $Y$  be real-valued random variables. Then*

$$|\text{Cov}(X, Y)| \leq 4\alpha(\sigma(X), \sigma(Y))\|X\|_\infty\|Y\|_\infty. \quad (1.11)$$

**Proof** Without loss of generality we may assume that  $\|X\|_\infty = \|Y\|_\infty = 1$ . Set  $\mathcal{A} = \sigma(X)$  and  $\mathcal{B} = \sigma(Y)$ , and observe that

$$\begin{aligned} |\text{Cov}(X, Y)| &= |\mathbb{E}[X(\mathbb{E}[Y|\mathcal{A}] - \mathbb{E}[Y])]| \leq \mathbb{E}[|\mathbb{E}[Y|\mathcal{A}] - \mathbb{E}[Y]|] \\ &= \mathbb{E}[X_1(\mathbb{E}[Y|\mathcal{A}] - \mathbb{E}[Y])] = \text{Cov}(X_1, Y), \end{aligned}$$

where  $X_1 = \text{sign}(\mathbb{E}[Y|\mathcal{A}] - \mathbb{E}[Y])$  is a  $\mathcal{A}$ -measurable random variable and we can apply the same argument again, now with roles of  $X_1$  and  $Y$  interchanged. In this way, we can get

$$|\text{Cov}(X_1, Y)| \leq |\mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1]| = \text{Cov}(X_1, Y_1),$$

where  $Y_1 = \text{sign}(\mathbb{E}[X_1|\mathcal{B}] - \mathbb{E}[X_1])$ . Now set  $A = \mathbb{1}_{\{X_1=1\}}$  and  $B = \mathbb{1}_{\{Y_1=1\}}$ . Then

$$\begin{aligned} |\text{Cov}(X_1, Y_1)| &\leq |\mathbb{P}\{A \cap B\} - \mathbb{P}\{A^c \cap B\} - \mathbb{P}\{A \cap B^c\} + \mathbb{P}\{A^c \cap B^c\} \\ &\quad - \mathbb{P}\{A\}\mathbb{P}\{B\} + \mathbb{P}\{A^c\}\mathbb{P}\{B\} + \mathbb{P}\{A\}\mathbb{P}\{B^c\} - \mathbb{P}\{A^c\}\mathbb{P}\{B^c\}| \\ &\leq 4\alpha(\mathcal{A}, \mathcal{B}). \end{aligned}$$

The three inequalities together yields (1.11). □

The following Lemma is due to [Davydov \(1970\)](#), the special case  $r = s$  to [Ibragimov \(1962\)](#).

**Lemma 1.1.4.** *Let  $1 \leq p, q, t \leq \infty$  satisfy  $p^{-1} + q^{-1} + r^{-1} = 1$  and let  $X$  and  $Y$  be real valued random variables in  $\mathbb{L}^p(\mathcal{A})$  and  $\mathbb{L}^q(\mathcal{B})$ , respectively. Then*

$$|\text{Cov}(X, Y)| \leq 10\|X\|_p\|Y\|_q\alpha(\sigma(X), \sigma(Y))^{1/r}.$$

For  $\varphi$ -mixing sequences, [Ibragimov \(1962\)](#) has given slightly stronger inequalities. In the following, we state such an inequality given by [Peligrad \(1983\)](#).



## Introduction

---

**Lemma 1.1.5.** *Let  $1 \leq p, q \leq \infty$  satisfy  $p^{-1} + q^{-1} = 1$  and let  $X$  and  $Y$  be real-valued random variable in  $\mathbb{L}^p(\mathcal{A})$  and  $\mathbb{L}^q(\mathcal{B})$ , respectively. Then*

$$|\text{Cov}(X, Y)| \leq 2\varphi(\mathcal{A}, \mathcal{B})^{1/p} \varphi(\mathcal{B}, \mathcal{A})^{1/q} \|X\|_p \|Y\|_q. \quad (1.12)$$

**Proof** The proof of (1.12) follows from classical approximation used in the construction of Lebesgue's integral. We approximate  $X$  and  $Y$  by  $X = \sum_i a_i \mathbb{1}_{A_i}$  and  $Y = \sum_j b_j \mathbb{1}_{B_j}$ , where  $(A_i)_i$  and  $(B_j)_j$  are respectively, finite decompositions of  $\Omega$  into disjoint elements of  $\mathcal{A}$  and  $\mathcal{B}$ . For notational conveniency, let us denote  $c_{ij} = \mathbb{P}\{B_j|A_i\} - \mathbb{P}\{B_j\}$  and  $d_{ij} = \mathbb{P}\{A_i|B_j\} - \mathbb{P}\{A_i\}$ . Using Hölder's inequality, we obtain

$$|\text{Cov}(X, Y)| \leq \left( \sum_i |a_i|^p \mathbb{1}_{A_i} \right)^{1/p} \left[ \sum_i \mathbb{P}\{A_i\} \left( \sum_j |b_j| |c_{ij}| \right)^q \right]^{1/q}.$$

Using that

$$\sum_j |b_j| |c_{ij}| = \sum_j |b_j| |c_{ij}|^{1/q} |c_{ij}|^{1/p}$$

and Hölder's inequality, it stems down

$$\begin{aligned} |\text{Cov}(X, Y)| &\leq (\mathbb{E}|X|^p)^{1/p} \left[ \sum_i \mathbb{P}\{A_i\} \left( \sum_j |b_j|^q |c_{ij}| \right) \sum_j |c_{ij}|^{q/p} \right]^{1/q} \\ &\leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q} \max_i \left( \sum_j |c_{ij}| \right)^{1/p} \max_j \left( \sum_i |d_{ij}| \right)^{1/q}. \end{aligned}$$

If  $C_i^+$  (or  $C_i^-$ ) is the union of those  $B_j$  for which  $c_{ij}$  is positive (or nonpositive) then

$$\sum_j |c_{ij}| \leq |\mathbb{P}\{C_i^+|A_i\} - \mathbb{P}\{C_i^+\}| + |\mathbb{P}\{C_i^-|A_i\} - \mathbb{P}\{C_i^-\}| \leq 2\varphi(\mathcal{A}, \mathcal{B}).$$

Similarly,

$$\sum_i |d_{ij}| \leq 2\varphi(\mathcal{B}, \mathcal{A}).$$

So (1.12) holds for simple random variable, and by passing to the limit the inequality remains valid for every  $X \in \mathbb{L}^p(\mathcal{A})$  and  $Y \in \mathbb{L}^q(\mathcal{B})$ .  $\square$

For most applications, the inequalities above are classical arguments. However, when striving for results that are optimal or nearly optimal, shaper versions are necessary. Such inequalities below are attributed to [Rio \(1993, 1999\)](#).

**Theorem 1.1.11.** *Let  $X$  and  $Y$  be integrable real-valued random variables. Assume that  $XY$  is integrable and let  $\alpha = \alpha(\sigma(X), \sigma(Y))$  be defined by (1.10). Then*

$$|\text{Cov}(X, Y)| \leq 2 \int_0^\alpha Q_{|X|}(u) Q_{|Y|}(u) du \leq 4 \int_0^{\alpha/2} Q_{|X|}(u) Q_{|Y|}(u) du.$$

## 1.1 A survival guide for high-dimensional extremal dependence modeling

Conversely, for any symmetric distribution function  $F_X$  and any  $\alpha \in [0, 1/4]$ , one can construct random variables  $X$  and  $Y$  with respective distribution function  $F_X$  such that  $\alpha(\sigma(X), \sigma(Y)) \leq \alpha$  and

$$\text{Cov}(X, Y) \geq \frac{1}{2} \int_0^{\alpha/2} Q_{|X|}^2(u) du, \quad (1.13)$$

provided that  $Q_{|X|}$  is square integrable on  $[0, 1]$ .

**Proof** We only prove the lower bound of the covariance in (1.13). To do it, let us construct a pair  $(U, V)$  of random variables with marginal distributions the uniform law over  $[0, 1]$ , satisfying  $\alpha(\sigma(U), \sigma(V)) \leq \alpha$  and such that (1.13) holds true for  $(X, Y) = (F_X^{\leftarrow}(U), F_X^{\leftarrow}(V))$ . Let  $a$  be any real in  $[0, 1/4]$ , and  $(Z, T)$  be a random variable with the uniform distribution over  $[0, 1]^2$ . Set

$$(U, V) = \mathbf{1}_{\{Z \in [0, 1-a]\}}(Z, (1-a)T) + \mathbf{1}_{\{Z \in ]1-a, 1\}}(Z, Z).$$

Let  $u \in [0, 1]$ , then  $\mathbb{P}\{U \leq u\} = \mathbb{P}\{Z \leq u\} = u$ . Now take  $v \in [0, 1]$ , one can write

$$\mathbb{P}\{V \leq v\} = \mathbb{P}\{T \leq v/(1-a)\} \mathbb{P}\{Z \leq 1-a\} + \mathbb{P}\{1-a < Z \leq v\}.$$

If  $v \leq 1-a$ , then  $\{1-a < Z \leq v\} = \emptyset$  and we obtain  $\mathbb{P}\{V \leq v\} = v$ . If  $v > 1-a$ , then  $\{T \leq v/(1-a)\} = \{T \leq 1\}$  and

$$\mathbb{P}\{V \leq v\} = 1-a + v - (1-a) = v,$$

hence  $U, V$  are distributed uniformly over  $[0, 1]$ . We now prove that

$$\alpha(\sigma(U), \sigma(V)) \leq a - a^2/2. \quad (1.14)$$

Clearly,

$$\|\mathbb{P}_{(U,V)} - \mathbb{P}_U \otimes \mathbb{P}_V\|_{TV} = 4a - 2a^2.$$

Since it is well known that  $\|\mathbb{P}_{(U,V)} - \mathbb{P}_U \otimes \mathbb{P}_V\|_{TV} \geq 4\alpha(\sigma(U), \sigma(V))$ . Hence (1.14) holds true. Next, let  $(X, Y) = (F_X^{\leftarrow}(U), F_X^{\leftarrow}(V))$ . Since  $X$  (resp.  $Y$ ) is a measurable function of  $U$  (resp.  $V$ ),  $\alpha(\sigma(X), \sigma(Y)) \leq \alpha$ . Now

$$XY = F_X^{\leftarrow}(Z)F_X^{\leftarrow}(Z)\mathbf{1}_{\{Z \in ]1-a, 1\}} + F_X^{\leftarrow}(Z)F_X^{\leftarrow}(T)\mathbf{1}_{\{Z \in [0, 1-a]\}}.$$

Taking the expectation of the formula and we recall that  $Z$  and  $T$  are independent, we get that

$$\mathbb{E}[XY] = \int_{1-a}^1 (F_X^{\leftarrow}(u))^2 du + \frac{1}{1-a} \left( \int_0^{1-a} F_X^{\leftarrow}(u) du \right)^2.$$

Using

$$\begin{aligned} \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] &= \int_{1-a}^1 (F_X^{\leftarrow}(u))^2 du + \frac{1}{1-a} \left( \int_0^{1-a} F_X^{\leftarrow}(u) du \right)^2 - \left( \int_0^1 F_X^{\leftarrow}(u) du \right)^2 \\ &\geq \int_{1-a}^1 (F_X^{\leftarrow}(u))^2 du + \left( \int_0^{1-a} F_X^{\leftarrow}(u) du \right)^2 - \left( \int_0^1 F_X^{\leftarrow}(u) du \right)^2 \end{aligned}$$

and since  $\int_0^1 F_X^{\leftarrow}(u)du = 0$

$$\begin{aligned} \left( \int_0^{1-a} F_X^{\leftarrow}(u)du \right)^2 &= \left( \int_0^1 F_X^{\leftarrow}(u)du - \int_{1-a}^1 F_X^{\leftarrow}(u)du \right)^2 \\ &= \left( \int_0^1 F_X^{\leftarrow}(u)du \right)^2 + \left( \int_{1-a}^1 F_X^{\leftarrow}(u)du \right)^2. \end{aligned}$$

Since  $X$  has a symmetric law,  $F_X^{\leftarrow}(1-u) = -F_X^{\leftarrow}(u) = Q_{|X|}(2u)$  for almost every  $u \in [0, 1/2[$ . Hence, since

$$\text{Cov}(X, Y) \geq \int_0^a Q_{|X|}^2(u)du \geq \int_0^\alpha Q_{|X|}^2(u)du.$$

□

Let  $X$  and  $Y$  be real random variables on the same probability space. We say that  $X$  and  $Y$  are partially coupled with probability  $p$  if

$$\mathbb{P}\{X = Y\} = p. \tag{1.15}$$

One of the most popular techniques to derive limit theorems for dependent process is to replace the original sequence with one exhibiting finite range dependence. In this regard coupling lemmas enable the substitution of the initial sequence after time zero with a new sequence that is independent of the past before time zero. Below, we present coupling theorems for mixing sequences. The complexity of the coupling hinges on the mixing condition. Here, we provide coupling results for strongly mixing or absolutely regular sequences. For sequences of a random variables satisfying  $\beta$ -mixing conditions, the new sequence is predominantly equal to the initial sequence after time  $n$ . This result was independently obtained by Berbee (1979) and Goldstein (1979). This result fails in the strong mixing case. Nevertheless, one can still derive weaker results that are effective for real-valued random variables. We first state this result below.

**Lemma 1.1.6.** *Let  $\mathcal{A}$  be a  $\sigma$ -field on  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $X$  be a real valued random variable with values in  $[a, b]$ . Let  $\delta$  be a random uniform distribution over  $[0, 1]$ , independent of the  $\sigma$ -field generated by  $X$  and  $\mathcal{A}$ . Then there exists a random variable  $X^*$ , with the same law as  $X$ , independent of  $X$  such that*

$$\mathbb{E}[|X - X^*|] \leq (b - a)\alpha(\mathcal{A}, \sigma(X)).$$

Furthermore,  $X^*$  is measurable.

**Proof** Let  $F_X$  be the distribution of  $X$  and  $F_{\mathcal{A}}$  be the conditional distribution function of  $X$  given  $\mathcal{A}$ , which is defined by  $F_{\mathcal{A}}(t) = \mathbb{P}\{X \leq t | \mathcal{A}\}$ . Since  $\delta$  is independent of  $\mathcal{A} \vee \sigma(X)$  and has the uniform distribution over  $[0, 1]$ , the random variable

$$V = \ell^-(F_{\mathcal{A}}(X)) + \delta (F_{\mathcal{A}}(X) - \ell^-(F_{\mathcal{A}}(X))),$$

has the uniform distribution over  $[0, 1]$ , conditionnally to  $\mathcal{A}$  (the proof of this statement is similar to the one given in Theorem 1.1.4). Hence  $V$  is independent of  $\mathcal{A}$  and has the uniform distribution over  $[0, 1]$ . Therefore

$$X^* = F_X^{\leftarrow}(V),$$

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

is independent of  $\mathcal{A}$  and has the same distribution function as  $X$ . Furthermore

$$X = F_{\mathcal{A}}^{\leftarrow}(V), \text{ a.s.},$$

whence

$$\mathbb{E}[|X - X^*|] = \mathbb{E} \left[ \int_0^1 |F_{\mathcal{A}}^{\leftarrow}(v) - F_X^{\leftarrow}(v)| dv \right].$$

Since  $X$  takes values in  $[a, b]$ ,

$$\int_0^1 |F_{\mathcal{A}}^{\leftarrow}(v) - F_X^{\leftarrow}(v)| dv = \int_a^b |F_{\mathcal{A}}(t) - F_X(t)| dt.$$

Interchanging the integrals, we infer that

$$\mathbb{E}[|X - X^*|] = \int_a^b \mathbb{E}[|F_{\mathcal{A}}(t) - F_X(t)|] dt,$$

and, one can prove that

$$\alpha(\mathcal{A}, \sigma(X)) = \sup_x \mathbb{E}[|\mathbb{P}\{X \leq x | \mathcal{A}\} - \mathbb{P}\{X \leq x\}|],$$

and we obtain the result. □

To obtain a coupling, we want to construct random variables  $X$  and  $Y$  on the same probability space, under the condition that their dispersions are  $\mathbb{P}_X$  and  $\mathbb{P}_Y$ , respectively, in such a way that the probability  $p$  of partial coupling in (1.15) is as large as possible. We shall show that  $p$  can be maximised. If this probability  $p$  is maximal, we say that  $X$  and  $Y$  are maximally couples. We state the coupling lemma of Berbee (1979) for random variables satisfying a  $\beta$ -mixing condition.

**Theorem 1.1.12.** *Suppose on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , there is defined a pair  $(X, Y)$  of random variables, with values in Borel space  $\mathcal{A}, \mathcal{B}$ . The probability space can be extended with a random variable  $X^*$ , independent of  $X$ , with the same distribution as  $X$ , such that*

$$\mathbb{P}\{X \neq X^*\} = \beta(\sigma(X), \sigma(Y))$$

while the dependence structure is not affected in the sense that

$$\mathbb{P}_{Z|X, X^*, Y} = \mathbb{P}_{Z|X, Y},$$

for any random variable  $Z$  defined on the original probability space with values in a Borel space.

Considering  $(X_k, k \in \mathbb{Z})$  a random sequence, starting from Theorem 1.1.12 we can construct by induction a sequence of random variables  $(\bar{X}_i, i \geq 0)$  such that

- (a) For any  $i \geq 0$ , the random variable  $\bar{Z}_i = (\bar{X}_{iq+1}, \dots, \bar{X}_{iq+q})$  has the same distribution as  $U_i = (X_{iq+1}, \dots, X_{iq+q})$ .
- (b) The sequence  $(\bar{Z}_{2i})_{i \geq 0}$  is i.i.d. and so is  $(\bar{Z}_{2i+1})_{i \geq 0}$ .
- (c) For any  $i \geq 0$ ,  $\mathbb{P}\{Z_i \neq \bar{Z}_i\} \leq \beta(q)$ .

This construction is a classical technique extensively used in the literature see, e.g., [Bücher and Segers \(2014\)](#); [Doukhan et al. \(1995a\)](#) or Chapter 3 and Chapter 4 in this thesis.

### 1.1.5 Clustering

In the framework of high-dimensional statistics, observations are more likely to come from heterogeneous processes. A recipe for dealing with such heterogeneous data is to consider them as an assemblage of several homogeneous datasets, corresponding to homogeneous “subpopulations”. Then each subpopulation can be treated either independently or jointly. The main hurdle in this approach is to recover the unknown subpopulations, which is the main goal of clustering algorithms. Then each subpopulation can be treated either independently jointly. The main hurdle in this approach is to recover the unknown subpopulations, which is the main goal of clustering algorithms. The goal of cluster analysis is to find meaningful groups in data. Typically, these groups will be internally cohesive and separated from one another within the data. The purpose is to identify groups whose members share common characteristics that distinguish them from members of other groups.

The methodology for clustering can be based on either a proximity-separation paradigm or on statistical models. The proximity-separation paradigm is model-free and offers some easy-to-understand algorithms. However, defining a clear “ground truth” objective in this perspective and evaluating the performance of a given algorithm can be challenging. These developments occurred largely independently from mainstream statistics, which often relied on probability distribution on the observations. Meanwhile, they left several practical questions unresolved, such as specifying the number of clusters, and assessing uncertainty about a partition. The statistical paradigm, based on probabilistic modeling, is more amenable to interpretation and statistical analysis, and has the potential to address these questions. This thesis focus on this approach. Assume that we have  $n$  data points  $\mathbf{X}_1, \dots, \mathbf{X}_n$  in a vectorial normed space  $(\mathbb{R}^d, \|\cdot\|)$  and let us denote

$$\forall i \in \{1, \dots, n\}, \mathbf{X}_i = (X_i^{(1)}, \dots, X_i^{(d)}) \in \mathbb{R}^d, \forall j \in \{1, \dots, d\}, \mathbf{X}^{(j)} = (X_1^{(j)}, \dots, X_n^{(j)}) \in \mathbb{R}^n,$$

and let the matrix of observations:

$$M = (X_i^{(j)})_{i=1, \dots, n, j=1, \dots, d} \in \mathbb{R}^{n \times d}.$$

Informally, the goal of clustering is to find a partition  $\mathcal{G} = \{G_1, \dots, G_K\}$  of the indices  $\{1, \dots, n\}$  or  $\{1, \dots, d\}$ . We will refer to row clustering as the problem of clustering applied to  $\{\mathbf{X}_1, \dots, \mathbf{X}_n\} \in \mathbb{R}^d$  such that data points with indices within a group are similar, and those with indices in different groups are different. The following simplified model, which stated that points within a cluster have the same mean, gives a clear illustration of the above consideration.

For any point  $\mathbf{X}_i$  belonging to a cluster  $G_a \in \mathcal{G}$ , we have

$$\mathbb{E}[\mathbf{X}_i] = \mu_a \in \mathbb{R}^d.$$

The alternative framework will be designated as variable clustering, which concerns grouping variables of  $\mathbf{X}$  through  $\mathbf{X}_1, \dots, \mathbf{X}_n$ ,  $n$  observations. In this setting, we aim to cluster entities that might exhibit strong dependence within a cluster.

Given the prevalence of proximity-separation in row clustering as observed in clustering literature within extreme value theory, we shall provide a concise overview of two prominent algorithms within this domain: the  $K$ -means algorithm and hierarchical clustering. Additionally, we will introduce a model-based clustering approach within the Gaussian context and elucidate its relationship with the  $K$ -means algorithm.

The  $K$ -means algorithm method prescribes a criterion for partitioning a collection of points into  $K$  distinct groups. To achieve this, we start by selecting  $K$  cluster centres, denoted as  $\theta_1, \dots, \theta_K$  in such a way that the overall sum of squared distances from each point to its nearest cluster center is minimised

$$W_n = \frac{1}{n} \sum_{i=1}^n \min_{k=1, \dots, K} \|\mathbf{X}_i - \theta_k\|^2. \quad (1.16)$$

The  $K$ -means algorithm assigns each data point to its nearest cluster centre and aims to solve the minimisation problem:

$$(\hat{\theta}_1, \dots, \hat{\theta}_K) \in \arg \min_{(\theta_1, \dots, \theta_K) \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \min_{k=1, \dots, K} \|\mathbf{X}_i - \theta_k\|^2.$$

Then, the partition  $\hat{\mathcal{G}}^{K\text{-means}} = \{\hat{G}_1^{K\text{-means}}, \dots, \hat{G}_K^{K\text{-means}}\}$  is defined by

$$\hat{G}_k^{K\text{-means}} = \left\{ i \in \{1, \dots, n\} : d(\mathbf{x}_i, \theta_k) = \min_{k=1, \dots, K} d(\mathbf{x}_i, \theta_k) \right\}.$$

Consider  $\mathbf{X}_1, \dots, \mathbf{X}_n$  a sample comprising i.i.d. observations drawn from an (unknown) distribution  $\mathbb{P}_{\mathbf{X}}$ . Pollard (1981) provides conditions that guarantee the almost sure convergence of the cluster centres as the sample size  $n$  grows. Let  $W_n$  be a function dependent on a set of cluster centres  $\Theta$  and the empirical measure. In essence, the objective is to minimise  $W$

$$W(\Theta, \mathbb{P}_n) = \int \min_{\theta \in \Theta} \|\mathbf{x} - \theta\|^2 \mathbb{P}_{\mathbf{X}}(d\mathbf{x}),$$

are all possible choices of the set  $\Theta$  containing  $K$  (or fewer) elements. Let  $\Theta_n$  be the set of optimal clusters centres for the sample, then the following hold.

**Theorem 1.1.13.** *Suppose that  $\int \|\mathbf{x}\|^2 \mathbb{P}_{\mathbf{X}}(d\mathbf{x}) < \infty$  and that for each  $k = 1, \dots, K$  there is a unique set  $\bar{\Theta}(k)$  for which  $W(\bar{\Theta}(k), \mathbb{P}) = m_k(\mathbb{P})$  where*

$$m_k(\mathbb{P}) := \inf\{W(\Theta, \mathbb{P}) : \Theta \text{ contains } k \text{ or fewer points}\}.$$

*Then  $\Theta_n \rightarrow \bar{\Theta}(k)$  a.s., and  $W_n(\Theta_n, \mathbb{P}_n) \rightarrow m_k(\mathbb{P})$  a.s. as  $n \rightarrow \infty$ .*

In general, solving the minimisation problem in  $K$ -means is NP-hard and even hard to approximate.

As demonstrated, the principle behind  $K$ -means is driven by optimisation. However, the strategy in hierarchical clustering differs. The principle involves merging data points step-by-step by merging the two closest groups of points at each iteration. Specifically, hierarchical clustering algorithms proceeds by sequentially clustering data points, starting with each data point forming its own singleton cluster and then iteratively merging them until a single cluster containing

## Introduction

---

all data points is achieved. This process yields a series of nested clusterings. In hierarchical clustering, the merging of points is straightforward: at each step, the algorithm merges the two closest centres of the current clustering while keep the other clusters unchanged (see Algorithm (HC) for more details). This necessitates the definition of a “distance”  $\ell(G, G')$  between clusters  $G$  and  $G'$ , typically referred to as linkage. Some classical examples of linkage methods include:

- Single linkage: refers to the smallest distance between the points of two clusters

$$\ell_{\text{single}}(G, G') = \min\{\|\mathbf{x}_i - \mathbf{x}_j\|, i \in G, j \in G'\}.$$

- Complete linkage: corresponds to the largest distance between the point of two clusters:

$$\ell_{\text{complete}}(G, G') = \max\{\|\mathbf{x}_i - \mathbf{x}_j\|, i \in G, j \in G'\}.$$

- Average linkage: corresponds to the average distance between the points of the clusters  $G, G'$ :

$$\ell_{\text{average}}(G, G') = \frac{1}{|G||G'|} \sum_{i \in G, j \in G'} \|\mathbf{x}_i - \mathbf{x}_j\|.$$

---

### Algorithm (HC)

---

- 1: **procedure** HC( $\mathbf{X}_1, \dots, \mathbf{X}_n, \ell$ )
  - 2:   Initialisation:  $G^{(n)} = \{\{1\}, \dots, \{n\}\}$ .
  - 3:   **for**  $t = n, \dots, 2$  **do**
  - 4:     Find  $(\hat{a}, \hat{b}) \in \arg \min_{(a,b)} \ell(G_a^{(n)}, G_b^{(n)})$ ;
  - 5:     Build  $G^{(t-1)}$  by merging  $G_{\hat{a}}^{(t)}$  and  $G_{\hat{b}}^{(t)}$ . The other clusters are left unchanged
  - 6:   Return the  $n$  partitions  $G^{(1)}, \dots, G^{(n)}$  of  $\{1, \dots, n\}$ .
- 

We now confine our analysis to the Gaussian setting and let us consider the following model

**Definition 1.1.10.** Let  $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$  be  $n$  observations such that:

- (i) The observations  $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$  are independent.
- (ii) There exists a partition  $\mathcal{G}^* = \{G_1^*, \dots, G_K^*\}$  of  $\{1, \dots, n\}$ , and  $\theta_1, \dots, \theta_K \in \mathbb{R}^d$ ,  $\Sigma_1, \dots, \Sigma_K \in \mathbb{R}^{d \times d}$  such that

$$\forall i \in G_k^*, \mathbf{X}_i \sim \mathcal{N}(\theta_k, \Sigma_k).$$

The negative log-likelihood of the distribution  $\mathcal{N}(\theta_k, \Sigma_k)$  with respect to the observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is

$$\frac{1}{2}(\mathbf{X}_i - \theta_k)^\top \Sigma_k^{-1}(\mathbf{X}_i - \theta_k) + \frac{1}{2} \ln(|\Sigma_k|) + \frac{d}{2} \ln(2\pi)$$

so the maximum likelihood estimator of the partition  $G^*$  is

$$\hat{\mathcal{G}}^{\text{MV}} \in \arg \min \sum_{k=1}^K \min_{\Sigma_k \in \mathcal{S}_d^+} \min_{\theta_k \in \mathbb{R}^d} \sum_{i \in G_k} \left( (\mathbf{X}_i - \theta_k)^\top \Sigma_k^{-1}(\mathbf{X}_i - \theta_k) + \ln(|\Sigma_k|) \right), \quad (1.17)$$

where  $\mathcal{S}_d^+$  is the set of  $d \times d$  positive semi definite matrices and where the first minima is over all partitions  $G$  of  $\{1, \dots, n\}$  into  $K$  groups.

The maximum likelihood estimator faces several drawbacks. One concern relates to the exponential growth in the cardinality of the set of partitions  $\{1, \dots, n\}$  into  $K$  groups, which grows exponentially fast with  $n$ , approximately as  $K^n/K!$ . Consequently, the computational expense of scanning the set of partitions of  $\{1, \dots, n\}$  into  $K$  groups becomes prohibitive, rendering estimation unfeasible for larger sample size. From a statistical perspective, the estimation of  $\Sigma_k$  becomes unstable in high-dimensional settings, and even degenerates when  $d$  exceeds  $n$ . To motivate the latter issue, a common approach is to consider setting all  $\Sigma_k$  to  $\sigma^2 I_n$  in (1.17). Consequently (1.17) simplifies to  $K$ -means criterion in (1.16).

### 1.1.6 Mathematics of high dimension

Classical statistics provide a comprehensive theoretical framework for analysing data with a small dimension  $d$  and a large number of observations  $n$ . Classical results meticulously describe the asymptotic properties of estimators as  $n$  goes to infinity, while  $d$  remains constant, a context where such analysis is relevant. In modern statistics, current data exhibit the opposite scenario with a substantial number of dimensions alongside a sample size  $n$  either comparable to  $d$  or significantly smaller. The traditional asymptotic analysis, where  $d$  is fixed and  $n$  approaches to infinity, loses its relevance and can yield misleading results. An alternative approach is to treat both  $n$  and  $d$  as they are and conduct a non-asymptotic analysis of the estimators. This approach remains valid for any value of  $n$  and  $d$ , circumventing the pitfalls associated with asymptotic analysis. As an illustration, let us consider the linear regression problem where we are given observations  $Y$  and  $X$  that satisfy the following relationship:

$$Y = X\theta + \epsilon,$$

where  $X \in \mathbb{R}^{n \times d}$  is a deterministic matrix,  $\theta \in \mathbb{R}^d$  is the unknown parameter that we want to estimate and  $\epsilon$  is an uncorrelated random vector with  $\mathbb{E}[\epsilon_i] \leq \sigma_i^2$ ,  $i = 1, \dots, n$ . We consider the global least squares estimator defined by

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^d} \|Y - X\theta\|^2.$$

One can show, if the model is not misspecified, the 2-expected loss of this estimator is bounded by

$$\mathbb{E} \left[ \|X\hat{\theta} - X\theta\|^2 \right] \leq \frac{\sigma^2 \text{Rank}(X)}{n}, \quad \text{Rank}(X) \leq \min(n, d),$$

see, e.g., (Rigollet and Tsybakov, 2011, Lemma 3.1). When  $d > n$ , the above bound is equal to  $\sigma^2$  whenever  $X$  is of full rank. Furthermore, it follows from the bound  $\sigma^2$  cannot be improved when employing the least squares estimator for  $d > n$ . Consequently, the estimator lacks statistical significance in high-dimensional scenarios, with the risk potentially failing to decrease as the sample size  $n$  increases, a phenomenon commonly termed as the ‘‘curse of dimensionality’’. In extreme value analysis, estimation becomes even more challenging as it relies on the largest observations of a sample, reducing the number of data points. This could elucidate why multivariate extreme value theory has been limited to small dimensions (see, e.g., Einmahl et al. (2018, 2012); Genest and Segers (2009)).

In order to quantify non-asymptotically the performance of an estimator, we need some tools to replace the classical convergence theorems used in classical statistics. A typical example of



## Introduction

---

convergence theorem is the central limit theorem, which describes the asymptotic convergence of an empirical mean toward its expected value : for  $f : \mathbb{R} \rightarrow \mathbb{R}$  and  $X_1, \dots, X_n$  i.i.d. such that  $\sigma^2 = \text{Var}(f(X_1)) < \infty$ , we have as  $n \rightarrow \infty$

$$\sqrt{\frac{n}{\sigma^2}} \left( \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[f(X_1)] \right) \xrightarrow{d} Z, \quad Z \sim \mathcal{N}(0, 1).$$

Informally, the gap between the empirical mean and the statistical mean tends to behave approximately like  $\sqrt{\sigma^2/n}$  when  $n$  is sufficiently large. By considering that  $f$  is Lipschitz continuous with a Lipschitz constant  $L$  and  $X_1, \dots, X_n$  i.i.d. random variables with finite variance  $\sigma^2$ , we proceed with the following analysis

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \geq \frac{L\sigma}{\sqrt{n}} x \right\} \leq \mathbb{P} \{ Z \geq x \} \leq e^{-x^2/2}.$$

Such a Gaussian tail inequality provides much more precise control of the fluctuations of an estimator. In this section, we begin the investigation of such concentration inequalities. A simple, yet powerful method for bounding tail probabilities relies on Markov's inequality for positive random variables  $X$ :

$$\mathbb{P} \{ X \geq t \} \leq \frac{\mathbb{E}[X \mathbf{1}_{\{X \geq t\}}]}{t} \leq \frac{\mathbb{E}[X]}{t}.$$

The relevance of this inequality is contingent upon  $\mathbb{E}[X] < \infty$ , that is,  $X$  is integrable. Through a more sophisticated approach, Markov's inequality can be boosted, resulting in much more precise estimates. Such improvements becomes feasible when  $X$  satisfies stronger integrability criteria. If  $\phi$  represents a non-decreasing and non-negative function defined on a (potentially infinite) interval  $I \subset \mathbb{R}$ , and  $X$  denotes a random variable with values in  $I$ , then Markov's inequality implies that for every  $t \in I$ , where  $\phi(t) > 0$

$$\mathbb{P} \{ X \geq t \} \leq \mathbb{P} \{ \phi(X) \geq \phi(t) \} \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}. \quad (1.18)$$

The most prevalent applications of this principle is Chebyshev's inequality derived by setting  $\phi(t) = t^2$  over  $I = (0, \infty)$  and considering the random variable  $|X - \mathbb{E}[X]|$ . In this scenario, we obtain, provided that we have an upper bound on the variance

$$\mathbb{P} \{ |X - \mathbb{E}[X]| > t \} \leq \frac{\text{Var}(X)}{t^2}.$$

However, this bound only diminishes at a rate of  $t^{-2}$ , and we cannot achieve a Gaussian tail bound using this approach. The Cramér-Chernoff method identifies the most optimal bound for a tail probability through Markov's inequality by employing an exponential  $\phi(t) = e^{\lambda t}$ .

**Lemma 1.1.7.** *Define the log-moment generating function of a random variable  $X$  and its Legendre dual  $\psi^*$  as*

$$\psi(\lambda) := \ln \mathbb{E} \left[ e^{\lambda(X - \mathbb{E}X)} \right], \quad \psi^*(t) = \sup_{\lambda > 0} \{ \lambda t - \psi(\lambda) \}.$$

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

Then  $\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq e^{-\psi^*(t)}$  for all  $t \geq 0$ .

**Proof** Using Markov's inequality with an exponential function  $\phi(t) = e^{\lambda t}$  in (1.18) gives

$$\mathbb{P}\{X - \mathbb{E}X \geq t\} = \mathbb{P}\left\{e^{\lambda(X - \mathbb{E}X)} \geq e^{\lambda t}\right\} \leq e^{-\lambda t} \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] = e^{-(\lambda t - \psi(\lambda))}.$$

As the LHS does not depend on the choice of  $\lambda > 0$ , we can optimise the RHS over  $\lambda$  to obtain the statement of the lemma.  $\square$

**Remark 1.1.2.** The usefulness of the Chernoff bound extends far beyond proving Gaussian tails, as we will do below. One can derive many different tails behaviors in this manner. However, the approach is only applicable if  $\psi(\lambda)$  remains finite, at least for  $\lambda$  in neighborhood of 0. Consequently, to employ the Chernoff's bound, the random method use powers instead of exponentials in Markov's inequality:

$$\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq \inf_{p \in \mathbb{N}} \frac{\mathbb{E}[(X - \mathbb{E}X)_+^p]}{t^p}.$$

**Example 1.1.8 (Gaussian law).** Let  $X \sim \mathcal{N}(0, \sigma^2)$ , one can compute  $\mathbb{E}[e^{\lambda X}] = e^{\lambda^2 \sigma^2 / 2}$  and deduce that

$$\psi^*(t) = \sup_{\lambda > 0} \left\{ \lambda t - \frac{\lambda^2 \sigma^2}{2} \right\} = -\frac{t^2}{2\sigma^2}.$$

In particular, if  $X_1, \dots, X_n$  are i.i.d. and distributed according to a Gaussian law with expectancy  $\mu$  and variance  $\sigma^2$ , we can compute

$$\forall t > 0 \quad \mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n X_i - \mu > \sqrt{\frac{2\sigma^2 t}{n}} \right\} \leq e^{-t}.$$

**Example 1.1.9 (Poisson Law).** Let  $X \sim \mathcal{P}(\theta)$ , then  $\mathbb{E}[e^{\lambda X}] = e^{-\theta} \sum_{k=0}^{\infty} \frac{e^{\lambda k} \theta^k}{k!} = e^{\theta(e^\lambda - 1)}$ . Since  $\mathbb{E}[X] = \theta = \text{Var}(X)$ , we obtain  $\mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] = e^{\theta(e^\lambda - 1 - \lambda)}$ . By denoting  $f(\lambda) = \lambda t - \theta(e^\lambda - 1 - \lambda)$ , we obtain that

$$\psi^*(t) = \sup_{\lambda > 0} f(\lambda) = (\theta + t) \ln \left( 1 + \frac{t}{\theta} \right) - t = \theta h \left( \frac{t}{\theta} \right),$$

with  $h(u) = (1 + u) \ln(1 + u) - u$ . If  $X_1, \dots, X_n$  are i.i.d. distributed according to a Poisson random variable with parameter  $\theta$ , then

$$\forall t > 0, \mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n X_i - \theta > t \right\} \leq e^{-n\theta h(t/\theta)}.$$

We can verify that

$$\forall t > 0, h(t) \geq \frac{t^2}{2(1 + t/3)},$$

so we obtain

$$\forall t > 0, \mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n X_i - \theta > t \right\} \leq e^{-\frac{nt^2}{2(1+t/3)}},$$

## Introduction

---

which also implies

$$\forall t > 0, \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - \theta > \sqrt{\frac{2\sigma^2 t}{n}} + \frac{2t}{3n} \right\} \leq e^{-t}.$$

**Example 1.1.10 (Gamma law).** Let  $X \sim \Gamma(a, b)$ , we have  $\mathbb{E}X = ab$ ,  $\text{Var}X = ab^2$  and  $\forall \lambda \in (0, 1/b)$ ,  $\mathbb{E}[e^{\lambda X}] = 1/(1 - b\lambda)^a$ . Then

$$\forall \lambda \in (0, 1/b), \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] = e^{-\lambda ab}/(1 - b\lambda)^a.$$

Let  $f(\lambda) = \lambda t + \lambda ab + a \ln(1 - b\lambda)$  and the value of  $f$  at this maximum gives  $t/b - ab(1 + t/(ab))$ . If  $X_1, \dots, X_n$  are i.i.d. with distribution  $\Gamma(a, b)$ , we have  $\sum_{i=1}^n X_i \sim \Gamma(na, b)$ , then

$$\forall t > 0, \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - ab > abt \right\} \leq e^{-na(t - \ln(1+t))}.$$

One can show that

$$t - \ln(1+t) = \sum_{k=2}^{\infty} \frac{(-1)^k t^k}{k} \leq \frac{t^2}{2(1-t)}.$$

So we finally obtain

$$\forall t \in (0, 1), \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - ab > abt \right\} \leq e^{-\frac{nat^2}{2(1-t)}},$$

which can be also rewritten as:

$$\forall t > 0, \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X \geq \sqrt{\frac{2\sigma^2 t}{n}} + \frac{2bt}{n} \right\} \leq e^{-t}.$$

**Example 1.1.11 (Binomial law).** Let  $X \sim \text{Bin}(r, \theta)$ , we thus have  $\forall \lambda > 0$ ,  $\mathbb{E}[e^{\lambda X}] = (1 - \theta + \theta e^\lambda)^r$  and  $\mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] = (e^{-\lambda\theta}(1 - \theta + \theta e^\lambda))^r$ . By setting  $f(\lambda) = \lambda(t + r\theta) - r \ln(1 + \theta + \theta e^\lambda)$  and computing its derivative, we have

$$f'(\lambda) = t + r\theta - \frac{\theta r e^\lambda}{1 - \theta + \theta e^\lambda}.$$

Hence  $f$  reaches its maximum at  $\lambda = \ln[((r\theta + t)(1 - \theta))/(\theta(r - r\theta - t))]$  and  $f$  equals at this value

$$(r\theta + t) \ln \left[ \frac{r\theta + t}{r\theta} \right] + (r(1 - \theta) - t) \ln \left[ \frac{r(1 - \theta) - t}{r(1 - \theta)} \right].$$

We introduce the Kullback-Leibler divergence between two Bernoulli with respective parameters  $p, q$ , as  $\text{KL}(p, q) = p \ln(p/q) + (1 - p) \ln[(1 - p)/(1 - q)]$ , we deduce that for any  $t \in (0, r(1 - \theta))$ ,  $\psi^*(t) = r \text{KL}(\theta + t/r, \theta)$ . Let  $X_1, \dots, X_n$  be i.i.d. random variables with law  $\text{Bin}(r, \theta)$ , we have  $\sum_{i=1}^n X_i \sim \text{Bin}(nr, \theta)$  and

$$\forall t > 0, \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - r\theta > t \right\} \leq e^{-nr \text{KL}(\theta + t/r, \theta)},$$

and one can show

$$\text{KL}(\theta + t/r, \theta) \geq \frac{(t/r)^2}{2\theta(1-\theta) + t/r},$$

hence

$$\forall t > 0, \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - r\theta > t \right\} \leq e^{-\frac{nr t^2}{2\theta(1-\theta) + r}}.$$

So we can conclude that, for any  $t > 0$ ,

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X > \sqrt{\frac{2\sigma^2 t}{n}} + \frac{t}{n} \right\} \leq e^{-t}.$$

In any of these examples mentioned above, we observe that we obtain a deviation of the following form:

$$\forall t > 0, \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X > \sqrt{\frac{2\sigma^2 t}{n}} + C \frac{t}{n} \right\} \leq e^{-t}, \quad C > 0.$$

Equivalently,

$$\forall t > 0, \mathbb{P} \left\{ \sqrt{n} \frac{\sum_{i=1}^n X_i - \mathbb{E}[X]}{\sqrt{\sigma^2}} > \sqrt{2t} + \frac{Ct}{\sqrt{\sigma^2 n}} \right\} \leq e^{-t}.$$

This type of result refines the central limit theorem by showing that the distribution of this statistic deviate from 0 akin to a Gaussian, with a corrective term of order  $\frac{Ct}{\sigma\sqrt{n}}$ ,  $C > 0$ . We now introduce a well-known condition that is sufficient to obtain concentration bounds behaving as those of Gaussians.

**Definition 1.1.11.** Let  $X$  be a random variable. This random variable is called  $\sigma^2$ -subGaussian if its log-moment generating function satisfies the inequality

$$\psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall \lambda \in \mathbb{R}.$$

So far, the sole example of subGaussian variables are Gaussians. One of the fundamental findings regarding subGaussian is that every bounded random variable falls under the category of subGaussian. This statement is made precise by Hoeffding's Lemma. Even in this simple setting, the proof provides a nontrivial illustration of the important roles of calculus in bounding moment generating functions.

**Lemma 1.1.8 (Hoeffding's Lemma).** *Let  $a \leq X \leq b$  a.s. for some  $a, b \in \mathbb{R}$ . Then  $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq e^{\lambda^2(b-a)^2/8}$ , i.e.,  $X$  is  $(b-a)^2/4$ -subGaussian.*

**Proof** We can assume without loss of generality that  $\mathbb{E}X = 0$ . In this case, we have  $\psi(\lambda) = \ln \mathbb{E}[e^{\lambda X}]$ , and we can readily compute

$$\psi'(\lambda) = \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}, \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left[ \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \right]^2.$$

## Introduction

---

Thus  $\psi''(\lambda)$  can be interpreted as the variance of the random variable  $X$  under the contorted probability measures  $d\mathbb{Q} = e^{\lambda X} / \mathbb{E}[e^{\lambda X}] d\mathbb{P}$ . But

$$\psi''(\lambda) = \text{Var}_{\mathbb{Q}}(X) \leq \mathbb{E}_{\mathbb{P}} \left[ \left( X - \frac{(b-a)}{2} \right)^2 \right] \leq \frac{(b-a)^2}{4},$$

and the fundamental theorem of calculus yields

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(s) ds \leq \lambda^2 (b-a)^2 / 8, \text{ using } \psi(0) = 0 \text{ and } \psi'(0) = \mathbb{E}X = 0.$$

□

Hoeffding's Lemma relies solely on the knowledge that the random variables are bounded and does not require any additional information about them. However, when the variance of  $X_i$  is small, then we get a sharper inequality. Before stating it, let us fix  $h_1(x) = 1 + x + \sqrt{1 + 2x}$ .

**Theorem 1.1.14.** *If  $X$  verifies the Bernstein's condition, i.e., there exist  $v^2 > 0$ ,  $b \geq 0$  such that*

$$\forall \lambda \in (0, 1/b), \quad \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \leq \frac{v^2 s^2}{2(1 - bs)},$$

then

$$\forall t > 0, \quad \mathbb{P}\{X - \mathbb{E}X > t\} \leq e^{-\frac{v^2}{b^2} h_1\left(\frac{bt}{v^2}\right)}, \quad \mathbb{P}\left\{X - \mathbb{E}X > \sqrt{2v^2 t} + bt\right\} \leq e^{-t}.$$

**Proof** Let  $t > 0$ , we will apply Chernoff's method and we define the function

$$\psi(\lambda) = \lambda t - \frac{v^2 \lambda^2}{2(1 - b\lambda)} = \lambda \left( t + \frac{v^2}{2b} \right) + \frac{v^2}{2b^2} - \frac{v^2}{2b^2(1 - b\lambda)}$$

which its maximum is equal to

$$\psi^*(t) = \sup_{\lambda > 0} \psi(\lambda) = \frac{v^2}{b^2} \left( 1 + \frac{bt}{v^2} - \sqrt{1 + \frac{2bt}{v^2}} \right) = \frac{v^2}{b^2} h_1\left(\frac{bt}{v^2}\right).$$

So the first result is a consequence of Chernoff's method. For the second results, we write

$$h_1(x) = \frac{1 + 2x}{3} - \sqrt{1 + 2x} + \frac{1}{2} = \left( \frac{\sqrt{1 + 2x} - 1}{2} \right)^2,$$

such that for any  $u > 0$ , we have  $h_1(x) = u$  if  $x \in [(1 + \sqrt{2u})^2 - 1]/2 = \sqrt{2u} + u$ . Then by considering  $h_1^{-1}(u) = \sqrt{2u} + u$ ,

$$u = \frac{v^2}{b} h_1^{-1}\left(\frac{b^2 u}{v^2}\right) = \sqrt{2v^2 u} + bu$$

and using the first result

$$\forall u > 0, \quad \mathbb{P}\left\{X - \mathbb{E}X > \sqrt{2v^2 u} + bu\right\} \leq e^{-u}.$$

□

We will consider a direct improvement of these above inequalities that allow some dependence between the random variables. Write

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^n \Delta_k,$$

where

$$\Delta_k = \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_k] - \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_{k-1}]$$

are martingale differences. The following simple result, which exploits the nice behavior of the exponential of a sum, could be viewed as an illustration of the entropy method (see (Boucheron et al., 2013, Chapter 6)). This approach is commonly referred to as the martingale method and the following lemma is due to Azuma (1967).

**Lemma 1.1.9.** *Let  $\{\mathcal{F}_k\}_{k \leq n}$  be any filtration, and let  $\Delta_1, \dots, \Delta_n$  be random variables that satisfy the following properties for  $k = 1, \dots, n$ :*

- (i) *Martingale difference property:  $\Delta_k$  is  $\mathcal{F}_k$  and  $\mathbb{E}[\Delta_k | \mathcal{F}_{k-1}] = 0$ .*
- (ii) *Conditional subGaussian property:  $\mathbb{E}[e^{\lambda \Delta_k} | \mathcal{F}_{k-1}] \leq e^{\lambda^2 \sigma_k^2 / 2}$  a.s. for  $\lambda \geq 0$ .*

Then the sum  $\sum_{k=1}^n \Delta_k$  is subGaussian with variance proxy  $\sum_{k=1}^n \sigma_k^2$ .

**Proof** For any any  $1 \leq k \leq n$ , we can compute,

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^k \Delta_i}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{k-1} \Delta_i}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{k-1} \Delta_i} \mathbb{E}\left[e^{\lambda \Delta_k} | \mathcal{F}_{k-1}\right]\right] \leq e^{\lambda^2 \sigma_k^2 / 2} \mathbb{E}\left[e^{\lambda \sum_{i=1}^{k-1} \Delta_i}\right].$$

It follows by induction that  $\mathbb{E}\left[e^{\lambda \sum_{i=1}^n \Delta_i}\right] \leq e^{\lambda^2 \sum_{i=1}^n \sigma_i^2 / 2}$ . □

Combined with Hoeffding's Lemma, we now obtain a classical result on the tail behavior of sums of martingale differences.

**Corollary 1.1.2 (Azuma-Hoeffding inequality).** *Let  $\{\mathcal{F}_k\}_{k \leq n}$  be any filtration and  $\Delta_k, A_k, B_k$  satisfy the following properties for  $k = 1, \dots, n$ :*

- (i) *Martingale difference property:  $\Delta_k$  is  $\mathcal{F}_k$  and  $\mathbb{E}[\Delta_k | \mathcal{F}_{k-1}] = 0$ .*
- (ii) *Predictable bounds:  $A_k, B_k$  are  $\mathcal{F}_{k-1}$  measurable and  $A_k \leq \Delta_k \leq B_k$  a.s.*

Then  $\sum_{k=1}^n \Delta_k$  is subGaussian with variance proxy  $\frac{1}{4} \sum_{k=1}^n \|B_k - A_k\|_\infty^2$ . In particular, we obtain for every  $t > 0$  the tail bound

$$\mathbb{P}\left\{\sum_{k=1}^n \Delta_k \geq t\right\} \leq \exp\left\{-\frac{2t^2}{\sum_{k=1}^n \|B_k - A_k\|_\infty^2}\right\}.$$

**Proof** Applying Hoeffding's Lemma (see Lemma 1.1.8) to  $\Delta_k$  conditionally on  $\mathcal{F}_{k-1}$  implies

$$\mathbb{E}\left[e^{\lambda \Delta_k} | \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 (B_k - A_k)^2 / 8}.$$

The result now follows from Lemma 1.1.8. □

## Introduction

---

Let us return to the case of functions  $f(X_1, \dots, X_n)$  of independent random variables  $X_1, \dots, X_n$ . Using the Azuma-Hoeffding inequality, we readily obtain our first and simple subGaussian concentration inequality. Denote by

$$D_i f(x) := \sup_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n) - \inf_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

the “discrete derivatives”.

**Theorem 1.1.15 (McDiarmid).** *For  $X_1, \dots, X_n$  independent,  $f(X_1, \dots, X_n)$  is subGaussian with variance proxy  $\frac{1}{n} \sum_{k=1}^n \|D_k f\|_\infty^2$ . In particular*

$$\mathbb{P}\{f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t\} \leq e^{-\frac{2t^2}{\sum_{k=1}^n \|D_k f\|_\infty^2}}.$$

**Proof** We write

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^n \Delta_k.$$

Note that  $A_k \leq \Delta_k \leq B_k$  with

$$A_k = \mathbb{E} \left[ \inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_n) \middle| X_1, \dots, X_{k-1} \right],$$

$$B_k = \mathbb{E} \left[ \sup_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_n) \middle| X_1, \dots, X_{k-1} \right],$$

where we have used the independence of  $X_k$  and  $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n$ . The results now follow immediately from the Azuma-Hoeffding inequality of Corollary 1.1.2 once we note that  $|B_k - A_k| \leq \|D_k f\|_\infty$ .  $\square$

The main objective of the subsequent discussion is to broaden these inequalities to larger classes of dependence sequences, such as  $\alpha$ -mixing and  $\varphi$ -mixing processes. A Hoeffding type inequality, which also extends the Azuma inequality Azuma (1967) for martingales to dependent sequences an can be found in Rio (1999) and stated below:

**Theorem 1.1.16 (Theorem 2.4 in Rio (1999)).** *Let  $(X_i, i \in \mathbb{Z})$  be a sequence of real-valued bounded random variables and let  $(m_1, \dots, m_n)$  be an  $n$ -tuple positive reals such that*

$$\sup_{j \in \{1, \dots, n\}} \left( \|X_j\|_\infty^2 + 2 \|X_j \sum_{k=i+1}^j \mathbb{E}[X_k | \mathcal{F}_i]\|_\infty \right) \leq m_i \quad \text{for every } i \in \{1, \dots, m\} \quad (\text{a})$$

with the convention  $\sum_{k=i+1}^i \mathbb{E}[X_k | \mathcal{F}_i] = 0$ . Then, for any nonnegative integer  $p$ ,

$$\mathbb{E} \left[ S_n^{2p} \right] \leq \frac{(2p)!}{2^{pp} p!} \left( \sum_{i=1}^n m_u \right)^p. \quad (\text{b})$$

Consequently, for any positive  $x$ ,

$$\mathbb{P}\{|S_n| \geq x\} \leq \sqrt{e} \exp \left\{ -x^2 / (2m_1 + \dots + 2m_n) \right\}. \quad (\text{c})$$

## 1.1 A survival guide for high-dimensional extremal dependence modeling

---

Using this theorem, we can provide a Hoeffding type inequality for uniformly mixing sequences of bounded random variables.

**Corollary 1.1.3.** *Let  $(X_k, k \in \mathbb{Z})$  be a sequence of centered and real-valued bounded random variables. Set  $\theta_n = 1 + \sum_{i=1}^{k-1} \varphi_i$  and  $M_i = \|X_i\|_\infty^2$ . Then for any positive integer  $p$ ,*

$$\mathbb{E} [S_n^{2p}] \leq \frac{(2p)!}{p!} \left(\frac{\theta_n}{2}\right)^p (M_1 + \cdots + M_n)^p. \quad (\text{a})$$

Next, for any positive  $x$ ,

$$\mathbb{P} \{|S_n| \geq x\} \leq \sqrt{e} \exp \left\{ -x^2 / (2\theta_n M_1 + \cdots + 2\theta_n M_n) \right\}. \quad (\text{b})$$

**Proof** Let us apply Theorem 1.1.16 to the sequence  $(X_k, k \in \mathbb{Z})$ . By the Riesz-Fisher theorem, there exists a random variable  $Y \in \mathbb{L}^1(\mathcal{F}_i)$  such that  $\|Y\|_1 = 1$  and

$$\|\mathbb{E}[X_k | \mathcal{F}_i]\|_\infty = |\mathbb{E}[Y \mathbb{E}[X_k | \mathcal{F}_i]]| = |\mathbb{E}[\mathbb{E}[X_k Y | \mathcal{F}_i]]|.$$

Using Equation (1.12) in Lemma 1.1.5 to obtain

$$\|\mathbb{E}[X_k | \mathcal{F}_i]\|_\infty \leq 2\varphi(\sigma(X_k), \mathcal{F}_i) \|X\|_\infty \|Y\|_1 = 2\varphi_{k-i} \|X\|_\infty.$$

Hence, we may apply Theorem 1.1.16 with

$$m_i = M_i + 4 \sum_{k=i+1}^n \sqrt{M_i M_k} \varphi_{k-i}.$$

Summing on  $i$ , we have

$$m_1 + \cdots + m_n \leq \sum_{i=1}^n M_i + 4 \sum_{1 \leq i < k \leq n} \sqrt{M_i M_k} \varphi_{k-i} \leq \sum_{i=1}^n M_i + 2 \sum_{1 \leq i < k \leq n} (M_i + M_k) \varphi_{k-i} \leq \theta_n \sum_{i=1}^n M_i.$$

The corollary follows from both Theorem 1.1.16 and the above upper bound.  $\square$

We also state a Bernstein type inequality for strongly mixing sequence of centered and bounded random variables satisfying for a certain  $c > 0$

$$\alpha(n) \leq \exp\{-2cn\}. \quad (1.19)$$

**Theorem 1.1.17** (Theorem 1 of Merlevède et al. (2009)). *Let  $(X_k, k \in \mathbb{Z})$  be a sequence of centered real valued random variables. Suppose that the sequence satisfies (1.19) and that there exists a positive  $M$  such that  $\sup_{i \geq 1} \|X_i\|_\infty \leq M$ . Then there is positive constants  $C_1$  and  $C_2$  depending only on  $c$  such that for  $n \geq 4$  and  $t$  such that  $0 < tC_1 M(\ln n)(\ln \ln n) < 1$ , we have*

$$\ln \mathbb{E} [\exp\{tS_n\}] \leq \frac{C_2 t^2 n M^2}{1 - C_1 t M(\ln n)(\ln \ln n)}.$$



## Introduction

---

In terms of probabilities, there is a constant  $C_3$  depending on  $c$  such that for all  $n \geq 4$  and  $x \geq 0$

$$\mathbb{P}\{|S_n| \geq x\} \leq \exp\left\{-\frac{C_3 x^2}{nM^2 + Mx(\ln n)(\ln \ln n)}\right\}.$$

We restate here the concentration bound by [Kontorovich and Ramanan \(2008\)](#), as adapted by [Mohri and Rostamizadeh \(2010\)](#), to make it readily applicable to  $\varphi$ -mixing sequences.

**Theorem 1.1.18** (Theorem 8 of [Mohri and Rostamizadeh \(2010\)](#)). *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a measurable function. If  $f$  is  $\ell$ -Lipschitz with respect to the Hamming distance for some  $\ell > 0$ , then the following holds for all  $\epsilon > 0$*

$$\mathbb{P}\{|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq t\} \leq 2 \exp\left\{-\frac{2\epsilon^2}{n\ell^2 \|\Delta_n\|_\infty^2}\right\},$$

where  $\|\Delta_n\|_\infty \leq 1 + 2 \sum_{i=1}^n \varphi_i$ .

We present a new concentration inequality on the supremum of the uniform empirical process  $\mathbb{G}_n(t)$ , i.e.,

$$\mathbb{G}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\xi_i \leq t\}},$$

subject to a suitable condition on the dependence of the sequence  $(\xi_k, k \in \mathbb{Z})$  where each is uniformly distributed over the unit interval  $[0, 1]$ . This concentration inequality is significant as it leads to a concentration result for empirical quantiles, which is commonly used in the i.i.d. case in the peak-over-threshold framework (see [Shorack and Wellner, 2009](#), Inequality 1, Chapter 11.3) for a statement of this concentration result and [Engelke and Volgushev \(2022\)](#); [Goix et al. \(2015\)](#) for application of this concentration result in the extreme value literature). To our knowledge, such a result has not yet been established for mixing sequences. The specific mixing condition under consideration is outlined below where the inspiration is taken from [Wintenberger \(2010\)](#).

**Condition (A).** For all  $r \geq 1$  a coupling scheme in  $L^\infty(\mathcal{F})$  for  $(\xi_k, r+k \leq k \leq 2r+j-1)$ ,  $k \geq 1$ , exists when we can construct  $(\xi_k^*, r+j \leq k \leq 2r+j-1)$  distributed as  $(\xi_k, r+k \leq k \leq 2r+j-1)$  and independent of  $\mathcal{F}_j$  such that:

$$\sup_{1 \leq j \leq n-2r+1} \sum_{i=r+j}^{2r+j-1} \mathbb{1}_{\{\xi_i^* \neq \xi_i\}} \leq r\delta_r, \quad \text{a.s.,} \quad \forall r \geq 1. \quad (*)$$

**Theorem 1.1.19.** *For any  $n \geq 1$ , if there exists  $\delta_k$  as in Condition (A) then for  $a \in [0, 1]$ ,  $\lambda \geq \max\{2, \delta_k/a\}$  and  $1 \leq k \leq n$ ,*

$$\mathbb{P}\left\{\sup_{t \in [a, 1]} \frac{\mathbb{G}_n(t)}{t} \geq \lambda\right\} \leq e^{-nah\left(\frac{1}{2}\left(\lambda - \frac{\delta_k}{a}\right)\right)},$$

where  $h(x) = x(\ln(x) - 1) + 1$ .

**Proof** We want to prove that

$$\ln \mathbb{E} [\exp \{t\mathbb{G}_n(a)\}] \leq \delta_k t + na(e^{2t/n} - 1). \quad (1.20)$$

To deal with the dependence, we first use the Bernstein's block technique and Bernstein type of estimates on the partial sums  $\sum_{i=1}^n \mathbb{1}_{\{\xi_i \leq a\}}$ . Let us denote by  $I_j$  the  $j$ -th block of indices of size  $k$ , i.e.,  $\{(j-1)k+1, jk\}$  except the last block and let  $p$  be an integer such that  $2p-1 \leq k^{-1}n \leq 2p$ . Denote  $S_1$  and  $S_2$  the sums of even and odd blocks defined as

$$S_1 = \sum_{i \in I_{2j}, 1 \leq j \leq p} \mathbb{1}_{\{\xi_i \leq a\}}, \text{ and } S_2 = \sum_{i \in I_{2j-1}, 1 \leq j \leq p} \mathbb{1}_{\{\xi_i \leq a\}}.$$

From Cauchy-Schwarz inequality, it holds

$$\ln \mathbb{E} [\exp \{t\mathbb{G}_n(a)\}] \leq \frac{1}{2} \left( \ln \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} S_1 \right\} \right] + \ln \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} S_2 \right\} \right] \right).$$

Now let us treat in detail the term depending on  $S_1$ , the same argument applies identically to  $S_2$ . To prove (1.20), let us use the  $\mathbb{L}^\infty$ -coupling scheme and (A) to derive for all  $1 \leq m \leq p$ :

$$\left\| \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_i \leq a\}} - \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_i^* \leq a\}} \right\|_\infty \leq \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_i \neq \xi_i^*\}} \leq k\delta_k,$$

where  $|I_j| = k$  for all  $1 \leq j \leq 2p$  with  $2p-1 \leq nk^{-1} \leq 2p$ . Then for any  $t \geq 0$ , we have

$$\exp \left\{ \frac{2t}{n} \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_i \leq a\}} \right\} \leq e^{\frac{2tk}{n} \delta_k} \exp \left\{ \frac{2t}{n} \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_i^* \leq a\}} \right\}, \text{ a.s.}$$

for all  $1 \leq m \leq p$ . Applying this inequality for  $m = p$ , we have

$$\begin{aligned} \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} S_1 \right\} \right] &= \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} \sum_{1 \leq m \leq p-1} \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_i \leq a\}} \right\} \right] \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} \sum_{i \in I_{2p}} \mathbb{1}_{\{\xi_i \leq a\}} \mid \mathcal{F}_{2(p-1)} \right\} \right] \\ &\leq e^{\frac{2tk}{n} \delta_k} \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} \sum_{i \in I_1} \mathbb{1}_{\{\xi_i^* \leq a\}} \right\} \right] \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} \sum_{1 \leq m \leq p-1} \sum_{i \in I_{2m}} \mathbb{1}_{\{\xi_a \leq a\}} \right\} \right]. \end{aligned}$$

Let us do the same reasoning recursively on  $m = p-1, \dots, 2$ , to obtain finally

$$\ln \mathbb{E} \left[ \exp \left\{ \frac{2t}{n} S_1 \right\} \right] \leq 2(p-1)k\delta_k \frac{t}{n} + apk \left( e^{2t/n} - 1 \right)$$

and the inequality follows from  $2(p-1)k \leq n$  and  $pk \leq n$  since  $nk^{-1} \geq 2p-1$  and  $2p-1 \geq p$ . We hence obtain (1.20). Now using Chernoff's bound and martingale inequality (see [Shorack and Wellner \(2009\)](#) for details), then for every  $n$  and  $r \geq 0$

$$\mathbb{P} \left\{ \sup_{t \in [a, 1]} \mathbb{G}_n(t)/t \geq \lambda \right\} \leq e^{-r\lambda} \mathbb{E} \left[ e^{\frac{r}{a} \mathbb{G}_n(a)} \right] \leq e^{-r\lambda} e^{\delta_k \frac{r}{a} + na \left( e^{\frac{2r}{na}} - 1 \right)},$$

where the last inequality stems down from (1.20). By optimising with respect to  $r$  gives the result.  $\square$

Having defined all the necessary tools used in this thesis and having drawn some classical developments stemming from these definitions, we now proceed to outline the thesis by presenting its contribution within the literature.

## 1.2 Outline and contributions

Below we briefly outline and summarise the contributions of this thesis.

**Chapter 2** The problem of missing data is pervasive across various fields, particularly in environmental research, often stemming from instrument, communication and processing errors. In this chapter, we explore nonparametric approaches for evaluating extremal dependence when variables are incompletely observed, following a missing mechanism dictated by the Missing Completely At Random (MCAR) condition. Several methods for handling missing values within the context of extremes have been proposed for univariate time series. However, handling missing values when  $d \geq 2$  is still in early stages of development. The primary contribution of this chapter is to provide estimators of the  $\mathbf{w}$ -madogram, i.e.,

$$\nu(\mathbf{w}) = \mathbb{E} \left[ \bigvee_{j=1}^d \{F^{(j)}(X^{(j)})\}^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d \{F^{(j)}(X^{(j)})\}^{1/w^{(j)}} \right], \mathbf{w} \in \Delta_{d-1}$$

which involves variables that are partially observed. We then examine its theoretical properties through an asymptotic analysis using the concept of the so-called hybrid copula estimator  $\hat{C}_n^{\mathcal{H}}$  introduced by Segers (2015). Under certain high-level conditions, the hybrid copula estimator is demonstrated to satisfy the following functional central limit theorem:

$$\left( \sqrt{n}(\hat{C}_n^{\mathcal{H}}(\mathbf{u}) - C(\mathbf{u})) \right)_{\mathbf{u} \in [0,1]^d} \rightsquigarrow (\mathbb{G}_n(\mathbf{u}))_{\mathbf{u} \in [0,1]^d} \text{ in } \ell^\infty([0,1]^d),$$

where  $\mathbb{G}$  is a tight Gaussian process. Using the above convergence result, we generalise the result for the madogram estimator, and hence, the madogram-based estimator of the Pickands dependence function under the framework of missing data. We also demonstrate that a central limit theorem holds when  $\mathbf{w}$  is fixed. Furthermore, by extending the techniques Genest and Segers (2009) to an arbitrary dimension  $d \geq 2$ , we derive the asymptotic variance of this limiting Gaussian which was unknown even in the framework of fully observed data.

The findings of this chapter have been considered as a publication for the *Journal of Multivariate Analysis*, and the reference can be found below.



Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde, Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework, *Journal of Multivariate Analysis*, Volume 192, November 2022.

**Chapter 3** One of the challenging issues in multivariate extreme values theory is the high-dimensional setting, thereby calling for the use of learning methods to reduce dimensionality.

The general idea of the proposed methods is to identify subsets of variables that can take their largest values simultaneously, while the others do not. One of the first approaches proposed in this is by [Goix et al. \(2016\)](#), who focus on the subsets  $R_\alpha$  defined by

$$R_\alpha := \{\mathbf{v} \geq 0, \|\mathbf{v}\|_\infty \geq 1, v^{(j)} > 0 \text{ for } j \in \alpha, v^{(j)} = 0 \text{ for } j \notin \alpha\},$$

with  $\alpha$  a nonempty subset of  $\{1, \dots, d\}$ . This is done using  $\epsilon$ -thickened rectangles  $R_\alpha^\epsilon$  defined as:

$$R_\alpha^\epsilon := \left\{ \mathbf{v} \geq 0, \|\mathbf{v}\|_\infty, v^{(j)} > \epsilon \text{ for } j \in \alpha, v^{(j)} \leq \epsilon \text{ for } j \notin \alpha \right\}.$$

The authors propose to estimate the quantity  $\Lambda(R_\alpha)$  by

$$\Lambda_n(R_\alpha^\epsilon) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{\hat{\mathbf{V}}_i \in (\frac{n}{k})R_\alpha^\epsilon\}},$$

where  $k = k_n$  satisfies  $k \rightarrow \infty$ ,  $k/n \rightarrow 0$ , when  $n \rightarrow 0$ , and

$$\hat{\mathbf{V}}_i = \left( (1 - \hat{F}_n^{(j)}(X_i^{(j)}))^{-1} \right)_{1 \leq j \leq d},$$

with  $\hat{F}_n^{(j)}(x) = (1/n) \sum_{i=1}^n \mathbb{1}_{\{X_i^{(j)} \leq x\}}$ . This procedure thus obtain an estimate  $\mathcal{M}$  of the representation

$$\mathcal{M} = \{\Phi(C_\alpha) : \emptyset \neq \alpha \subset \{1, \dots, d\}\}, \quad C_\alpha = \mathbb{S}_{\mathcal{E}_+} \cap R_\alpha,$$

where  $\Phi$  is the spectral measure. The latter verifies the following non-asymptotic bound

$$\sup_{\emptyset \neq \alpha \subset \{1, \dots, d\}} |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \leq Cd \left( \sqrt{\frac{\ln(d/\delta)}{\epsilon k}} + M d \epsilon \right) + \text{biais}(\epsilon, k, n) \quad (1.21)$$

with probability greater than  $1 - \delta$  holds.

The sparse representation by [Goix et al. \(2017\)](#) may result in a very large number of subsets  $C_\alpha$ . The idea proposed by [Chiapino and Sabourin \(2017\)](#) is an incremental-type algorithm called CLEF, aimed at grouping together components that maybe large together. This algorithm requires a tolerance parameter  $\kappa$ . Several variants of the CLEF algorithm have been proposed by [Chiapino et al. \(2019\)](#). These approaches differ in their stopping criteria, which are based on asymptotic results of the coefficient of tail dependence. [Janßen and Wan \(2020\)](#) propose an approach based on  $k$ -means clustering by adaptating the spherical  $k$ -means clustering algorithm to the extremal setting and construct a nonparametric estimator for the theoretical cluster centres. While they provide a consistency result, a major point in this procedure remains the choice of the number of clusters and a non-asymptotic analysis to better understand the effect of the dimension  $d$  in the estimation process. To reduce dimension, [Meyer and Wintenberger \(2021\)](#) propose a method based on the Euclidean projection onto the unit simplex which introduces sparsity in the considered vector. These considerations lead to the definition of sparse regular variation. The theoretical context established, [Meyer and Wintenberger \(2023\)](#) provide a learning approach to identify subspaces  $C_\alpha$  on which extreme events appear. For a Borel subset of  $\mathbb{S}_{\mathcal{E}_+}$ , we set

$$p_n(A) = \mathbb{P} \{ \pi(\mathbf{X}/a_n) \in A \mid \|\mathbf{X}\| > a_n \}, \quad T_n(A) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{ \pi(\mathbf{X}_i/a_n) \in A, \|\mathbf{X}_i\| > a_n \}},$$

where  $\pi$  is the projection onto  $\mathbb{S}_{\mathcal{E}_+}$ , it is shown that, in the worst cases, the following holds

$$\sup_{\emptyset \neq \alpha \subset \{1, \dots, d\}} |T_n(C_\alpha) - p_n(C_\alpha)| \leq C \sqrt{\frac{d \ln(1/\delta)}{k}} + \frac{d \ln(1/\delta)}{3k}, \quad (1.22)$$

with probability greater than  $1 - \delta$ .

For the non-asymptotic bounds provided (see Equations (1.21) and (1.22)), even though these procedures are computationally efficient, we observe that in the high-dimensional setting, i.e., when  $d$  vary with  $n$  and  $d > n$ , the bound does not decrease as  $n$  grows in the worst cases, hence loosing for their statistical efficiency. Even though such information is not yet available for spherical  $k$ -means in [Janßen and Wan \(2020\)](#), this concern can be raised intuitively since this procedure relies entirely of the spectral measure, which the estimation may suffer in the high-dimensional setting (see ([Cléménçon et al., 2023](#), Theorem 3.1)).

In this chapter, we introduce a probabilistic framework called Asymptotic Independent block models (AI-block) to tackle the problem of variable clustering towards extreme values of a random vector. These models are built on the assumption that clusters of components of a multivariate random vector are independent relative to their extremes. This approach offers the advantage of being amenable to theoretical analysis, and we demonstrate that those models are identifiable.

Subsequently, we motivate and develop an algorithm specifically tailored for these models, where a tuning parameter need to be specified. We provide theoretical results and a data-driven approach to calibrate this tuning parameter, which has been shown to be effective in numerical experiments and applications. We analyse its performance in terms of exact cluster recovery for sufficiently separated clusters, using a cluster separation metric. This metric quantifies the difficulty of the statistical problem and is shown to decrease with  $k$ , the number of block maxima, as long as  $\ln(d) = o(k)$  indicating that our approach can handle high dimensions in the worst cases. We investigate the issue in the context of non-parametric estimation over block maxima of a multivariate stationary mixing random process, where the block length serves as a tuning parameter.

By studying the asymptotic independence between random vectors, this chapter hints at a new divergence measure that highlights the differences in extremal dependence structures for asymptotically dependent and independent random vectors. This divergence will be of prime interest for the next chapter, Chapter 4.

The content of this chapter derives from an ArXiv preprint and has been submitted for consideration as an original article in the *Journal of the American Statistical Association*. Currently, it has been resubmitted after a revision, and the link for access is provided below.



[Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde \(2023\), High-dimensional variable clustering based on maxima of a weakly dependent random process.](#)

**Chapter 4** The occurrence of extreme weather events is often exacerbated by the convergence of distinct geographic factors and concurrent weather patterns. When these various processes coalesce to yield a substantial impact, it is referred to as a compound event. In this chapter, our main goal is to adapt clustering techniques to handle compound extreme events involving both wind speed and precipitation in gridded climate data across Europe. To accomplish this, we utilise daily precipitation totals and maximum wind speed data obtained from the ERA5 reanalysis dataset covering the period from 1979 to 2022. The resulting dataset comprises 6655 daily precipitation totals and maximum wind speed measurements, covering  $91 \times 116$  grid cells with the chosen spatial resolution, totaling 10556 grid cells for clustering.


In the field of high-dimensional extremes, researchers have made significant contributions to identify hidden dependence structure of extremes of a random vector. While learning dependence between univariate climate extreme events is a well-studied area, multivariate compound extreme event at larger scales have received less attention. In this chapter, our objective is to expand upon the AI block model given in Chapter 3 to tackle the challenge posed by the considered environmental dataset. We hence introduced the concept of *constrained* AI block model, compelling grid cells to represent a collection of univariate time series, i.e., a random vector.

Our objective is the following: cluster a number of  $d = 10556$  pixels across Europe based on their asymptotic independence on compound precipitation and wind speed extremes where data are relatively scarce, i.e., the sample size  $n = 6655$ . To efficiently implement a fast algorithm designed for this model-based approach in such a high-dimensional setting, we employ a divergence measure that highlights the differences in extremal dependence structures for asymptotically dependent and independent random vectors. This divergence is linked to a well-known quantity in Extreme Value Theory and can be consistently estimated, under some mixing conditions, without the need of parametric assumptions.

When applied to our environmental dataset, this clustering procedure is efficient and produces clusters that are spatially concentrated, which is a pattern commonly observed in spatial processes. We also propose a simple method to better understand how the clustering is influenced by both wind speed and precipitation. To further analyze the results, we make use of a straightforward modification of our dissimilarity measure which allows us to comment on the different clusterings obtained through various algorithms.

The two previous chapters considered only hard clustering, where each variable belongs to a unique cluster and no others. However, when studying spatial processes, this hypothesis could be too restrictive because extreme events at one location could be driven by different spatial processes. This is the focus of the next chapter, where we propose to estimate the well-known max-linear model by introducing a model-based clustering approach allowing for overlapping clusters.

The outcomes of this chapter are accessible through a preprint on ArXiv and have been submitted for publication in the *Journal of the Royal Statistical Society, Series C*. Presently, the manuscript is undergoing revision. You can find the link below for accessing the preprint.

 Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde (2023), Identifying regions of concomitant compound precipitation and wind speed extremes over Europe.

**Chapter 5** In this chapter, our aim is to estimate the  $d \times K$  loading matrix  $A$ , which might exhibit sparsity and serves as the parameter for the decomposition of an observable random vector  $\mathbf{X}$ . This can be expressed as

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}.$$

In this equation,  $\mathbf{Z}$  represents an unobservable asymptotically independent random vector, serving as an underlying factor.  $\mathbf{E} \in \mathbb{R}^d$  serves as a noise vector with a tail that is lighter than that of the associated factors. Furthermore, it exhibits independence from these factors. Per the construction, the exponent measure  $\Lambda_{\mathbf{Z}}$  is

$$\Lambda_{\mathbf{Z}} = \sum_{a=1}^K \delta_0 \otimes \cdots \otimes \Lambda_{Z^{(a)}} \otimes \cdots \otimes \delta_0, \quad \Lambda_{Z^{(a)}}(dy) = y^{-2} dy.$$

Hence  $\mathbf{X}$  is also regularly varying sharing the spectral measure  $\Phi$  of the max-linear model (see Lemma 1.1.1, i.e.,

$$\Phi(\cdot) = \sum_{a=1}^K \|A_{\cdot a}\| \frac{\delta_{A_{\cdot a}}}{\|A_{\cdot a}\|}.$$

Estimating parameters in linear factor models poses a difficult task, primarily because there is no spectral density that rules out standard maximum likelihood procedures. [Janßen and Wan \(2020\)](#) introduce the spherical  $k$ -means designed for extremes employing its output to estimate  $A_{\cdot 1}/\|A_{\cdot 1}\|, \dots, A_{\cdot K}/\|A_{\cdot K}\|$  and  $\|A_{\cdot 1}\|/w, \dots, \|A_{\cdot K}\|/w$ . This method faces limitations in higher dimensions, grappling with running time difficulties or curse of dimension. Moreover, these methods also assume that the number  $K$  is known a priori, a requirement that is often scarcely fulfilled in practical scenarios. Addressing this hurdle, additional methods, as proposed by [Avella-Medina et al. \(2021, 2022\)](#), introduce a procedure coupled with the so-called screeplot to aid in the selection of the elusive number  $K$ . Despite the practical utility of such an approach, the theoretical underpinnings supporting these findings are still in their early stage of development. To our current understanding, methods for estimating  $A$  in higher dimensions have emerged specifically under the condition of a squared matrix  $A \in \mathbb{R}^{d \times d}$ . Notably, these methods have found fruitful application in contexts characterised by moderate dimensions. For instance, in Directed Acyclic Graph, [KlÜppelberg and Krali \(2021\)](#) have made noteworthy contributions, while [Kiriliouk and Zhou \(2022\)](#) have demonstrated successful applications of their estimator in environmental and financial datasets. Foremost, a critical lens on the theoretical foundations reveals a reliance on an i.i.d. sample and the asymptotic framework in the mentioned literature. The assumption of serial independence may face scrutiny when these methods are extended to environmental datasets, where deviations from serial independence are legitimately suspected. Moreover, the asymptotic framework, with a fixed arbitrary dimension  $d$  while the sample size  $n \rightarrow \infty$ , may offer limited insights into the performance of estimation processes in high-dimensional settings, i.e.,  $d$  varies with  $n$  and might even surpass the sample size.

We propose a model-based clustering via  $A$  with the crucial distinction that the covariance matrix of  $\mathbf{X}$  does not exist in our model. Within the framework of model (5.1), we consider two components, namely  $X^{(i)}$  and  $X^{(j)}$  belonging to the vector  $\mathbf{X}$ , as akin if they share a non-zero association. This association is established through the intermediary of the matrix  $A$ , connecting them to a common latent factor  $Z^{(a)}$ . Variables exhibiting this similarity are

grouped together within the cluster denoted as  $G_a$ :

$$G_a = \{j \in \{1, \dots, d\} : A_{ja} \neq 0\}, \quad \text{for each } a \in \{1, \dots, K\}.$$

Given that  $X^{(j)}$  can potentially be linked to multiple latent factors, the resulting clusters are characterised by overlap. Under two conditions, the matrix  $A$  can be recovered solely through bivariate measures, namely extremal correlation coefficients. We provide a sparse estimator  $\hat{A}$  of  $A$  that is tailored to our model specification. Our approach follows the constructive techniques used in our identifiability proofs. We place the theoretical study under considering exponentially decaying strong mixing coefficients processes. The method, under the setting  $\ln(d) = o(k)$  with  $k$ , the number of block maxima, large enough and for an appropriate choice of the tuning parameter, recovers the number of latent variables with high probability. We establish an upper bound on  $\|\hat{A} - A\|_2$  and we give guarantees for recovering the set of overlapping clusters  $\{G_a\}_{1 \leq a \leq K}$ .

All the findings of this chapter are available on a ArXiv preprint and are soon to be submitted, the link is made accessible below:

 [Alexis Boulin \(2024\), Estimating Max-Stable Random Vectors with Discrete Spectral Measure using Model-Based Clustering.](#)

**Appendices** All chapters in this thesis include numerical results considering a wide variety of dependencies. While software implementation of copulae has been extensively studied in R, methods for working with copulae in Python are still limited. In Appendix E, we introduce the package `clayton`, which provides an intuitive, user-friendly, and efficient way to sample from copulae. This package is implemented in pure Python, making it easy to install and use. The `clayton` package serves as a cornerstone for each numerical section in this manuscript.

This chapter has been recognised as an original contribution to the journal *Computo*. You can access the link provided below for further details.

 [Alexis Boulin. 2023. “A Python Package for Sampling from Copulae: Clayton.” \*Computo\*, January.](#)





## CHAPTER 2

# NON-PARAMETRIC ESTIMATOR OF A MULTIVARIATE MADOGRAM FOR MISSING-DATA AND EXTREME VALUE FRAMEWORK

The findings of this chapter are based on the following work



Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde, Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework, *Journal of Multivariate Analysis*, Volume 192, November 2022.

### Abstract.

The modeling of dependence between maxima is an important subject in several applications in risk analysis. To this aim, the extreme value copula function, characterised via the madogram, can be used as a margin-free description of the dependence structure. From a practical point of view, the family of extreme value distributions is very rich and arise naturally as the limiting distribution of properly normalised component-wise maxima. In this chapter, we investigate the nonparametric estimation of the madogram where data are completely missing at random. We provide the functional central limit theorem for the considered multivariate madogram correctly normalised, towards a tight Gaussian process for which the covariance function depends on the probabilities of missing. Explicit formula for the asymptotic variance is also given. Our results are illustrated in a finite sample setting with a simulation study. Our method is also illustrated on a sparse dataset of annual maxima rainfall in Central Eastern Canada.

## 2.1 Introduction

Management of environmental resources often requires the analysis of multivariate extreme values. In climate studies, extreme events represent a major challenge due to their consequences. The problem of missing data is present in many fields in particular in environmental research (see [Xia et al. \(1999\)](#), or ([Saunders et al., 2021](#), Section 2)), usually due to instruments, communication and processing errors. In a time series setting, the observation periods of a multivariate series could be different and overlap only partially. The problem of estimating when unequal amounts of data are available to each variable is meaningful in many applications for financial economics where data cannot be generated as neatly overlapping samples (see [Patton and Wiley \(2006\)](#)). Missing values in dependence modeling is of a prime interest as the nonparametric estimation of the empirical copula process has been tackled by [Segers \(2015\)](#)

under the **Missing Completely At Random (MCAR)** condition. In this paper, we consider nonparametric methods for assessing extremal dependencies involving variables with missing values under **MCAR** condition. We are particularly interested in the dependence structure of multivariate extreme value distribution. Formally, this concept is defined as follows.

Let  $(\Omega, \mathcal{A}, \mathbb{P})$  be a probability space and  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  be a  $d$ -dimensional random vector with values in  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , with  $d \geq 2$ . This random vector has a joint distribution function  $F$  and its margins are denoted by  $F^{(j)}(x) = \mathbb{P}\{X^{(j)} \leq x\}$  for all  $x \in \mathbb{R}$  and  $j \in \{1, \dots, d\}$ . A function  $C_\infty : [0, 1]^d \rightarrow [0, 1]$  is called a  $d$ -dimensional copula if it is the restriction to  $[0, 1]^d$  of a distribution function whose margins are given by the uniform distribution on the interval  $[0, 1]$ . Since the work of [Sklar \(1959\)](#), it is well known that every distribution function  $F$  can be decomposed as  $F(\mathbf{x}) = C(F^{(1)}(x_1), \dots, F^{(d)}(x^{(d)}))$ , for all  $\mathbf{x} \in \mathbb{R}^d$  and the copula  $C$  is unique if the margins are continuous. Under the framework of extreme, the notion of copulas leads to the so-called extreme value copulas. We will consider in the rest of the paper a  $d$ -dimensional random vector  $\mathbf{X}$  which distribution is a multivariate extreme value distribution  $F$ , i.e., its one dimensional distributions are Generalised Extreme-Value (GEV) distributions and the copula  $C_\infty$  is an extreme value copula (see [Gudendorf and Segers \(2010\)](#) or Section 1.1.2 in Chapter 1), defined by

$$C_\infty(\mathbf{u}) = \exp\left(-L(-\ln(u^{(1)}), \dots, -\ln(u^{(d)}))\right), \quad \mathbf{u} \in (0, 1]^d, \quad (2.1)$$

with  $L : [0, \infty)^d \rightarrow [0, \infty)$  the stable tail dependence function which is convex, homogeneous of order one, namely  $L(cx^{(1)}, \dots, cx^{(d)}) = cL(x^{(1)}, \dots, x^{(d)})$  for  $c > 0$  and satisfies  $\max(x^{(1)}, \dots, x^{(d)}) \leq L(x^{(1)}, \dots, x^{(d)}) \leq x^{(1)} + \dots + x^{(d)}$ ,  $\forall (x^{(1)}, \dots, x^{(d)}) \in [0, \infty)^d$ . Denote by  $\Delta_{d-1} = \{(w^{(1)}, \dots, w^{(d)}) \in [0, 1]^d : w^{(1)} + \dots + w^{(d)} = 1\}$  the unit simplex. By homogeneity,  $L$  is characterised by the *Pickands dependence function*  $\mathcal{A} : \Delta_{d-1} \rightarrow [1/d, 1]$ , which is the restriction of  $L$  to the unit simplex  $\Delta_{d-1}$ :

$$L(x^{(1)}, \dots, x^{(d)}) = (x^{(1)} + \dots + x^{(d)})\mathcal{A}(w^{(1)}, \dots, w^{(d)}), \quad w^{(j)} = \frac{x^{(j)}}{x^{(1)} + \dots + x^{(d)}}, \quad (2.2)$$

for  $j \in \{2, \dots, d\}$  and  $w^{(1)} = 1 - w^{(2)} - \dots - w^{(d)}$  with  $(x^{(1)}, \dots, x^{(d)}) \in [0, \infty)^d \setminus \{\mathbf{0}\}$ . Notice that, for every  $\mathbf{w} \in \Delta_{d-1}$  and  $u \in ]0, 1[$

$$C_\infty(u^{w^{(1)}}, \dots, u^{w^{(d)}}) = u^{\mathcal{A}(\mathbf{w})}. \quad (2.3)$$

Based on the madogram concept from geostatistics, the  $\lambda$ -madogram is introduced in [Naveau et al. \(2009\)](#) to capture bivariate extremal dependencies. The generalisation of the  $\lambda$ -madogram was previously proposed by [Fonseca et al. \(2015\)](#) and [Marcon et al. \(2017\)](#), this quantity is defined in the latter as:

$$\nu(\mathbf{w}) = \mathbb{E} \left[ \bigvee_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right], \quad (2.4)$$

if  $w^{(j)} = 0$  and  $0 < u < 1$ , then  $u^{1/w^{(j)}} = 0$  by convention. The  $\mathbf{w}$ -madogram can be interpreted as the  $L_1$ -distance between the maximum and the average of the uniform margins  $F^{(1)}(X^{(1)}), \dots, F^{(d)}(X^{(d)})$  elevated to the inverse of the corresponding weights  $w^{(1)}, \dots, w^{(d)}$ .

This quantity describes the dependence structure between extremes by its relation with the Pickands dependence function as stated by the Proposition 2.2 of [Marcon et al. \(2017\)](#), namely

$$\mathfrak{A}(\mathbf{w}) = \frac{\nu(\mathbf{w}) + c(\mathbf{w})}{1 - \nu(\mathbf{w}) - c(\mathbf{w})}, \quad (2.5)$$

with  $c(\mathbf{w}) = d^{-1} \sum_{j=1}^d w^{(j)} / (1 + w^{(j)})$ . Through this relation, it contributes to the vast literature of the estimation of the Pickands dependence function for bivariate extreme value copula (see [Pickands \(1981\)](#), [Deheuvels \(1991\)](#), [Capéraà et al. \(1997\)](#) or [Hall and Tajvidi \(2000\)](#)) but also multivariate extreme value copula, e.g., [Gudendorf and Segers \(2010\)](#). Also, a test for assessing asymptotic independence in dimension  $d \geq 2$  has been designed based on the  $\mathbf{w}$ -madogram (see [Guillou et al. \(2018\)](#)). Several methods for handling missing values in the framework of extremes have been proposed for univariate time series (see, e.g., [Ferreira et al. \(2021\)](#); [Hall and Scotto \(2008\)](#)). However, handling missing values in the context of multivariate extreme values with  $d \geq 2$  is still in their infancy.

**Main results** The main contribution of this paper is to give an estimator of the  $\mathbf{w}$ -madogram in (2.4) involving variables with missing values and to study its asymptotic properties. As far as we know, only [Guillou et al. \(2014\)](#) detailed the variance for the madogram of a bivariate random vector while taking the independent copula and found 1/90. In this paper we propose improvements in three directions : we consider a general multidimensional case ( $d \geq 2$ ), we deal with missing data and we consider a dependence structure given by an extreme value copula. Thus, we present in Theorem 2.2.1 a functional central limit theorem that gives the weak convergence for the considered multivariate madogram towards a tight Gaussian process for which the covariance function depends on the probabilities of missing. When the trajectory of our empirical process is fixed, we show in Proposition 2.2.1 the asymptotic normality of the estimator of the multivariate madogram where explicit formula for the asymptotic variance is also given. These results are transposed to the estimation of the Pickands dependence function with missing data in Corollary 2.2.2 by the use of the functional delta method.

**Notations** The symbol  $\triangleq$  means to be equal to. In order to shorten formulas, notations

$$\begin{aligned} \mathbf{u}^{(j)}(t) &\triangleq (u^{(1)}, \dots, u^{(j-1)}, t, u^{(j+1)}, \dots, u^{(d)}), \\ \mathbf{u}^{(jk)}(s, t) &\triangleq (u^{(1)}, \dots, u^{(j-1)}, s, u^{(j+1)}, \dots, u_{k-1}, t, u_{k+1}, \dots, u^{(d)}), \end{aligned}$$

will be adopted for  $s, t \in [0, 1]$ ,  $(u^{(1)}, \dots, u^{(j-1)}, u^{(j+1)}, \dots, u^{(d)}) \in [0, 1]^{d-1}$  and  $j, k \in \{1, \dots, d\}$  with  $j < k$ . The notation  $\mathbf{1}$  (resp.  $\mathbf{0}$ ) corresponds to the  $d$ -dimensional vector composed out of 1 (resp. 0). Similarly, we define  $\mathbf{1}^{(j)}(s)$ ,  $\mathbf{0}^{(j)}(s)$ ,  $\mathbf{1}^{(jk)}(s, t)$  and  $\mathbf{0}^{(jk)}(s, t)$  with the same idea of previous notations of this paragraph.

The following notations are also used. Given  $\mathcal{X}$  an arbitrary set, let  $\ell^\infty(\mathcal{X})$  denote the space of bounded real-valued functions on  $\mathcal{X}$ . For  $f : \mathcal{X} \rightarrow \mathbb{R}$ , let  $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ . Here, we use the abbreviation  $Q(f) = \int f dQ$  for a given measurable function  $f$  and signed measure  $Q$ . The arrows  $\xrightarrow{a.s.}$ ,  $\xrightarrow{d}$  denote almost sure convergence and convergence in distribution of random vectors. Weak convergence of a sequence of maps will be understood in the sense of J.Hoffman-Jørgensen (see Part 1 in [van der Vaart and Wellner \(1996\)](#)). Given that  $n \in \mathbb{N}^*$ ,  $X, X_n$  are maps from  $(\Omega, \mathcal{A}, \mathbb{P})$  into a metric space  $\mathcal{X}$  and that  $X$  is Borel measurable,  $(X_n)_{n \geq 1}$  is said to converge

weakly to  $X$  if  $\mathbb{E}^* f(X_n) \rightarrow \mathbb{E} f(X)$  for every bounded continuous real-valued function  $f$  defined on  $\mathcal{X}$ , where  $\mathbb{E}^*$  denotes outer expectation in the event that  $X_n$  may not be Borel measurable. In what follows, weak convergence is denoted by  $X_n \rightsquigarrow X$ .

The paper is organised as follows: We propose in Section 2.2 estimators of the  $\mathbf{w}$ -madogram suitable to the missing data framework. We state the weak convergence of the depicted estimators. Explicit formula for the asymptotic variance are also given. In Section 2.3, we illustrate the performance of the considered estimator in the finite-sample framework. Section 2.4 is devoted to apply our method on a dataset with missing data and non-concomittant record periods of annual maxima rainfall in Central Eastern Canada. A discussion on our assumptions and possible extensions of this work are presented in Section 2.4. All the proofs are postponed to Appendix A.1.

## 2.2 Non parametric estimation of the Madogram with missing data

We consider independent and identically distributed (i.i.d.) copies  $\mathbf{X}_1, \dots, \mathbf{X}_n$  of  $\mathbf{X}$ . In presence of missing data, we do not observe a complete vector  $\mathbf{X}_i$  for  $i \in \{1, \dots, n\}$ . We introduce  $\mathbf{I}_i \in \{0, 1\}^d$  which satisfies,  $\forall j \in \{1, \dots, d\}$ ,  $I_i^{(j)} = 0$  if  $X_i^{(j)}$  is not observed. To formalise incomplete observations, we introduce the incomplete vector  $\tilde{\mathbf{X}}_i$  with values in the product space  $\otimes_{j=1}^d (\mathbb{R} \cup \{\text{NA}\})$  (where NA denotes a missing data) such as

$$\tilde{X}_i^{(j)} = X_i^{(j)} I_i^{(j)} + \text{NA}(1 - I_i^{(j)}), \quad i \in \{1, \dots, n\}, j \in \{1, \dots, d\}.$$

We thus suppose that we observe a  $2d$ -tuple such as

$$(\mathbf{I}_i, \tilde{\mathbf{X}}_i), \quad i \in \{1, \dots, n\}, \quad (2.6)$$

i.e., at each  $i \in \{1, \dots, n\}$ , several entries may be missing. We also suppose that for all  $i \in \{1, \dots, n\}$ ,  $\mathbf{I}_i$  are i.i.d copies from  $\mathbf{I} = (I^{(1)}, \dots, I^{(d)})$  where  $I^{(j)}$  is distributed according to a Bernoulli random variable  $\mathcal{B}(p^{(j)})$  with  $p^{(j)} = \mathbb{P}(I^{(j)} = 1)$  for  $j \in \{1, \dots, d\}$ . We denote by  $p$  the probability of observing completely a realization from  $\mathbf{X}$ , that is  $p = \mathbb{P}(I^{(1)} = 1, \dots, I^{(d)} = 1)$ . Let us now define the empirical cumulative distribution in case of missing data, we write for notational convenience  $\{\tilde{\mathbf{X}}_i \leq \mathbf{x}\} \triangleq \{\tilde{X}_i^{(1)} \leq x^{(1)}, \dots, \tilde{X}_i^{(d)} \leq x^{(d)}\}$  and  $n^{(j)} = \sum_{i=1}^n I_i^{(j)}$ ,

$$\hat{F}_n^{(j)}(x) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\tilde{X}_i^{(j)} \leq x\}} I_i^{(j)}}{n^{(j)}}, \quad \forall x \in \mathbb{R}, \quad \hat{F}_n(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\tilde{\mathbf{X}}_i \leq \mathbf{x}\}} \prod_{j=1}^d I_i^{(j)}}{\sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}}, \quad \forall \mathbf{x} \in \mathbb{R}^d, \quad (2.7)$$

where  $\{\tilde{X}_i^{(j)} \leq x\} = \emptyset$  (resp.  $\{\tilde{\mathbf{X}}_i \leq \mathbf{x}\} = \emptyset$ ) if  $\tilde{X}_i^{(j)} = \text{NA}$  (resp. if there exists  $j \in \{1, \dots, d\}$  such that  $\tilde{X}_i^{(j)} = \text{NA}$ ). The idea raised here is to estimate non parametrically the margins using all available data of the corresponding series. To avoid dealing with points at the boundary of the unit square, it is more convenient to work with scaled ranks (see for example Genest and

Segers (2009)) defined explicitly by

$$\tilde{U}_{i,j} = \frac{n^{(j)}}{n^{(j)} + 1} \hat{F}_n^{(j)}(\tilde{X}_i^{(j)}) = \frac{1}{n^{(j)} + 1} \sum_{k=1}^n \mathbb{1}_{\{\tilde{X}_k^{(j)} \leq \tilde{X}_i^{(j)}\}} I_i^{(j)}, \quad j \in \{1, \dots, d\}. \quad (2.8)$$

We recall the definition of the *hybrid copula estimator* introduced by Segers (2015)

$$\hat{C}_n^{\mathcal{H}}(\mathbf{u}) = \hat{F}_n((\hat{F}_n^{(1)})^{\leftarrow}(u^{(1)}), \dots, (\hat{F}_n^{(d)})^{\leftarrow}(u^{(d)})), \quad \mathbf{u} \in [0, 1]^d,$$

where  $(\hat{F}_n^{(j)})^{\leftarrow}$  is the generalised inverse function of  $\hat{F}_n^{(j)}$  for  $j \in \{1, \dots, d\}$ , i.e.,  $(\hat{F}_n^{(j)})^{\leftarrow}(u) = \inf\{x \in \mathbb{R} | \hat{F}_n^{(j)}(x) \geq u\}$  with  $0 < u < 1$ . The normalised estimation error of the hybrid copula estimator is

$$\mathbb{C}_n^{\mathcal{H}}(\mathbf{u}) = \sqrt{n} \left( \hat{C}_n^{\mathcal{H}}(\mathbf{u}) - C_{\infty}(\mathbf{u}) \right), \quad \mathbf{u} \in [0, 1]^d. \quad (2.9)$$

On the condition that the first-order partial derivatives of the copula function  $C_{\infty}$  exists and are continuous on a subset of the unit hypercube, Segers (2012) obtained weak convergence of the normalised estimation error of the classical empirical copula process (see Galambos (1977)). To satisfy this condition, we introduce the following assumption as suggested in Segers (2012) (see Example 5.3).

**Condition A.**

- (i) The distribution function  $F$  has continuous margins  $F^{(1)}, \dots, F^{(d)}$ .
- (ii) For every  $j \in \{1, \dots, d\}$ , the first-order partial derivative  $\dot{\ell}_j$  of  $\ell$  with respect to  $x^{(j)}$  exists and is continuous on the set  $\{x \in [0, \infty)^d : x^{(j)} > 0\}$ .

The Condition A(i) guarantees that the representation  $F(\mathbf{x}) = C_{\infty}(F^{(1)}(x^{(1)}), \dots, F^{(d)}(x^{(d)}))$  is unique. Under the Condition A, the first-order partial derivatives of  $C_{\infty}$  with respect to  $u^{(j)}$  denoted as  $\dot{C}_{\infty}^{(j)}$  exists and are continuous on the set  $\{\mathbf{u} \in [0, 1]^d : 0 < u^{(j)} < 1\}$ . We now propose an estimator of the  $\mathbf{w}$ -madogram defined in Equation (2.4) under a general context with possible missing data.

**Definition 2.2.1.** Let  $(\mathbf{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$  be a sample given by Equation (2.6), we define the hybrid nonparametric estimator of the  $\mathbf{w}$ -madogram in Equation (2.4) by

$$\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) = \frac{1}{\sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}} \sum_{i=1}^n \left[ \left( \bigvee_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} \right) \prod_{j=1}^d I_i^{(j)} \right], \quad (2.10)$$

where  $\tilde{U}_{i,j}$  are scaled ranks defined as in Equation (2.8).

The intuitive idea here is to estimate the margins using all available data from the corresponding variables and estimate  $\nu(\mathbf{w})$  using only the overlapping data. Notice that in the complete data framework, i.e. when  $p = 1$  we retrieve a variation of the  $\mathbf{w}$ -madogram such as defined in Marcon et al. (2017), namely

$$\hat{\nu}_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \left[ \bigvee_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} \right],$$

with  $\tilde{U}_i^{(j)}$  in  $\{1/(n+1), \dots, n/(n+1)\}$ .

Note that the theoretical quantity defined in (2.4) does verify endpoint constraints, i.e.  $\nu(\mathbf{e}^{(j)}) = (d-1)/2d$  for all  $j \in \{1, \dots, d\}$  where  $\mathbf{e}^{(j)}$  is the  $j$ th vector of the canonical basis.

**Remark 2.2.1.** Unlike  $\nu$ , the estimator defined in (2.10) does not verify the endpoints constraints. In addition, the variance at  $\mathbf{e}_j$  does not equal 0. Indeed, suppose that we evaluate this statistic at  $\mathbf{w} = \mathbf{e}^{(j)}$  as  $\tilde{U}_i^{(j)} \in (0, 1)$  for every  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, d\}$  we obtain the following estimator

$$\hat{\nu}_n^{\mathcal{H}}(\mathbf{e}_j) = \frac{1}{\sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}} \sum_{i=1}^n \left[ \tilde{U}_i^{(j)} - \frac{1}{d} \tilde{U}_i^{(j)} \right] \prod_{j=1}^d I_i^{(j)}.$$

In this situation, the sample  $(\tilde{U}_i^{(1)}, \dots, \tilde{U}_i^{(j-1)}, \tilde{U}_i^{(j+1)}, \dots, \tilde{U}_i^{(d)})_{i=1}^n$  is taken into account through the indicators sequence  $(I_i^{(1)}, \dots, I_i^{(j-1)}, I_i^{(j+1)}, \dots, I_i^{(d)})_{i=1}^n$  and induces a supplementary variance when estimating.

Proceeding as in Naveau et al. (2009) for the bivariate case and complete data framework, we propose below a modified estimator which satisfies the endpoint constraints in the general multivariate framework with possible missing data.

**Definition 2.2.2.** Let  $(\mathbf{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$  be a sample given by Equation (2.6) and  $\hat{\nu}_n^{\mathcal{H}}(\mathbf{w})$  be as in (2.10). Given continuous functions  $\lambda^{(1)}, \dots, \lambda^{(d)} : \Delta_{d-1} \rightarrow \mathbb{R}$  verifying  $\lambda^{(j)}(\mathbf{e}^{(k)}) = \delta_{jk}$  (the Kronecker delta) for  $j, k \in \{1, \dots, d\}$ , we define the hybrid corrected estimator of the  $\mathbf{w}$ -madogram by

$$\hat{\nu}_n^{\mathcal{H}*}(\mathbf{w}) = \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \sum_{j=1}^d \frac{\lambda^{(j)}(\mathbf{w})(d-1)}{d} \left[ \frac{1}{\sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}} \sum_{i=1}^n \left( (\tilde{U}_i^{(j)})^{1/w^{(j)}} \prod_{j=1}^d I_i^{(j)} \right) - \frac{w^{(j)}}{1+w^{(j)}} \right]. \quad (2.11)$$

**Remark 2.2.2.** One has often that endpoint corrections do not have an impact to the asymptotic behavior with complete data framework and unknown margins (see Section 2.3 and 2.4 of Genest and Segers (2009)). That is not always the case in the missing data framework and this feature is of interest as discussed in Remark 2.2.1.

In the following we prove a functional central limit theorem (see Theorem 2.2.1) concerning the weak convergence of the following processes

$$\sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}) \right)_{\mathbf{w} \in \Delta_{d-1}}, \quad \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}*}(\mathbf{w}) - \nu(\mathbf{w}) \right)_{\mathbf{w} \in \Delta_{d-1}}. \quad (2.12)$$

Before presenting this result, we introduce below a specific assumption on the missing mechanism.

**Condition B.** We suppose that for all  $i \in \{1, \dots, n\}$ , the vector  $\mathbf{I}_i$  and  $\mathbf{X}_i$  are independent, i.e., the data are missing completely at random (**MCAR**).

Without missing data, the weak convergence of the normalised estimation error of the empirical copula process has been proved by Fermanian et al. (2004) under a more restrictive condition than Condition A. The difference being that  $C_\infty$  should be continuously differentiable on the closed hypercube. Denoting by  $D([0, 1]^2)$  the Skorokhod space, this statement makes use of previous results on the Hadamard differentiability of the map  $\phi : D([0, 1]^2) \rightarrow \ell^\infty([0, 1]^2)$  which transforms the cumulative distribution function  $F$  into its copula function  $C_\infty$  (see also Lemma

## 2.2 Non parametric estimation of the Madogram with missing data

3.9.28 from [van der Vaart and Wellner \(1996\)](#)). With the hybrid copula estimator, we need a technical assumption in order to guarantee the weak convergence of the process  $\mathbb{C}_n^{\mathcal{H}}$  in (2.9) (see [Segers \(2015\)](#)). We note for convenience marginal distributions and quantile functions into vector valued functions  $\mathbf{F}^{(d)}$  and  $(\mathbf{F}^{(d)})^{\leftarrow}$ :

$$\begin{aligned}\mathbf{F}^{(d)}(\mathbf{x}) &= (F^{(1)}(x^{(1)}), \dots, F^{(d)}(x^{(d)})), \quad \mathbf{x} \in \mathbb{R}^d, \\ (\mathbf{F}^{(d)})^{\leftarrow}(\mathbf{u}) &= ((F^{(1)})^{\leftarrow}(u^{(1)}), \dots, (F^{(d)})^{\leftarrow}(u^{(d)})), \quad \mathbf{u} \in [0, 1]^d.\end{aligned}$$

**Condition C.** In the space  $\ell^\infty(\mathbb{R}^d) \otimes (\ell^\infty(\mathbb{R}), \dots, \ell^\infty(\mathbb{R}))$  equipped with the topology of uniform convergence, we have the joint weak convergence

$$\left(\sqrt{n}(\hat{F}_n - F); \sqrt{n}(\hat{F}_n^{(1)} - F^{(1)}), \dots, \sqrt{n}(\hat{F}_n^{(d)} - F^{(d)})\right) \rightsquigarrow \left(\alpha \circ \mathbf{F}^{(d)}, \beta^{(1)} \circ F^{(1)}, \dots, \beta^{(d)} \circ F^{(d)}\right)$$

where the stochastic processes  $\alpha$  and  $\beta^{(j)}$ ,  $j \in \{1, \dots, d\}$  take values in  $\ell^\infty([0, 1]^d)$  and  $\ell^\infty([0, 1])$  respectively, and are such that  $\alpha \circ F$  and  $\beta^{(j)} \circ F^{(j)}$  have continuous trajectories on  $[-\infty, \infty]^d$  and  $[-\infty, \infty]$  almost surely.

Under Conditions [A](#) and [C](#), the stochastic process  $\mathbb{C}_n^{\mathcal{H}}$  in (2.9) converges weakly to the tight Gaussian process  $S_{C_\infty}$  defined by

$$S_{C_\infty}(\mathbf{u}) = \alpha(\mathbf{u}) - \sum_{j=1}^d \dot{C}_\infty^{(j)}(\mathbf{u})\beta^{(j)}(u^{(j)}), \quad \forall \mathbf{u} \in [0, 1]^d. \quad (2.13)$$

Lemma [A.1.1](#) in Appendix [A.1](#) states that the estimator  $\hat{F}_n$  of the joint distribution and estimators of margins  $\hat{F}_n^{(j)}$  defined in Equation (2.7) verify Condition [C](#) (see Appendix [A.1](#) for details). We now have all tools in hand to consider the weak convergence of the stochastic processes in Equation (2.12). We note by  $\{\mathbf{X} \leq (\mathbf{F}^{(d)})^{\leftarrow}(\mathbf{u})\} = \{X^{(1)} \leq (F^{(1)})^{\leftarrow}(u^{(1)}), \dots, X^{(d)} \leq (F^{(d)})^{\leftarrow}(u^{(d)})\}$ .

**Theorem 2.2.1.** *Let  $\mathbb{G}$  to be a tight Gaussian process and continuous functions  $\lambda^{(1)}, \dots, \lambda^{(d)} : \Delta_{d-1} \rightarrow \mathbb{R}$  verifying  $\lambda^{(j)}(\mathbf{e}_k) = \delta_{jk}$ . If  $C_\infty$  is an extreme value copula with Pickands dependence function  $\mathcal{A}$  and under Conditions [A](#) and [B](#), we have the weak convergence in  $\ell^\infty(\Delta_{d-1})$  for hybrid estimators defined in Equations (2.10) and (2.11), as  $n \rightarrow \infty$ ,*

$$\begin{aligned}\sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}) \right)_{\mathbf{w} \in \Delta_{d-1}} &\rightsquigarrow \left( \frac{1}{d} \sum_{j=1}^d \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx \right. \\ &\quad \left. - \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right)_{\mathbf{w} \in \Delta_{d-1}}, \\ \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) - \nu(\mathbf{w}) \right)_{\mathbf{w} \in \Delta_{d-1}} &\rightsquigarrow \left( \frac{1}{d} \sum_{j=1}^d (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx \right. \\ &\quad \left. - \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right)_{\mathbf{w} \in \Delta_{d-1}},\end{aligned}$$



where  $S_{C_\infty}$  is defined in (2.13),  $\alpha(\mathbf{u}) = p^{-1}\mathbb{G}(\mathbb{1}_{\{\mathbf{X} \leq \mathbf{F}_d^{\leftarrow}(\mathbf{u}), I=1\}} - C_\infty(\mathbf{u})\mathbb{1}_{\{I=1\}})$  and  $\beta^{(j)}(u^{(j)}) = (p^{(j)})^{-1}\mathbb{G}(\mathbb{1}_{\{X^{(j)} \leq (F^{(j)})^{\leftarrow}(u^{(j)}), I^{(j)}=1\}} - u^{(j)}\mathbb{1}_{\{I^{(j)}=1\}})$  for  $j \in \{1, \dots, d\}$  and  $\mathbf{u} \in [0, 1]^d$ . For  $(\mathbf{u}, \mathbf{v}, v^{(k)}) \in [0, 1]^{2d+1}$ , for  $j \in \{1, \dots, d\}$  and  $j < k$  the covariance functions of the processes  $\alpha$  and  $\beta^{(j)}$  are given by

$$\begin{aligned} \text{cov}(\beta^{(j)}(u^{(j)}), \beta^{(j)}(v^{(j)})) &= (p^{(j)})^{-1} (u^{(j)} \wedge v^{(j)} - u^{(j)}v^{(j)}), \\ \text{cov}(\beta^{(j)}(u^{(j)}), \beta^{(k)}(v^{(k)})) &= \frac{p^{(jk)}}{p^{(j)}p^{(k)}} (C_\infty(\mathbf{1}^{(j,k)}(u^{(j)}, v^{(k)})) - u^{(j)}v^{(k)}), \end{aligned}$$

and

$$\begin{aligned} \text{cov}(\alpha(\mathbf{u}), \alpha(\mathbf{v})) &= p^{-1} (C_\infty(\mathbf{u} \wedge \mathbf{v}) - C_\infty(\mathbf{u})C_\infty(\mathbf{v})), \\ \text{cov}(\alpha(\mathbf{u}), \beta^{(j)}(v^{(j)})) &= (p^{(j)})^{-1} (C_\infty(\mathbf{u}_j(u^{(j)} \wedge v^{(j)})) - C_\infty(\mathbf{u})v^{(j)}), \end{aligned}$$

where  $\mathbf{u} \wedge \mathbf{v}$  denotes the vector of componentwise minima and  $p^{(jk)} = \mathbb{P}(I^{(j)} = 1, I^{(k)} = 1)$ .

We use empirical process arguments formulated in van der Vaart and Wellner (1996) to establish such a result. The following proposition states the asymptotic distribution of the estimators and gives explicit formula for the asymptotic variances for a fixed element of the unit simplex  $\Delta_{d-1}$ .

**Proposition 2.2.1.** *Let  $\mathbf{p} = (p^{(1)}, \dots, p^{(d)}, p)$  and  $\mathbf{w} \in \Delta_{d-1}$ , under the framework of Theorem 2.2.1, we have*

$$\sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left( 0, \mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w}) \right), \quad \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) - \nu(\mathbf{w}) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left( 0, \mathcal{S}^{\mathcal{H}^*}(\mathbf{p}, \mathbf{w}) \right).$$

Moreover the asymptotic variances are given by

$$\begin{aligned} \mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w}) &= \frac{1}{d^2} \sum_{j=1}^d (p^{-1} - (p^{(j)})^{-1}) \sigma_j^2(\mathbf{w}) + \sigma_{d+1}^2(\mathbf{p}, \mathbf{w}) \\ &\quad + \frac{2}{d^2} \sum_{j < k} \left( p^{-1} - (p^{(j)})^{-1} - (p^{(k)})^{-1} + \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \right) \sigma_{jk}(\mathbf{w}) \\ &\quad - \frac{2}{d} \sum_{j=1}^d (p^{-1} - (p^{(j)})^{-1}) \sigma_j^{(1)}(\mathbf{w}) + \frac{2}{d} \sum_{j=1}^d \sum_{k=1}^d \left( (p^{(k)})^{-1} - \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \right) \sigma_{jk}^{(2)}(\mathbf{w}), \end{aligned}$$

and

$$\begin{aligned} \mathcal{S}^{\mathcal{H}^*}(\mathbf{p}, \mathbf{w}) &= \frac{1}{d^2} \sum_{j=1}^d (p^{-1} - (p^{(j)})^{-1}) (1 + \lambda^{(j)}(\mathbf{w})(d-1))^2 \sigma_j^2(\mathbf{w}) + \sigma_{d+1}^2(\mathbf{p}, \mathbf{w}) \\ &\quad + \frac{2}{d^2} \sum_{j < k} \left( p^{-1} - (p^{(j)})^{-1} - (p^{(k)})^{-1} + \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \right) (1 + \lambda^{(j)}(\mathbf{w})(d-1))(1 + \lambda^{(k)}(\mathbf{w})(d-1)) \sigma_{jk}(\mathbf{w}) \\ &\quad - \frac{2}{d} \sum_{j=1}^d (p^{-1} - (p^{(j)})^{-1}) (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \sigma_j^{(1)}(\mathbf{w}) \\ &\quad + \frac{2}{d} \sum_{j=1}^d \sum_{k=1}^d \left( (p^{(k)})^{-1} - \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \right) (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \sigma_{jk}^{(2)}(\mathbf{w}), \end{aligned}$$

## 2.2 Non parametric estimation of the Madogram with missing data

---

where explicit expressions of the functions  $\sigma_j^2$  for  $j \in \{1, \dots, d\}$ ,  $\sigma_{d+1}^2$ ,  $\sigma_{jk}$  with  $j < k$ ,  $\sigma_j^{(1)}$  with  $j \in \{1, \dots, d\}$ ,  $\sigma_{jk}^{(2)}$  for  $j, k \in \{1, \dots, d\}$  are detailed in the proof for the sake of readability.

Considering the special case of independent copula, Corollary 2.2.1 below gives a closed form of the limit variance which no longer depends on the Pickands dependence function.

**Corollary 2.2.1.** *In the framework of Theorem 2.2.1 and if  $C_\infty(\mathbf{u}) = \prod_{j=1}^d u^{(j)}$ , then the functions  $\sigma_{d+1}^2$ ,  $\sigma_j^{(1)}$  with  $j \in \{1, \dots, d\}$ , have the following forms, for  $\mathbf{w} \in \Delta_{d-1}$  :*

$$\begin{aligned}\sigma_{d+1}^2(\mathbf{p}, \mathbf{w}) &= \frac{1}{4} \left( \frac{1}{3p} - \sum_{j=1}^d (p^{(j)})^{-1} \frac{w^{(j)}}{4 - w^{(j)}} \right), \\ \sigma_j^{(1)}(\mathbf{w}) &= \frac{1}{2} \left[ \frac{1}{3} - \frac{1}{1 + w^{(j)}} \right] + \frac{w^{(j)}}{3(1 + w^{(j)})(3 + w^{(j)})},\end{aligned}$$

and  $\sigma_{jk}$  for  $j < k$ ,  $\sigma_{jk}^{(2)}$  for  $j < k$  and  $\sigma_{kj}^{(2)}$  with  $k < j$  are constants and equal to 0.

**Remark 2.2.3.** From our knowledge, only [Guillou et al. \(2014\)](#) gave an explicit value of the variance for the madogram of a bivariate random vector considering the independent copula. The result stated in Corollary 2.2.1 is not an extension of this result because the hypothesis  $\mathbf{w} \in \Delta_{d-1}$  is crucial. Nevertheless, the same techniques used to prove Proposition 2.2.1 can be applied to show a similar explicit formula of the asymptotic variance for an extension of the madogram in [Guillou et al. \(2014\)](#) for  $d \geq 2$ .

Weak consistency of our estimators directly comes down from Proposition 2.2.1. We are nonetheless able to state the strong consistency only under Condition B.

**Proposition 2.2.2 (Strong consistency).** *Let  $(\mathbf{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$  an i.i.d sample given by Equation (2.6). Under Condition B for a fixed  $\mathbf{w} \in \Delta_{d-1}$ , it holds that*

$$\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) \xrightarrow[n \rightarrow \infty]{a.s.} \nu(\mathbf{w}), \quad \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) \xrightarrow[n \rightarrow \infty]{a.s.} \nu(\mathbf{w}).$$

For the rest of this section, we use our previous results to state some properties of the Pickands estimator in the missing data framework.

It is a common knowledge that the  $\mathbf{w}$ -madogram is of main interest to construct of the Pickands dependence function. Indeed, given Equation (2.5), one can define an estimator of the Pickands dependence function by estimating the  $\mathbf{w}$ -madogram and using it as a plug-in estimator. Most interesting properties of the  $\mathbf{w}$ -madogram such as strong consistency and the weak convergence are thus translated for the Pickands estimator using continuous mapping theorem and the Delta method. In the missing data framework we define the following estimator.

**Definition 2.2.3.** Let  $(\mathbf{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$  be a sample given by (2.6), the hybrid nonparametric estimator of the Pickands dependence function is defined as

$$\hat{\mathcal{A}}_n^{\mathcal{H}^*}(\mathbf{w}) = \frac{\hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) + c(\mathbf{w})}{1 - \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) - c(\mathbf{w})}, \quad (2.14)$$

where  $\hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w})$  defined in Equation (2.11) and  $c(\mathbf{w}) = d^{-1} \sum_{j=1}^d w^{(j)} / (1 + w^{(j)})$ .

Using the results of [Marcon et al. \(2017\)](#) (namely, Theorem 2.4), Proposition 2.2.1 and Proposition 2.2.2 of this paper, we state the following corollary.

**Corollary 2.2.2.** *Let  $\mathbf{p} = (p^{(1)}, \dots, p^{(d)}, p)$  and  $(\mathbf{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$  be a sample given by (2.6). For  $\mathbf{w} \in \Delta_{d-1}$ , if  $C_\infty$  is an extreme value copula with Pickands dependence function and under Condition B, it holds that*

$$\hat{\mathcal{A}}_n^{\mathcal{H}^*}(\mathbf{w}) \xrightarrow[n \rightarrow \infty]{a.s.} \mathcal{A}(\mathbf{w}).$$

Furthermore, if  $C_\infty$  additionally verifies Conditions A (i) and A(ii), we obtain

$$\sqrt{n} \left( \hat{\mathcal{A}}_n^{\mathcal{H}^*}(\mathbf{w}) - \mathcal{A}(\mathbf{w}) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \mathcal{V}(\mathbf{p}, \mathbf{w})),$$

where the closed formula of the asymptotic variance is  $\mathcal{V}(\mathbf{p}, \mathbf{w}) = (1 + \mathcal{A}(\mathbf{w}))^4 \mathcal{S}^{\mathcal{H}^*}(\mathbf{p}, \mathbf{w})$ , with  $\mathcal{S}^{\mathcal{H}^*}(\mathbf{p}, \mathbf{w})$  as in Proposition 2.2.1.

## 2.3 Numerical results

In this section we verify our findings concerning the closed formula of the asymptotic variances through a simulation study. To do so, we compare empirical counterparts of the asymptotic variances computed out with Monte Carlo simulations with the explicit asymptotic variances given by Proposition 2.2.1. Our simulation studies are implemented using Python programming language and all the codes are available online via the link <https://github.com/Aleboul/missing> heading to a Github repository.

### 2.3.1 Presentation of the models

We present here the six models (**M1** to **M6**) used for this simulation study. The  $d$ -dimensional Gumbel and the asymmetric logistic models are considered in models **M1** and **M2** below, the remaining ones (models **M3** to **M6**) concern only the bivariate case.

**M1** **The symmetric logistic, or Gumbel model** [Gumbel \(1960a\)](#) is defined by the following Pickands dependence function

$$\mathcal{A}(w^{(1)}, \dots, w^{(d)}) = \left( \sum_{j=1}^d (w^{(j)})^\theta \right)^{1/\theta},$$

with  $\theta \in [1, \infty)$ . We retrieve the independent case when  $\theta = 1$  and the dependence between the variables is stronger as  $\theta$  goes to infinity. The restriction to  $d = 2$  is immediate from the definition.

**M2** Let  $B$  be the set of all nonempty subsets of  $\{1, \dots, d\}$  and  $B_1 = \{b \in B, |b| = 1\}$ , where  $|b|$  denotes the number of elements in the set  $b$ . **The asymmetric logistic model** in [Tawn \(1990\)](#) is defined by the following Pickands dependence function

$$\mathcal{A}(w^{(1)}, \dots, w^{(d)}) = \sum_{b \in B} \left( \sum_{j \in b} (\theta_{j,b} w^{(j)})^{\theta_b} \right)^{1/\theta_b},$$

where  $\theta_b \in [1, \infty)$  for all  $b \in B \setminus B_1$ , and the asymmetry parameters  $\theta_{j,b} \in [0, 1]$  for all  $b \in B$  and  $j \in b$ . The model should verify the following constraints  $\sum_{b \in B(j)} \theta_{j,b} = 1$  for  $j \in \{1, \dots, d\}$  where  $B(j) = \{b \in B, j \in b\}$  and if  $\theta_b = 1$  for every  $b \in B \setminus B_1$ , then  $\theta_{j,b} = 0 \forall j \in b$ . The model contains  $2^d - d - 1$  dependence parameters and  $d(2^{d-1} - 1)$  asymmetry parameters. In case of  $d = 2$ , we go back to the asymmetric logistic model in [Tawn \(1988\)](#), namely

$$\mathcal{A}(w) = (1 - \psi_1)w + (1 - \psi_2)(1 - w) + \left[ (\psi_1 w)^\theta + (\psi_2(1 - w))^\theta \right]^{1/\theta},$$

with  $\theta \in [1, \infty)$ ,  $\psi_1, \psi_2 \in [0, 1]$ . For  $d = 3$ , the Pickands dependence function is expressed as

$$\begin{aligned} \mathcal{A}(\mathbf{w}) = & \alpha_1 w^{(1)} + \psi_1 w^{(2)} + \phi_1 w^{(3)} + \left( (\alpha_2 w^{(1)})^{\theta_1} + (\psi_2 w^{(2)})^{\theta_1} \right)^{1/\theta_1} \\ & + \left( (\alpha_3 w^{(2)})^{\theta_2} + (\phi_2 w^{(3)})^{\theta_2} \right)^{1/\theta_2} + \left( (\psi_3 w^{(2)})^{\theta_3} + (\phi_3 w^{(3)})^{\theta_3} \right)^{1/\theta_3} \\ & + \left( (\alpha_4 w^{(1)})^{\theta_4} + (\psi_4 w^{(2)})^{\theta_4} + (\phi_4 w^{(3)})^{\theta_4} \right)^{1/\theta_4}, \end{aligned}$$

where  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_4)$ ,  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_4)$ ,  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_4)$  are all elements of  $\Delta^3$ .

**M3** The asymmetric negative logistic model in [Joe \(1990\)](#) is defined via

$$\mathcal{A}(w) = 1 - \left[ (\psi_1(1 - w))^{-\theta} + (\psi_2 w)^{-\theta} \right]^{-1/\theta},$$

with parameters  $\theta \in (0, \infty)$ ,  $\psi_1, \psi_2 \in (0, 1]$ . The special case  $\psi_1 = \psi_2 = 1$  returns the Galambos model [Oliveira and Galambos \(1977\)](#).

**M4** The asymmetric mixed model in [Tawn \(1988\)](#) corresponds to

$$\mathcal{A}(w) = 1 - (\theta + \kappa)w + \theta w^2 + \kappa w^3,$$

with parameters  $\theta$  and  $\kappa$  satisfying  $\theta \geq 0$ ,  $\theta + 3\kappa \geq 0$ ,  $\theta + \kappa \leq 1$ ,  $\theta + 2\kappa \leq 1$ . The special case  $\kappa = 0$  and  $\theta \in [0, 1]$  yields the symmetric mixed model. In the symmetric mixed model, when  $\theta = 0$ , we recover the independent copula.

**M5** The model of [Hüsler and Reiss](#) in [Hüsler and Reiss \(1989\)](#) is given by the Pickands dependence function

$$\mathcal{A}(t) = (1 - t)\Phi\left(\theta + \frac{1}{2\theta} \ln\left(\frac{1-t}{t}\right)\right) + t\Phi\left(\theta + \frac{1}{2\theta} \ln\left(\frac{t}{1-t}\right)\right),$$

where  $\theta \in (0, \infty)$  and  $\Phi$  is the standard normal distribution function. As  $\theta$  goes to  $0^+$ , the dependence between the two variables increases. When  $\theta$  goes to infinity, we are in case of near independence.

**M6** The Student  $t$ -EV model in [Demarta and McNeil \(2005\)](#) is given by

$$\begin{aligned} \mathcal{A}(w) = & wt_{\nu+1}(z_w) + (1 - w)t_{\nu+1}(z_{1-w}), \\ \text{with } z_w = & (1 + \nu)^{1/2} \left[ \{w/(1 - w)\}^{1/\nu} - \theta \right] (1 - \theta^2)^{-1/2}, \end{aligned}$$

and parameters  $\nu > 0$ , and  $\theta \in (-1, 1)$ , where  $t_{\nu+1}$  is the distribution function of a Student- $t$  random variable with  $\nu + 1$  degrees of freedom.

### 2.3.2 Description of numerical experiments

For each numerical experiment, the endpoint-corrected  $\mathbf{w}$ -madogram estimator in (2.11) is computed using  $\lambda^{(j)}(\mathbf{w}) = w^{(j)}$ . The study consists in three different experiments (**E1**, **E2** and **E3**). For all experiments, the empirical counterpart of the asymptotic variance given by Proposition 2.2.1 is computed out through a given grid of the simplex  $\Delta_{d-1}$ . For a given element  $\mathbf{w}$  of this grid,  $n_{iter} \in \mathbb{N} \setminus \{0\}$  random samples of size  $n$  are generated from the models **M1** to **M6** given above. By using these samples we estimate the associated  $\mathbf{w}$ -madogram. We thus compute the empirical variance of the normalised estimation error namely,

$$\mathcal{E}_n^{\mathcal{H}}(\mathbf{w}) \triangleq \widehat{Var} \left( \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}) \right) \right), \quad \mathcal{E}_n^{\mathcal{H}^*}(\mathbf{w}) \triangleq \widehat{Var} \left( \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) - \nu(\mathbf{w}) \right) \right), \quad (2.15)$$

where  $\hat{\nu}_n^{\mathcal{H}}$  and  $\hat{\nu}_n^{\mathcal{H}^*}$  are the vectors composed out of the  $n_{iter}$  hybrid and corrected estimators (see Equations (2.10) and (2.11)) of the  $\mathbf{w}$ -madogram, respectively. We also define the Mean Integrated Squared Error (MISE) between  $\mathcal{E}_n^{\mathcal{H}}$  and  $\mathcal{S}^{\mathcal{H}}$  the asymptotic variance computed in Proposition 2.2.1 (resp. between  $\mathcal{E}_n^{\mathcal{H}^*}$  and  $\mathcal{S}^{\mathcal{H}^*}$ ), that is

$$MISE^{\mathcal{H}} \triangleq \mathbb{E} \left[ \int_{\Delta_{d-1}} \left( \mathcal{E}_n^{\mathcal{H}}(\mathbf{w}) - \mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w}) \right)^2 d\mathbf{w} \right], \quad (2.16)$$

$$MISE^{\mathcal{H}^*} \triangleq \mathbb{E} \left[ \int_{\Delta_{d-1}} \left( \mathcal{E}_n^{\mathcal{H}^*}(\mathbf{w}) - \mathcal{S}^{\mathcal{H}^*}(\mathbf{p}, \mathbf{w}) \right)^2 d\mathbf{w} \right]. \quad (2.17)$$

**E1** We set  $d = 2$ . A Monte Carlo study is implemented here to illustrate Proposition 2.2.1 in finite-sample setting with missing data. We consider **M2**, **M3**, **M4**, **M5** and **M6** where we fix  $n_{iter} = 300$  and  $n = 1024$ . The chosen grid is  $\{1/200, \dots, 199/200\}$  and we take  $p^{(1)} = p_2 = 0.75$ . We estimate  $MISE^{\mathcal{H}}$  in (2.16) by

$$\widehat{MISE}_n^{\mathcal{H}} = \frac{1}{10} \sum_{l=1}^{10} \frac{1}{199} \sum_{k=1}^{199} \left( \mathcal{E}_{n,l}^{\mathcal{H}} \left( \frac{k}{200} \right) - \mathcal{S}^{\mathcal{H}} \left( \mathbf{p}, \frac{k}{200} \right) \right)^2,$$

with  $\mathcal{E}_{n,l}^{\mathcal{H}}$ ,  $l \in \{1, \dots, 10\}$  is the empirical counterpart of  $\mathcal{S}^{\mathcal{H}}$  taking the empirical variance of 30 estimators  $\hat{\nu}_n^{\mathcal{H}}(\mathbf{w})$  where  $\mathbf{w} = (k/200, 1 - k/200)$  and  $k \in \{1, \dots, 199\}$ . Each estimator of the  $\mathbf{w}$ -madogram is computed out through a random sample with  $n = 1024$ . By using the second equation in (2.16), the estimator  $\widehat{MISE}_n^{\mathcal{H}^*}$  is defined similarly.

**E2** We fix  $d = 3$  and we consider **M1** and **M2** with  $n_{iter} = 100$  and  $n = 512$ . We set the dependence parameter as  $\theta = 1$  and  $\theta = 2$  for the first model. For the second one we take  $\boldsymbol{\alpha} = (0.4, 0.3, 0.1, 0.2)$ ,  $\boldsymbol{\psi} = (0.1, 0.2, 0.4, 0.3)$ ,  $\boldsymbol{\phi} = (0.6, 0.1, 0.1, 0.2)$  and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_4) = (0.6, 0.5, 0.8, 0.3)$  as the dependence parameter. We take  $p^{(1)} = p_2 = p_3 = 0.9$  and thus  $p = 0.729$ ,  $p_{ij} = 0.81$  with  $i, j \in \{1, 2, 3\}$  and  $i < j$ . We grid the  $[0, 1]^2$  cube into 10000 points at same distance from each other and we only keep those with  $w^{(2)} + w^{(3)} < 1.0$  where  $w^{(2)}$  and  $w^{(3)}$  are in the grid of the cube, we set  $w^{(1)} = 1 - w^{(2)} - w^{(3)}$ . Let  $\Delta_{d-1}^n$  be

199 points uniformly sampled from  $\Delta^2$  and  $n_{iter} = 300$ , Equation (2.16) is estimated with

$$\widehat{MISE}_n^{\mathcal{H}} = \frac{1}{10} \sum_{l=1}^{10} \frac{1}{199} \sum_{k \in \Delta_{d-1}^n} \left( \mathcal{E}_{n,l}^{\mathcal{H}}(k) - \mathcal{S}^{\mathcal{H}}(\mathbf{p}, k) \right)^2,$$

where  $\mathcal{E}_{n,l}^{\mathcal{H}}, l \in \{1, \dots, 10\}$  is the empirical counterpart of  $\mathcal{S}^{\mathcal{H}}$  taking the empirical variance of 30 estimators  $\hat{\nu}_n^{\mathcal{H}}(\mathbf{w})$  with  $\mathbf{w} \in \Delta_{d-1}^n$ . Each estimator of the  $\mathbf{w}$ -madogram is computed out through a random sample with  $n = 512$ . Again,  $\widehat{MISE}_n^{\mathcal{H}*}$  is defined in a similar way.

**E3** In this experiment, we aim to show that our conclusions are verified in a high dimension setting. We compute empirical counterpart of the asymptotic variance for a varying dimension  $d$  and we compare its value to the theoretical one given by Proposition 2.2.1. Furthermore, as the probability of observing a complete row decrease quickly with respect to the dimension  $d$ , i.e.  $p = (p^{(1)})^{-d}$ , we set that there is no missing data. We consider the symmetric logistic model with dependence parameter  $\theta = 2$ . We sample 300 points from the unit simplex  $\Delta_{d-1}$  and we compute the following quantity

$$\delta_n^{\mathcal{H}}(\mathbf{w}) \triangleq \frac{|\mathcal{E}_n^{\mathcal{H}}(\mathbf{w}) - \mathcal{S}^{\mathcal{H}}(\mathbf{1}, \mathbf{w})|}{\mathcal{S}^{\mathcal{H}}(\mathbf{1}, \mathbf{w})}, \quad (2.18)$$

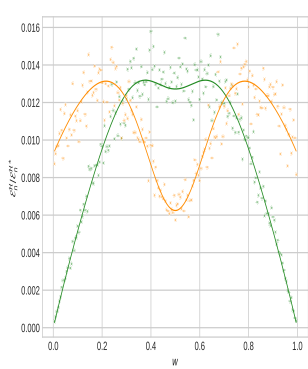
where  $\mathcal{E}_n^{\mathcal{H}}$  is computed from  $n_{iter} = 100$  estimators of the  $\mathbf{w}$ -madogram with sample size  $n \in \{216, 512, 1024\}$ . The results are collected for several values of  $d \in \{5, 10, \dots, 40\}$ .

Note that for Experiments **E1** and **E2**, the missing mechanism is such as  $I^{(1)}, \dots, I^{(d)}$  are pairwise independent and  $p^{(j)} = p^{(1)}, \forall j \in \{1, \dots, d\}$ . The independence setup corresponds to the worst scenario where the missingness of one variable does not influence the missingness of the other variables. *A contrario*, if we suppose that  $I^{(1)}, \dots, I^{(d)}$  are strongly dependent, i.e. none or all entries are missing, we then estimate a statistic on a sample of average length  $p \times n$  and we are turning back to inference in a complete data framework with a reduced sample size. This is also readily seen from the closed formula in Proposition 2.2.1, indeed in a strongly dependent setting we have  $p = p^{(1)}$ , so the asymptotic variance is reduced to the complete data framework up to a multiplicative factor.

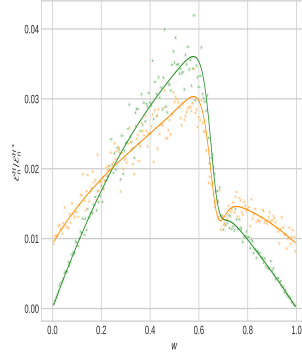
### 2.3.3 Results of experiments

Results of Experiment **E1** are depicted in Figure 2.1. For all panels, empirical counterparts given by Equation (2.15) (points) fit the theoretical values exhibited from Proposition 2.2.1 (solid lines). For the hybrid estimator, as discussed in Remark 2.2.1, both empirical and theoretical values of the asymptotic variance are different from zero for each  $w \in \{\{0\}, \{1\}\}$ . The corrected version provides this feature and also modifies the shape of the curve (see Remark 2.2.2). Indeed the asymptotic behavior of the hybrid and the corrected estimators are different in the missing data framework. Notice that, in terms of variance, we do not have a strict dominance from one estimator to another.

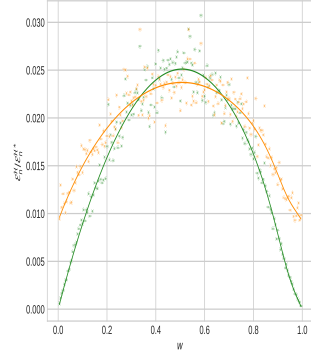
Results for Experiment **E2** are depicted in Figures 2.2 and 2.3. In Figure 2.2, empirical counterparts given by Equation (2.15) are depicted with points and closed expressions of the asymptotic variance given by Proposition 2.2.1 are drawn by a surface. Figure 2.3 presents the same studies differently by showing the level sets associated to the surfaces of Figure 2.2. As in



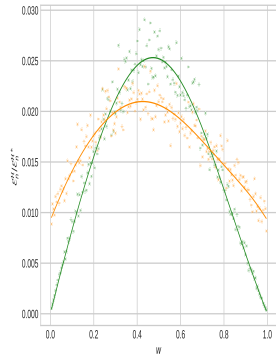
(a) GAL (**M3**,  $\theta = 2.5$ )



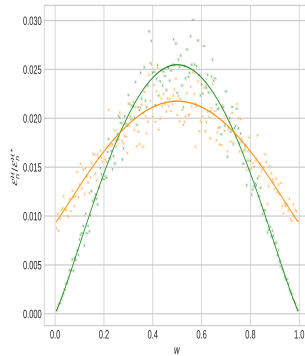
(b) ANL (**M3**,  $\theta = 10$ ,  $\psi_1 = .5$ ,  $\psi_2 = 1$ )



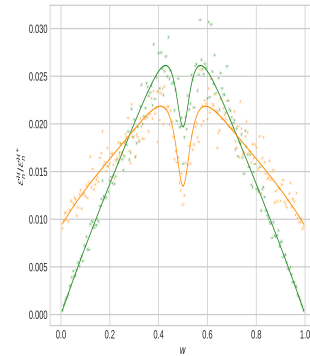
(c) ASL (**M2**,  $\theta = \frac{5}{2}$ ,  $\psi_1 = .1$ ,  $\psi_2 = 1$ )



(d) ASM (**M4**,  $\theta = \frac{4}{3}$ ,  $\kappa = -\frac{1}{3}$ )



(e) HR (**M5**,  $\theta = 1.0$ )



(f) tEV (**M6**,  $\theta = 0.8$ ,  $\nu = 0.2$ )

Fig. 2.1  $\mathcal{E}_n^{\mathcal{H}}$  in red and  $\mathcal{E}_n^{\mathcal{H}^*}$  in green (see (2.15)) as a function of  $w$ , of the asymptotic variances of the estimators of the  $\mathbf{w}$ -madogram for six extreme-value copula models. The empirical variances are based on 300 samples of size  $n = 1024$ . Solid lines are the theoretical value given by Proposition 2.2.1.

Experiment **E1**, empirical counterparts given by the points fits the surface. Also, for the first row of Figure 2.2, we see that if  $\mathbf{w} \in \{\{\mathbf{e}^{(1)}\}, \{\mathbf{e}^{(2)}\}, \{\mathbf{e}^{(3)}\}\}$  then both theoretical and empirical counterparts are different from zero while this feature no longer applies in the second row with the introduction of the corrected version. In this two figures, we see that  $\mathcal{E}_n^{\mathcal{H}}$  and  $\mathcal{E}_n^{\mathcal{H}*}$  and their empirical counterparts are close.

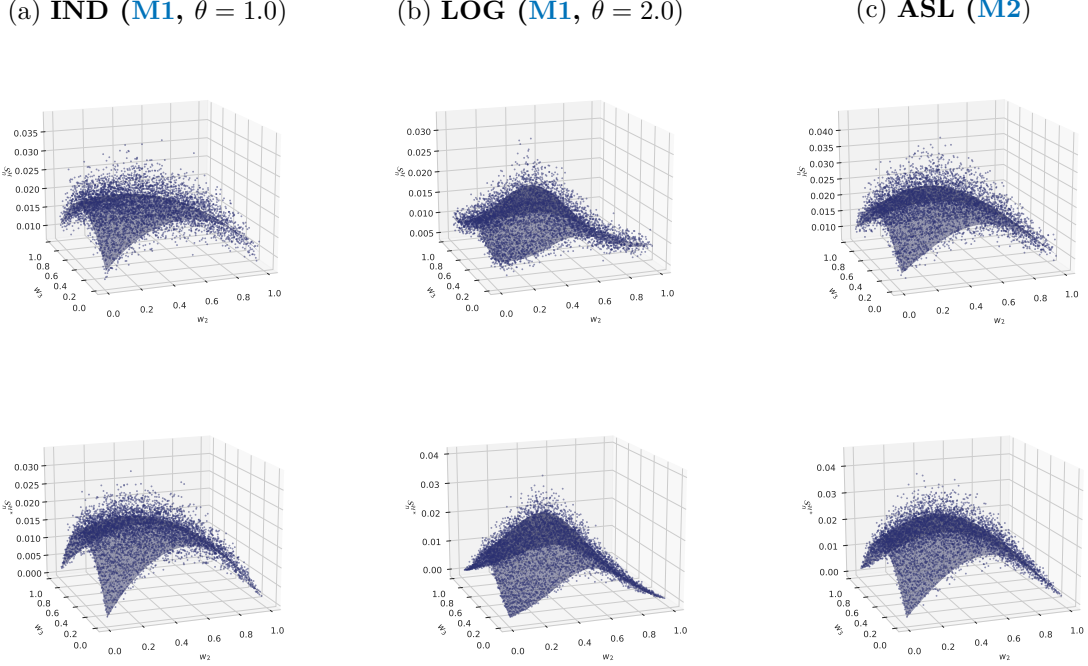


Fig. 2.2  $\mathcal{E}_n^{\mathcal{H}}$  (first row) and  $\mathcal{E}_n^{\mathcal{H}*}$  (second row) given by (2.15) as a function of  $\mathbf{w}$ -madogram. The empirical variances are based on 100 samples of size  $n = 512$ . Empirical counterparts are represented with points and theoretical values given by Proposition 2.2.1 are drawn by a surface.

In order to quantify errors in Figures 2.1 and 2.2, in Table 2.1 are displayed  $\widehat{MISE}_n^{\mathcal{H}}$  and  $\widehat{MISE}_n^{\mathcal{H}*}$  for the corresponding models in Experiments **E1** and **E2** to appreciate the proximity between the terms  $\mathcal{E}_n^{\mathcal{H}}$  and  $\mathcal{S}^{\mathcal{H}}$  (respectively for the corrected terms  $\mathcal{E}_n^{\mathcal{H}*}$  and  $\mathcal{S}^{\mathcal{H}*}$ ). As indicated by Figures 2.1 and 2.2, errors in Table 2.1 are close to zero.

$MISE (\times 10^{-5})$	<b>E1</b>						<b>E2</b>		
	GAL	ANL	ASL	ASM	HR	tEV	IND	LOG	ASL
$\widehat{MISE}_n^{\mathcal{H}}$	2.49	8.10	2.43	1.85	1.89	1.93	2.93	1.31	3.40
$\widehat{MISE}_n^{\mathcal{H}*}$	2.77	7.02	2.04	1.94	1.96	1.93	1.95	1.57	2.91

Table 2.1 Estimation of  $MISE^{\mathcal{H}}$  and  $MISE^{\mathcal{H}*}$  ( $\times 10^{-5}$ ) defined in (2.16) for Experiment **E1** in the sixth first columns and **E2** in the last three columns.



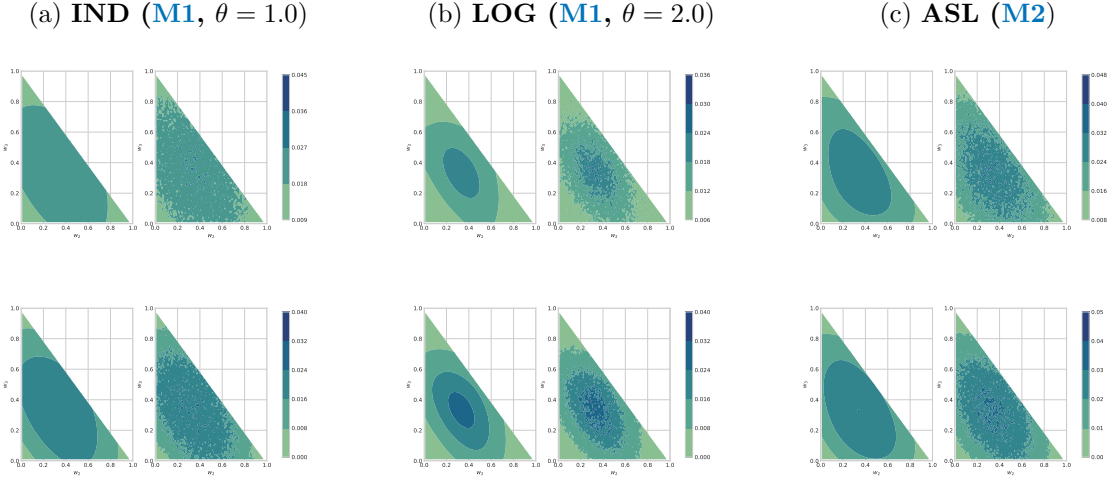


Fig. 2.3 Level sets of  $\mathcal{E}_n^H$  and  $\mathcal{E}_n^{H*}$ , as a function of  $\mathbf{w}$ , of the asymptotic variances of the estimators of the  $\mathbf{w}$ -madogram. We present the level sets corresponding to surfaces of Fig 2.2. On the left panel is represented the theoretical value given by Proposition 2.2.1 while on the right the empirical counterpart is given.

Figure 2.4 illustrates the results of Experiment E3 where we have drawn boxplots for Equation (2.18). Not surprisingly, we observe that both the size of the boxplots and the median value are increasing with  $d$ . However, this augmentation drops as the sample size  $n$  increases and seems to appear reasonable. A limitation (due to computation time issues) of this figure is that the number of points on the simplex is constant ( $= 300$ ) as a function of the dimension.

## 2.4 Extremal dependence rainfall analysis via hybrid madogram

In climate studies, extreme events such as heavy precipitations represent major challenge since damages from extreme weather events may have heavy consequences in both economic and human terms. Their spatial characteristics are of a prime interest and  $\mathbf{w}$ -madogram and its estimator studied in this paper (see Equation (2.10)) are able to capture those characteristics. A seminal application which bridges extreme value theory and geostatistics is the study of extreme rainfall since we expect spatial dependence among the recording weather stations. Precisely, we observe daily precipitation at station  $j \in \{1, \dots, d\}$  over  $n$  years. Concerning extreme events, one cannot use directly the observation for inference and we focus on block maxima. The block maxima approach is based on the observation of a sample of block maxima  $\mathbf{X}_i = (X_i^{(1)}, \dots, X_i^{(d)})$  where  $X_i^{(j)}$  corresponds to the maximum at station  $j \in \{1, \dots, d\}$  within the  $i$ th disjoint block of observation. A block could be either hourly, daily or annual for example. Consistent to our approach, we do not observe  $\mathbf{X}_i$  but an incomplete vector  $\tilde{\mathbf{X}}_i \in \otimes_{j=1}^d (\mathbb{R}_+ \cup \text{NA})$ . Our main goal is to estimate the extremal dependence between maxima of groups of station. This will be done for several clusters within which similar climate characteristics are envisaged leading to dependence among extremes.

For each cluster, we compute the corrected hybrid madogram in Equation (2.10). This quantity is used to estimate the extremal coefficient (see for instance Smith (1990)), using the relation

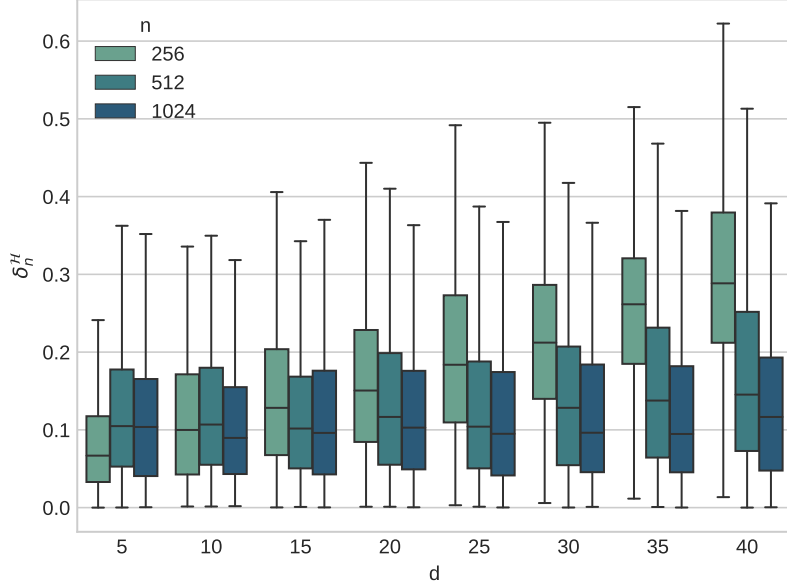


Fig. 2.4 Boxplots for  $\delta_n^H$  for different values of  $d$  and  $n$ .

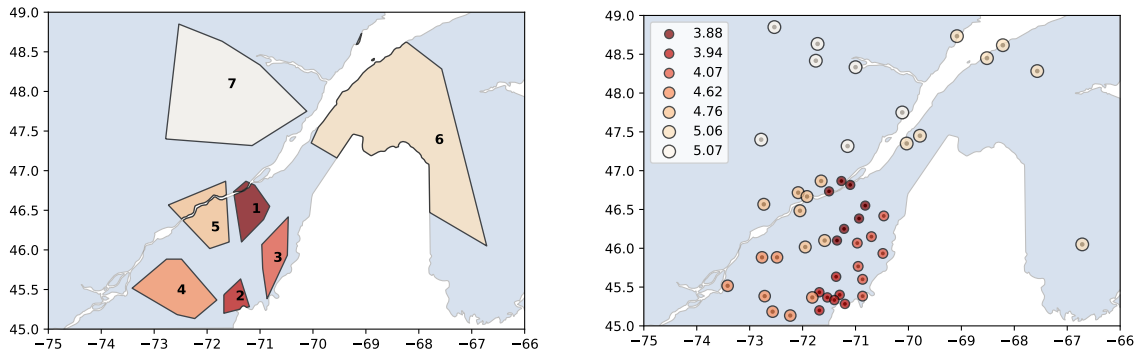
between the Pickands and the madogram given in Equation (2.5), defined by

$$\theta = d \mathfrak{A} \left( \frac{1}{d}, \dots, \frac{1}{d} \right). \quad (2.19)$$

This satisfies the condition  $1 \leq \theta \leq d$ , where the lower and upper bounds represent the case of complete positive dependence and independence among the extremes, respectively. Since its upper bound depends on  $d$ , the extremal coefficient can, alas, only be used to compare clusters of the same size. In each cluster, the extremal coefficient in (2.19) is estimated by  $\hat{\theta}_n = d \hat{\mathfrak{A}}_n^{H*}$  where  $\hat{\mathfrak{A}}_n^{H*}$  is given in Definition 2.2.3.

We illustrate the proposed methodology on rainfall data measured in millimeter registered in 95 stations in Center Eastern Canada for a duration of 24 hours publicly available in the section [engineering climate datasets](#) of the Government of Canada website. Annual maxima precipitations for a 24-hour duration are recorded from 1914 to 2017. The location of stations in Fig. 2.5 are given in the WGS84 coordinate space in order to have Euclidean distance between the stations and taking account of the geodesic geometry of the Earth. A specific characteristic of the considered rainfall data is the sparsity of the recorded data, i.e., a lot of recordings are missing (see [Palacios-Rodriguez et al. \(2023\)](#) for details). Four stations were removed of the analysis due to a tiny coverage of the observation period. As the measurements are maxima over a long period of time, it is reasonable to assume that they come from a multivariate extreme value distribution see Equation (2.1). The dataset we consider in this section and codes are available in <https://github.com/Aleboul/missing>.

With the remaining 91 stations, we compare the extremal dependence between several groups of stations as it has been done by Marcon et al. (2017) (see Section 5) for France using a dataset with complete observations. We emphasize that the comparison of the extremal coefficient is solely relevant when clusters are of the same size. Thus, clusters were obtained by running the constrained  $k$ -means algorithm on the station coordinates (see for instance Bradley et al. (2000)) by forcing clusters of the same size :  $d = 7$  or  $d = 13$ , i.e., 13 groups of 7 stations and *vice versa*. As overlapping data naturally decrease as the size of clusters increases, the case of cluster size  $d = 13$  cannot be considered here. Among the 13 clusters of size  $d = 7$ , we only keep those having at least 10 overlapping annual maxima within the cluster which results on 7 remaining clusters depicted in Figure 2.5a. The estimated coefficient range is between 3.88, indicating strong dependence, and 5.07, indicating medium dependence (see Figure 2.5b).



(a) Resulting clusters using constrained  $k$ -means (b) Values of the extremal coefficient for each cluster

Fig. 2.5 Analysis of Canadian annual rainfall maxima in the period 1914-2017. (a) Spatial representation of the 7 selected clusters obtained via the constrained  $k$ -means algorithm. (b) Clusters of 49 weather stations and their estimated extremal coefficients (with  $d = 7$ ) obtained with the corrected version of the hybrid madogram.

Our estimations suggest an acute dependence among extremes in clusters 1-3 in Figure 2.5a. We can observe in Figure 2.5b that extreme precipitations are more likely to be dependent in the central coastal Atlantic region, *a contrario*, one can notice a weak dependence among extreme values in the scattered clusters in the north of the region.

## 2.5 Conclusions

A method based on madograms to estimate multivariate extremal dependencies with allowing missing data has been developed in this paper. Under the **MCAR** hypothesis, we studied the asymptotic behaviour for the proposed estimators. This approach is of interest to study spatio-temporal process punctually observed as observations may not overlap. Moreover, we have derived closed expressions of their respective asymptotic variances for a fixed element in the simplex. Numerical results in a finite sample setting give further evidences to our theoretical results and on performances of the proposed estimators of the madogram in the missing data

setting. Finally, we applied our approach to the study of extremal dependencies of annual maxima of daily rainfall in Central Eastern Canada.

As for future work, an interesting improvement could be to lower the **MCAR** assumption on the missing data. Indeed, estimating nonparametrically the empirical copula process with missing data outside this framework is still unexplored. As a starting point, semiparametric inference for copula and copula based-regression allowing missing data under **Missing At Random (MAR)** mechanism have been studied by Hamori et al. (2019) and Hamori et al. (2020).

Another interesting direction could also be to build a dissimilarity measure based on the bivariate  $\mathbf{w}$ -madogram for clustering. This approach was already tackled by Bernard et al. (2013), Bador et al. (2015) and Saunders et al. (2021) to partition respectively France, Europe and Australia with respect to extreme observations using the sole madogram. The idea here could be to use the infimum or the integral over  $\mathbf{w} \in (0, 1)$  of the bivariate  $\mathbf{w}$ -madogram as a dissimilarity measure and to show its strong consistency in the sense formulated by Pollard (1981). One limitation of our application is that clusters of same size is mandatory to compare the estimated extremal coefficient between clusters in Equation (2.19). This feature stems from the bounds of the Pickands dependence function which depends on the dimension of the extremal random vector. Further investigations are thus needed to interpret extremal coefficient between clusters of different sizes, e.g., to assess asymptotic independence between two extremal random vectors.



# APPENDIX A

## PROOFS OF CHAPTER 2

### A.1 Proofs

#### A.1.1 Proofs of main results

For the rest of this section, we will write, for notational convenience,  $N = \sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}$ . The following proof gives arguments used to establish the functional central limit theorem of our processes defined in Equation (2.12). Before going into details, we need an intermediary lemma to assert that the empirical cumulative distribution functions in case of missing data verify Assumption C and give covariance functions of the asymptotic processes  $\alpha$  and  $\beta^{(j)}$  with  $j \in \{1, \dots, d\}$ . This result comes down from Segers (2015) (see Example 3.5) where the result was proved for bivariate random variables but the higher dimension is directly obtained using same arguments.

**Lemma A.1.1.** *Let  $(\sqrt{n}(\hat{F}_n - F); \sqrt{n}(\hat{F}_n^{(1)} - F^{(1)}), \dots, \sqrt{n}(\hat{F}_n^{(d)} - F^{(d)}))$  with  $\hat{F}_n$  and  $\hat{F}_n^{(j)}$  for  $j \in \{1, \dots, d\}$  as in (2.7). Then Assumption C is satisfied with*

$$\begin{aligned} \beta^{(j)}(u^{(j)}) &= (p^{(j)})^{-1} \mathbb{G} \left( \mathbf{1}_{\{X^{(j)} \leq (F^{(j)})^{\leftarrow}(u^{(j)})\}, I^{(j)}=1\}} - u^{(j)} \mathbf{1}_{\{I^{(j)}=1\}} \right), \quad j \in \{1, \dots, d\}, \\ \alpha(\mathbf{u}) &= p^{-1} \mathbb{G} \left( \mathbf{1}_{\{\mathbf{X} \leq (\mathbf{F}^{(d)})^{\leftarrow}(\mathbf{u}), \mathbf{I}=\mathbf{1}\}} - C_\infty(\mathbf{u}) \mathbf{1}_{\{\mathbf{I}=\mathbf{1}\}} \right), \end{aligned}$$

where  $\mathbb{G}$  is a tight Gaussian process. Furthermore the covariance functions of the processes  $\beta^{(j)}(u^{(j)})$ ,  $\alpha(\mathbf{u})$ , for  $(\mathbf{u}, \mathbf{v}, v^{(k)}) \in [0, 1]^{2d+1}$ ,  $j \in \{1, \dots, d\}$  and  $j < k$ , are given by

$$\begin{aligned} \text{cov} \left( \beta^{(j)}(u^{(j)}), \beta^{(j)}(v^{(j)}) \right) &= (p^{(j)})^{-1} \left( u^{(j)} \wedge v^{(j)} - u^{(j)} v^{(j)} \right), \\ \text{cov} \left( \beta^{(j)}(u^{(j)}), \beta^{(k)}(v^{(k)}) \right) &= \frac{p^{(jk)}}{p^{(j)} p^{(k)}} \left( C_\infty(\mathbf{1}^{(jk)}(u^{(j)}, v^{(k)})) - u^{(j)} v^{(k)} \right), \\ \text{cov} \left( \alpha(\mathbf{u}), \alpha(\mathbf{v}) \right) &= p^{-1} \left( C_\infty(\mathbf{u} \wedge \mathbf{v}) - C_\infty(\mathbf{u}) C_\infty(\mathbf{v}) \right), \\ \text{cov} \left( \alpha(\mathbf{u}), \beta^{(j)}(v^{(j)}) \right) &= (p^{(j)})^{-1} \left( C_\infty(\mathbf{u}^{(j)}(u^{(j)} \wedge v^{(j)})) - C_\infty(\mathbf{u}) v^{(j)} \right), \end{aligned}$$

where  $\mathbf{u} \wedge \mathbf{v}$  denotes the vector of componentwise minima and  $p^{(jk)} = \mathbb{P}(I^{(j)} = 1, I^{(k)} = 1)$ .

Proof of Lemma A.1.1 is postponed to A.1.2.

**Proof of Theorem 2.2.1** First, let us define the rank-corrected hybrid copula process suited with our estimator and its associated empirical copula process by

$$\hat{C}_n^{\mathcal{R}}(\mathbf{u}) = \frac{1}{\sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}} \sum_{i=1}^n \prod_{j=1}^d \mathbf{1}_{\{\tilde{U}_i^{(j)} \leq u^{(j)}\}} I_i^{(j)}, \quad \mathbb{C}_n^{\mathcal{R}} = \sqrt{n} \left( \hat{C}_n^{\mathcal{R}} - C_\infty \right).$$

One can show that

$$\sup_{\mathbf{u} \in [0,1]^d} \left| \hat{C}_n^{\mathcal{H}}(\mathbf{u}) - C_n^{\mathcal{R}}(\mathbf{u}) \right| \leq \frac{2d}{n\hat{p}_n},$$

with  $\hat{p}_n = n^{-1} \sum_{i=1}^n \prod_{j=1}^d I_i^{(j)}$ . Note that  $\hat{p}_n$  converges in probability to  $p \in ]0, 1]$  which implies that the difference between  $\hat{C}_n^{\mathcal{H}}$  and  $\hat{C}_n^{\mathcal{R}}$  is asymptotically negligible. Details for the proof are given solely for the estimator  $\hat{\nu}_n^{\mathcal{H}^*}$  as the weak convergence for  $\hat{\nu}_n^{\mathcal{H}}$  is obtained similarly via an adequate continuous transformation of  $\hat{\nu}_n^{\mathcal{H}}$  with  $\hat{C}_n^{\mathcal{R}}$ . Using that  $\mathbb{E}[F^{(j)}(X^{(j)})^\alpha] = (1 + \alpha)^{-1}$  for  $\alpha \neq -1$ , we can write  $\nu(\mathbf{w})$  as :

$$\begin{aligned} \nu(\mathbf{w}) &= \mathbb{E} \left[ \prod_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right] \\ &\quad + \sum_{j=1}^d \frac{\lambda^{(j)}(\mathbf{w})(d-1)}{d} \left( \frac{w^{(j)}}{1+w^{(j)}} - \mathbb{E} \left[ \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right] \right) \\ &= \mathbb{E} \left[ \prod_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right] - \frac{1}{d} \sum_{j=1}^d (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \mathbb{E} \left[ \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right] + a(\mathbf{w}), \end{aligned}$$

with  $a(\mathbf{w}) = (d-1)d^{-1} \sum_{j=1}^d \lambda^{(j)}(\mathbf{w})w^{(j)}/(1+w^{(j)})$ . Let us note by  $g_{\mathbf{w}}$  the function defined as

$$g_{\mathbf{w}} : [0, 1]^d \rightarrow [0, 1], \quad \mathbf{u} \mapsto \prod_{j=1}^d (u^{(j)})^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d (1 + \lambda^{(j)}(\mathbf{w})(d-1))(u^{(j)})^{1/w^{(j)}}.$$

One can write our estimator of the  $\mathbf{w}$ -madogram and the theoretical  $\mathbf{w}$ -madogram in missing data framework as an integral with respect to the rank-corrected hybrid copula estimator and the copula function, respectively. We thus have:

$$\begin{aligned} \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) &= \frac{1}{N} \sum_{i=1}^n g_{\mathbf{w}} \left( \tilde{U}_i^{(1)}, \dots, \tilde{U}_i^{(d)} \right) \prod_{j=1}^d I_i^{(j)} + a(\mathbf{w}) = \int_{[0,1]^d} g_{\mathbf{w}}(\mathbf{u}) d\hat{C}_n^{\mathcal{R}}(\mathbf{u}) + a(\mathbf{w}), \\ \nu(\mathbf{w}) &= \int_{[0,1]^d} g_{\mathbf{w}}(\mathbf{u}) dC(\mathbf{u}) + a(\mathbf{w}). \end{aligned}$$

We obtain, proceeding as in Theorem 2.4 of [Marcon et al. \(2017\)](#) :

$$\begin{aligned} \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w}) - \nu(\mathbf{w}) \right) &= \frac{1}{d} \sum_{j=1}^d \left( 1 + \lambda^{(j)}(\mathbf{w})(d-1) \right) \int_{[0,1]} \mathbb{C}_n^{\mathcal{R}}(\mathbf{1}^{(j)}(x^{w^{(j)}})) dx \\ &\quad - \int_{[0,1]} \mathbb{C}_n^{\mathcal{R}}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx, \end{aligned}$$

where  $\mathbf{1}^{(j)}(u)$  denotes the vector composed out of 1 except for the  $j$ th component where  $u$  does stand and with  $\mathbb{C}_n^{\mathcal{H}}$  in (2.9). Consider the function  $\phi : \ell^\infty([0, 1]^d) \rightarrow \ell^\infty(\Delta_{d-1})$ ,  $f \mapsto \phi(f)$ , defined by

$$(\phi)(f)(\mathbf{w}) = \frac{1}{d} \sum_{j=1}^d (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \int_{[0,1]} f(\mathbf{1}^{(j)}(x^{w^{(j)}})) dx - \int_{[0,1]} f(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx.$$

This function is linear and bounded thus continuous. The continuous mapping theorem (see, e.g., Theorem 1.3.6 of [van der Vaart and Wellner \(1996\)](#)) implies, as  $n \rightarrow \infty$

$$\sqrt{n}(\hat{\nu}_n^{\mathcal{H}^*} - \nu) = \phi(\mathbb{C}_n^{\mathcal{R}}) \rightsquigarrow \phi(S_{C_\infty}),$$

in  $\ell^\infty(\Delta_{d-1})$ . Recall that  $S_{C_\infty}$  is the asymptotic process where  $\mathbb{C}_n^{\mathcal{H}}$  does converge in the sense of the weak convergence in  $\ell^\infty(\Delta_{d-1})$  and is defined by  $S_{C_\infty}(\mathbf{u}) = \alpha(\mathbf{u}) - \sum_{j=1}^d \beta^{(j)}(u^{(j)}) \dot{C}_\infty^{(j)}(\mathbf{u})$  with  $\mathbf{u} \in [0, 1]^d$  and  $\alpha$  and  $\beta^{(j)}$  are processes defined in Lemma [A.1.1](#). We note that  $S_{C_\infty}(\mathbf{1}^{(j)}(x^{w^{(j)}})) = \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(u^{(j)})$  and we obtain our statement.  $\square$

The asymptotic normality of our estimators directly comes down from being a linear transformation of a tight Gaussian process for  $\mathbf{w} \in \Delta_{d-1}$ . The proof below uses technical arguments to exhibit the closed expressions of the asymptotic variances of the Gaussian limit distributions of our estimators in Equation [\(2.10\)](#) and [\(2.11\)](#). Furthermore, this proof strengthens our choice of the definition of the corrected estimator. Indeed, the chosen form of the corrected estimator makes computations more tractable as we only have to compute terms for the hybrid estimator and to multiply those by different factors. Two tools make the computation feasible. The first one is the form exhibited by Equation [\(2.2\)](#) which transforms a double integral with respect to the trajectory of the copula function as the double integral of a power function. When this trick is not possible, again the expression of the extreme value copula with respect to the Pickands dependence function is of main interest. Indeed, with some substitutions, we are able to express the double integrals as the integral with respect to the Pickands dependence function using the following equality :

$$-\int_{[0,1]} w^\alpha \ln(w) dw = \frac{1}{(\alpha + 1)^2},$$

where  $\alpha \neq -1$ .

**Proof of Proposition 2.2.1** Recall that  $\mathbf{p} = (p^{(1)}, \dots, p^{(d)}, p)$ . By definition the asymptotic variance  $\mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w})$  for a fixed  $\mathbf{w} \in \Delta_{d-1}$  is given by

$$\mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w}) \triangleq \text{Var} \left( \frac{1}{d} \sum_{j=1}^d \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx - \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right).$$

Using properties of the variance operator, we thus obtain

$$\begin{aligned} \mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w}) &= \frac{1}{d^2} \sum_{j=1}^d \text{Var} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx \right) + \text{Var} \left( \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right) \\ &+ \frac{2}{d^2} \sum_{j < k} \text{cov} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx, \int_{[0,1]} \alpha(\mathbf{1}^{(k)}(x^{w^{(k)}})) - \beta^{(k)}(x^{w^{(k)}}) dx \right) \\ &- \frac{2}{d} \sum_{j=1}^d \text{cov} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx, \int_{[0,1]} \alpha(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right) \\ &+ \frac{2}{d} \sum_{j=1}^d \sum_{k=1}^d \text{cov} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx, \int_{[0,1]} \beta^{(k)}(x^{w^{(k)}}) \dot{C}_\infty^{(k)}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right). \end{aligned}$$



By definition of the covariance functions of  $\alpha$ ,  $\beta^{(j)}$  with  $j \in \{1, \dots, d\}$  given in Lemma A.1.1, we have

$$\begin{aligned}
 \text{Var} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx \right) &= \left( p^{-1} - (p^{(j)})^{-1} \right) \sigma_j^2(\mathbf{w}), \\
 \text{Var} \left( \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right) &= \sigma_{d+1}^2(\mathbf{p}, \mathbf{w}), \\
 \text{cov} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx, \int_{[0,1]} \alpha(\mathbf{1}^{(k)}(x^{w^{(k)}})) - \beta^{(k)}(x^{w^{(k)}}) dx \right) &= \\
 &= \left( p^{-1} - (p^{(j)})^{-1} - (p^{(k)})^{-1} + \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \right) \sigma_{jk}(\mathbf{w}), \\
 \text{cov} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx, \int_{[0,1]} \alpha(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right) &= \\
 &= \left( p^{-1} - (p^{(j)})^{-1} \right) \sigma_j^{(1)}(\mathbf{w}), \\
 \text{cov} \left( \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx, \int_{[0,1]} \beta^{(k)}(x^{w^{(k)}}) \dot{C}_\infty^{(k)}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right) &= \\
 &= \left( (p^{(k)})^{-1} - \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \right) \sigma_{jk}^{(2)}(\mathbf{w}).
 \end{aligned}$$

We first show in details the closed form for  $\sigma_{d+1}^2$ , the other forms are given without explanations as the technical tools used are those used for  $\sigma_{d+1}^2$ . Proceeding as before, we decompose this quantity as a linear combination of the variance (the squared term  $\gamma_1^2$  and  $\gamma_j^2$  for  $j \in \{1, \dots, d\}$ ) and the covariance terms ( $\gamma_{1j}$  and  $\tau_{jk}$ ) with the probabilities of missing. The explicit formula of these quantities will be defined below. We set

$$\sigma_{d+1}^2(\mathbf{p}, \mathbf{w}) = p^{-1} \gamma_1^2(\mathbf{w}) + \sum_{j=1}^d (p^{(j)})^{-1} \gamma_j^2(\mathbf{w}) - 2 \sum_{j=1}^d (p^{(j)})^{-1} \gamma_{1j}(\mathbf{w}) + 2 \sum_{j < k} \frac{p^{(jk)}}{p^{(j)}p^{(k)}} \tau_{jk}(\mathbf{w}). \quad (\text{A.1})$$

Let us exhibit a useful form of the partial derivatives of the extreme value copula. We have  $\forall j \in \{1, \dots, d\}$ :

$$\dot{C}_\infty^{(j)}(\mathbf{u}) = \frac{C(\mathbf{u})}{u^{(j)}} \dot{L}_j(-\ln(u_1), \dots, -\ln(u_d)).$$

Furthermore, as  $L(x_1, \dots, x_d)$  is homogeneous of degree 1, the partial derivative  $\dot{L}_j(x_1, \dots, x_d)$  is homogeneous of degree 0 for  $j \in \{1, \dots, d\}$ . We thus obtain a suitable form of the partial derivatives of the extreme value copula for  $u \in ]0, 1[$  and  $\mathbf{w} \in \Delta_{d-1}$ :

$$\begin{aligned}
 \dot{C}_\infty^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) &= \frac{u^{\mathcal{A}(\mathbf{w})}}{u^{w^{(j)}}} \dot{L}_j(-w^{(1)} \ln(u), \dots, -w^{(d)} \ln(u)) = \frac{u^{\mathcal{A}(\mathbf{w})}}{u^{w^{(j)}}} \dot{L}_j(-w^{(1)}, \dots, -w^{(d)}) \\
 &= \frac{u^{\mathcal{A}(\mathbf{w})}}{u^{w^{(j)}}} \mu^{(j)}(\mathbf{w}),
 \end{aligned}$$

where  $\mu^{(j)}(\mathbf{w}) \triangleq \dot{L}_j(-w^{(1)}, \dots, -w^{(d)})$ . Now, using linearity of the integral and the definition of the covariance function of  $\alpha$ , we obtain

$$\begin{aligned} p^{-1}\gamma_1^2(\mathbf{w}) &\triangleq \mathbb{E} \left[ \int_{[0,1]} \alpha(u^{w^{(1)}}, \dots, u^{w^{(d)}}) du \int_{[0,1]} \alpha(v^{w^{(1)}}, \dots, v^{w^{(d)}}) dv \right] \\ &= \frac{2}{p} \int_{[0,1]} \int_{[0,v]} u^{\mathcal{A}(\mathbf{w})} (1 - v^{\mathcal{A}(\mathbf{w})}) duv. \end{aligned}$$

Let us compute

$$\gamma_1^2(\mathbf{w}) = 2 \int_{[0,1]} \int_{[0,v]} u^{\mathcal{A}(\mathbf{w})} (1 - v^{\mathcal{A}(\mathbf{w})}) duv = \frac{1}{(1 + \mathcal{A}(\mathbf{w}))^2} \frac{\mathcal{A}(\mathbf{w})}{2 + \mathcal{A}(\mathbf{w})}.$$

The quantity  $\gamma_j^2(\mathbf{w})$  is defined by the following

$$\begin{aligned} (p^{(j)})^{-1}\gamma_j^2(\mathbf{w}) &\triangleq \mathbb{E} \left[ \int_{[0,1]} \beta^{(j)}(u^{w^{(j)}}) \dot{C}_\infty^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) du \int_{[0,1]} \beta^{(j)}(v^{w^{(j)}}) \dot{C}_\infty^{(j)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) dv \right] \\ &= \frac{2}{p^{(j)}} \int_{[0,1]} \int_{[0,v]} u^{w^{(j)}} (1 - v^{w^{(j)}}) \mu^{(j)}(\mathbf{w}) \mu^{(j)}(\mathbf{w}) u^{\mathcal{A}(\mathbf{w}) - w^{(j)}} v^{\mathcal{A}(\mathbf{w}) - w^{(j)}} duv. \end{aligned}$$

It is clear that

$$\begin{aligned} \gamma_j^2(\mathbf{w}) &= 2 \int_{[0,1]} \int_{[0,v]} u^{w^{(j)}} (1 - v^{w^{(j)}}) \mu^{(j)}(\mathbf{w}) \mu^{(j)}(\mathbf{w}) u^{\mathcal{A}(\mathbf{w}) - w^{(j)}} v^{\mathcal{A}(\mathbf{w}) - w^{(j)}} duv \\ &= \left( \frac{\mu^{(j)}(\mathbf{w})}{1 + \mathcal{A}(\mathbf{w})} \right)^2 \frac{w^{(j)}}{2\mathcal{A}(\mathbf{w}) + 1 + 1 - w^{(j)}}. \end{aligned}$$

We now deal with cross product terms, the first we define is

$$\begin{aligned} (p^{(j)})^{-1}\gamma_{1j}(\mathbf{w}) &\triangleq \mathbb{E} \left[ \int_{[0,1]} \alpha(u^{w^{(1)}}, \dots, u^{w^{(d)}}) du \int_{[0,1]} \beta^{(j)}(v^{w^{(j)}}) \dot{C}_\infty^{(j)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) dv \right] \\ &= (p^{(j)})^{-1} \int_{[0,1]^2} \left( C_\infty(u^{w^{(1)}}, \dots, (u \wedge v)^{w^{(j)}}, \dots, u^{w^{(d)}}) - u^{\mathcal{A}(\mathbf{w})} v^{w^{(j)}} \right) \dot{C}_\infty^{(j)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) duv. \end{aligned}$$

Under the rectangle  $[0, 1] \times [0, v]$ , we have

$$\begin{aligned} \gamma_{1j}(\mathbf{w}) &= \int_{[0,1] \times [0,v]} \left( C_\infty(u^{w^{(1)}}, \dots, u^{w^{(j)}}, \dots, u^{w^{(d)}}) - u^{\mathcal{A}(\mathbf{w})} v^{w^{(j)}} \right) \dot{C}_\infty^{(j)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) duv \\ &= \int_{[0,1] \times [0,v]} u^{\mathcal{A}(\mathbf{w})} (1 - v^{w^{(j)}}) v^{\mathcal{A}(\mathbf{w}) - w^{(j)}} \mu^{(j)}(\mathbf{w}) duv \\ &= \frac{\mu^{(j)}(\mathbf{w})}{2(1 + \mathcal{A}(\mathbf{w}))^2} \frac{w^{(j)}}{2\mathcal{A}(\mathbf{w}) + 1 + (1 - w^{(j)})}. \end{aligned}$$

Under the rectangle  $[0, 1] \times [0, u]$ , we have for the right term

$$\int_{[0,1] \times [0,u]} u^{\mathcal{A}(\mathbf{w})} v^{w^{(j)}} v^{\mathcal{A}(\mathbf{w}) - w^{(j)}} \mu^{(j)}(\mathbf{w}) dvu = \frac{\mu^{(j)}(\mathbf{w})}{2(1 + \mathcal{A}(\mathbf{w}))^2}.$$

For the left term, by definition, we have

$$\int_{[0,1] \times [0,u]} C_\infty(u^{w^{(1)}}, \dots, v^{w^{(j)}}, \dots, u^{w^{(d)}}) \dot{C}_\infty^{(j)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) dvu.$$

Let us consider the substitution  $x = v^{w^{(j)}}$  and  $y = u^{1-w^{(j)}}$ , we obtain

$$\frac{1}{w^{(j)}(1-w^{(j)})} \int_{[0,1]} \int_{[0,y^{w^{(j)}/(1-w^{(j)})}]} \left[ C_\infty \left( y^{w^{(1)}/(1-w^{(j)})}, \dots, x, \dots, y^{w^{(d)}/(1-w^{(j)})} \right) \right] \times \\ \left[ \dot{C}_\infty^{(j)} \left( x^{w^{(1)}/w^{(j)}}, \dots, x^{w^{(d)}/w^{(j)}} \right) x^{(1-w^{(j)})/w^{(j)}} y^{w^{(j)}/(1-w^{(j)})} \right] dx y.$$

Let us compute the quantity

$$\dot{C}_\infty^{(j)}(x^{w^{(1)}/w^{(j)}}, \dots, x^{w^{(d)}/w^{(j)}}) = \frac{C_\infty(x^{w^{(1)}/w^{(j)}}, \dots, x^{w^{(d)}/w^{(j)}})}{x} \mu^{(j)}(\mathbf{w}).$$

Using Equation (2.1), we have

$$C_\infty(x^{w^{(1)}/w^{(j)}}, \dots, x^{w^{(d)}/w^{(j)}}) = \exp \left( -L \left( -\frac{\ln(x)}{w^{(j)}} w^{(1)}, \dots, \frac{\ln(x)}{w^{(j)}} w^{(d)} \right) \right) \\ = \exp \left( -\frac{\ln(x)}{w^{(j)}} L(-w^{(1)}, \dots, -w^{(d)}) \right) = x^{\mathcal{A}(\mathbf{w})/w^{(j)}} = x^{\mathcal{A}^{(j)}(\mathbf{w})},$$

where we use the homogeneity of order one of  $L$  and that  $-L(-w^{(1)}, \dots, -w^{(d)}) = \mathcal{A}(\mathbf{w})$  as stated by the identity of Equation (2.2) and that  $\mathbf{w} \in \Delta_{d-1}$ . Now, consider the substitution  $x = w^{1-s}$  and  $y = w^s$ , the jacobian of this transformation is given by  $-\ln(w)$ , we have

$$-\frac{\mu^{(j)}(\mathbf{w})}{w^{(j)}(1-w^{(j)})} \int_{[0,1]} \int_{[0,1-w^{(j)}]} \left[ C_\infty \left( w^{sw^{(1)}/(1-w^{(j)})}, \dots, w^{1-s}, \dots, w^{sw^{(d)}/(1-w^{(j)})} \right) \right] \\ \times \left[ w^{(1-s)} \left[ \mathcal{A}^{(j)}(\mathbf{w}) + \frac{1-w^{(j)}}{w^{(j)}} - 1 \right] + s \frac{w^{(j)}}{1-w^{(j)}} \ln(w) \right] ds w,$$

where we note by  $\mathcal{A}^{(j)}(\mathbf{w}) \triangleq \mathcal{A}(\mathbf{w})/w^{(j)}$  with  $j \in \{1, \dots, d\}$ . We now compute the quantity

$$C_\infty \left( w^{sw^{(1)}/(1-w^{(j)})}, \dots, w^{1-s}, \dots, w^{sw^{(d)}/(1-w^{(j)})} \right).$$

Using the same techniques as above, we have that the latter is equal to

$$\exp \left( -L \left( -\frac{sw^{(1)}}{1-w^{(j)}} \ln(w), \dots, -(1-s) \ln(w), \dots, -\frac{sw^{(d)}}{1-w^{(j)}} \ln(w) \right) \right) \\ = \exp \left( -\ln(w) L \left( -\frac{sw^{(1)}}{1-w^{(j)}}, \dots, -(1-s), \dots, -\frac{sw^{(d)}}{1-w^{(j)}} \right) \right).$$

Now, using that  $\mathbf{w} \in \Delta_{d-1}$ , remark that  $s \sum_{i \neq j} w_i / (1-w^{(j)}) = s$ , we have, using Equation (2.2)

$$-L \left( -\frac{sw^{(1)}}{1-w^{(j)}}, \dots, -(1-s), \dots, -\frac{sw^{(d)}}{1-w^{(j)}} \right) = \mathcal{A}(\mathbf{z}^{(j)}(1-s)),$$

where  $\mathbf{z} = (sw^{(1)}/(1-w^{(j)}), \dots, sw^{(d)}/(1-w^{(j)}))$ . So we have

$$\begin{aligned}\gamma_{1j}(\mathbf{w}) &= -\frac{\mu^{(j)}(\mathbf{w})}{w^{(j)}(1-w^{(j)})} \int_{[0,1-w^{(j)}]} \int_{[0,1]} w^{\mathcal{A}(\mathbf{z}^{(j)}(1-s))+(1-s)} \left( \mathcal{A}^{(j)}(\mathbf{w}) + \frac{1-w^{(j)}}{w^{(j)}} - 1 \right) + s \frac{w^{(j)}}{1-w^{(j)}} \ln(w) dw s \\ &= \frac{\mu^{(j)}(\mathbf{w})}{w^{(j)}(1-w^{(j)})} \int_{[0,1-w^{(j)}]} \left[ \mathcal{A}(\mathbf{z}^{(j)}(1-s)) + (1-s) \left( \mathcal{A}^{(j)}(\mathbf{w}) + \frac{1-w^{(j)}}{w^{(j)}} - 1 \right) + s \frac{w^{(j)}}{1-w^{(j)}} + 1 \right]^{-2} ds.\end{aligned}$$

No further simplifications can be obtained. For  $j < k$ , let us define the quantity  $\tau_{jk}$  such as

$$\frac{p^{(jk)}}{p^{(j)}p^{(k)}} \tau_{jk}(\mathbf{w}) \triangleq \mathbb{E} \left[ \int_{[0,1]} \beta^{(j)}(u^{w^{(j)}}) \dot{C}_{\infty}^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) du \int_{[0,1]} \beta^{(k)}(v^{w^{(k)}}) \dot{C}_{\infty}^{(k)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) dv \right]. \quad (\text{A.2})$$

Again, we have

$$\tau_{jk}(\mathbf{w}) = \int_{[0,1]^2} \left( C_{\infty}(\mathbf{1}^{(jk)}(u^{w^{(j)}}, v^{w^{(j)}})) - u^{w^{(j)}} v^{w^{(j)}} \right) \dot{C}_{\infty}^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) \dot{C}_{\infty}^{(k)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) duv.$$

We set  $x = u^{w^{(j)}}$  and  $y = v^{w^{(k)}}$ , the left side becomes

$$\begin{aligned}& \frac{1}{w^{(j)}w^{(k)}} \int_{[0,1]^2} \left[ C_{\infty}(\mathbf{1}^{(jk)}(x, y)) \right] \\ & \times \left[ \dot{C}_{\infty}^{(j)}(x^{w^{(1)}/w^{(j)}}, \dots, x^{w^{(d)}/w^{(j)}}) \right] \\ & \times \left[ \dot{C}_{\infty}^{(k)}(y^{w^{(1)}/w^{(k)}}, \dots, y^{w^{(d)}/w^{(k)}}) x^{(1-w^{(j)})/w^{(j)}} y^{(1-w^{(k)})/w^{(k)}} \right] dx y \\ & = \frac{\mu^{(j)}(\mathbf{w})\mu^{(k)}(\mathbf{w})}{w^{(j)}w^{(k)}} \int_{[0,1]^2} C_{\infty}(\mathbf{1}^{(jk)}(x, y)) x^{\mathcal{A}^{(j)}(\mathbf{w})+(1-w^{(j)})/w^{(j)}-1} y^{\mathcal{A}^{(k)}(\mathbf{w})+(1-w^{(k)})/w^{(k)}-1} dx y.\end{aligned}$$

Now, we set  $x = w^{1-s}$  and  $y = w^s$  and we obtain

$$\begin{aligned}\tau_{jk}(\mathbf{w}) &= \frac{\mu^{(j)}(\mathbf{w})\mu^{(k)}(\mathbf{w})}{w^{(j)}w^{(k)}} \int_{[0,1]} \left[ \mathcal{A}(\mathbf{0}^{(jk)}(1-s, s)) \right. \\ & \left. + (1-s) \left( \mathcal{A}^{(j)}(\mathbf{w}) + \frac{1-w^{(j)}}{w^{(j)}} - 1 \right) + s \left( \mathcal{A}^{(k)}(\mathbf{w}) + \frac{1-w^{(k)}}{w^{(k)}} - 1 \right) + 1 \right]^{-2} ds.\end{aligned}$$

The right side of Equation (A.2) is given by

$$\int_{[0,1]^2} u^{w^{(j)}} v^{w^{(k)}} \dot{C}_{\infty}^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) \dot{C}_{\infty}^{(k)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) duv = \frac{\mu^{(j)}(\mathbf{w})\mu^{(k)}(\mathbf{w})}{(1+\mathcal{A}(\mathbf{w}))^2}.$$

Hence the result for  $\sigma_{d+1}^2(\mathbf{w})$ . Using the same techniques, we show that for  $j \in \{1, \dots, d\}$

$$\sigma_j^2(\mathbf{w}) = \int_{[0,1]^2} (u \wedge v)^{w^{(j)}} - u^{w^{(j)}} v^{w^{(j)}} duv = \frac{1}{(1+w^{(j)})^2} \frac{w^{(j)}}{2+w^{(j)}}.$$

For  $j < k$ , we compute

$$\begin{aligned}\sigma_{jk}(\mathbf{w}) &= \int_{[0,1]^2} C_\infty(\mathbf{1}^{(jk)}(u^{w^{(j)}}, v^{w^{(k)}})) - u^{w^{(j)}} v^{w^{(k)}} duv \\ &= \frac{1}{w^{(j)}w^{(k)}} \int_{[0,1]} \left[ \mathcal{A}(\mathbf{0}^{(jk)}(1-s, s)) + (1-s) \frac{1-w^{(j)}}{w^{(j)}} + s \frac{1-w^{(k)}}{w^{(k)}} + 1 \right]^{-2} ds - \frac{1}{1+w^{(j)}} \frac{1}{1+w^{(k)}}.\end{aligned}$$

Let  $j \in \{1, \dots, d\}$ , thus

$$\begin{aligned}\sigma_j^{(1)}(\mathbf{w}) &= \int_{[0,1]^2} C_\infty(u^{w^{(1)}}, \dots, (u \wedge v)^{w^{(j)}}, \dots, u^{w^{(d)}}) - C_\infty(u^{w^{(1)}}, \dots, u^{w^{(d)}}) v^{w^{(j)}} ds \\ &= \frac{1}{w^{(j)}(1-w^{(j)})} \int_{[0,1]} \left[ \mathcal{A}(\mathbf{z}^{(j)}(1-s)) + (1-s) \frac{1-w^{(j)}}{w^{(j)}} + s \frac{w^{(j)}}{1-w^{(j)}} + 1 \right]^{-2} ds \\ &\quad + \frac{1}{1+\mathcal{A}(\mathbf{w})} \left[ \frac{1}{2+\mathcal{A}(\mathbf{w})} - \frac{1}{1+w^{(j)}} \right].\end{aligned}$$

Now, for  $\sigma_{jk}^{(2)}$ , we have to consider three cases :

- if  $j = k$ , we directly have

$$\sigma_{jk}^{(2)}(\mathbf{w}) = 0,$$

- if  $j < k$ , we obtain

$$\sigma_{jk}^{(2)}(\mathbf{w}) = \frac{\mu^{(k)}(\mathbf{w})}{w^{(j)}w^{(k)}} \int_{[0,1]} \left[ \mathcal{A}(\mathbf{0}^{(jk)}(1-s, s)) + (1-s) \frac{1-w^{(j)}}{w^{(j)}} + s \left( \mathcal{A}^{(k)}(\mathbf{w}) + \frac{1-w^{(k)}}{w^{(k)}} - 1 \right) + 1 \right]^{-2} ds - \frac{\mu^{(k)}(\mathbf{w})}{1+\mathcal{A}(\mathbf{w})} \frac{1}{1+w^{(j)}},$$

- if  $j > k$ , we have

$$\sigma_{jk}^{(2)}(\mathbf{w}) = \frac{\mu^{(k)}(\mathbf{w})}{w^{(j)}w^{(k)}} \int_{[0,1]} \left[ \mathcal{A}(\mathbf{0}^{(kj)}(1-s, s)) + s \frac{1-w^{(j)}}{w^{(j)}} + (1-s) \left( \mathcal{A}^{(k)}(\mathbf{w}) + \frac{1-w^{(k)}}{w^{(k)}} - 1 \right) + 1 \right]^{-2} ds - \frac{\mu^{(k)}(\mathbf{w})}{1+\mathcal{A}(\mathbf{w})} \frac{1}{1+w^{(j)}}.$$

Hence the statement.  $\square$

The following lines will give some details to establish the explicit formula of the asymptotic variance when we suppose that components of the random vector  $\mathbf{X}$  are independent. In this framework, we have that  $\mu^{(j)}(\mathbf{w}) = 1$  for every  $j \in \{1, \dots, d\}$  and thus  $\dot{C}_\infty^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) = u^{1-w^{(j)}}$ . Furthermore, in the independent case, most of the integrals are reduced to zero.

**Proof of Corollary 2.2.1** In the term  $\sigma_{d+1}^2$  given in Equation (A.1), only the terms  $\gamma_1^2$ ,  $\gamma_j^2$  and  $\gamma_{1j}$  matter because, in the independent case :

$$\tau_{jk}(\mathbf{w}) = \int_{[0,1]^2} \left( u^{w^{(j)}} v^{w^{(k)}} - u^{w^{(j)}} v^{w^{(k)}} \right) \dot{C}_\infty^{(j)}(u^{w^{(1)}}, \dots, u^{w^{(d)}}) \dot{C}_\infty^{(k)}(v^{w^{(1)}}, \dots, v^{w^{(d)}}) duv = 0.$$

For  $\gamma_{1j}$ , we have to compute

$$\gamma_{1j}(\mathbf{w}) = 2 \int_{[0,1] \times [0,v]} u(1-v^{w^{(j)}}) v^{1-w^{(j)}} duv = \frac{1}{4} \frac{w^{(j)}}{4-w^{(j)}}.$$

For  $\gamma_1^2$  and  $\gamma_j^2$ , we just have to set  $\mathcal{A}(\mathbf{w}) = 1$  in their respective expressions to obtain :

$$\gamma_1^2(\mathbf{w}) = \frac{1}{12}, \quad \gamma_j^2 = \frac{1}{4} \frac{w^{(j)}}{4 - w^{(j)}}.$$

We thus have

$$\sigma_{d+1}^2(\mathbf{p}, \mathbf{w}) = \frac{1}{4} \left( \frac{1}{3p} - \sum_{j=1}^d (p^{(j)})^{-1} \frac{w^{(j)}}{4 - w^{(j)}} \right).$$

Other computations follow from the same arguments.  $\square$

We are now going to prove Proposition 2.2.2. The strong consistency of the our estimators will be established in a two-step process : first, we prove the strong consistency of the estimator  $\nu_n(\mathbf{w})$  which is the nonparametric estimator of the  $\mathbf{w}$ -madogram with known margins and, second, we show that the limit of

$$\sup_{j \in \{1, \dots, d\}} \sup_{i \in \{1, \dots, n\}} \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \{F^{(j)}(\tilde{X}_i^{(j)})\}^{1/w^{(j)}} \right|,$$

is zero almost surely. Before going into the main arguments, we need the following lemma.

**Lemma A.1.2.** *We have,  $\forall i \in \{1, \dots, n\}$*

$$\left| \bigvee_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \bigvee_{j=1}^d \{F^{(j)}(X^{(j)})\}^{1/w^{(j)}} \right| \leq \sup_{j \in \{1, \dots, d\}} \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \{F^{(j)}(X^{(j)})\}^{1/w^{(j)}} \right|.$$

The proof of Lemma A.1.2 is postponed to Section A.1.2.

**Proof of Proposition 2.2.2** We prove it for  $\hat{\nu}_n^{\mathcal{H}}(\mathbf{w})$  as the strong consistency for  $\hat{\nu}_n^{\mathcal{H}^*}(\mathbf{w})$  uses the same arguments. The estimator  $\hat{\nu}_n^{\mathcal{H}}(\mathbf{w})$  in (2.10) is strongly consistent since it holds

$$\left| \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}) \right| = \left| \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu_n(\mathbf{w}) + \nu_n(\mathbf{w}) - \nu(\mathbf{w}) \right| \leq \left| \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu_n(\mathbf{w}) \right| + |\nu_n(\mathbf{w}) - \nu(\mathbf{w})|,$$

where

$$\nu_n(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^n \left[ \left( \bigvee_{j=1}^d \{F^{(j)}(\tilde{X}_i^{(j)})\}^{1/w^{(j)}} - \frac{1}{d} \sum_{j=1}^d \{F^{(j)}(\tilde{X}_i^{(j)})\}^{1/w^{(j)}} \right) \prod_{j=1}^d I_i^{(j)} \right].$$

By direct application of Condition B and the law of large number, we have that

$$|\nu_n(\mathbf{w}) - \nu(\mathbf{w})| \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

For the second term, we write :

$$\begin{aligned} \left| \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}) \right| &\leq \frac{1}{N} \sum_{i=1}^n \left| \bigvee_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \bigvee_{j=1}^d \{F^{(j)}(\tilde{X}_i^{(j)})\}^{1/w^{(j)}} \right| \prod_{j=1}^d I_i^{(j)} \\ &\quad + \frac{1}{Nd} \sum_{i=1}^n \sum_{j=1}^d \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \{F^{(j)}(\tilde{X}_i^{(j)})\}^{1/w^{(j)}} \right| \prod_{j=1}^d I_i^{(j)} \\ &\leq 2 \sup_{j \in \{1, \dots, d\}} \sup_{i \in \{1, \dots, n\}} \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \{F^{(j)}(\tilde{X}_i^{(j)})\}^{1/w^{(j)}} \right|, \end{aligned}$$

where we used Lemma A.1.2 to obtain the second inequality. The right term converges almost surely to zero by Glivenko-Cantelli Theorem and the uniform continuity of  $x \mapsto x^{1/w^{(j)}}$  on  $[0, 1]$ .  $\square$

Finally, we give some elements to establish Corollary 2.2.2. The strong consistency follows directly from the stability of the almost surely convergence through a continuous function. The weak convergence comes down from the functional Delta method (see, *e.g.*, Theorem 3.9.4 of van der Vaart and Wellner (1996)) and from result in Proposition 2.2.1.

**Proof of Corollary 2.2.2** Applying the functional Delta method, we have as  $n \rightarrow \infty$ ,

$$\begin{aligned} \sqrt{n} \left( \hat{\mathcal{A}}_n^{\mathcal{H}^*}(\mathbf{w}) - \mathcal{A}(\mathbf{w}) \right) &\rightsquigarrow - (1 + \mathcal{A}(\mathbf{w}))^2 \left\{ \frac{1}{d} \sum_{j=1}^d (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx \right. \\ &\quad \left. - \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right\}_{\mathbf{w} \in \Delta_{d-1}}. \end{aligned}$$

For a fixed  $\mathbf{w} \in \Delta_{d-1}$ , as a linear transformation of a tight Gaussian process, it follows that

$$\sqrt{n} \left( \hat{\mathcal{A}}_n^{\mathcal{H}^*}(\mathbf{w}) - \mathcal{A}(\mathbf{w}) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \mathcal{V}(\mathbf{p}, \mathbf{w})),$$

with, by definition

$$\begin{aligned} \mathcal{V}(\mathbf{p}, \mathbf{w}) &\triangleq \text{Var} \left( - (1 + \mathcal{A}(\mathbf{w}))^2 \left\{ \frac{1}{d} \sum_{j=1}^d (1 + \lambda^{(j)}(\mathbf{w})(d-1)) \int_{[0,1]} \alpha(\mathbf{1}^{(j)}(x^{w^{(j)}})) - \beta^{(j)}(x^{w^{(j)}}) dx \right. \right. \\ &\quad \left. \left. - \int_{[0,1]} S_{C_\infty}(x^{w^{(1)}}, \dots, x^{w^{(d)}}) dx \right\} \right) \\ &= (1 + \mathcal{A}(\mathbf{w}))^4 \mathcal{S}^{\mathcal{H}^*}(\mathbf{p}, \mathbf{w}), \end{aligned}$$

where we used Proposition 2.2.1 to conclude.  $\square$

## A.1.2 Proofs of auxiliary results

**Proof of Lemma A.1.1** Following Segers (2015) Example 3.5, we consider the functions from  $\{0, 1\}^d \times \mathbb{R}^d$  into  $\mathbb{R}$  : for  $\mathbf{x} \in \mathbb{R}^d$ , and  $j \in \{1, \dots, d\}$

$$F^{(j)}(\mathbf{I}, \mathbf{X}) = \mathbf{1}_{\{I^{(j)}=1\}}, g_{j,x^{(j)}}(\mathbf{I}, \mathbf{X}) = \mathbf{1}_{\{X^{(j)} \leq x^{(j)}, I^{(j)}=1\}}, f_{d+1} = \prod_{j=1}^d F^{(j)}, g_{d+1,\mathbf{x}} = \prod_{j=1}^d g_{j,x^{(j)}}.$$

Let  $P$  denote the common distribution of the tuple  $(\mathbf{I}, \mathbf{X})$ . The collection of functions

$$\mathcal{F} = \{f_1, \dots, f_d, f_{d+1}\} \cup \bigcup_{j=1}^d \{g_{j, \mathbf{x}^{(j)}}, x^{(j)} \in \mathbb{R}\} \cup \{g_{d+1, \mathbf{x}}, \mathbf{x} \in \mathbb{R}^d\}$$

is a finite union of VC-classes and thus  $P$ -Donsker (for more information, see Chapter 2.6 of [van der Vaart and Wellner \(1996\)](#)). The empirical process  $\mathbb{G}_n$  defined by

$$\mathbb{G}_n(f) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n f(\mathbf{I}_i, \mathbf{X}_i) - \mathbb{E}[f(\mathbf{I}_i, \mathbf{X}_i)] \right), \quad f \in \mathcal{F},$$

converges in  $\ell^\infty(\mathcal{F})$  to a  $P$ -brownian bridge  $\mathbb{G}$ . For  $\mathbf{x} \in \mathbb{R}^d$ ,

$$\begin{aligned} \hat{F}_n^{(j)}(x^{(j)}) &= \frac{p^{(j)} F^{(j)}(x^{(j)}) + n^{-1/2} \mathbb{G}_n g_{j, \mathbf{x}^{(j)}}}{p^{(j)} + n^{-1/2} \mathbb{G}_n f_j}, \\ \hat{F}_n(\mathbf{x}) &= \frac{p F(\mathbf{x}) + n^{-1/2} \mathbb{G}_n g_{d+1, \mathbf{x}}}{p + n^{-1/2} \mathbb{G}_n f_{d+1}}. \end{aligned}$$

We obtain for the second one

$$\begin{aligned} p(\hat{F}_n(\mathbf{x}) - F(\mathbf{x})) &= n^{-1/2} (\mathbb{G}_n(g_{d+1, \mathbf{x}}) - \hat{F}_n(\mathbf{x}) \mathbb{G}_n(f_{d+1})) \\ &= n^{-1/2} (\mathbb{G}_n(g_{d+1, \mathbf{x}} - F(\mathbf{x}) f_{d+1})) - n^{-1/2} \mathbb{G}_n(f_{d+1})(\hat{F}_n(\mathbf{x}) - F(\mathbf{x})). \end{aligned}$$

We thus have

$$\sqrt{n} (\hat{F}_n(\mathbf{x}) - F(\mathbf{x})) = p^{-1} (\mathbb{G}_n(g_{d+1, \mathbf{x}} - F(\mathbf{x}) f_{d+1})) - p^{-1} \mathbb{G}_n(f_{d+1})(\hat{F}_n(\mathbf{x}) - F(\mathbf{x})).$$

Applying the central limit theorem and Condition **B** gives that  $\mathbb{G}_n(f_{d+1}) \xrightarrow{d} \mathcal{N}(0, \mathbb{P}(f_{d+1} - \mathbb{P}f_{d+1})^2)$ , the law of large numbers gives also  $\hat{F}_n(\mathbf{x}) - F(\mathbf{x}) = o_{\mathbb{P}}(1)$ . Using Slutsky's lemma gives us

$$\sqrt{n} (\hat{F}_n(\mathbf{x}) - F(\mathbf{x})) = p^{-1} (\mathbb{G}_n(g_{d+1, \mathbf{x}} - F(\mathbf{x}) f_{d+1})) + o_{\mathbb{P}}(1).$$

Similar reasoning might be applied to the margins, as a consequence, Condition **C** is fulfilled with for  $\mathbf{u} \in [0, 1]^d$ ,

$$\begin{aligned} \beta^{(j)}(u^{(j)}) &= (p^{(j)})^{-1} \mathbb{G} \left( g_{j, (F^{(j)})^{-1}(u^{(j)})} - u^{(j)} F^{(j)} \right), \\ \alpha(\mathbf{u}) &= p^{-1} \mathbb{G} \left( g_{d+1, (F^{(d)})^{-1}(\mathbf{u})} - C_\infty(\mathbf{u}) f_{d+1} \right). \end{aligned}$$

Let us compute one covariance function, the method still the same for the others, without loss of generality, suppose that  $j < k$ , we have for  $u^{(j)}, v^{(k)} \in [0, 1]$



$$\begin{aligned}
\text{cov}(\beta^{(j)}(u^{(j)}), \beta^{(k)}(v^{(k)})) &= \mathbb{E} \left[ (p^{(j)})^{-1} \mathbb{G} \left( g_{j, (F^{(j)})^{\leftarrow}(u^{(j)})} - u^{(j)} F^{(j)} \right) (p^{(k)})^{-1} \mathbb{G} \left( g_{k, (F^{(k)})^{\leftarrow}(v^{(k)})} - v^{(k)} f_k \right) \right] \\
&= \frac{1}{p^{(j)} p^{(k)}} \mathbb{E} \left[ \mathbb{G} \left( g_{j, (F^{(j)})^{\leftarrow}(u^{(j)})} - u^{(j)} F^{(j)} \right) \mathbb{G} \left( g_{k, (F^{(k)})^{\leftarrow}(v^{(k)})} - v^{(k)} f_k \right) \right] \\
&= \frac{1}{p^{(j)} p^{(k)}} \mathbb{P} \left\{ X^{(j)} \leq (F^{(j)})^{\leftarrow}(u^{(j)}), X^{(k)} \leq (F^{(k)})^{\leftarrow}(v^{(k)}), I^{(j)} = 1, I^{(k)} = 1 \right\} \\
&\quad - \frac{p^{(jk)}}{p^{(j)} p^{(k)}} u^{(j)} v^{(k)} \\
&= \frac{1}{p^{(j)} p^{(k)}} \mathbb{P} \left\{ X^{(j)} \leq (F^{(j)})^{\leftarrow}(u^{(j)}), X^{(k)} \leq (F^{(k)})^{\leftarrow}(v^{(k)}) \right\} \mathbb{P} \left\{ I^{(j)} = 1, I^{(k)} = 1 \right\} \\
&\quad - \frac{p^{(jk)}}{p^{(j)} p^{(k)}} u^{(j)} v^{(k)} \\
&= \frac{p^{(jk)}}{p^{(j)} p^{(k)}} \left( C_{\infty}(\mathbf{1}^{(jk)}(u^{(j)}, v^{(k)})) - u^{(j)} v^{(k)} \right).
\end{aligned}$$

Hence the result.  $\square$

**Proof of Lemma A.1.2** The lemma becomes trivial once we write,  $\forall i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, d\}$

$$\begin{aligned}
(\tilde{U}_i^{(j)})^{1/w^{(j)}} &= \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} + (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \\
&\leq \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} + \sup_{j \in \{1, \dots, d\}} \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right| \\
&\leq \bigvee_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} + \sup_{j \in \{1, \dots, d\}} \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right|.
\end{aligned}$$

Taking the max over  $j \in \{1, \dots, d\}$  gives


$$\bigvee_{j=1}^d (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \bigvee_{j=1}^d \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \leq \sup_{j \in \{1, \dots, d\}} \left| (\tilde{U}_i^{(j)})^{1/w^{(j)}} - \left\{ F^{(j)}(X^{(j)}) \right\}^{1/w^{(j)}} \right|.$$

Moreover, by symmetry of  $\tilde{U}_i^{(j)}$  and  $F^{(j)}$ , the second one follows similarly.  $\square$

## CHAPTER 3

# HIGH-DIMENSIONAL VARIABLE CLUSTERING BASED ON MAXIMA OF A WEAKLY DEPENDENT RANDOM PROCESS

This chapter is based on work currently under revision for publication in an international peer-reviewed journal.

 Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde (2023), High-dimensional variable clustering based on maxima of a weakly dependent random process.

### Abstract.

We propose a new class of models for variable clustering called Asymptotic Independent block (AI-block) models, which defines population-level clusters based on the independence of the maxima of a multivariate stationary mixing random process among clusters. This class of models is identifiable, meaning that there exists a maximal element with a partial order between partitions, allowing for statistical inference. We also present an algorithm depending on a tuning parameter that recovers the clusters of variables without specifying the number of clusters *a priori*. Our work provides some theoretical insights into the consistency of our algorithm, demonstrating that under certain conditions it can effectively identify clusters in the data with a computational complexity that is polynomial in the dimension. A data-driven selection method for the tuning parameter is also proposed. To further illustrate the significance of our work, we applied our method to neuroscience and environmental real-datasets. These applications highlight the potential and versatility of the proposed approach.

### 3.1 Introduction

**Motivation** Multivariate extremes arise when two or more extreme events occur simultaneously. These events are of prime interest to assess natural hazard, stemming from heavy rainfall, wind storms and earthquakes since they are driven by joint extremes of several of meteorological variables. Results from multivariate extreme value theory show that the possible dependence structure of extremes satisfy certain constraints. Indeed, the dependence structure may be described in various equivalent ways (Beirlant et al. (2004)): by the exponent measure (Balkema and Resnick (1977)), by the Pickands dependence function (Pickands (1981)), by the stable tail dependence function (Huang (1992)), by the madogram (Naveau et al. (2009), Chapter 2), and by the extreme value copula (Gudendorf and Segers (2010)).

While the modeling of univariate and low-dimensional extreme events has been well-studied, it remains a challenge to model multivariate extremes, particularly when multiple rare events

may occur simultaneously. Recent research in this area has focused on connecting the study of multivariate extremes to modern statistical and machine learning techniques. The general idea of the proposed methods is to identify groups of variables that may become large without affecting the others, also referred to as extreme direction. [Goix et al. \(2016\)](#) focus on identifying extreme directions, thus providing a sparse representation of the extremal dependence. [Chiapino et al. \(2019\)](#) proposed an incremental-type algorithm for scenarios with a high number of extreme directions. [Janßen and Wan \(2020\)](#) identify extreme directions by adapting the spherical  $K$ -means (sKmeans) clustering algorithm to the extremal setting and construct a nonparametric estimator for the theoretical cluster centers. Lastly, [Meyer and Wintenberger \(2021, 2023\)](#) frame extreme directions within what they call sparse regular variation. Our work is aligned with these directions of research as we propose a clustering algorithm for learning the dependence structure of multivariate extremes and, withal, to bridge important ideas from modern statistics and machine learning to the framework of extreme-value theory.

It is possible to perform clustering on  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , where  $n$  is the number of observations of a random vector  $\mathbf{X} \in \mathbb{R}^d$ , through two different approaches: by partitioning the set of row indices  $\{1, \dots, n\}$  or by partitioning the set of column indices  $\{1, \dots, d\}$ . The first problem is known as the data clustering problem, while the second is called the variable clustering problem, which is the focus of this paper. In data clustering, observations are drawn from a mixture distribution, and clusters correspond to different realizations of the mixing distribution, which is a distribution over all of  $\mathbb{R}^d$ .

The problem of variable clustering (see, e.g., [Bunea et al. \(2020\)](#)) involves grouping similar components of a random vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  into clusters. The goal is to recover these clusters from observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$ . Instead of clustering similar observations based on a dissimilarity measure, the focus is on defining cluster models that correspond to subsets of the components  $X^{(j)}$  of  $\mathbf{X} \in \mathbb{R}^d$ . The goal is to cluster similar variables such that variables within the same cluster are more similar to each other than they are to variables in other clusters. Variable clustering is of particular interest in the study of weather extremes, with examples in the literature on regionalization ([Bador et al. \(2015\)](#); [Bernard et al. \(2013\)](#); [Saunders et al. \(2021\)](#)), where spatial phenomena are observed at a limited number of sites. A specific case of interest is clustering these sites according to their extremal dependencies. This can be done using techniques such as  $k$ -means or hierarchical clustering with a dissimilarity measure designed for extremes. However, the statistical properties of these procedures have not been extensively studied, and it is not currently known which probabilistic models on  $\mathbf{X}$  can be estimated using these techniques. In this paper, we consider model-based clustering, where the population-level clusters are well-defined, offering interpretability and a benchmark to evaluate the performance of a specific clustering algorithm.

The assumption that data are realizations of independent and identically distributed (i.i.d.) random variables is a fundamental assumption in statistical theory and modeling. However, this assumption is often unrealistic for modern datasets or the study of time series. Developing methods and theory to handle departures from this assumption is an important area of research in statistics. One common approach is to assume that the data are drawn from a multivariate stationary and mixing random process, which implies that the dependence between observations weakens over the trajectory. This assumption is widely used in the study of non-i.i.d. processes.

Our contribution is twofold. First, we develop a probabilistic setting for Asymptotic Independent block (AI-block) models to address the problem of clustering extreme values of the target vector. These models are based on the assumption that clusters of components of a multivariate random process are independent relative to their extremes. This approach has the added benefit of being amenable to theoretical analysis, and we show that these models are identifiable (see Theorem 3.2.1). Second, we motivate and derive an algorithm specifically designed for these models (see Algorithm (ECO)). We analyze its performance in terms of exact cluster recovery for minimally separated clusters, using a cluster separation metric (see Theorem 3.3.1). The issue is investigated in the context of nonparametric estimation over block maxima of a multivariate stationary mixing random process, where the block length is a tuning parameter.

**Notations** All bold letters  $\mathbf{x}$  correspond to vectors in  $\mathbb{R}^d$ . Let  $O = \{O_g\}_{g=1,\dots,G}$  be a partition of  $\{1, \dots, d\}$  into  $G$  groups and let  $s : \{1, \dots, d\} \rightarrow \{1, \dots, G\}$  be a variable index assignment function, thus  $O_g = \{a \in 1, \dots, d : s(a) = g\} = \{i_{g,1}, \dots, i_{g,d_g}\}$  with  $d_1 + \dots + d_G = d$ . Using these notations, the variable  $X^{(i_{g,\ell})}$  should be read as the  $\ell$ th element from the  $g$ th cluster. By considering  $B \subseteq \{1, \dots, d\}$ , we denote the  $|B|$ -subvector of  $\mathbf{x}$  by  $\mathbf{x}^{(B)} = (x^{(j)})_{j \in B}$ . We denote by  $\mathbf{X} \in \mathbb{R}^d$  a random vector with cumulative distribution function  $H$  and  $\mathbf{X}^{(B)}$  a random subvector of  $\mathbf{X}$  with marginal distribution  $H^{(B)}$  whose domain is  $\mathbb{R}^{|B|}$ . Remark that when  $B = \{1, \dots, d\}$ , one has  $H = H^{(B)}$ . Classical inequalities of vectors such as  $\mathbf{x} > 0$  should be understood componentwise. The notation  $\delta_x$  corresponds to the Dirac measure at  $x$ . Let  $\mathbf{X}^{(O_g)}$ ,  $g \in \{1, \dots, G\}$  be random vectors with  $\mathbf{X} = (\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)})$ , we recall that  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  are independent if and only if  $H(\mathbf{x}) = \prod_{g=1}^G H^{(O_g)}(\mathbf{x}^{(O_g)})$ ,  $\mathbf{x} \in \mathbb{R}^d$ .

**Structure of the chapter** In Section 3.2, we provide background on extreme-value theory and describe the probabilistic framework of AI-block models. We show that these models are identifiable and provide a series of equivalent characterizations. In Section 3.3, we develop a new clustering algorithm for AI-block models and prove that it can recover the target partition with high probability under mixing conditions over the random process. We provide a process that satisfies our probabilistic and statistical assumptions in Section 3.4. We illustrate the finite sample performance of our approach on simulated datasets in Section 3.5. To exemplify further motivation for our research, we applied our method to real-data from neuroscience and environmental sciences, as discussed in Section 3.6.

## 3.2 A model for variable clustering

### 3.2.1 Background setting

Consider  $\mathbf{Z}_t = (Z_t^{(1)}, \dots, Z_t^{(d)})$ , where  $t \in \mathbb{Z}$  is a strictly stationary multivariate random process. Let  $\mathbf{M}_m = (M_m^{(1)}, \dots, M_m^{(d)})$  be the vector of component-wise maxima, where  $M_m^{(j)} = \max_{i=1, \dots, m} Z_i^{(j)}$ . Consider a random vector  $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$  with cumulative distribution function  $H$ . A normalizing function  $a$  on  $\mathbb{R}$  is a non-decreasing, right continuous function that goes to  $\pm\infty$  as  $x \rightarrow \pm\infty$ . In extreme value theory (see, for example, the monograph of Beirlant et al. (2004)), a fundamental problem is to characterize the limit distribution  $H$  in the following limit:

$$\lim_{m \rightarrow \infty} \mathbb{P}\{\mathbf{M}_m \leq \mathbf{a}_m(\mathbf{x})\} = H(\mathbf{x}), \quad (3.1)$$

where  $\mathbf{a}_m = (a_m^{(1)}, \dots, a_m^{(d)})$  with  $a_m^{(j)}, 1 \leq j \leq d$  are normalizing functions and  $H$  is a non-degenerate distribution. Typically,  $H$  is an extreme value distribution, and  $\mathbf{X}$  is a max-stable random vector with generalized extreme value margins. In this case, we can write:

$$\mathbb{P}\{\mathbf{X} \leq \mathbf{x}\} = \exp\{-\Lambda(E \setminus [0, \mathbf{x}])\},$$

where  $\Lambda$  is a Radon measure on the punctured cone  $E = [0, \infty)^d \setminus \mathbf{0}$ . When (3.1) holds with  $H$  an extreme value distribution, the process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is said to be in the max-domain of attraction of the random vector  $\mathbf{X}$  with cumulative distribution function  $H$ , denoted as  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \in \mathcal{D}(H)$ , where  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z}))$  is the law of the stationary time series  $(\mathbf{Z}_t, t \in \mathbb{Z})$  on  $(\mathbb{R}^d)^\mathbb{Z}$ . In our context of a dependent process  $(\mathbf{Z}_t, t \in \mathbb{Z})$ , the limit in (3.1) will in general be different from a multivariate extreme value distribution, see, e.g., (Bücher and Segers, 2014, Section 4.1), and further conditions over the regularity (or mixing conditions, please refer to Section 1.1.4 for definitions) are thus needed to obtain a multivariate extreme value distribution. In particular, if the random process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is  $\beta$ -mixing, then (3.1) holds with  $H$  a multivariate extreme value distribution.

The max-domain of attraction can be described in the language of copulae. Subsequently, we assume that the marginals of  $Z_1^{(1)}, \dots, Z_1^{(d)}$  are continuous and we denote by  $C_m$  the unique copula associated with  $\mathbf{M}_m$ . More precisely, the max-domain of attraction condition in Equation (3.1) is equivalent to a max-domain of attraction condition on the levels of copulae (see Condition  $\mathcal{A}$  below) and a max-domain of attraction for each margin.

**Condition  $\mathcal{A}$ .** There exists a copula  $C_\infty$  such that

$$\lim_{m \rightarrow \infty} C_m(\mathbf{u}) = C_\infty(\mathbf{u}), \quad \mathbf{u} \in [0, 1]^d.$$

Specifically, when Equation (3.1) holds, Condition  $\mathcal{A}$  is satisfied, and consequently, the copula associated with  $H$  is  $C_\infty$ . Typically, the limit  $C_\infty$  is an extreme value copula, that is, the copula  $C_\infty$  is max-stable  $C_\infty(\mathbf{u}^{1/s})^s = C_\infty(\mathbf{u})$ , for all  $s > 0$  and it can be expressed as follows for  $\mathbf{u} \in [0, 1]^d$ :

$$C_\infty(\mathbf{u}) = \exp\left\{-L\left(-\ln(u^{(1)}), \dots, -\ln(u^{(d)})\right)\right\},$$

where  $L : [0, \infty]^d \rightarrow [0, \infty]$  is the associated stable tail dependence function (see Gudendorf and Segers (2010) for an overview of extreme value copulae). However,  $C_\infty$  is in general different from the extreme value copula, denoted  $C_\infty^{\text{iid}}$ , obtained when the process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is serially independent (see, e.g., (Bücher and Segers, 2014, Section 4.1)).

As  $L$  is an homogeneous function of order 1, i.e.,  $L(a\mathbf{z}) = aL(\mathbf{z})$  for all  $a > 0$ , we have, for all  $\mathbf{z} \in [0, \infty)^d$ ,

$$L(\mathbf{z}) = (z^{(1)} + \dots + z^{(d)})A(\mathbf{t}),$$

with  $t^{(j)} = z^{(j)} / (z^{(1)} + \dots + z^{(d)})$  for  $j \in \{2, \dots, d\}$ ,  $t^{(1)} = 1 - (t^{(2)} + \dots + t^{(d)})$ , and  $A$  is the restriction of  $L$  into the  $d$ -dimensional unit simplex, viz.

$$\Delta_{d-1} = \{(v^{(1)}, \dots, v^{(d)}) \in [0, 1]^d : v^{(1)} + \dots + v^{(d)} = 1\}.$$

The function  $A$  is known as the Pickands dependence function and is often used to quantify the extremal dependence among the elements of  $\mathbf{X}$ . Indeed,  $A$  satisfies the constraints  $1/d \leq$

$\max(t^{(1)}, \dots, t^{(d)}) \leq A(\mathbf{t}) \leq 1$  for all  $\mathbf{t} \in \Delta_{d-1}$ , with lower and upper bounds corresponding to the complete dependence and independence among maxima. For the latter, it is commonly said that the stationary random process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  exhibits asymptotic independence, i.e., the multivariate extreme value distribution  $H$  in the max-domain of attraction is equal to the product of its marginal extreme value distributions.

### 3.2.2 Proposed AI-block models

In this paper, our main focus is to identify disjoint groups of variables that may simultaneously be large without affecting the other groups. We thus introduce a novel class of models called AI-block models for variable clustering. These models define population-level clusters as groups of variables that exhibit dependence within clusters but extremes are independent from variables in other clusters. Formally, these variables can be partitioned into an unknown number, denoted as  $G$ , of clusters represented by  $O = \{O_1, \dots, O_G\}$ . Within each cluster, the variables display dependence, while the clusters themselves are asymptotically independent. In this section, our primary focus is on the identifiability of the model, specifically addressing the existence of a unique maximal element according to a specific partial order on the partition. We provide an explicit construction of this maximal element, which represents the thinnest partition where the desired property holds. This maximal element serves as a target for statistical inference within our framework.

In a different framework, consider  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  be arbitrary random subvectors with marginal copulae  $C^{(O_1)}, \dots, C^{(O_G)}$  respectively. Independence between random vectors holds if and only if the underlying copula of  $\mathbf{X} = (\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)})$  is the product of the marginal copulae. This statement also holds for marginal extreme value copulae  $C_\infty^{(O_1)}, \dots, C_\infty^{(O_G)}$  with the property that the copula of  $\mathbf{X}$  is again an extreme value copula.

**Proposition 3.2.1.** *Let  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  be independent extreme value random vectors with extreme value copulae  $C_\infty^{(O_1)}, \dots, C_\infty^{(O_G)}$ . Then the function  $C_\infty$  defined as*

$$\begin{aligned} C_\infty : [0, 1]^d &\longrightarrow [0, 1] \\ \mathbf{u} &\longmapsto \prod_{g=1}^G C_\infty^{(O_g)}(u^{(i_{g,1})}, \dots, u^{(i_{g,d_g})}), \end{aligned}$$

*is an extreme value copula associated to the random vector  $\mathbf{X} = (\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)})$ .*

As a result, a random vector  $\mathbf{X}$  that exhibits (asymptotic) independence between extreme-valued subvectors therefore inherits this extreme-valued property. Using the definitions and notations so far introduced in this work, we now present the definition of our model.

**Definition 3.2.1 (Asymptotic Independent-block model).** Let  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a  $d$ -variate stationary random process and  $\mathbf{X}$  a random vector with cumulative distribution function  $H$ , a multivariate extreme value distribution with copula  $C_\infty$ . The random process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is said to follow an AI-block model if  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \in D(H)$  and there exists a partition  $O = \{O_1, \dots, O_G\}$  of  $\{1, \dots, d\}$  with  $C_\infty(\mathbf{u}) = \prod_{g=1}^G C_\infty^{(O_g)}(\mathbf{u}^{(O_g)})$ .

Notice that, when  $G = 1$ , the definition of AI-block models thus reduces to  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \in D(H)$ .

Following [Bunea et al. \(2020\)](#), we introduce the following notation in our framework. We say that  $(\mathbf{Z}_t, t \in \mathbb{Z})$  follows an AI-block model with a partition  $O$ , denoted  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O$ . We define the set  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z})) = \{O : O \text{ is a partition of } \{1, \dots, d\} \text{ and } \mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O\}$ , which is nonempty and finite, and therefore has maximal elements. We introduce a partial order on partitions as follows: let  $O = \{O_g\}_g$  and  $\{S_{g'}\}_{g'}$  be two partitions of  $\{1, \dots, d\}$ . We say that  $S$  is a sub-partition of  $O$  if, for each  $g'$ , there exists  $g$  such that  $S_{g'} \subseteq O_g$ . We define the partial order  $\leq$  between two partitions  $O$  and  $S$  of  $\{1, \dots, d\}$  as follows:

$$O \leq S, \text{ if } S \text{ is a sub-partition of } O. \quad (3.2)$$

For any partition  $O = \{O_g\}_{1 \leq g \leq G}$ , we write  $a \stackrel{O}{\sim} b$  where  $a, b \in \{1, \dots, d\}$  if there exists  $g \in \{1, \dots, G\}$  such that  $a, b \in O_g$ .

**Definition 3.2.2.** For any two partitions  $O, S$  of  $\{1, \dots, d\}$ , we define  $O \cap S$  as the partition induced by the equivalence relation  $a \stackrel{O \cap S}{\sim} b$  if and only if  $a \stackrel{O}{\sim} b$  and  $a \stackrel{S}{\sim} b$ .

Checking that  $a \stackrel{O \cap S}{\sim} b$  is an equivalence relation is straightforward. With this definition, we have the following interesting properties that lead to the desired result, the identifiability of AI-block models.

**Theorem 3.2.1.** *Let  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a stationary random process. The following properties hold:*

- (i) *Consider  $O \leq S$ . Then  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim S$  implies  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O$ ,*
- (ii)  *$O \leq O \cap S$  and  $S \leq O \cap S$ ,*
- (iii)  *$\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O$  and  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim S$  is equivalent to  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O \cap S$ ,*
- (iv) *The set  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$  has a unique maximum  $\bar{O}$ , with respect to the partition partial order  $\leq$  in (3.2).*

The proof demonstrates that, for any partition such that  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z}))$  follows an AI-block model, there exists a maximal partition, denoted by  $\bar{O}$ , and its structure is intrinsic to the definition of the extreme random vector  $\mathbf{X}$ . This partition, which represents the thinnest partition where  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z}))$  is asymptotically independent per block, matches our expectations for a reasonable clustering target in these models. Also, a careful reading of the proof shows that this statement can also hold for the setting of mutually independent random vectors.

### 3.2.3 Extremal dependence structure for AI-block models

In extreme value theory, independence between the components  $X^{(1)}, \dots, X^{(d)}$  of a random vector with extreme value distribution  $H$  can be characterized in a useful way: according to ([Takahashi, 1994](#), Theorem 2.2), total independence of  $\mathbf{X}$  is equivalent to the existence of a vector  $\mathbf{p} = (p^{(1)}, \dots, p^{(d)}) \in \mathbb{R}^d$  such that  $H(\mathbf{p}) = H^{(1)}(p^{(1)}) \dots H^{(d)}(p^{(d)})$ . This characterization was extended for the independence of a multivariate extreme value distribution to its multivariate marginals from ([Ferreira, 2011](#), Proposition 2.1), i.e., it holds that  $H(\mathbf{x}) = \prod_{g=1}^G H^{(O_g)}(\mathbf{x}^{(O_g)})$  for every  $\mathbf{x} \in \mathbb{R}^d$  if and only if there exists  $\mathbf{p} \in \mathbb{R}^d$  such that  $0 < H^{(O_g)}(\mathbf{p}^{(O_g)}) < 1$  for every  $g \in \{1, \dots, G\}$  and  $H(\mathbf{p}) = \prod_{g=1}^G H^{(O_g)}(\mathbf{p}^{(O_g)})$ . An alternative proof of this result, which involves the spectral measure, along with additional characterizations of extremal dependence structures in AI-block models, is presented in [Appendix B.2.1](#). One direct application of this result in AI-block models is that  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  are independent if and only if  $L(1, \dots, 1) = \sum_{g=1}^G L^{(O_g)}(\mathbf{1}^{(O_g)})$ .

**Definition 3.2.3** (Sum of Extremal COefficients (SECO)). The extremal coefficient of a random vector  $\mathbf{X}$  with copula  $C_\infty$  is defined as (see [Smith \(1990\)](#)):

$$\theta := \theta^{\{\{1, \dots, d\}\}} = L(1, \dots, 1), \quad (3.3)$$

where  $L$  is the stable tail dependence function. For a partition  $O = \{O_1, \dots, O_G\}$  of  $\{1, \dots, d\}$ , we define  $\theta^{(O_g)} = L^{(O_g)}(\mathbf{1}^{(O_g)})$ , as the extremal coefficient of the subvectors  $\mathbf{X}^{(O_g)}$  where  $d_g = |O_g|$  is the size of the set  $O_g$  and  $L^{(O_g)}$  is the stable tail dependence function associated to  $C_\infty^{(O_g)}$ . Using these coefficients, we define the following quantity SECO as

$$\text{SECO}(O) = \sum_{g=1}^G \theta^{(O_g)} - \theta. \quad (3.4)$$

The extremal coefficient satisfies  $1 \leq \theta \leq d$  where the lower and upper bounds correspond to the complete dependence and independence among maxima, respectively. The Sum of Extremal Coefficient (SECO) serves as a quantitative measure that assesses how much the sum of extremal coefficients for subvectors  $\mathbf{X}^{(O_g)}$  deviates from the extremal coefficient of the full vector  $\mathbf{X}$ . When the SECO equals 0, it signifies that the subvectors  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  form an independent partition (see [Ferreira, 2011](#), Proposition 2.1). In other words, these subvectors exhibit asymptotic independence, irrespective of any underlying distributional assumptions. Therefore, the SECO, as defined in Equation (3.4), is a valuable tool for capturing the asymptotic independent block structure of the random vector  $\mathbf{X}$ , and it offers the dual advantages of computational feasibility and being free from parametric assumptions, as discussed in Section 3.3.4.

Additionally, we establish a condition based on the extremal dependence of each cluster, which allows us to introduce a straightforward yet robust algorithm. This algorithm facilitates the comparison of pairwise extreme dependence between vector components, enabling us to draw informed conclusions about the dependence structures using only pairwise comparisons. It provides a practical means of assessing and quantifying the relationships among the various components of the vector, aiding in the analysis of complex high-dimensional data.

**Condition B.** For every  $g \in \{1, \dots, G\}$ , the extreme value random subvector  $\mathbf{X}^{(\bar{O}_g)}$  of  $\mathbf{X}$  where the latter is given in Definition 3.2.1 and  $\bar{O}_g$  is the maximal element of  $\mathcal{O}(\{\mathbf{Z}_t, t \in \mathbb{Z}\})$  in Theorem 3.2.1, exhibits dependence between all of its components.

One sufficient condition to satisfy Condition B is to suppose that the exponent measure of the random subvector  $\mathbf{X}^{(\bar{O}_g)}$  has nonnegative Lebesgue densities on the nonnegative orthant  $[0, \infty)^{d_g} \setminus \{\mathbf{0}^{(\bar{O}_g)}\}$ , for every  $g \in \{1, \dots, G\}$  (see, e.g., [Engelke and Hitz \(2020\)](#) and the associated discussions). This condition implies that components within a cluster are simultaneously large. Various classes of tractable extreme value distributions satisfy Condition B. These popular models, commonly used for statistical inference, include the asymmetric logistic model ([Tawn \(1990\)](#)), the asymmetric Dirichlet model ([Coles and Tawn \(1991\)](#)), the pairwise Beta model ([Cooley et al. \(2010\)](#)) or the Hüsler Reiss model ([Hüsler and Reiss \(1989\)](#)).



### 3.3 Consistent estimation of minimally separated clusters

#### 3.3.1 Multivariate tail coefficient

Throughout this section, assume that we observe one excerpt  $\mathbf{Z}_1 \dots, \mathbf{Z}_n$  from a  $d$ -dimensional stationary random process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  that satisfies Definition 3.2.1. The sample of size  $n$  of  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is divided into  $k$  blocks of length  $m$ , so that  $k = \lfloor n/m \rfloor$ , the integer part of  $n/m$  and there may be a remaining block of length  $n - km$ . For the  $i$ -th block, the maximum value in the  $j$ -th component is denoted by

$$M_{m,i}^{(j)} = \max \left\{ Z_t^{(j)} : t \in (im - m, im] \cap \mathbb{Z} \right\}.$$

Let us denote by  $\mathbf{M}_{m,i} = (M_{m,i}^{(1)}, \dots, M_{m,i}^{(d)})$  the vector of the componentwise maxima in the  $i$ -th block. For a fixed block length  $m$ , the sequence of block maxima  $(\mathbf{M}_{m,i})_i$  forms a stationary process that exhibits the same regularity of the process  $(\mathbf{Z}_t, t \in \mathbb{Z})$ . The distribution functions of block maxima are denoted by

$$F_m(\mathbf{x}) = \mathbb{P} \{ \mathbf{M}_{m,1} \leq \mathbf{x} \}, \quad F_m^{(j)}(X^{(j)}) = \mathbb{P} \{ M_{m,1}^{(j)} \leq X^{(j)} \},$$

with  $\mathbf{x} \in \mathbb{R}^d$  and  $j \in \{1, \dots, d\}$ . Denote by  $U_{m,1}^{(j)} = F_m^{(j)}(M_{m,1}^{(j)})$  the unobservable uniform margin of  $M_{m,1}^{(j)}$  with  $j \in \{1, \dots, d\}$ . Let  $C_m$  be the unique (as the margins of  $\mathbf{M}_{m,1}$  are continuous) copula of  $F_m$ . Then, from Condition  $\mathcal{A}$ ,  $C_m$  is in the domain-of-attraction of a copula  $C_\infty$ . By (Hsing, 1989, Theorem 4.2),  $C_\infty$  is an extreme value copula if the time series  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is  $\beta$ -mixing.

One way to measure tail dependence for a  $d$ -dimensional extreme value random vector is through the use of the extremal coefficient, as defined in Equation (3.3). According to Schlather and Tawn (2002), the coefficient  $\theta$  can be interpreted as the number of independent variables that are involved in the given random vector. Let  $x \in \mathbb{R}$  and  $\theta_m(x)$  be the extremal coefficient for the vector of maxima  $\mathbf{M}_{m,1}$ , which is defined by the following relation:

$$\mathbb{P} \left\{ \bigvee_{j=1}^d U_{m,1}^{(j)} \leq x \right\} = \mathbb{P} \{ U_{m,1}^{(1)} \leq x \}^{\theta_m(x)}.$$

Under Condition  $\mathcal{A}$ , the coefficient  $\theta_m(x)$  of the componentwise maxima  $\mathbf{M}_{m,1}$  converges to the extremal coefficient  $\theta$  of the random vector  $\mathbf{X}$ , that is:

$$\theta_m(x) \xrightarrow{m \rightarrow \infty} \theta, \quad \forall x \in \mathbb{R}.$$

It is worth noting that  $\theta$  is a constant since  $\mathbf{X}$  is a multivariate extreme value distribution. To generalize the bivariate madogram for the random vectors  $\mathbf{M}_{m,1}$  we follow the same approach as in Chapter 2 and define:

$$\nu_m = \mathbb{E} \left[ \bigvee_{j=1}^d U_{m,1}^{(j)} - \frac{1}{d} \sum_{j=1}^d U_{m,1}^{(j)} \right], \quad \nu = \mathbb{E} \left[ \bigvee_{j=1}^d H^{(j)}(X^{(j)}) - \frac{1}{d} \sum_{j=1}^d H^{(j)}(X^{(j)}) \right]. \quad (3.5)$$

Condition  $\mathcal{A}$  implies that the distribution of  $\mathbf{M}_{m,1}$  converges to a multivariate extreme distribution with copula  $C_\infty$ . A common approach for estimating the extremal coefficient in this scenario consists of supposing that the sample follows exactly the extreme value distribution and to consider  $\theta_m(x) := \theta_m$  where the latter quantity is defined as the *pre-asymptotic* extremal coefficient (see, for example, Engelke and Volgushev (2022) for a similar terminology) which is constant for every  $x$ . Thus, we have

$$\theta_m = \frac{1/2 + \nu_m}{1/2 - \nu_m}, \quad 1 \leq \theta_m \leq d.$$

One issue with the *pre-asymptotic* extremal coefficient is that it is misspecified, as extreme value distributions only arise in the limit as the block size  $m$  tends to infinity, while in practice we must use a finite sample size. We study this misspecification error in Section 3.3.3. A plug-in estimation process can be obtained using:

$$\hat{\theta}_{n,m} = \frac{1/2 + \hat{\nu}_{n,m}}{1/2 - \hat{\nu}_{n,m}}, \quad (3.6)$$

where  $\hat{\nu}_{n,m}$  is an estimate of  $\nu_m$  obtained using:

$$\hat{\nu}_{n,m} = \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d \hat{U}_{n,m,i}^{(j)} - \frac{1}{d} \sum_{j=1}^d \hat{U}_{n,m,i}^{(j)} \right], \quad (3.7)$$

and  $(\hat{U}_{n,m,1}^{(j)}, \dots, \hat{U}_{n,m,k}^{(j)})$  are the empirical counterparts of  $(U_{m,1}^{(j)}, \dots, U_{m,k}^{(j)})$  or, equivalently, scaled ranks of the sample. A data-driven method for selection the block size  $m$  is still lacking in the literature. To the best of our knowledge, only Zou et al. (2021) propose a method in the multivariate time series setting for selecting  $m$  through bias correction using sliding-block maxima, which is out of the scope of the paper. In the following, we provide non-asymptotic bounds for the error  $|\hat{\nu}_{n,m} - \nu_m|$ .

**Proposition 3.3.1.** *Let  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a stationary process with algebraic  $\varphi$ -mixing distribution,  $\varphi(n) \leq \lambda n^{-\zeta}$  where  $\lambda > 0$ , and  $\zeta > 1$ . Then the following concentration bound holds*

$$\mathbb{P} \left\{ |\hat{\nu}_{n,m} - \nu_m| \geq C_1 k^{-1/2} + C_2 k^{-1} + t \right\} \leq (d + 2\sqrt{e}) \exp \left\{ -\frac{t^2 k}{C_3} \right\},$$

where  $k$  is the number of block maxima and  $C_1$ ,  $C_2$  and  $C_3$  are constants depending only on  $\zeta$  and  $\lambda$ .

The proof of Proposition 3.3.1, along with all proofs of the mathematical results derived in Section 3.3 may be found in Appendix B.1.2 in the supplementary material. The non-asymptotic analysis in Proposition 3.3.1 is stringent and requires the use of  $\varphi$ -mixing in order to apply Hoeffding and McDiarmid inequalities in a setting where observations are not serially independent (see (Boucheron et al., 2013, Section 2)). However, tail bounds can also be established under  $\beta$ -mixing coefficients. One can also use Bernstein inequalities for  $\alpha$ -mixing sequences with a more stringent condition, namely exponentially decaying  $\alpha$ -mixing, using the main theorem in Merlevède et al. (2009).

### 3.3.2 Inference in AI-block models

In this section, we present an adapted version of the algorithm developed in [Bunea et al. \(2020\)](#) for clustering variables based on a metric on their covariances, named as **CORD**. Our adaptation involves the use of the extremal correlation as a measure of dependence between the extremes of two variables.

The **SECO** in Equation (3.4) can be written in the bivariate setting as

$$\chi(a, b) := \text{SECO}(\{a, b\}) = 2 - \theta(a, b), \quad (3.8)$$

where for notational convenience,  $\theta(a, b) := \theta(\{a, b\})$  is the bivariate extremal coefficient between  $X^{(a)}$  and  $X^{(b)}$  as defined in Equation (3.3). In fact, the bivariate **SECO** is exactly equal to the extremal correlation  $\chi$  defined in [Coles et al. \(1999\)](#). This metric has a range between 0 and 1, with the boundary cases representing asymptotic independence and comonotonic extremal dependence, respectively. In an AI-block model, the statement

$$\mathbf{X}^{(O_g)} \perp\!\!\!\perp \mathbf{X}^{(O_h)}, \quad g \neq h,$$

is equivalent to

$$\chi(a, b) = \chi(b, a) = 0, \quad \forall a \in O_g, \forall b \in O_h, \quad g \neq h. \quad (3.9)$$

Thus using Condition **B** and Equation (3.9), where the first condition can be equivalently stated using extremal correlation as:

$$a \stackrel{\bar{O}}{\sim} b \implies \chi(a, s) > 0, \chi(b, s) > 0, \text{ where } s \in \{1, \dots, d\} \text{ such that } a \stackrel{\bar{O}}{\sim} s \text{ and } b \stackrel{\bar{O}}{\sim} s,$$

the extremal correlation is a sufficient statistic to recover clusters in an AI-block model. Indeed, Equation (3.9) reveals:

$$a \not\stackrel{\bar{O}}{\sim} b \implies \chi(a, b) = 0.$$

Consequently, in an AI-block model, two variables  $X^{(a)}$  and  $X^{(b)}$  are considered part of the same cluster under Condition **B** if and only if  $\chi(a, b) > 0$ . For the estimation procedure, using tools introduced in the previous section, we give a sample version of the extremal correlation associated to  $M_{m,1}^{(a)}$  and  $M_{m,1}^{(b)}$  by

$$\hat{\chi}_{n,m}(a, b) = 2 - \hat{\theta}_{n,m}(a, b), \quad a, b \in \{1, \dots, d\},$$

where  $\hat{\theta}_{n,m}(a, b)$  is the sampling version defined in (3.6) of  $\theta(a, b)$ . With some technical arguments, a concentration result estimate follows directly from Proposition 3.3.1.

We can represent the matrix of all extremal correlations as  $\mathcal{X} = [\chi(a, b)]_{a=1, \dots, d, b=1, \dots, d}$ . Additionally, we introduce its empirical counterpart, denoted as  $\hat{\mathcal{X}}$ . This version,  $\hat{\mathcal{X}}$  incorporates elements  $\hat{\chi}_{n,m}(a, b)$  for pairs  $(a, b) \in \{1, \dots, d\}^2$ . We present an algorithm, named **ECO** (Extremal COrrrelation), which estimates the partition  $\bar{O}$  using a dissimilarity metric based on the extremal correlation. This algorithm, outlined in Algorithm (ECO), does not require the specification of the number of groups  $G$ , as it is automatically estimated by the procedure. The algorithm complexity for computing the  $k$  vectors  $\hat{\mathbf{U}}_{n,m,i} = (\hat{U}_{n,m,i}^{(1)}, \dots, \hat{U}_{n,m,i}^{(d)})$  for  $i \in \{1, \dots, k\}$  is of order  $O(dk \ln(k))$ . Given the empirical ranks, computing  $\hat{\mathcal{X}}$  and performing the algorithm

### 3.3 Consistent estimation of minimally separated clusters

---

require  $O(d^2 \vee dk \ln(k))$  and  $O(d^3)$  computations, respectively. So the overall complexity of the estimation procedure is  $O(d^2(d \vee k \ln(k)))$ .

---

**Algorithm (ECO)** Clustering procedure for AI-block models

---

```

1: procedure ECO( $S, \tau, \hat{\mathcal{X}}$ )
2:   Initialize:  $S = \{1, \dots, d\}$ ,  $\hat{\chi}_{n,m}(a, b)$  for  $a, b \in \{1, \dots, d\}$  and  $l = 0$ 
3:   while  $S \neq \emptyset$  do
4:      $l = l + 1$ 
5:     if  $|S| = 1$  then
6:        $\hat{O}_l = S$ 
7:     if  $|S| > 1$  then
8:        $(a_l, b_l) = \arg \max_{a, b \in S} \hat{\chi}_{n,m}(a, b)$ 
9:       if  $\hat{\chi}_{n,m}(a_l, b_l) \leq \tau$  then
10:         $\hat{O}_l = \{a_l\}$ 
11:       if  $\hat{\chi}_{n,m}(a_l, b_l) > \tau$  then
12:         $\hat{O}_l = \{s \in S : \hat{\chi}_{n,m}(a_l, s) \wedge \hat{\chi}_{n,m}(b_l, s) \geq \tau\}$ 
13:        $S = S \setminus \hat{O}_l$ 
14:   return  $\hat{O} = (\hat{O}_l)_l$ 

```

---

In B.2.2, we provide conditions under the regularity of the process ensuring that our algorithm is asymptotically consistent. These conditions involve  $\beta$ -mixing coefficients which are less stringent than  $\varphi$ -mixing used in the next section. Unlike in asymptotic analysis where the choice of the threshold becomes trivial, in a non-asymptotic framework, the algorithm's performance is influenced by the parameter  $\tau$ . In a non-asymptotic framework, when  $\tau \approx 0$ , the algorithm is prone to identifying the sole cluster as  $\{1, \dots, d\}$ , while a value of  $\tau \approx 1$  suggests that the algorithm is likely to return the largest partition  $\{\{1\}, \dots, \{d\}\}$ . Thus, the parameter  $\tau$  serves as a threshold that determines the algorithm's tolerance to differentiate between the noise in the inference and the signal indicating asymptotic dependence. This discriminatory capability depends on factors such as the sample size  $n$ , the dimension  $d$ , and the proximity between the sub-asymptotic framework and the maximum domain of attraction. Consequently, selecting an appropriate threshold  $\tau$  becomes a critical consideration. However, this challenge can be addressed through a non-asymptotic analysis of the algorithm, which we will discuss in the following section.

#### 3.3.3 Estimation in growing dimensions

We provide consistency results for our algorithm, allowing estimation in the case of growing dimensions, by adding non asymptotic bounds on the probability of consistently estimating the maximal element  $\hat{O}$  of an AI-block model. Furthermore, this result provides an answer for how to leverage  $\tau$  in Algorithm (ECO). The difficulty of clustering in AI-block models can be assessed via the size of the Minimal Extremal COrrrelation (MECO) separation between two variables in a same cluster:

$$\text{MECO}(\mathcal{X}) := \min_{a \overset{O}{\sim} b} \chi(a, b).$$

In AI-block models, with Condition  $\mathcal{B}$ , we always have  $\text{MECO}(\mathcal{X}) > \eta$  with  $\eta = 0$ . However, a large value of  $\eta$  will be needed for retrieving consistently the partition  $\bar{O}$  stationary observations. We are now ready to state the main result of this section.

**Theorem 3.3.1.** *We consider  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a  $d$ -multivariate stationary process following a AI-block model given in Definition 3.2.1 satisfying Condition  $\mathcal{B}$  and algebraic  $\varphi$ -mixing distribution,  $\varphi(n) \leq \lambda n^{-\zeta}$  where  $\lambda > 0$  and  $\zeta > 1$ . Define*

$$d_m = \max_{a \neq b} |\chi_m(a, b) - \chi(a, b)|.$$

Let  $(\tau, \eta)$  be parameters fulfilling

$$\begin{aligned} \tau &\geq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{(1 + \gamma) \ln(d)}{k}}, \\ \eta &\geq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{(1 + \gamma) \ln(d)}{k}} + \tau, \end{aligned}$$

where  $C_1, C_2, C_3$  are universal constants depending only on  $\lambda$  and  $\zeta$ ,  $k$  is the number of block maxima, and  $\gamma > 0$ . For a given  $\mathcal{X}$  and its corresponding estimator  $\hat{\mathcal{X}}$ , if  $\text{MECO}(\mathcal{X}) > \eta$ , then the output of Algorithm (ECO) is consistent, i.e.,

$$\mathbb{P} \left\{ \hat{O} = \bar{O} \right\} \geq 1 - 2(1 + \sqrt{e})d^{-2\gamma}.$$

The analysis of Algorithm (ECO) can be separated into two distinct components: an analytic part that provides conditions ensuring  $\hat{O} = \bar{O}$ , as detailed in Lemma B.1.2, and a stochastic part that deals with concentration results for  $\hat{\chi}_{n,m}$  in Proposition 3.3.1, directly stated in the proof of Theorem 3.3.1. In Section 3.4, we provide an example of a mixing process that satisfies all the conditions stated in Theorem 3.3.1. As Theorem 3.3.1 is not concerned with asymptotics, we did not actually assume Condition  $\mathcal{A}$ . A link between  $\mathbf{M}_m$  and  $\mathbf{X}$  is implicitly provided through the bias term  $d_m$  which measures the distance between  $\chi_m(a, b)$  and  $\chi(a, b)$ . This quantity vanishes when Condition  $\mathcal{A}$  holds as  $m \rightarrow \infty$ .

Some comments on the implications of Theorem 3.3.1 are in order. On a high level, larger dimension  $d$  and bias  $d_m$  lead to a higher threshold  $\tau$ . The effects of the dimension  $d$  and the bias  $d_m$  are intuitive: larger dimension or more bias make the partition recovery problem more difficult. It is clear that the partition recovery problem becomes more difficult as the dimension or bias increases. This is reflected in the bound of the MECO value below which distinguish between noise and asymptotic independence is impossible by our algorithm. Thus, whereas the dimension  $d$  increases, the dependence between each component should be stronger in order to distinguish between the two. In other words, for alternatives that are sufficiently separated from the asymptotic independence case, the algorithm will be able to distinguish between asymptotic independence and noise at the  $\sqrt{\ln(d)k^{-1}}$  scale. For a more quantitative discussion, our algorithm is able to recover clusters when the data dimension scales at a polynomial rate, i.e.,  $d = o(n^p)$ , with  $p > 0$  as  $\eta$  in Theorem 3.3.1 decreases with increasing  $n$ .

The order of the threshold  $\tau$  involves known quantity such as  $d$  and  $k$  and a unknown parameter  $d_m$ . For the latter, there is no simple manner to choose optimally this parameter, as there is no simple way to determine how fast is the convergence to the asymptotic extreme behavior, or how far into the tail the asymptotic block dependence structure appears. In particular, Condition  $\mathcal{A}$  does not contain any information about the rate of convergence of  $C_m$  to  $C_\infty$ . More precise statements about this rate can be made with second order conditions. Let a regularly varying function  $\Psi : \mathbb{N} \rightarrow (0, \infty)$  with coefficient of regular variation  $\rho_\Psi < 0$  and a continuous non-zero function  $S$  on  $[0, 1]^d$  such that

$$C_m(\mathbf{u}) - C_\infty(\mathbf{u}) = \Psi(m)S(\mathbf{u}) + o(\Psi(m)), \quad \text{for } m \rightarrow \infty, \quad (3.10)$$

uniformly in  $\mathbf{u} \in [0, 1]^d$  (see, e.g., [Bücher et al. \(2019\)](#); [Zou et al. \(2021\)](#) for a proper introduction to this condition). In this case, we can show that  $d_m = O(\Psi(m))$ . In the typical case  $\Psi(m) = cm^{\rho_\Psi}$  with  $c > 0$ , choosing  $m$  proportional to  $n^{1/(1-\rho_\Psi)}$  leads to the optimal convergence rate  $n^{\rho_\Psi/(1-2\rho_\Psi)}$  (see [Drees and Huang \(1998\)](#)). However, there is no simple way to know in advance or infer the value of  $\rho_\Psi$  and, in practice, it is advisable to use a data-driven procedure to select the threshold.

#### 3.3.4 Data-driven selection of the threshold parameter

The performance of Algorithm [\(ECO\)](#) depends crucially on the value of the threshold parameter  $\tau$ . This threshold involves known quantities such as  $d$  and  $k$  and a unknown parameter  $d_m$  (see [Theorem 3.3.1](#)). For the latter, there is no simple manner to choose optimally this parameter, as there is no simple way to determine how fast is the convergence to the asymptotic extreme behavior, or how far into the tail the asymptotic block dependence structure appears. Second order conditions, which are commonly used in the literature to ensure convergence to the stable tail dependence function at a certain rate, are theoretically relevant (see [Dombry and Ferreira \(2019\)](#); [Einmahl et al. \(2012\)](#); [Fougères et al. \(2015\)](#) for examples). However, finding the optimal value for the block length parameter remains a challenging task.

In practice, it is advisable to use a data-driven procedure to select the threshold in Algorithm [\(ECO\)](#). The idea is to use the SECO criteria presented in [Equation \(3.4\)](#). Let  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O$ , given a partition  $\hat{O} = \{\hat{O}_g\}_g$ , we know from [Ferreira \(2011\)](#) that the SECO similarity given by [\(3.4\)](#) is equal to 0 if and only if  $\hat{O} \leq \bar{O}$ . We thus construct a loss function given by the SECO where we evaluate its value over a grid of the  $\tau$  values. The value of  $\tau$  for which the SECO similarity has minimum values is also the value of  $\tau$  for which we have consistent recovery of our clusters. The based estimator of the SECO in [\(3.4\)](#) is thus defined as

$$\widehat{\text{SECO}}_{n,m}(\hat{O}) = \sum_g \hat{\theta}_{n,m}^{(\hat{O}_g)} - \hat{\theta}_{n,m}. \quad (3.11)$$

Let  $\hat{\mathcal{O}}$  be a collection of partitions computed with Algorithm [\(ECO\)](#), by varying  $\tau$  around its theoretical optimal value, of order  $(d_m + \sqrt{\ln(d)k^{-1}})$ , on a fine grid. For any  $\hat{O} \in \hat{\mathcal{O}}$ , we evaluate  $\widehat{\text{SECO}}_{n,m}$  in [\(3.11\)](#). In practice, the  $\widehat{\text{SECO}}(\hat{O})$  could be minimal for several values of  $\tau$ . For example, if we incorrectly group all the components of the random vector into a single cluster. Therefore, we recommend retaining the partition obtained for the minimal value of  $\widehat{\text{SECO}}(\hat{O})$  associated with the largest parameter  $\tau$ , which results in the thinnest partition of the variables of the random vector. [Proposition 3.3.2](#) offers theoretical support for this procedure.

**Proposition 3.3.2.** *We consider  $(\mathbf{Z}_t, t \in \mathbb{Z})$  to be a  $d$ -multivariate stationary process following an AI-block model given in Definition 3.2.1 with algebraic  $\varphi$ -mixing distribution,  $\varphi(n) \leq \lambda n^{-\zeta}$  where  $\lambda > 0$  and  $\zeta > 1$ . Let  $\bar{O} = \{\bar{O}_1, \dots, \bar{O}_G\}$  be the thinnest partition given by Theorem 3.2.1 with corresponding sizes  $d_1, \dots, d_G$ . Let  $\hat{O} = \{\hat{O}_1, \dots, \hat{O}_I\}$  be any partition of  $\{1, \dots, d\}$  with corresponding sizes  $d_1, \dots, d_I$ . Define*

$$D_m = \max \left\{ \left| \sum_{g=1}^G \theta_m^{(\bar{O}_g)} - \sum_{g=1}^G \theta^{(\bar{O}_g)} \right|, \left| \sum_{i=1}^I \theta_m^{(\hat{O}_i)} - \sum_{i=1}^I \theta^{(\hat{O}_i)} \right| \right\},$$

Then, there exists a constant  $c > 0$ , such that, if  $\hat{O} \not\leq \bar{O}$  and

$$\text{SECO}(\hat{O}) > 2 \left( D_m + c \sqrt{\frac{\ln(d)}{k}} \max(G, I) \max(\sqrt[2]{\sum_{g=1}^G d_g^2}, \sqrt[2]{\sum_{i=1}^I d_i^2}) \right), \quad (3.12)$$

it holds that

$$\mathbb{E}[\widehat{\text{SECO}}_{n,m}(\bar{O})] < \mathbb{E}[\widehat{\text{SECO}}_{n,m}(\hat{O})].$$

However, the bound presented in Equation (3.12) is overly pessimistic since it exhibits polynomial growth with respect to cluster sizes. Nevertheless, when we consider the scenario where  $n \rightarrow \infty$  with  $d$  fixed, then under Condition  $\mathcal{A}$ , this condition simplifies to  $\text{SECO}(\hat{O}) > 0$ , which holds true for every  $\hat{O} \not\leq \bar{O}$  (see Appendix B.2.2 in the supplementary material). Therefore, despite the pessimistic nature of this bound, the asymptotic relevance of choosing the threshold parameter based on data-driven approaches remains intact. Additionally, numerical studies provide support for the effectiveness of SECO as an appropriate criterion for determining the threshold parameter for a suitable number of data and for important cluster sizes (see Section 5.5). Furthermore, we establish the weak convergence of an estimator for  $\text{SECO}(O)$  when  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim O$  (we refer to Appendix B.3.2 for detailed information).

### 3.4 Hypotheses discussion for a multivariate random persistent process

A trivial example of an AI-block model is given by a partition  $O$  such that  $\mathcal{L}((\mathbf{Z}_t^{(O_g)}, t \in \mathbb{Z})) \in D(H^{(O_g)})$  for  $g \in \{1, \dots, G\}$  and  $\mathcal{L}((\mathbf{Z}_t^{(O_1)}, t \in \mathbb{Z})), \dots, \mathcal{L}((\mathbf{Z}_t^{(O_G)}, t \in \mathbb{Z}))$  are independent. In this simple model, the peculiar dependence structure under study is not inherent of large values of the stationary law of the process.

More interestingly, in this section we will focus on a process where the dependence between clusters disappears in the distribution tails. To this aim, we recall here a  $\varphi$ -algebraically mixing process. The interested reader is referred for instance to Bücher and Segers (2014). We show that Conditions  $\mathcal{A}$  and  $\mathcal{B}$  hold with a bit more work.

Let  $D$  denote a copula and consider i.i.d  $d$ -dimensional random vectors  $\mathbf{Z}_0, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots$  from  $D$  and independent Bernoulli random variables  $I_1, I_2, \dots$  i.i.d. with  $\mathbb{P}\{I_t = 1\} = p \in (0, 1]$ . For  $t = 1, 2, \dots$ , define the stationary random process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  by

$$\mathbf{Z}_t = \boldsymbol{\xi}_t \delta_1(I_t) + \mathbf{Z}_{t-1} \delta_0(I_t), \quad (3.13)$$

### 3.4 Hypotheses discussion for a multivariate random persistent process

---

where we suppose without loss of generality that the process is defined for all  $t \in \mathbb{Z}$  using stationarity. The persistence of the process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  arises from repeatable values in (3.13). From this persistence,  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is  $\varphi$ -mixing with coefficient of order  $O((1-p)^n)$  (Bücher and Segers, 2014, Lemma B.1), hence algebraically mixing.

Assuming that the copula  $D$  belongs to the (i.i.d.) copula domain of attraction of an extreme value copula  $D_\infty^{(iid)}$ , denoted as

$$D_m(\mathbf{u}) = \{D(\mathbf{u}^{1/m})\}^m \longrightarrow D_\infty^{(iid)}(\mathbf{u}), \quad (m \rightarrow \infty).$$

Here,  $D_m$  represents the copula of the componentwise block maximum of size  $m$  based on the serially independent sequence  $(\xi_t, t \in \mathbb{N})$ .

According to (Bücher and Segers, 2014, Proposition 4.1), if  $C_m$  denotes the copula of the componentwise block maximum of size  $m$  based on the sequence  $(\mathbf{Z}_t, t \in \mathbb{N})$ , then

$$C_m(\mathbf{u}) \xrightarrow{m \rightarrow \infty} D_\infty^{(iid)}(\mathbf{u}), \quad \mathbf{u} \in [0, 1]^d.$$

This implies that Condition  $\mathcal{A}$  is satisfied.

Consider the multivariate outer power transform of a Clayton copula with parameters  $\theta > 0$  and  $\beta \geq 1$ , defined as:

$$D(\mathbf{u}; \theta, \beta) = \left[ 1 + \left\{ \sum_{j=1}^d (\{u^{(j)}\}^{-\theta} - 1)^\beta \right\}^{1/\beta} \right]^{-1/\theta}, \quad \mathbf{u} \in [0, 1]^d.$$

The copula of multivariate componentwise maxima of an i.i.d. sample of size  $m$  from a continuous distribution with copula  $D(\cdot; \theta, \beta)$  is given by:

$$\left\{ D \left( \{u^{(1)}\}^{1/m}, \dots, \{u^{(d)}\}^{1/m}; \theta, \beta \right) \right\}^m = D \left( u^{(1)}, \dots, u^{(d)}; \theta/m, \beta \right), \quad (3.14)$$

As  $m \rightarrow \infty$ , this copula converges to the Logistic copula with shape parameter  $\beta \geq 1$ :

$$D_\infty^{(iid)}(\mathbf{u}) = D(\mathbf{u}; \beta) = \lim_{m \rightarrow \infty} D \left( u^{(1)}, \dots, u^{(d)}; \theta/m, \beta \right) = \exp \left[ - \left\{ \sum_{j=1}^d (-\ln u^{(j)})^\beta \right\}^{1/\beta} \right],$$

uniformly in  $\mathbf{u} \in [0, 1]^d$ . This result, originally stated in (Bücher and Segers, 2014, Proposition 4.3) for the bivariate case, can be extended to an arbitrary dimension without further arguments. Now, consider the following nested Archimedean copula given by:

$$D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \theta, \beta_1), \dots, D^{(O_g)}(\mathbf{u}^{(O_g)}; \theta, \beta_g); \theta, \beta_0 \right). \quad (3.15)$$

We aim to show that this copula is in the domain of attraction of an AI-block model. That is the purpose of the proposition stated below.



**Proposition 3.4.1.** *Consider  $1 \leq \beta_0 \leq \min\{\beta_1, \dots, \beta_G\}$ , then the nested Archimedean copula given in (3.15) is in the copula domain of attraction of an extreme value copula given by*

$$D\left(D^{(O_1)}(\mathbf{u}^{(O_1)}; \beta_1), \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \beta_G); \beta_0\right).$$

*In particular, taking  $\beta_0 = 1$  gives an AI-block model where extreme value random vectors  $\mathbf{X}^{(O_g)}$  correspond to a Logistic copula with parameter shape  $\beta_g$ .*

From the last conclusion of Proposition 3.4.1, we obtain Condition  $\mathcal{A}$ , that is  $(\mathbf{Z}_t, t \in \mathbb{Z})$  in (3.13) is in max-domain of attraction of an AI-block model. Noticing that the exponent measure of each cluster is absolutely continuous with respect to the Lebesgue measure, Condition  $\mathcal{B}$  is thus valid.

**Remark 3.4.1.** Notice that, using results from Bücher and Segers (2014); Zou et al. (2021), in the i.i.d. case, i.e.  $p = 1$ , there exists an auxiliary function  $\Psi_D$  for  $D_m$  with  $\Psi_D(m) = O(m^{-1})$ . By using considerations after Equation (3.10), we thus obtain  $d_m = O(m^{-1})$ .

## 3.5 Numerical examples

### 3.5.1 Numerical results

In this section, we investigate the finite-sample performance of our algorithm to retrieve clusters in AI-block models. The results in this section can be reproduced using the code made available at [https://github.com/Aleboul/ai\\_block\\_model](https://github.com/Aleboul/ai_block_model). We consider a number of AI-block models of increasing complexity. We design three resulting partitions in the limit model  $C_\infty$ :

- E1**  $C_\infty$  is composed of two blocks  $O_1$  and  $O_2$ , of equal lengths where  $C_\infty^{(O_1)}$  and  $C_\infty^{(O_2)}$  are Logistic extreme value copulae with parameters set to  $\beta_1 = \beta_2 = 10/7$ .
- E2**  $C_\infty$  is composed of  $G = 5$  blocks of random sample sizes  $d_1, \dots, d_5$  from a multinomial distribution with parameter  $q_g = 0.5^g$  for  $g \in \{1, \dots, 4\}$  and  $q_5 = 1 - \sum_{g=1}^4 q_g$ . Each random vector is distributed according to a Logistic distribution where parameters  $\beta_g = 10/7$  for  $g \in \{1, \dots, 5\}$ .
- E3** We consider the same model as **E2** where we add 5 singletons. Then we have 10 resulting clusters. Model with singletons are known to be the hardest model to recover in the clustering literature.

We consider here observations from the model described in Equation (3.13) in Section 3.4. Here, the copula  $D$  is derived from a nested Archimedean copula, as indicated in Equation (3.15). Specifically, the outer Power Clayton copula with a parameter  $\beta_0 = 1$  serves as the “mother” copula, while the outer Power Clayton copula with parameters  $\beta_1 = \dots = \beta_G = 10/7$  act as the “child” copulae. It is worth noting that the copula  $D_m$  does not fall under the category of an extreme value copula. This can be observed by considering two observations,  $u^{(i)}$  and  $u^{(j)}$ , belonging to the same cluster  $O_1$ . In this case, the nested Archimedean copula presented in Equation (3.15) takes the following form:

$$D^{(O_1)}(\mathbf{1}, u^{(i)}, u^{(j)}, \mathbf{1}; \theta, \beta_1),$$

where the margins for the indices outside of  $i$  and  $j$  are considered as 1. Consequently, the dependence is determined by an outer Power Clayton copula that does not exhibit max-stability. Similarly, when  $i$  and  $j$  belong to different clusters, the nested Archimedean copula in Equation (3.15) follows the expression:

$$D(\mathbf{1}, u^{(i)}, u^{(j)}, \mathbf{1}; \theta, 1),$$

representing a Clayton copula. It is worth noting that indices in different clusters exhibit dependence when the max-domain of attraction is not yet reached. This framework is particularly relevant as it allows us to evaluate the effectiveness of the proposed method in estimating the extremal dependence structure. We set  $\theta = 1$  for every copula, as it does not alter the domain of attraction. Based on Proposition 3.4.1 and Proposition 4.1 of Bücher and Segers (2014), we know that  $C_m$  falls within the max domain of attraction of the corresponding copula  $C_\infty$  defined in Experiments E1-E3. In other words, it represents an AI-block model with a Logistic dependence structure for the marginals. We simulate them using the method proposed by the copula R package (Marius Hofert and Martin Mächler (2011)). The goal of our algorithm is to cluster  $d$  variables in  $\mathbb{R}^n$ . Several simulation frameworks are considered and detailed in the following.

- F1** We first investigate the choice of the intermediate sequence  $m$  of the block length used for estimation. We let  $m \in \{3, 6, \dots, 30\}$  with a fixed sample size  $n = 10000$  and  $k = \lfloor n/m \rfloor$ .
- F2** We compute the performance of the structure learning method for varying sample size  $n$ . Since the value of  $m$  which is required for consistent estimation is unknown in practice we choose  $m = 20$ .
- F3** We show the relationship between the average SECO and exact recovery rate of the method presented in Section 3.3.4. We use the case  $n = 16000$ ,  $k = 800$  and  $d = 1600$  to study the “large  $k$ , large  $d$ ” of our approach.

In the simulation study, we use the fixed threshold  $\alpha = 2 \times (1/m + \sqrt{\ln(d)/k})$  for F1 and F2 since our theoretical results given in Theorem 3.3.1 suggest the usage of a threshold proportional  $d_m + \sqrt{\ln(d)/k}$  and we can show, in the i.i.d. settings (where  $p = 1$ ) that  $d_m = O(1/m)$  (see details in Section B.1.2). For Framework F3, we vary  $\alpha$  around its theoretical optimal value, on a fine grid. The specific parameter setting we employ involves setting  $p = 0.9$ , which is further detailed below and illustrated in Figure 3.1.

**Results.** Figure 3.1 states all the results we obtain from each experiment and framework considered in this numerical section. We plot the exact recovery rate for Algorithm (ECO) with dimensions  $d = 200$  and  $d = 1600$ . Each experiment is performed using  $p = 0.9$ . As expected, the performance of our algorithm in Framework F1 (see Figure 3.1, first row) is initially increasing in  $m$ , reaches a peak, and then decreases. This phenomenon depicts a trade-off between bias and the accuracy of inference. Indeed, a large block’s length  $m$  induces a lesser bias as we reach the domain of attraction. However, the number of blocks  $k$  is consequently decreasing and implies a high variance for the inference process. These joint phenomena explain the parabolic form of the exact recovery rate for our algorithms for  $d \in \{200, 1600\}$ . Considering the Framework F2 the performance of our algorithm is better as the number of block-maxima increases (see Figure 3.1, second row).

A classical pitfall for learning algorithms is high dimensional settings. Here, when the dimension increases from 200 to 1600, our algorithm consistently reports the maximal element  $\bar{O}$  with

a reasonable number of blocks. This is in accordance with our theoretical findings, as the difficulty of clustering in AI-block models, as quantified by  $\eta$  in Theorem 3.3.1, scales at a rate of  $\sqrt{\ln(d)k^{-1}}$ . This rate has a moderate impact on the dimension  $d$ . In Framework F3, the numerical studies in Figure 3.1 (third row) show that the optimal ranges of  $\tau$  value, for high exact recovery percentages, are also associated with low average SECO losses. This supports our data-driven choice of  $\tau$  provided in Section 3.3.4.

### 3.5.2 Comparison with competitors

In this section, we examine the performance of approximate recovery of clusters of (ECO) compared to DAMEX (Goix et al. (2016)), CLEF (Chiapino et al. (2019)), sKmeans (Janßen and Wan (2020)), MUSCLE (Meyer and Wintenberger (2023)) in terms of the Adjusted Rand Index (ARI). The ARI is a continuous metric ranging from -1 to 1 used to compare two partitions of a set. An ARI value of 1 indicates identical partitions, while random partitions typically yield a value close to zero. Negative values occur for adversarial partitions, indicating that two elements that should be together fall into different groups more often than expected at random. The results in this section can be reproduced using the code made available at [https://github.com/Aleboul/ai\\_block\\_model](https://github.com/Aleboul/ai_block_model).

**The setup** We consider the discrete-time  $d$ -variate moving maxima process  $(\mathbf{Y}_t, t \in \mathbb{Z})$  of order  $p \in \mathbb{N}$  given by

$$Y_t^{(a)} = \bigvee_{\ell=0}^p \rho^\ell \epsilon_{t+\ell}^{(a)}, \quad (t \in \mathbb{Z}, a = 1, \dots, K), \quad \rho \in (0, 1). \quad (3.16)$$

Here  $(\epsilon_t, t \in \mathbb{Z})$  is an i.i.d. sequence of  $K$ -dimensional random vectors having a Clayton copula dependence function with parameter equal to unity and standard Pareto margins. Let us consider  $(\mathbf{Z}_t, t \in \mathbb{Z})$  as  $\mathbf{Z}_t = A\mathbf{Y}_t + \mathbf{E}_t$ , where  $A = (A_{ja})_{j=1, \dots, d, a=1, \dots, K} \in [0, 1]^{d \times K}$  be a coefficient matrix with rows sums to  $\sum_{a=1}^K A_{ja} = 1$  for all  $j = 1, \dots, d$  and  $\mathbf{E}_t$  serves as a vector of noise, independent of  $\mathbf{Y}_t$  with a tail that is lighter than  $\mathbf{Y}_t$ , for any  $t \in \mathbb{Z}$ . Specifically, taking  $\mathbf{E}_t$  to be a multivariate Gaussian vector with the identity as its covariance matrix verifies this tail condition and is considered in this section. Then, the considered process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is in the max-domain of attraction of a max-linear model (see Chapter 5 for details). The extreme directions of this model are the sets  $J_a = \{j \in \{1, \dots, d\}, A_{ja} > 0\}$ , for  $a = 1, \dots, K$ . Moreover, this model can also be linked to AI-block models through the matrix  $A$  by considering  $\mathcal{L} = \{L_1, \dots, L_G\}$  a partition of  $\{1, \dots, K\}$ , then the clusters

$$O_g = \{j \in \{1, \dots, d\}, \exists! g \in \{1, \dots, G\}, A_{ja} \neq 0, a \in L_g\},$$

constitute an asymptotic independent partition of  $\{1, \dots, d\}$ , hence an AI-block model. Moreover, we specifically have in this setting  $\bigcup_{a \in L_g} J_a = O_g$ . This equation also supports a merging step for procedures that learn extreme directions to achieve clustering in AI-block models. In the experiments, we specifically merge two extreme directions if they share a common variable. We design the extremal dependence using the matrix  $A$  in two Experiments E4 and E5. In each of these experiments, we consider two different frameworks F4 and F5. They are described below:

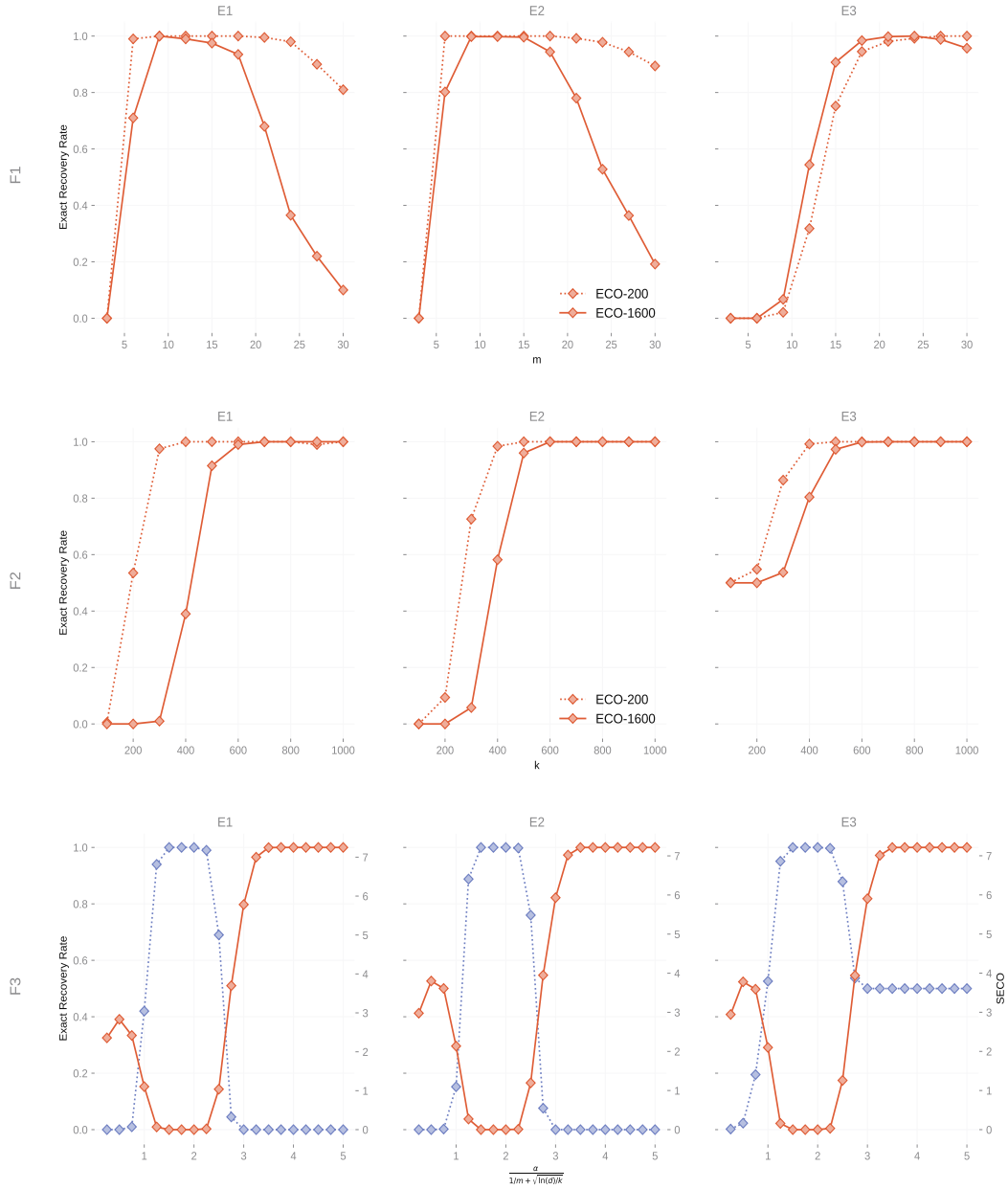


Fig. 3.1 Simulation results with  $p = 0.9$ . From top to bottom: Framework **F1**, Framework **F2**, Framework **F3**. From left to right: Experiment **E1**, Experiment **E2**, Experiment **E3**. Exact recovery rate for our algorithm (red, diamond points) for Frameworks **F1** and **F2** across 100 runs. Dotted lines correspond to  $d = 200$ , solid lines to  $d = 1600$ . The threshold  $\tau$  is taken as  $2 \times (1/m + \sqrt{\ln(d)/k})$ . For Framework **F3**, average SECO losses (red solid lines, diamonds points) and exact recovery percentages (blue dotted lines, diamond points) across 100 simulations. For better illustration, the SECO losses are standardized first by subtracting the minimal SECO loss in each figure, and the standardized SECO losses plus 1 are then plotted on the logarithmic scale.

- E4** *Few large clusters:* We set  $K = 100$ , with 5 clusters associated with groups of columns  $L_g = \{20 \times g + 1, \dots, 20 \times (g + 1)\}$  where  $g = 0, \dots, 4$ . These groups contain respectively  $(6, 5, 4, 3, 2) \times C$  entities, where  $C$  is a positive integer.
- E5** *Many small clusters:* We set  $K = 100 \times C$ , with  $5 \times C$  clusters corresponding to the group of columns  $L_g = \{20 \times k \times c + 1, \dots, 20 \times (k + 1) \times c\}$  where  $k \in \{0, 1, 2, 3, 4\}$ ,  $c \in \{1, \dots, C\}$ ,  $g = (k + 1) \times C$  so that  $G$  equals  $5 \times C$  with  $C$  is a positive integer.
- F4** We consider a framework where *Condition B holds*: rows of  $A$ , denoted as  $A_j$ , with  $j \in O_g$ , are sampled uniformly over the unit simplex  $\mathbb{R}_+^{(L_g)}$ . We investigate the performance of the algorithms with varying  $d$  and  $n$ . We let  $C$  range over  $\{1, 2, 4, 8, 16, 32\}$ , resulting in  $d \in \{20, 40, 80, 160, 320\}$  and using  $n \in \{2000, 3000, \dots, 10000\}$ .
- F5** In this scenario, we explore a framework where *Condition B fails*. Let  $s \in \{3, \dots, 20\}$  represents the sparsity index. Then, for  $j \in O_g$ , the rows of the matrix  $A$  are uniformly sampled from a random subset of  $L_g$  of size  $s$  over the unit simplex in  $\mathbb{R}_+^s$ . In this setup, we enforce clusters to be asymptotically dependent by ensuring that at least one association is shared between any pairs of variables, not necessarily the same association, so that Condition **B** fails. We let  $C$  range over  $\{1, 2, 4, 8, 16, 32\}$ , resulting in  $d \in \{20, 40, 80, 160, 320\}$  and using  $s \in \{3, 4, \dots, 20\}$  with a fixed  $n = 5000$ .

We present and provide commentary on the results for specific values of  $d$  and  $n$ ; results for other values are available upon request.

**Calibrating parameters.** The tuning parameter  $\tau$  of (ECO) is selected by the data-driven approach described in Section 3.3.4 where the block size is taken to be  $m = 20$ . In CLEF and DAMEX, the threshold was chosen by trial and error using the associated Adjusted Rand Index (ARI) with respect to the ground truth (which is unknown in practice) in the interval  $(0, 1)$ . Thus,  $\epsilon = 0.3$  and  $\kappa = 0.2$  were selected for CLEF and DAMEX, respectively. The selected number of extremes is the one used by the authors, i.e.,  $k = \lfloor \sqrt{n} \rfloor$ . The MUSCLE algorithm is fully adaptative and does not require specifying any parameters. We exclude the first extreme direction from the merging step because it is always associated with the trivial direction  $\{1, \dots, d\}$ , a phenomenon previously observed in Meyer and Wintenberger (2023) (Appendix 2).

Since sKmeans does not directly perform variable clustering, we gather the estimated centroids  $\hat{\mathbf{w}}_a \in \mathbb{R}^d$ ,  $a = 1, \dots, G$ . We then threshold them by  $\tau$ . Variables that remain positive represent groups of variables that are extremes together. Since this threshold parameter changes with the structure of  $A$ , several values of  $\tau$  must be chosen. Specifically,  $\tau$  was selected from  $\{0.15, 0.1, 0.05, 0.05, 0.04, 0.025, 0.02\}$  for Experiment E4 and set to  $\tau = 0.15$  for Experiment E5 where, for each, we set the true number of clusters (unknown in practice) to  $G = 5$  in Experiment E4 and  $G = 5 \times C$  in Experiment E5.

Figure 3.2 provides a diagnostic plot to set the threshold  $\tau$  for the sorted estimated centroids  $\hat{w}_a^{(j)}$  with  $j = 1, \dots, d$  in Experiment E5. In cases where  $d = 20$ , the gap between components that are extreme together and those that are not is clear, but it narrows as the dimension increases.

**Results and discussion** Figure 3.3 illustrates the numerical results on the approximate recovery of clusters using ARI in Framework F4, considering Experiments E4 and E5. We were able to

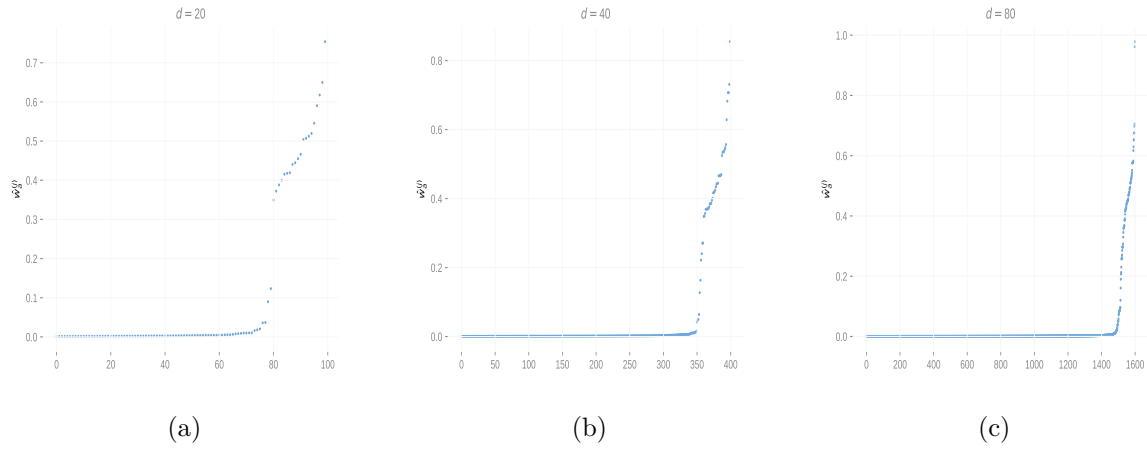


Fig. 3.2 Sorted centroids  $\hat{w}_a^{(j)}$  in Experiment E5 with  $j = 1, \dots, d$  for  $d \in \{20, 40, 80\}$ ,  $a = 1, \dots, G$  with  $G \in \{5, 10, 20\}$  and  $n = 10000$ .

run the CLEF algorithm for small values of  $d$  in Experiment E4, specifically for  $d \in \{20, 40\}$ , before encountering memory limitations for larger dimensions. As sKmeans cannot be performed when there are fewer extreme observations than the desired number of clusters, some data are missing in Experiment E5.

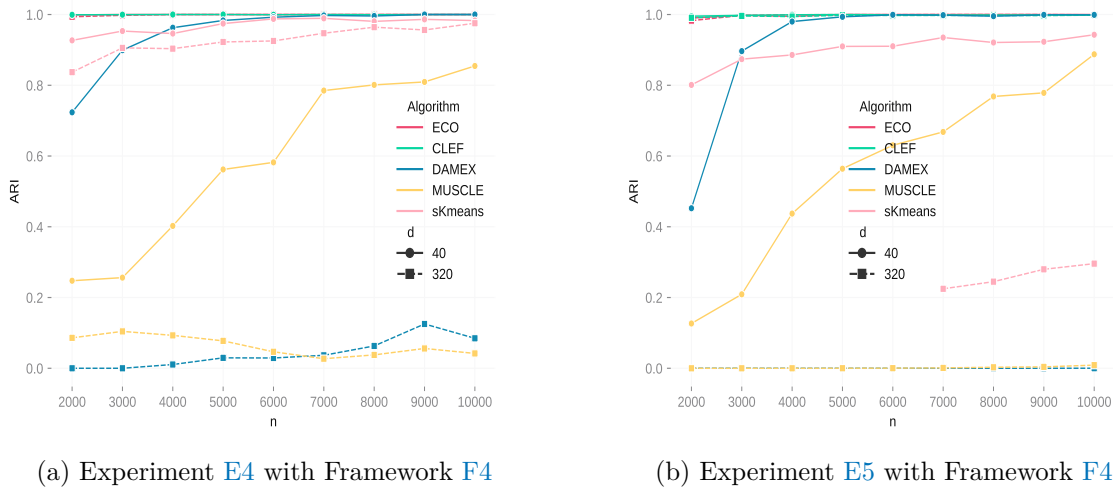


Fig. 3.3 Panel a (resp. Panel b) depicts numerical results for Experiment E4 (resp. E5) coupled with Framework F4 for  $n \in \{2000, 3000, \dots, 10000\}$  and  $d \in \{40, 320\}$ .

All algorithms demonstrate an increase in performance as the number  $n$  of observations increases. However, with increasing dimensionality, we observe decreasing performance for DAMEX, MUSCLE, and sKmeans, indicating difficulties in recovering extreme directions in higher dimensions. As expected, Algorithm (ECO) remains robust to the rise in dimensionality, even for smaller values of  $n$ . Since the CLEF algorithm constructs asymptotically dependent

pairs, triplets, quadruplets, and so on, it is anticipated that in Experiment E5 the procedure operates without memory limitations, given that the maximum cluster size is 6. Figure 3.4 presents the numerical results on approximate recovery of clusters using Adjusted Rand Index (ARI) in Framework F5, considering both Experiments E4 and E5. The selected threshold for sKmeans is directly linked to the structure of the matrix A. Due to the complexity of determining this threshold within this context, this procedure is excluded in Framework F5. Additionally, the CLEF algorithm requires a large amount of memory, and the procedure fails to run for a sparsity index greater than 11 when  $d = 160$  in Experiment E4, which explains missing points in panel a of Figure 3.4. As anticipated, our procedure demonstrates decreasing

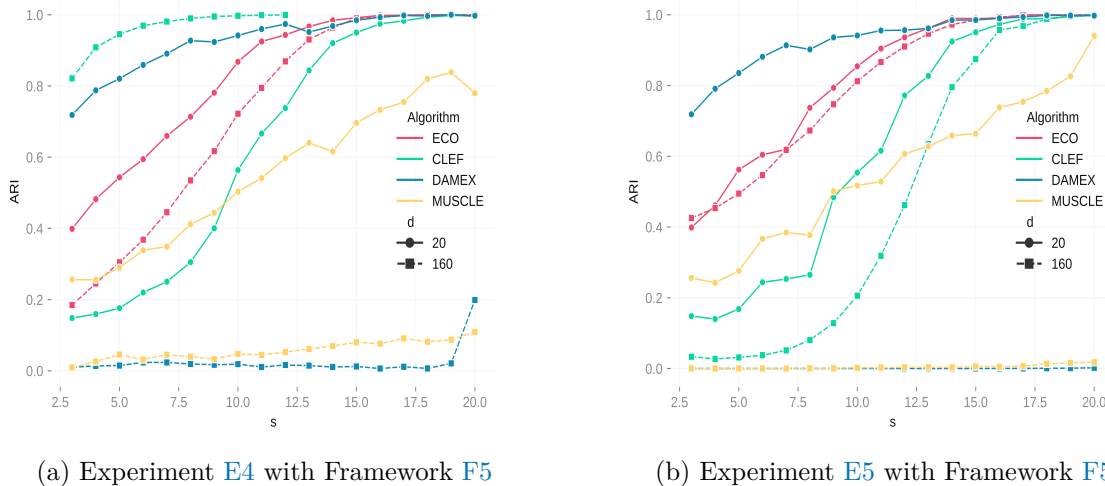


Fig. 3.4 Panel a (resp. Panel b) depicts numerical results for Experiment E4 (resp. E5) coupled with Framework F5 for  $n = 5000$  and  $d \in \{20, 160\}$ , with the sparsity index  $s \in \{3, 4, \dots, 20\}$ .

performance as the sparsity index decreases, given its heavy reliance on Condition B. When this conditions fails, our procedure recovers clusters that are too sparse. Surprisingly, other algorithms also exhibit sensitivity to the sparsity index  $s$  and display a similar declining trend. Notably, the (ECO) algorithm remains the most robust procedure to increasing dimensions in both experiments, while both DAMEX and MUSCLE show declining performance. We now provide a more nuanced discussion of the CLEF algorithm. In Experiment E4, both in Framework F4 and F5, the CLEF algorithm demonstrates better performance in higher dimensions. This phenomenon can be explained by considering that one cluster may contain many variables that exhibit asymptotic dependence. Consequently, by construction, the CLEF algorithm is more likely to identify “good candidates” of pairs, triplets, quadruplets, and so on, that are indeed asymptotically dependent. Thus, the merging step we introduce to construct the cluster is more likely to yield the desired outcome. This explanation is coherent with Experiment E4, where clusters have a constant size. In this case, we observe that CLEF shows a decreasing performance in higher dimensions.

## 3.6 Real-data applications

### 3.6.1 Clustering brain extreme from EEG channel data

Epilepsy, a significant neurological disorder, manifests as recurring unprovoked seizures. These seizures represent uncontrolled and abnormal electricity activity in the brain, posing a negative impact on one’s quality of life and potentially triggering comorbid conditions like depression and anxiety. During a seizure episode, the patient may experience a loss of muscle control, which can result in accidents and injuries (see [Strzelczyk et al. \(2023\)](#)).

One essential tool used in the diagnosis of epilepsy is the electroencephalogram (EEGs). EEGs are utilized to measure the electrical activity of the brain by employing a uniform array of electrodes. Each EEG channel is formed by calculating the potential difference between two electrodes and captures the combined potential of millions of neurons. The EEG plays a crucial role in capturing the intricate brain activity, especially during epileptic seizures, and requires analysis using statistical models. Currently, most analysis methods rely on Gaussian models that focus on the central tendencies of the data distribution (see, for example, [Embleton et al. \(2020\)](#); [Ombao et al. \(2005\)](#)). However, a significant limitation of these approaches is their disregard for the fact that neuronal oscillations exhibit non-Gaussian probability distributions with heavy tails. To address this limitation, we employ AI-block models as a comprehensive framework to overcome the limitations of light-tailed Gaussian models and investigate the extreme neural behavior during an epileptic seizure.

The dataset used to evaluate our method comprises of 916 hours of continuous scalp EEG data sampled at a rate of 256 Hz. This dataset were recorded from a total of 23 pediatric patients at Children’s Hospital Boston, see, e.g., [Shoeb \(2009\)](#). We focus the analysis on the Patient number 5 which is the first patient where 40 hours of continuous scalp EEG were sampled without interruption. Throughout the recordings, the patient experienced a total of five events that were identified as clinical seizures by medical professionals. The pediatric EEG data used in this paper is contained within the CHB-MIT database, which can be downloaded from: <https://physionet.org/content/chbmit/1.0.0/>.

For each non-seizure and seizure events, we follow the same specific processing pipeline. First, we calculate the block maxima, then calibrate the threshold using the SECO metric, as is suggested in Section 3.3.4. Finally, we perform the clustering task (see Algorithm (ECO)) using this adjusted threshold.

In the case of non-seizure records, we compute the block maxima using a block duration of 4 minutes. Figure 3.5a illustrates the relationship between the SECO and the threshold  $\tau$ . Two notable local minima are observed at  $\tau = 0.24$  and  $\tau = 0.4$ . We execute the algorithm for both values and present the results for  $\tau = 0.4$  since these results are better suited to AI-block models. Indeed, we obtain three clusters that demonstrate extreme dependence within the clusters while displaying weak extreme dependence in the block’s off-diagonal (refer to Figure 3.5b). The spatial organisation of channel clusters is depicted in Figure 3.5c.

Regarding seizure events, as the time series spans only 558 seconds, we compute block maxima with a length of 5 seconds. Considering the heavy-tailed nature of oscillations during a seizure, we believe that the limited length of the block used would not introduce a significant bias with respect to the domain of attraction. Figure 3.5d shows that the SECO is monotonically



increasing. Thus, the optimal selected threshold is the lowest value (in this case,  $\tau = 0.1$ ), which results in the minimal cluster  $\{1, \dots, d\}$ . This phenomenon is also reflected, in the extremal correlation matrix, where each channel exhibits strong pairwise extremal dependence with other channels. Consequently, the neurological disorder of the studied Patient 5 manifests simultaneous extremes across all channels, indicating generalized seizures with inter-channel communication.

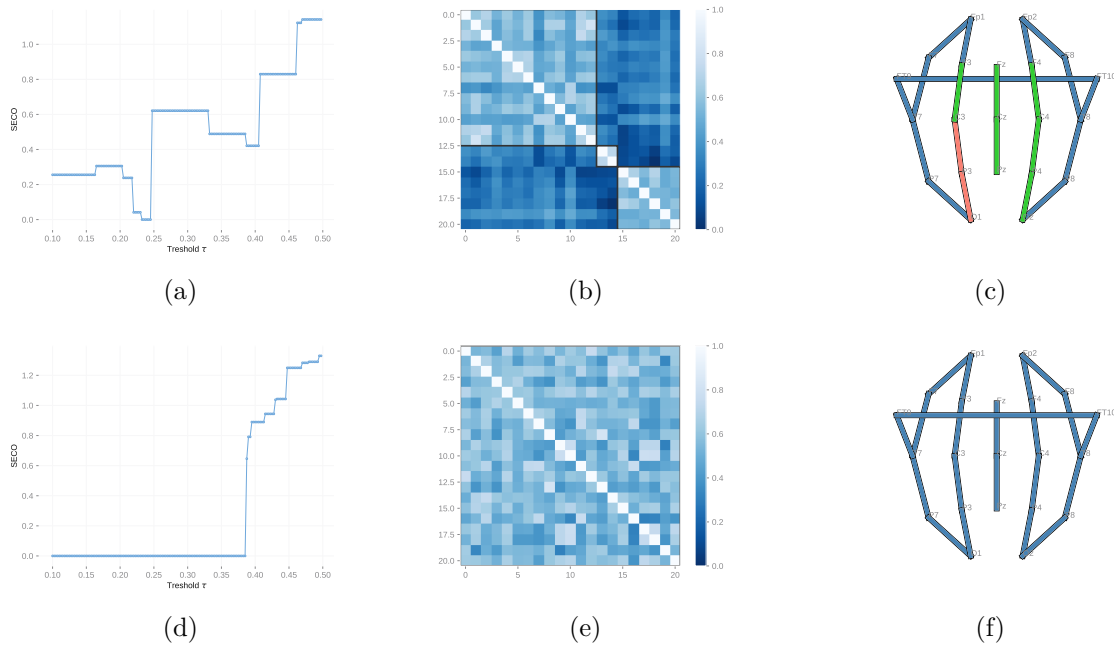


Fig. 3.5 Clustering analysis on extreme brain activity derived from EEG channel data. The results are presented in the first and second rows, representing non-seizure and seizure events, respectively. The first column illustrates the behavior of the SECO metric as it relates to the threshold level,  $\tau$ . The second column showcases the resulting clustering performed on the extremal correlation matrix using the optimal value of  $\tau$ . Finally, the third column provides a spatial organisation of the clustered channels.

### 3.6.2 Extremes on river network

To demonstrate the novel regionalization method described in this paper, we employed bi-weekly maximum river discharge data, specifically, records collected over 14-day intervals, measured in  $(m^3/s)$ . This dataset were sourced from a network of 1123 gauging stations strategically positioned across European rivers. The European Flood Awareness System (EFAS) provided these data, and they are accessible free of charge via the following website <https://cds.climate.copernicus.eu/>. EFAS primarily relies on a distributed hydrological model that operates on a grid-based system, focusing on extreme river basins. The model integrates various medium-range weather forecasts, including comprehensive sets from the Ensemble Prediction System (EPS). The dataset was generated by inputting gridded observational precipitation data, with a resolution of  $5 \times 5$  km, into the LISFLOOD hydrological model across

the EFAS domain. The temporal resolution utilized was a 24-hour time step, covering a span over 50 years.

For the calibration of the LISFLOOD within the EFAS framework, a total of 1137 stations from 215 different catchments across the Pan-European EFAS domain were used. From this list of stations with available coordinates, we extracted time-series data from the nearest cell where EFAS data were accessible. However, in this pre-processing step, stations from Albania had to be excluded as the extracted time series were identical for those stations. Additionally, calibration stations from Iceland and Israel were removed since they were located far outside the domain. As a result, we were left with 1123 gauging stations, covering 10898 observed days of river discharge between 1991 and 2020. The biweekly block maxima approach yielded 783 observations.

Following the pipeline described in Section 3.6.1, in Figure 3.6a, the SECO is depicted as it evolves in relation to the threshold  $\tau$ . The minimum value is attained at  $\tau = 0.25$ . Using this data-driven threshold, the Algorithm (ECO) is applied, resulting in 17 clusters, with 11 clusters comprising fewer than 20 stations. Figure 3.6b presents the resulting extremal correlation matrix, with clusters visually highlighted by squares. Within the clusters, there is evidence of asymptotic dependence, while moderate asymptotic dependence is observed in the off block-diagonal. Figure 3.6c provides a spatial representation of three main clusters. Notably, the clusters exhibit spatial concentration, despite the algorithm being unaware of their spatial dispersion. Overall, distinct clusters representing western, central, and northern Europe can be identified. It is crucial to emphasize that the northern Europe cluster includes stations situated in the Alps and the Pyrenees, which are geographically distant from the Scandinavian peninsula. Despite the geographical separation, these regions share mountainous terrain, and the simultaneous occurrence of extreme river discharges may be attributed to snow melting.

### 3.7 Conclusions

Our main focus in this work was to develop and analyze an algorithm for recovering clusters in AI-block models, and to understand how the dependence structure of maxima impacts the difficulty of clustering in these models. This is particularly challenging when we are dealing with high-dimensional data and weakly dependent observations that are sub-asymptotically distributed. In order to better understand these phenomena, we ask stronger assumptions about the extremal dependence structure in our theoretical analysis. Specifically, we assume the asymptotic independence between blocks, which is the central assumption of AI-block models. This assumption enables us to examine the impact of the dependence structure and develop an efficient algorithm for recovering clusters in AI-block models. By employing this procedure, we can recover the clusters with high probability by employing a threshold that scales logarithmically with the dimension  $d$ . However, it remains important to explore the optimal achievable rate for recovering AI-block models.

In this paper, we find a bound for the minimal extremal correlation separation  $\eta > 0$ . A further goal is to find the minimum value  $\eta^*$  below which it is impossible, with high probability, to exactly recover  $\bar{O}$  by any method. This question can be formally expressed using Le Cam's theory as follows:

$$\inf_{\bar{O}} \sup_{\mathcal{X} \in \mathbb{X}(\eta)} \mathbb{P}_{\mathcal{X}}(\hat{O} \neq \bar{O}) \geq \text{constant} > 0, \quad \forall \eta < \eta^*,$$

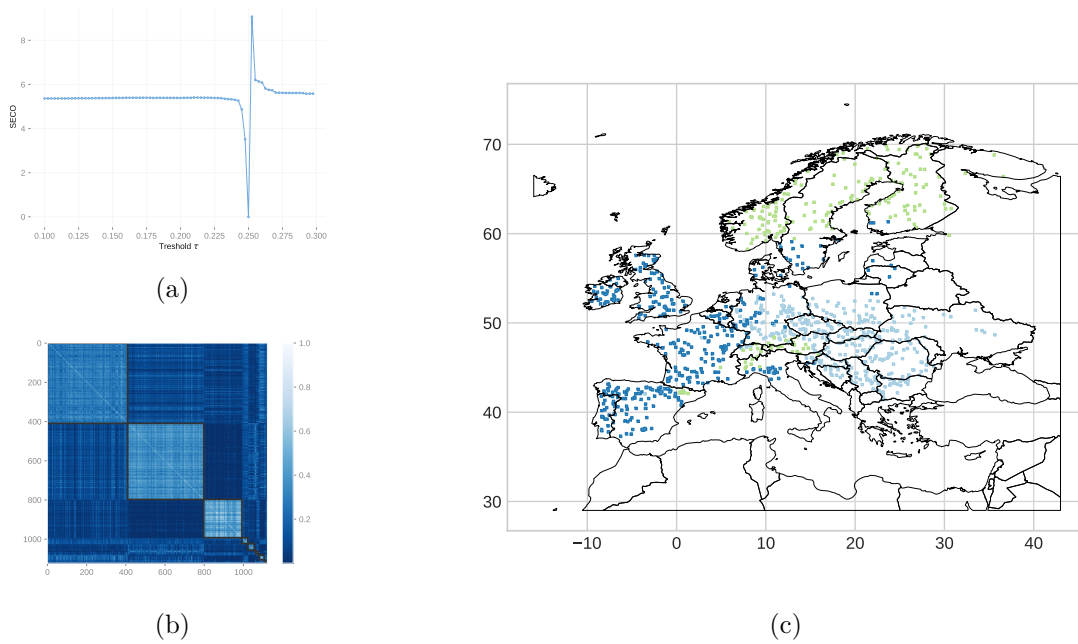


Fig. 3.6 Clustering analysis on extreme river discharges on EFAS data. The first panel illustrates the behavior of the SECO metric as it relates to the threshold level,  $\tau$ . The second panel showcases the resulting clustering performed on the extremal correlation matrix using the optimal value of  $\tau$ . Finally, the third one provides a spatial representation of the clustered stations.

with  $\mathbb{X}(\eta) = \{\mathcal{X}, \text{MECO}(\mathcal{X}) > \eta\}$  and the infimum is taken over all possible estimators. One possible direction to obtain such a result is to follow methods introduced by [Drees \(2001\)](#) for risk bounds of extreme value index. An interesting consequence of this result is to determine whether our procedure is optimal (in a minimax sense), i.e., whether the order of  $\eta^*$  and the one found in [Theorem 3.3.1](#) are the same.

# APPENDIX B

## PROOFS OF CHAPTER 3

### B.1 Proofs of main results

In the subsequent section of our materials, we employ the notation  $(\mathbf{1}, \mathbf{x}^{(B)}, \mathbf{1})$  having its  $j$ th component equal to  $x^{(j)} \mathbb{1}_{\{j \in B\}} + \mathbb{1}_{\{j \notin B\}}$ . In a similar way, we note  $(\mathbf{0}, \mathbf{x}^{(B)}, \mathbf{0})$  the vector in  $\mathbb{R}^d$  which equals  $x^{(j)}$  if  $j \in B$  and 0 otherwise.

In the subsequent section of our materials, we employ the notation  $(\mathbf{1}, \mathbf{x}^{(B)}, \mathbf{1})$  having its  $j$ th component equal to  $x^{(j)} \mathbb{1}_{\{j \in B\}} + \mathbb{1}_{\{j \notin B\}}$ . In a similar way, we note  $(\mathbf{0}, \mathbf{x}^{(B)}, \mathbf{0})$  the vector in  $\mathbb{R}^d$  which equals  $x^{(j)}$  if  $j \in B$  and 0 otherwise.

#### B.1.1 Proofs of Section 3.2

In Proposition 3.2.1, we prove that the function introduced in Section 3.2.2 is an extreme value copula. We do this by showing that its margins are distributed uniformly on the unit interval  $[0,1]$  and that it is max-stable, which is a defining characteristic of extreme value copulae.

**Proof of Proposition 3.2.1** We first show that  $C_\infty$  is a copula function. It is clear that  $C_\infty(\mathbf{u}) \in [0,1]$  for every  $\mathbf{u} \in [0,1]^d$ . We check that its univariate margins are uniformly distributed on  $[0,1]$ . Without loss of generality, take  $u^{(i_1,1)} \in [0,1]$  and let us compute

$$C_\infty(1, \dots, u^{(i_1,1)}, \dots, 1) = C_\infty^{(O_1)}(u^{(i_1,1)}, 1, \dots, 1) = u^{(i_1,1)}.$$

So  $C_\infty$  is a copula function. We now have to prove that  $C_\infty$  is an extreme value copula. We recall that  $C_\infty$  is an extreme value copula if and only if  $C_\infty$  is max-stable, that is for every  $m \geq 1$

$$C_\infty(u^{(1)}, \dots, u^{(d)}) = C_\infty(\{u^{(1)}\}^{1/m}, \dots, \{u^{(d)}\}^{1/m})^m.$$

By definition, we have

$$C_\infty(\{u^{(1)}\}^{1/m}, \dots, \{u^{(d)}\}^{1/m})^m = \prod_{g=1}^G \left\{ C_\infty^{(O_g)} \left( \{u^{(i_{g,1})}\}^{1/m}, \dots, \{u^{(i_{g,d_g})}\}^{1/m} \right) \right\}^m.$$

Using that  $C_\infty^{(O_1)}, \dots, C_\infty^{(O_G)}$  are extreme value copulae, thus max stable, we obtain

$$C_\infty(\{u^{(1)}\}^{1/m}, \dots, \{u^{(d)}\}^{1/m})^m = \prod_{g=1}^G C_\infty^{(O_g)} \left( u^{(i_{g,1})}, \dots, u^{(i_{g,d_g})} \right) = C_\infty(u^{(1)}, \dots, u^{(d)}).$$

Thus  $C_\infty$  is an extreme value copula. Finally, we prove that  $C_\infty$  is the copula of the random vector  $\mathbf{X} = (\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)})$ , that is

$$\mathbb{P}\{\mathbf{X} \leq \mathbf{x}\} = C_\infty(H^{(1)}(x^{(1)}), \dots, H^{(d)}(x^{(d)})), \quad \mathbf{x} \in \mathbb{R}^d.$$

Using mutual independence between random vectors, we have

$$\begin{aligned}\mathbb{P}\{\mathbf{X} \leq \mathbf{x}\} &= \prod_{g=1}^G \mathbb{P}\left\{X^{(i_{g,1})} \leq x^{(i_{g,1})}, \dots, X^{(i_{g,d_g})} \leq x^{(i_{g,d_g})}\right\} \\ &= \prod_{g=1}^G C_{\infty}^{(O_g)}\left(H^{(i_{g,1})}(x^{(i_{g,1})}), \dots, H^{(i_{g,d_g})}(x^{(i_{g,d_g})})\right) \\ &= C_{\infty}(H^{(1)}(x^{(1)}), \dots, H^{(d)}(x^{(d)})).\end{aligned}$$

Hence the result.  $\square$

Theorem 3.2.1, proved below, establishes several fundamental properties of the set  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$ , including the fact that subpartitions of an element  $O \in \mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$  also belong to  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$  (item (i)), the ordering of partitions and their intersections (item (ii)) and the stability of the intersection of two elements  $O, S \in \mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$  (item (iii)). Using these results, the theorem also provides an explicit construction of the unique maximal element  $\bar{O}$  of  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$  (see item (iv)).

**Proof of Theorem 3.2.1** For (i), if  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \sim S$ , then there exist a random vector  $\mathbf{X}$  with extreme value distribution  $H$  such that  $\mathcal{L}((\mathbf{Z}_t, t \in \mathbb{Z})) \in D(H)$  and a partition  $S = \{S_1, \dots, S_G\}$  of  $\{1, \dots, d\}$  which induces mutually independent random vectors  $\mathbf{X}^{(S_1)}, \dots, \mathbf{X}^{(S_G)}$ . As  $S$  is a sub-partition of  $O$ , it also generates a partition where vectors are mutually independent.

Now let us prove (ii), take  $g \in \{1, \dots, G\}$  and  $a, b \in (O \cap S)_g$ , in particular  $a \stackrel{O}{\sim} b$ , thus there exists  $g' \in \{1, \dots, G'\}$  such that  $a, b \in O_{g'}$ . The following inclusion  $(O \cap S)_g \subseteq O_{g'}$  is hence obtained and the second statement follows.

The third result (iii) comes down from the definition for the direct sense and by (i) and (ii) for the reverse one. We now go to the last item of the theorem, i.e. item (iv). The set  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$  is non-empty since the trivial partition  $O = \{1, \dots, d\}$  belongs to  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$ . It is also a finite set, and we can enumerate it  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z})) = \{O_1, \dots, O_M\}$ . Define the sequence  $O'_1, \dots, O'_M$  recursively according to

- $O'_1 = O_1$ ,
- $O'_g = O_g \cap O'_{g-1}$  for  $g = 2, \dots, M$ .

According to (iii), we have that by induction  $O'_1, \dots, O'_M \in \mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$ . In addition, we have both  $O'_{g-1} \leq O'_g$  and  $O_g \leq O'_g$ , so by induction  $O_1, \dots, O_g \leq O'_g$ . Hence the partition  $\bar{O} := O'_M = O_1 \cap \dots \cap O_{M-1}$  is the maximum of  $\mathcal{O}((\mathbf{Z}_t, t \in \mathbb{Z}))$ .  $\square$

**Remark B.1.1.** The examination of the proof of Theorem 3.2.1 reveals that many arguments may also apply to the scenario of mutually independent random vectors.

### B.1.2 Proofs of Section 3.3

Denote by  $C_{n,m}^o$  the empirical estimator of the copula  $C_m$  based on the (unobservable) sample  $(U_{m,1}^{(j)}, \dots, U_{m,k}^{(j)})$  for  $j \in \{1, \dots, d\}$ . In Proposition 3.3.1 we state a concentration inequality for the madogram estimator. This inequality is obtained through two main steps, that are using classical concentration inequalities, such as Hoeffding and McDiarmid inequalities and chaining arguments in our specific framework of multivariate mixing random process. In the following,  $C_1, C_2$  and  $C_3$  denote universal constants whose values could change from line to line of the proof.

**Proof of Proposition 3.3.1** Let us define the following quantity

$$\hat{\nu}_{n,m}^o = \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d U_{m,i}^{(j)} - \frac{1}{d} \sum_{j=1}^d U_{m,i}^{(j)} \right], \quad (\text{B.1})$$

that is the madogram estimated through the sample  $\mathbf{U}_{m,1}, \dots, \mathbf{U}_{m,k}$ . Then, the following bound is given:

$$|\hat{\nu}_{n,m} - \nu_m| \leq |\hat{\nu}_{n,m} - \hat{\nu}_{n,m}^o| + |\hat{\nu}_{n,m}^o - \nu_m|.$$

For the second term, using the triangle inequality, we obtain

$$\begin{aligned} |\hat{\nu}_{n,m}^o - \nu_m| &\leq \left| \frac{1}{k} \sum_{i=1}^k \left\{ \bigvee_{j=1}^d U_{m,i}^{(j)} - \mathbb{E} \left[ \bigvee_{j=1}^d U_{m,i}^{(j)} \right] \right\} \right| + \left| \frac{1}{k} \sum_{i=1}^k \left\{ \frac{1}{d} \sum_{j=1}^d U_{m,i}^{(j)} - \mathbb{E} \left[ \frac{1}{d} \sum_{j=1}^d U_{m,i}^{(j)} \right] \right\} \right| \\ &\triangleq E_1 + E_2, \end{aligned}$$

and for the first term,

$$|\hat{\nu}_{n,m} - \hat{\nu}_{n,m}^o| \leq 2 \sup_{j \in \{1, \dots, d\}} \sup_{x \in \mathbb{R}} |\hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x)| \triangleq E_3.$$

The rest of this proof is devoted to control each term:  $E_1$ ,  $E_2$  and  $E_3$ . Notice that the sequences  $(\bigvee_{j=1}^d U_{n,m,i}^{(j)})_{i=1}^k$ ,  $(\frac{1}{d} \sum_{j=1}^d U_{n,m,i}^{(j)})_{i=1}^k$  and  $(\mathbf{1}_{\{\bigvee_{j=1}^d U_{n,m,i}^{(j)} \leq x\}})_{i=1}^k$  share the same mixing regularity as  $(\mathbf{Z}_t)_{t \in \mathbb{Z}}$  as measurable transformation of this process. Thus, they are in particular algebraically  $\varphi$ -mixing.

**Control of the term  $E_1$ .** For every  $i \in \{1, \dots, k\}$ , we have that  $\|\bigvee_{j=1}^d U_{n,m,i}^{(j)}\|_\infty \leq 1$ , by applying the Hoeffding's inequality for algebraically  $\varphi$ -mixing sequences (see (Rio, 2017, Corollary 2.1)) we can control the following event, for  $t > 0$ ,

$$\mathbb{P}\{E_1 \geq t\} \leq \sqrt{e} \exp \left\{ -\frac{t^2 k}{2(1 + 4 \sum_{i=1}^{k-1} \varphi(i))} \right\}.$$

The term in the numerator can be bounded as

$$1 + 4 \sum_{i=1}^k \varphi(i) \leq 1 + 4 \sum_{i=1}^k \lambda i^{-\zeta} \leq 1 + 4\lambda \left( 1 + \int_1^k x^{-\zeta} dx \right) = 1 + 4\lambda \left( 1 + \frac{k^{1-\zeta} - 1}{1-\zeta} \right).$$

Using the assumption  $\zeta > 1$ , we can upper bound  $k^{1-\zeta}$  by 1 and obtain

$$1 + 4\lambda \left( 1 + \frac{k^{1-\zeta} - 1}{1-\zeta} \right) \leq 1 + 4\lambda \left( 1 + \frac{1}{\zeta - 1} \right) = 1 + \frac{4\lambda\zeta}{\zeta - 1}.$$

We thus obtain

$$\mathbb{P}\left\{E_1 \geq \frac{t}{3}\right\} \leq \sqrt{e} \exp \left\{ -\frac{t^2 k}{C_3} \right\},$$

where  $C_3 > 0$  is a constant depending on  $\zeta$  and  $\lambda$ .

**Control of the term  $E_2$ .** This control is obtained with the same arguments used for  $E_1$ . Thus, we obtain, for  $t > 0$ ,

$$\mathbb{P} \left\{ E_2 \geq \frac{t}{3} \right\} \leq \sqrt{e} \exp \left\{ -\frac{t^2 k}{C_3} \right\}.$$

**Control of the term  $E_3$ .** This bound is more technical. Before proceeding, we introduce some notations. For every  $j \in \{1, \dots, d\}$ , we define

$$\alpha_{n,m}^{(j)} = \left( \mathbb{P}_{n,m}^{(j)} - \mathbb{P}_m^{(j)} \right), \quad \beta_{n,m}^{(j)}(x) = \alpha_{n,m}^{(j)}(\cdot - \infty, x], \quad x \in \mathbb{R},$$

where  $\mathbb{P}_{n,m}^{(j)}$  corresponds to the empirical measure for the sample  $(M_{m,1}^{(j)}, \dots, M_{m,k}^{(j)})$  and  $\mathbb{P}_m^{(j)}$  is the law of the random variable  $M_m^{(j)}$ . To control the term  $E_3$ , we introduce chaining arguments as used in the proof of Proposition 7.1 of Rio (2017). Let be  $j \in \{1, \dots, d\}$  fixed and  $N$  be some positive integer to be chosen later. For any real  $x$  such that  $F_m^{(j)}(x) \neq 0$  and  $F_m^{(j)}(x) \neq 1$ , let us write  $F_m^{(j)}(x)$  in base 2 :

$$F_m^{(j)}(x) = \sum_{l=1}^N b_l(x) 2^{-l} + r_N(x), \quad \text{with } r_N(x) \in [0, 2^{-N}]$$

where  $b_l = 0$  or  $b_l = 1$ . For any  $L$  in  $[1, \dots, N]$ , set

$$\Pi_L(x) = \sum_{l=1}^L b_l(x) 2^{-l} \quad \text{and} \quad i_L = \Pi_L(x) 2^L.$$

Let the reals  $(x_L)_L$  be chosen in such a way that  $F_m^{(j)}(x_L) = \Pi_L(x)$ . With these notations

$$\begin{aligned} \beta_{n,m}^{(j)}(x) &= \beta_{n,m}^{(j)}(\Pi_1(x)) + \beta_{n,m}^{(j)}(x) - \beta_{n,m}^{(j)}(\Pi_N(x)) \\ &\quad + \sum_{L=2}^N \left[ \beta_{n,m}^{(j)}(\Pi_L(x)) - \beta_{n,m}^{(j)}(\Pi_{L-1}(x)) \right]. \end{aligned}$$

Let the reals  $x_{L,i}$  be defined by  $F_m^{(j)}(x_{L,i}) = i 2^{-L}$ . Using the above equality, we get that

$$\sup_{x \in \mathbb{R}} \left| \beta_{n,m}^{(j)}(x) \right| \leq \sum_{L=1}^N \Delta_L + \Delta_N^*,$$

with

$$\Delta_L = \sup_{i \in [1, 2^L]} \left| \alpha_{n,m}^{(j)}(\cdot - x_{L,i-1}, x_{L,i}) \right| \quad \text{and} \quad \Delta_N^* = \sup_{x \in \mathbb{R}} \left| \alpha_{n,m}^{(j)}(\cdot - \Pi_N(x), x) \right|.$$

From the inequalities

$$-2^{-N} \leq \alpha_{n,m}^{(j)}(\cdot - \Pi_N(x), x) \leq \alpha_{n,m}^{(j)}(\cdot - \Pi_N(x), \Pi_N(x) + 2^{-N}) + 2^{-N},$$

we get that

$$\Delta_N^* \leq \Delta_N + 2^{-N} \text{ and } \mathbb{E} \left[ \sup_{x \in \mathbb{R}} |\beta_{n,m}^{(j)}(x)| \right] \leq 2 \sum_{L=1}^N \|\Delta_L\|_1 + 2^{-N},$$

where  $\|\Delta_L\|_1$  is the  $L^1$ -norm of  $\Delta_L$ . Let  $N$  be the natural number such that  $2^{N-1} < k \leq 2^N$ . For this choice of  $N$ , we obtain

$$\mathbb{E} \left[ \sup_{x \in \mathbb{R}} |\beta_{n,m}^{(j)}(x)| \right] \leq 2 \sum_{L=1}^N \|\Delta_L\|_1 + k^{-1}.$$

Hence, using (Rio, 2017, Lemma 7.1) (where we divide by  $\sqrt{k}$  the considering inequality in the lemma), we obtain that

$$\begin{aligned} \mathbb{E} \left[ \sup_{x \in \mathbb{R}} |\beta_{n,m}^{(j)}(x)| \right] &\leq 2 \frac{C_0}{\sqrt{k}} \sum_{L=1}^N \left( 2^{-\frac{(\zeta-1)^2}{(4\zeta)^2}} \right)^L + k^{-1} \\ &\leq \frac{2}{\sqrt{k}} \frac{C_0}{1 - 2^{-\frac{(\zeta-1)^2}{(4\zeta)^2}}} + k^{-1} \triangleq C_1 k^{-1/2} + k^{-1}, \end{aligned}$$

where  $C_0$  and  $C_1$  are constants depending on  $\zeta$  and  $\lambda$ .

Now, fix  $x \in \mathbb{R}$  and denote by  $\Phi : \mathbb{R}^k \mapsto [0, 1]$ , the function defined by

$$\Phi(x_1, \dots, x_k) = \sup_{x \in \mathbb{R}} \left| \frac{1}{k} \sum_{i=1}^k \mathbf{1}_{\{x_i \leq x\}} - F_m^{(j)}(x) \right|.$$

For  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ , we obtain with some calculations:

$$|\Phi(\mathbf{x}) - \Phi(\mathbf{y})| \leq \sup_{x \in \mathbb{R}} \frac{1}{k} \sum_{i=1}^k \left| \mathbf{1}_{\{x_i \leq x\}} - \mathbf{1}_{\{y_i \leq x\}} \right| \leq \frac{1}{k} \sum_{i=1}^k \mathbf{1}_{\{x_i \neq y_i\}}.$$

Thus,  $\Phi$  is  $k^{-1}$ -Lipschitz with respect to the Hamming distance. Under algebraically  $\varphi$ -mixing process, we may apply (Mohri and Rostamizadeh, 2010, Theorem 8) with  $(M_{m,1}^{(j)}, \dots, M_{m,k}^{(j)})$ , we obtain with probability at least  $1 - \exp\{-2t^2k/\|\Delta_k\|_\infty^2\}$  where  $\|\Delta_k\|_\infty \leq 1 + 4 \sum_{i=1}^k \varphi(i)$

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x) \right| \leq \mathbb{E} \left[ \sup_{x \in \mathbb{R}} |\beta_{n,m}^{(j)}(x)| \right] + \frac{t}{3} \leq C_1 k^{-1/2} + C_2 k^{-1} + \frac{t}{3}.$$

Thus, for a sufficiently large  $C_3$ , with probability at most  $\exp\{-t^2k/C_3\}$

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x) \right| \geq C_1 k^{-1/2} + k^{-1} + \frac{t}{3}.$$

Using Bonferroni inequality

$$\mathbb{P} \left\{ E_3 \geq \frac{t}{3} \right\} \leq d\mathbb{P} \left\{ \sup_{x \in \mathbb{R}} \left| \hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x) \right| \geq t \right\},$$



we thus obtain a control bound for  $E_3$ . Assembling all the controls obtained for  $E_1$ ,  $E_2$  and  $E_3$ , we obtain the desired result. □

The proof of Theorem 3.3.1 needs the following results : (1) an upper bound over the quantity  $|\hat{\theta}_{n,m}(a, b) - \theta_m(a, b)|$  with respect to  $|\hat{\nu}_{n,m}(a, b) - \nu_m(a, b)|$  to use the concentration inequality introduced in Proposition 3.3.1, (2) exhibit an event such that  $\{\hat{O} = \bar{O}\}$ . Lemmas B.1.1 and B.1.2 below address these two questions. Then, taking benefits of these results, we show that the probability of the exhibited event such that  $\{\hat{O} = \bar{O}\}$  holds with high probability, as stated in Theorem 3.3.1.

**Lemma B.1.1.** *Consider a pair  $(a, b) \in \{1, \dots, d\}^2$ , the following inequality holds:*

$$|\hat{\theta}_{n,m}(a, b) - \theta_m(a, b)| \leq 9|\hat{\nu}_{n,m}(a, b) - \nu_m(a, b)|.$$

**Proof of Lemma B.1.1** We may write the respective quantities as  $\theta = f(\nu(a, b))$  and  $\hat{\theta}_{n,m} = f(\hat{\nu}_{n,m}(a, b))$  where  $f$  is a function defined as follows,

$$\begin{aligned} f : [0, 1/6] &\rightarrow [1, 2] \\ x &\mapsto \frac{1/2+x}{1/2-x}, \end{aligned}$$

with  $f(x) \in [1, 2]$  by definition of the pre-asymptotic extremal coefficient  $\theta_m$ . The domain of this function is restricted to the interval  $[0, 1/6]$  because we have  $f(x) \leq 2$ , or

$$x + \frac{1}{2} \leq 1 - 2x,$$

which holds if  $x \leq 1/6$ . The inequality  $f(x) \geq 1$  gives the positivity of the domain. In particular,  $x < 1/2$  and thus  $2^{-1} - x \geq 3^{-1} > 0$ . Taking derivative of  $f$ , we find that

$$|f'(x)| = \frac{1}{(1/2 - x)^2} \leq 3^2, \quad x \in [0, 1/6].$$

Therefore,  $f$  is 9-Lipschitz continuous and we have

$$|\hat{\theta}_{n,m}(a, b) - \theta_m(a, b)| = |f(\hat{\nu}_{n,m}(a, b)) - f(\nu_m(a, b))| \leq 9|\hat{\nu}_{n,m}(a, b) - \nu_m(a, b)|.$$

This completes the proof. □

**Lemma B.1.2.** *Consider the AI-block model in Definition 3.2.1. Define*

$$\kappa = \sup_{a, b \in \{1, \dots, d\}} |\hat{\chi}_{n,m}(a, b) - \chi(a, b)|.$$

*Consider parameters  $(\tau, \eta)$  fulfilling*

$$\tau \geq \kappa, \quad \eta \geq \kappa + \tau. \tag{B.2}$$

*If  $\text{MECO}(\mathcal{X}) > \eta$ , then Algorithm (ECO) yields  $\hat{O} = \bar{O}$ .*

**Proof of Lemma B.1.2** If  $a \stackrel{\bar{O}}{\not\sim} b$ , then  $\chi(a, b) = 0$  and

$$\hat{\chi}_{n,m}(a, b) = \hat{\chi}_{n,m}(a, b) - \chi(a, b) \leq \kappa \leq \tau.$$

Now, if  $a \stackrel{\bar{O}}{\sim} b$ , if  $\mathcal{X} \in \mathbb{X}(\eta)$  then  $\chi(a, b) > \kappa + \tau$  and

$$\kappa + \tau < \chi(a, b) - \hat{\chi}_{n,m}(a, b) + \hat{\chi}_{n,m}(a, b),$$

and thus  $\hat{\chi}_{n,m}(a, b) > \tau$ . In particular, under (B.2) and the separation condition  $\text{MECO}(\mathcal{X}) > \eta$ , we have

$$a \stackrel{\bar{O}}{\sim} b \iff \hat{\chi}_{n,m}(a, b) > \tau. \quad (\text{B.3})$$

Let us prove the lemma by induction on the algorithm step  $l$ . We consider the algorithm at some step  $l - 1$  and assume that the algorithm was consistent up to this step, i.e.  $\hat{O}_j = \bar{O}_j$  for  $j = 1, \dots, l - 1$ .

If  $\hat{\chi}_{n,m}(a_l, b_l) \leq \tau$ , then according to (B.3), no  $b \in S$  is in the same group of  $a_l$ . Since the algorithm has been consistent up to this step  $l$ , it means that  $a_l$  is a singleton and  $\hat{O}_l = \{a_l\}$ .

If  $\hat{\chi}_{n,m}(a_l, b_l) > \tau$ , then  $a_l \stackrel{\bar{O}}{\sim} b$  according to (B.3). Furthermore, the equivalence implies that  $\hat{O}_l = S \cap \bar{O}_l$ . Since the algorithm has been consistent up to this step, we have  $\hat{O}_l = \bar{O}_l$ . To conclude, the algorithm remains consistent at the step  $l$  and the result follows by induction.  $\square$

**Proof of Theorem 3.3.1** We have that for  $t > 0$  :

$$\mathbb{P} \left\{ \sup_{a,b \in \{1, \dots, d\}} |\hat{\theta}_{n,m}(a, b) - \theta_m(a, b)| \geq t \right\} \leq d^2 \mathbb{P} \left\{ |\hat{\theta}_{n,m}(a, b) - \theta_m(a, b)| \geq t \right\}.$$

With probability at least  $1 - 2(1 + \sqrt{e})d^2 \exp\{-t^2 k / C_3\}$ , and by using Proposition 3.3.1 and Lemma B.1.1, one has

$$\sup_{a,b \in \{1, \dots, d\}} |\hat{\theta}_{n,m}(a, b) - \theta(a, b)| \leq d_m + C_1 k^{-1/2} + C_2 k^{-1} + t,$$

By considering  $\delta \in ]0, 1[$  and solve the following equation

$$\frac{\delta}{d^2} = 2(1 + \sqrt{e}) \exp \left\{ -\frac{kt^2}{C_3} \right\},$$

with respect to  $t$  gives that the event

$$\sup_{a,b \in \{1, \dots, d\}} |\hat{\theta}_{n,m}(a, b) - \theta(a, b)| \geq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{1}{k} \ln \left( \frac{2(1 + \sqrt{e})d^2}{\delta} \right)},$$

is of probability at most  $\delta$ . Now, taking  $\delta = 2(1 + \sqrt{e})d^{-2\gamma}$ , with  $\gamma > 0$ , we have

$$\sup_{a,b \in \{1, \dots, d\}} \left| \hat{\theta}_{n,m}(a, b) - \theta(a, b) \right| \leq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{(1 + \gamma) \ln(d)}{k}},$$

with probability at least  $1 - 2(1 + \sqrt{e})d^{-2\gamma}$  for  $C_3$  sufficiently large. The result then follows from Lemma B.1.2 along with Condition  $\mathcal{B}$  and algebraically  $\varphi$ -mixing random process, since

$$\mathbb{P} \left\{ \kappa \leq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{(1 + \gamma) \ln(d)}{k}} \right\} \geq 1 - 2(1 + \sqrt{e})d^{-2\gamma},$$

and  $\text{MECO}(\mathcal{X}) > \eta$  by assumption. □

Therein, we prove the argument that were stated without proof in the paragraph next to Theorem 3.3.1. A condition of order two were introduced and we have state that  $d_m = O(\Psi_m)$  can be shown. We propose a proof of this statement below.

**Proof of  $d_m = O(\Psi(m))$**  Take  $a \neq b$  fixed, we have, using Lemma B.1.1

$$|\chi_m(a, b) - \chi(a, b)| = |\theta_m(a, b) - \theta(a, b)| \leq 9 |\nu_m(a, b) - \nu(a, b)|,$$

where  $\nu_m(a, b)$  (resp.  $\nu(a, b)$ ) is the madogram computed between  $M_m^{(a)}$  and  $M_m^{(b)}$  (resp. between  $X^{(a)}$  and  $X^{(b)}$ ) and we use Lemma B.1.1 to obtain the inequality. Using the results of Lemma 1 of Marcon et al. (2017), we have

$$\begin{aligned} \nu_m(a, b) - \nu(a, b) &= \frac{1}{2} \left( \int_{[0,1]} (C_m - C_\infty)(\mathbf{1}, x^{(a)}, \mathbf{1}) dx^{(a)} + \int_{[0,1]} (C_m - C_\infty)(\mathbf{1}, x^{(b)}, \mathbf{1}) dx^{(b)} \right) \\ &\quad - \int_{[0,1]} (C_m - C_\infty)(1, \dots, \underbrace{x}_{a\text{th index}}, 1, \dots, 1, \underbrace{x}_{b\text{th index}}, \dots, 1) dx, \end{aligned}$$

where the integration is taken respectively for the  $a$ -th,  $b$ -th and  $a, b$ -th components. Hence

$$\begin{aligned} |\nu_m(a, b) - \nu(a, b)| &\leq \frac{1}{2} \int_{[0,1]} |(C_m - C_\infty)(\mathbf{1}, x^{(a)}, \mathbf{1})| dx^{(a)} \\ &\quad + \frac{1}{2} \int_{[0,1]} |(C_m - C_\infty)(\mathbf{1}, x^{(b)}, \mathbf{1})| dx^{(b)} \\ &\quad + \int_{[0,1]} |(C_m - C_\infty)(1, \dots, \underbrace{x}_{a\text{th index}}, 1, \dots, 1, \underbrace{x}_{b\text{th index}}, \dots, 1)| dx. \end{aligned}$$

Using the second order condition in Equation (3.10) we obtain that  $|C_m - C_\infty|(\mathbf{u}) = O(\Psi_m)$ , uniformly in  $\mathbf{u} \in [0, 1]^d$ . Hence the statement. □

Now, we prove the theoretical result giving support to our cross validation process.

**Proof of Proposition 3.3.2** Using triangle inequality several times, we may obtain the following bound

$$\begin{aligned} \widehat{\text{SECO}}_{n,m}(\bar{O}) - \widehat{\text{SECO}}_{n,m}(\hat{O}) &\leq 2D_m + \left| \sum_{g=1}^G \hat{\theta}_{n,m}^{(\bar{O}_g)} - \sum_{g=1}^G \theta_m^{(\bar{O}_g)} \right| \\ &\quad + \left| \sum_{i=1}^I \hat{\theta}_{n,m}^{(\hat{O}_i)} - \sum_{i=1}^I \theta_m^{(\hat{O}_i)} \right| + \text{SECO}(\bar{O}) - \text{SECO}(\hat{O}) \\ &=: 2D_m + E_1 + E_2 + \text{SECO}(\bar{O}) - \text{SECO}(\hat{O}). \end{aligned}$$

Taking expectancy, we now have

$$\mathbb{E}[\widehat{\text{SECO}}_{n,m}(\bar{O}) - \widehat{\text{SECO}}_{n,m}(\hat{O})] \leq 2D_m + \mathbb{E}[E_1] + \mathbb{E}[E_2] + \text{SECO}(\bar{O}) - \text{SECO}(\hat{O}).$$

Using the same tool involved in the proof of Lemma B.1.1, we can show

$$|\hat{\theta}_{n,m}^{(\bar{O}_g)} - \hat{\theta}_m^{(\bar{O}_g)}| \leq (d_g + 1)^2 |\hat{\nu}_{n,m}^{(\bar{O}_g)} - \hat{\nu}_m^{(\bar{O}_g)}|,$$

Thus, using concentration bounds in Proposition 3.3.1, there exists a universal constant  $K_1 > 0$  independent of  $n, k, m, t$  such that

$$\mathbb{P} \left\{ |\hat{\theta}_{n,m}^{(\bar{O}_g)} - \hat{\theta}_m^{(\bar{O}_g)}| \geq t \right\} \leq d_g \exp \left\{ -\frac{t^2 k}{K_1 d_g^4} \right\}.$$

Now,

$$\begin{aligned} \mathbb{P} \left\{ \left| \sum_{g=1}^G \hat{\theta}_{n,m}^{(\bar{O}_g)} - \sum_{g=1}^G \theta_m^{(\bar{O}_g)} \right| \geq t \right\} &\leq \sum_{g=1}^G \mathbb{P} \left\{ |\hat{\theta}_{n,m}^{(\bar{O}_g)} - \hat{\theta}_m^{(\bar{O}_g)}| \geq \frac{t}{G} \right\} \\ &\leq d \exp \left\{ -\frac{t^2 k}{K_1 G^2 \sqrt{g=1}^G d_g^4} \right\} \end{aligned}$$

Thus, for every  $\delta > 0$ , one obtains

$$\mathbb{E}[E_1]^2 \leq \mathbb{E}[E_1^2] \leq \delta + \int_{\delta}^{\infty} \mathbb{P} \left\{ E_1 > t^{1/2} \right\} dt \leq \delta + d \int_{\delta}^{\infty} \exp \left\{ -\frac{t}{2\sigma^2} \right\} dt,$$

where  $\sigma^2 = \frac{K_1 G^2 \sqrt{g=1}^G d_g^4}{2k}$ . Set  $\delta = 2\sigma^2 \ln(d)$ , we can obtain

$$\mathbb{E}[E_1]^2 \leq \delta + 2\sigma^2 = c^2 \frac{\ln(d) G^2 \sqrt{g=1}^G d_g^4}{k}$$

with  $c > 0$ . Same results hold for  $\mathbb{E}[E_2]$  with corresponding sizes, thus

$$\begin{aligned} \mathbb{E}[\widehat{\text{SECO}}_{n,m}(\bar{O}) - \widehat{\text{SECO}}_{n,m}(\hat{O})] &\leq 2 \left( D_m + c \sqrt{\frac{\ln(d)}{k}} \max(G, I) \max(\sqrt{g=1}^G d_g^2, \sqrt{i=1}^I d_i^2) \right) \\ &\quad + \text{SECO}(\bar{O}) - \text{SECO}(\hat{O}), \end{aligned}$$

which is strictly negative by assumption.  $\square$

### B.1.3 Proofs of Section 3.4

In the following we prove that the model introduced in Section 3.4 is in the domain of attraction of an AI-block model. This comes down from some elementary algebra where the fundamental argument is given by (Bücher and Segers, 2014, Proposition 4.2), from which the inspiration for the model was drawn thereof.

**Proof of Proposition 3.4.1** We aim to show that the following quantity

$$\left| D \left( D^{(O_1)}(\{\mathbf{u}^{(O_1)}\}^{1/m}; \theta, \beta_1), \dots, D(\{\mathbf{u}^{(O_G)}\}^{1/m}; \theta, \beta_G); \theta, \beta_0 \right)^m - D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \beta_1), \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \beta_G); \beta_0 \right) \right|,$$

converges to 0 uniformly in  $\mathbf{u} \in [0, 1]^d$ . Using Equation (3.14) in the main article, the latter term is equal to

$$E_{0,m} := \left| D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \theta/m, \beta_1)^{1/m}, \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \theta/m, \beta_G)^{1/m}; \theta, \beta_0 \right)^m - D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \beta_1), \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \beta_G); \beta_0 \right) \right|.$$

Thus

$$\begin{aligned} E_{0,m} &\leq \left| D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \theta/m, \beta_1)^{1/m}, \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \theta/m, \beta_G)^{1/m}; \theta, \beta_0 \right)^m - D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \theta/m, \beta_1), \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \theta/m, \beta_G); \beta_0 \right) \right| \\ &\quad + \left| D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \theta/m, \beta_1), \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \theta/m, \beta_G); \beta_0 \right) - D \left( D^{(O_1)}(\mathbf{u}^{(O_1)}; \beta_1), \dots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \beta_G); \beta_0 \right) \right| \\ &=: E_{1,m} + E_{2,m}. \end{aligned}$$

As  $D(\cdot; \theta/m, \beta_0)$  converges uniformly to  $D(\cdot, \beta_0)$ , then, uniformly in  $\mathbf{u} \in [0, 1]^d$ ,  $E_{1,m} \xrightarrow{m \rightarrow \infty} 0$ . Now, using Lipschitz property of the copula function, one has

$$E_{2,m} \leq \sum_{g=1}^G \left| D^{(O_g)}(\mathbf{u}^{(O_g)}; \theta/m, \beta_g) - D^{(O_g)}(\mathbf{u}^{(O_g)}; \beta_g) \right|,$$

which converges almost surely to 0 as  $m \rightarrow \infty$ . The limiting copula is an extreme value copula by  $\beta_0 \leq \min\{\beta_1, \dots, \beta_G\}$ , see Example 3.8 of Hofert et al. (2018). Hence the result.  $\square$

## B.2 Additional results

### B.2.1 Additional results of Section 3.2

Let  $\mathbf{Z} \geq \mathbf{0}$  be a random vector, and for simplicity, let's assume that it has heavy-tailed marginal distributions with a common tail-index  $\alpha > 0$ . There are two distinct yet closely related classical approaches for describing the extreme values of the multivariate distribution of  $\mathbf{Z}$ .

The first approach focuses on scale-normalized componentwise maxima:

$$c_n^{-1} \bigvee_{i=1}^n \mathbf{Z}_i,$$

where  $\mathbf{Z}_i$  are independent copies of  $\mathbf{Z}$ , and  $c_n$  is a scaling sequence. The limiting results are typically derived under the assumption of independence for the sake of consistency. However, they hold under more general conditions, such as mixing conditions (see, e.g., [Hsing \(1989\)](#)). The only possible limit laws for such maxima are max-stable distributions with the following distribution function:

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \bigvee_{i=1}^n \mathbf{Z}_i \leq c_n \mathbf{u} \right\} = e^{-\Lambda([0, \mathbf{u}]^c)}, \quad \mathbf{u} \in \mathbb{R}^d + \setminus \mathbf{0},$$

where the exponent measure  $\Lambda$  is  $(-\alpha)$ -homogeneous.

The second approach examines the distribution of scale-normalized exceedances:

$$u^{-1} \mathbf{Z} \mid \bigvee_{j=1}^d Z^{(j)} > u,$$

which considers conditioning on the event that at least one component  $Z^{(j)}$  exceeds a high threshold  $u$ . The only possible limits of these peak-over-thresholds as  $u \rightarrow \infty$  are multivariate Pareto distributions ([Rootzén and Tajvidi \(2006\)](#)). The probability laws of these distributions are induced by a homogeneous measure  $\Lambda$  on the set  $\mathcal{L} = E \setminus [0, 1]^d$ , where  $E = [0, \infty)^d \setminus \mathbf{0}$ . The probability measure takes the form:

$$\mathbb{P}_{\mathcal{L}}(dy) = \frac{\Lambda(dy)}{\Lambda(\mathcal{L})}.$$

The exponent measure serves as a clear connection between these two approaches, as it characterizes the distribution function for both cases. In fact, the connection arises from a fundamental limiting result that establishes a link between the two approaches through regular variation. This result has been elegantly presented in Theorem 2.1.6 and Equation (2.3.1) in [Kulik and Soulier \(2020\)](#). As in the main text, let us denote by  $\mathbf{X}$  the random vector with extreme value distribution  $H(\mathbf{x}) = e^{-\Lambda(E \setminus [0, \mathbf{x}])}$ . The following proposition provides the form of the exponent measure when the random vectors  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  are independent, and it establishes the connection between AI-block models for the two approaches.

**Proposition B.2.1.** *Suppose  $\mathbf{X}$  is a random vector having extreme value distribution  $H$  with exponent measure  $\Lambda$  concentrating on  $E \setminus [0, \mathbf{x}]$  where  $E = [0, \infty)^d \setminus \{\mathbf{0}\}$  and  $\mathbf{x} > \mathbf{0}$ . The following properties are equivalent:*

- (i) The vectors  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  are independent.  
(ii) The vectors are blockwise independent: for every  $1 \leq g < h \leq G$

$\mathbf{X}^{(O_g)}$  and  $\mathbf{X}^{(O_h)}$ , are independent random vectors.

- (iii) The exponent measure  $\Lambda$  concentrates on

$$\bigcup_{g=1}^G \{\mathbf{0}\}^{d_1} \times \dots \times ]0, \infty[^{d_g} \times \dots \times \{\mathbf{0}\}^{d_G}, \quad (\text{B.4})$$

so that for  $\mathbf{x} > \mathbf{0}$ ,

$$\Lambda \left( \bigcup_{1 \leq g < h \leq G} \left\{ \mathbf{y} \in E, \exists a \in O_g, \exists b \in O_h, y^{(a)} > x^{(a)}, y^{(b)} > x^{(b)} \right\} \right) = 0.$$

These conditions generalize straightforwardly those stated in Proposition 5.24 of Resnick (2008) (see Exercise 5.5.1 of the book aforementioned or the Lemma in Strokorb (2020)).

**Proof of Proposition B.2.1** We will establish the result proceeding as (iii)  $\implies$  (i)  $\implies$  (ii)  $\implies$  (iii) where we directly have (i)  $\implies$  (ii). Now for (iii)  $\implies$  (i), suppose  $\Lambda$  concentrates on the set (B.4). Then for  $\mathbf{x} > \mathbf{0}$ , noting  $A_g(\mathbf{x}) = \{\mathbf{y} \in E, \exists a \in O_g, y^{(a)} > x^{(a)}\}$  for  $g \in \{1, \dots, G\}$ , we obtain

$$\begin{aligned} -\ln H(\mathbf{x}) &= \Lambda(E \setminus [0, \mathbf{x}]) = \Lambda \left( \bigcup_{g=1}^G A_g(\mathbf{x}) \right) \\ &= \sum_{g=1}^G \Lambda(A_g(\mathbf{x})) + \sum_{g=2}^G (-1)^{g+1} \sum_{1 \leq i_1 < i_2 < \dots < i_g \leq G} \Lambda(A_{i_1}(\mathbf{x}) \cap \dots \cap A_{i_g}(\mathbf{x})), \end{aligned}$$

so that because of Equation (B.4),

$$-\ln H(\mathbf{x}) = \sum_{g=1}^G \Lambda(A_g(\mathbf{x})),$$

and we have  $H(\mathbf{x}) = \prod_{g=1}^G \exp \left\{ -\Lambda \left( \{\mathbf{y} \in E, \exists a \in O_g, y^{(a)} > x^{(a)}\} \right) \right\} = \prod_{g=1}^G H^{(O_g)}(\mathbf{x}^{(O_g)})$ .

Thus  $H$  is written as a product of the  $G$  distributions corresponding to random vectors  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$ , as desired.

It remains to show (ii)  $\implies$  (iii). Set  $Q^{(O_g)}(\mathbf{x}^{(O_g)}) = -\ln \mathbb{P}\{\mathbf{X}^{(O_g)} \leq \mathbf{x}^{(O_g)}\}$  for  $g \in \{1, \dots, G\}$ . We have for  $\mathbf{x} > \mathbf{0}$  that blockwise independence implies, with  $g \neq h$ ,

$$Q^{(O_g)}(\mathbf{x}^{(O_g)}) + Q^{(O_h)}(\mathbf{x}^{(O_h)}) = -\ln \mathbb{P}\{\mathbf{X}^{(O_g)} \leq \mathbf{x}^{(O_g)}, \mathbf{X}^{(O_h)} \leq \mathbf{x}^{(O_h)}\}.$$

Since  $H(\mathbf{x}) = \exp\{-\Lambda(E \setminus [0, \mathbf{x}])\}$  for  $\mathbf{x} > \mathbf{0}$ , we have

$$\begin{aligned}
 Q^{(O_g)}(\mathbf{x}^{(O_g)}) + Q^{(O_h)}(\mathbf{x}^{(O_h)}) &= \Lambda(\{\mathbf{y}, \exists a \in O_g, y^{(a)} > x^{(a)}\} \cup \{\mathbf{y}, \exists b \in O_h, y^{(b)} > x^{(b)}\}) \\
 &= \Lambda(\{\mathbf{y}, \exists a \in O_g, y^{(a)} > x^{(a)}\}) + \Lambda(\{\mathbf{x}, \exists b \in O_h, y^{(b)} > x^{(b)}\}) \\
 &\quad - \Lambda(\{\mathbf{y}, \exists a \in O_g, \exists b \in O_h, y^{(a)} > x^{(a)}, y^{(b)} > x^{(b)}\}) \\
 &= Q^{(O_g)}(\mathbf{x}^{(O_g)}) + Q^{(O_h)}(\mathbf{x}^{(O_h)}) \\
 &\quad - \Lambda(\{\mathbf{y}, \exists a \in O_g, \exists b \in O_h, y^{(a)} > x^{(a)}, y^{(b)} > x^{(b)}\}),
 \end{aligned}$$

and thus

$$\Lambda(\{\mathbf{y}, \exists a \in O_g, \exists b \in O_h, y^{(a)} > x^{(a)}, y^{(b)} > x^{(b)}\}) = 0,$$

so that (iii) holds. This is equivalent to  $\Lambda$  concentrates on the set in Equation (B.4).  $\square$

If  $\mathbf{X}$  is a random vector with multivariate extreme value distribution  $H$  then its extreme value copula, denoted as,  $C_\infty$  is written as:

$$C_\infty(\mathbf{u}) = \exp \left\{ -L \left( -\ln(u^{(1)}), \dots, -\ln(u^{(d)}) \right) \right\},$$

where  $L$  is the stable tail dependence function. This function captures the tail dependence structure of the random vector and can be expressed as a specific integral with respect to the exponent measure (we refer to Section 8 of [Beirlant et al. \(2004\)](#)). In the context of AI-block models, the tail dependence function takes the following form:

$$L(z^{(1)}, \dots, z^{(d)}) = \sum_{g=1}^G L^{(O_g)}(\mathbf{z}^{(O_g)}), \quad \mathbf{z} \in [0, \infty)^d, \quad (\text{B.5})$$

where  $L^{(O_1)}, \dots, L^{(O_G)}$  are the corresponding stable tail dependence functions with copulae  $C_\infty^{(O_1)}, \dots, C_\infty^{(O_G)}$ , respectively. This model is a specific form of the nested extreme value copula, as mentioned in the remark below and discussed in further detail in [Hofert et al. \(2018\)](#).

**Remark B.2.1.** Equation (B.5) can be rewritten as

$$L(\mathbf{z}) = L_\Pi \left( L^{(O_1)}(z^{(O_1)}), \dots, L^{(O_G)}(z^{(O_G)}) \right),$$

where  $L_\Pi(z^{(1)}, \dots, z^{(G)}) = \sum_{g=1}^G z^{(g)}$  is a stable tail dependence function corresponding to asymptotic independence. According to Proposition 3.2.1,  $C_\infty$  is an extreme value copula. Therefore, it follows that  $C_\infty$ , which has the representation

$$C_\infty(\mathbf{u}) = C_\Pi \left( C_\infty^{(O_1)}(\mathbf{u}^{(O_1)}), \dots, C_\infty^{(O_G)}(\mathbf{u}^{(O_G)}) \right), \quad C_\Pi = \prod_{g=1}^G u^{(g)},$$

is also a nested extreme value copula, as defined in [Hofert et al. \(2018\)](#).

Equation (B.5) can be restricted to the simplex, allowing us to express the stable tail dependence function in terms of the Pickands dependence function. Specifically, the Pickands dependence function  $A$  can be written as a convex combination of the Pickands dependence functions



$\mathfrak{A}^{(O_1)}, \dots, \mathfrak{A}^{(O_G)}$  as follows:

$$\begin{aligned} \mathfrak{A}(t^{(1)}, \dots, t^{(d)}) &= \frac{1}{z^{(1)} + \dots + z^{(d)}} \left[ \sum_{g=1}^G (z^{(i_{g,1})} + \dots + z^{(i_{g,d_g})}) \mathfrak{A}^{(O_g)}(\mathbf{t}^{(O_g)}) \right] \\ &= \sum_{g=1}^G w^{(O_g)}(\mathbf{t}) \mathfrak{A}^{(O_g)}(\mathbf{t}^{(O_g)}) =: \mathfrak{A}^{(O)}(t^{(1)}, \dots, t^{(d)}), \end{aligned} \quad (\text{B.6})$$

with  $t^{(j)} = z^{(j)} / (z^{(1)} + \dots + z^{(d)})$  for  $j \in \{2, \dots, d\}$  and  $t^{(1)} = 1 - (t^{(2)} + \dots + t^{(d)})$ ,  $w^{(O_g)}(\mathbf{t}) = (z^{(i_{g,1})} + \dots + z^{(i_{g,d_g})}) / (z^{(1)} + \dots + z^{(d)})$  for  $g \in \{2, \dots, G\}$  and  $w^{(O_1)}(\mathbf{t}) = 1 - (w^{(O_2)}(\mathbf{t}) + \dots + w^{(O_G)}(\mathbf{t}))$ ,  $\mathbf{t}^{(O_g)} = (t^{(i_{g,1})}, \dots, t^{(i_{g,d_g})})$  where  $t^{(i_{g,\ell})} = z^{(i_{g,\ell})} / (z^{(i_{g,1})} + \dots + z^{(i_{g,d_g})})$  and  $(i_{g,\ell})$  designates the  $\ell$ th variable in the  $g$ th cluster for  $\ell \in \{1, \dots, d_g\}$  and  $g \in \{1, \dots, G\}$ . As a convex combination of Pickands dependence functions,  $\mathfrak{A}$  is itself a Pickands dependence function (see (Falk et al., 2010, Page 123)).

In the context of independence between extreme random variables, it is well-known that the inequality  $\mathfrak{A}(\mathbf{t}) \leq 1$  holds for  $\mathbf{t} \in \Delta_{d-1}$ , where  $\mathfrak{A}$  is the Pickands dependence function and equality stands if and only if the random variables are independent. This result extends to the case of random vectors, with the former case being a special case where  $d_1 = \dots = d_G = 1$ .

**Proposition B.2.2.** *Consider a random vector  $\mathbf{X} \in \mathbb{R}^d$  with copula  $C_\infty$  and Pickands dependence function  $\mathfrak{A}$ . Let  $\mathfrak{A}^{(O)}$  be as defined in (B.6). For all  $\mathbf{t} \in \Delta_{d-1}$ , we have:*

$$\left( \mathfrak{A}^{(O)} - \mathfrak{A} \right) (\mathbf{t}) \geq 0,$$

with equality if and only if  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  are independent.

We provide two methods for establishing this result: the first leverages the convexity and homogeneity of order one of the stable tail dependence function, while the second takes advantage of the associativity of random vectors having extreme value distribution  $H$ .

**Proof of Proposition B.2.2** For the first method, the stable tail dependence function  $L$  is subadditive as an homogeneous convex function under a cone, i.e.,

$$L(\mathbf{x} + \mathbf{y}) \leq L(\mathbf{x}) + L(\mathbf{y}),$$

for every  $\mathbf{x}, \mathbf{y} \in [0, \infty)^d$ . In particular, we obtain by induction on  $G$

$$L \left( \sum_{g=1}^G \mathbf{x}^{(g)} \right) \leq \sum_{g=1}^G L(\mathbf{x}^{(g)}),$$

where  $\mathbf{x}^{(g)} \in [0, \infty)^d$  and  $g \in \{1, \dots, G\}$ . Consider now  $\mathbf{z}^{(O_g)} = (\mathbf{0}, z^{(i_{g,1})}, \dots, z^{(i_{g,d_g})}, \mathbf{0})$ , we directly obtain using the equation above

$$L(\mathbf{z}) = L \left( \sum_{g=1}^G \mathbf{z}^{(O_g)} \right) \leq \sum_{g=1}^G L(\mathbf{z}^{(O_g)}) = \sum_{g=1}^G L^{(O_g)}(z^{(i_{g,1})}, \dots, z^{(i_{g,d_g})}).$$

Translating the above inequality in terms of Pickands dependence function results on

$$\begin{aligned} \mathcal{A}(\mathbf{t}) &\leq \sum_{g=1}^G \frac{1}{z^{(1)} + \dots + z^{(d)}} L^{(O_g)}(z^{(i_{g,1})}, \dots, z^{(i_{g,d_g})}) \\ &= \sum_{g=1}^G \frac{z^{(i_{g,1})} + \dots + z^{(i_{g,d_g})}}{z^{(1)} + \dots + z^{(d)}} \mathcal{A}^{(O_g)}(t^{(i_{g,1})}, \dots, t^{(i_{g,d_g})}), \end{aligned}$$

where  $t^{(i)} = z^{(i)}/(z^{(1)} + \dots + z^{(d)})$ . Hence the result.

We can also prove this result by using the associativity of extreme-value distributions (see (Marshall and Olkin, 1983, Proposition 5.1) or (Resnick, 2008, Section 5.4.1)), i.e.,

$$\mathbb{E}[f(\mathbf{X})g(\mathbf{X})] \geq \mathbb{E}[f(\mathbf{X})] \mathbb{E}[g(\mathbf{X})],$$

for every increasing (or decreasing) functions  $f, g$ . By induction on  $G \in \mathbb{N}_*$ ,

$$\mathbb{E}\left[\prod_{g=1}^G f^{(g)}(\mathbf{X})\right] \geq \prod_{g=1}^G \mathbb{E}\left[f^{(g)}(\mathbf{X})\right]. \quad (\text{B.7})$$

Take  $f^{(g)}(\mathbf{x}) = \mathbb{1}_{\{\lfloor -\infty, \mathbf{x}^{(O_g)}\rfloor\}}$  for each  $g \in \{1, \dots, G\}$ , thus Equation (B.7) gives

$$C(H^{(1)}(x^{(1)}), \dots, H^{(d)}(x^{(d)})) \geq \prod_{g=1}^G C^{(O_g)}\left(H^{(O_g)}\left(\mathbf{x}^{(O_g)}\right)\right),$$

which can be restated in terms of stable tail dependence function as

$$L(\mathbf{z}) \leq \sum_{g=1}^G L^{(O_g)}(\mathbf{z}^{(O_g)}).$$

We obtain the statement expressing this inequality with Pickands dependence function. Finally, notice that (B.7) with  $f^{(g)}(\mathbf{x}) = \mathbb{1}_{\{\lfloor -\infty, \mathbf{x}^{(O_g)}\rfloor\}}$  for each  $g \in \{1, \dots, G\}$  holds as an equality if and only if  $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$  are independent random vectors.  $\square$

In the following paragraph, we give another proof of the extension of the results found in Takahashi (1987, 1994) made by (Ferreira, 2011, Proposition 2.1). Before going into details, we recall some useful expression of the dependence structure of extreme closely related to the notion of regular variation.

Let  $\mathbf{X}$  be a regularly varying random vector in  $\mathbb{R}_+^d$  with exponent measure  $\Lambda$  which is  $(-\alpha)$ -homogeneous, i.e. for  $y > 0$  and  $A$  separated from  $\mathbf{0}$ , that is there exists an open set  $U$  such that  $\mathbf{0} \in U$  and  $U^c \subset A$ , we have

$$\Lambda(yA) = y^{-\alpha} \Lambda(A).$$

Using the homogeneity of the exponent measure, we may define a probability measure  $\Phi$  on  $\Theta = S_d \cap [0, \infty)$  where  $S_d = \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| = 1\}$  called the spectral measure associated to the norm  $\|\cdot\|$  and defined by

$$\Phi(B) = \Lambda\left(\mathbf{z} \in E : \|\mathbf{z}\| > 1, \mathbf{z}\|\mathbf{z}\|^{-1} \in B\right)$$

## Proofs of Chapter 3

---

for any Borel subset  $B$  of  $\Theta$  (for a proper introduction to these notions, see (Resnick, 2008, Section 5.1) or (Kulik and Soulier, 2020, Section 2.2)). The measure  $\Phi$  is called the spectral measure. It is uniquely determined by the exponent measure  $\Lambda$  and the chosen norm. The homogeneity of  $\Lambda$  implies :

$$\Lambda\left(\mathbf{z} \in E : \|\mathbf{z}\| > r, \mathbf{z}\|\mathbf{z}\|^{-1} \in B\right) = r^{-1}\Phi(B),$$

for  $0 < r < \infty$ .

**Proposition B.2.3.** *Let  $\mathbf{X}$  be a regularly varying random vector in  $\mathbb{R}_+^d$  with exponent measure  $\Lambda$ . Consider  $O = \{O_1, \dots, O_g\}$  be a partition of  $\{1, \dots, d\}$ , then the following are equivalent:*

(i) Let  $\Lambda^{(O_g)}$  be the restriction of the exponent measure to  $\mathbb{R}_+^{(O_g)}$ , we have

$$\Lambda = \sum_{g=1}^G \delta_0 \otimes \dots \otimes \Lambda^{(O_g)} \otimes \dots \otimes \delta_0.$$

(ii) The spectral measure  $\Phi$  associated to the exponent measure  $\Lambda$  verifies

$$\Phi = \sum_{g=1}^G \delta_0 \otimes \dots \otimes \Phi^{(O_g)} \otimes \dots \otimes \delta_0 =: \Phi_{\Pi}, \quad (\text{B.8})$$

where  $\Phi^{(O_g)}(B) := \Phi(\Theta^{(O_g)} \cap B)$  where  $B$  is a borel set of  $\Theta$  and

$$\Theta^{(O_g)} = \left\{ \mathbf{w} \in \Theta, w^{(j)} > 0 \text{ if and only if } j \in O_g \right\}$$

for  $g \in \{1, \dots, G\}$ .

(iii) There exists a  $\mathbf{v} \in (0, \infty)^d$  such that

$$\int_{\Theta} \bigvee_{j=1}^d w^{(j)} v^{(j)} \Phi(d\mathbf{w}) = \sum_{g=1}^G \int_{\Theta^{(O_g)}} \bigvee_{j \in O_g} w^{(j)} v^{(j)} \Phi^{(O_g)}(d\mathbf{w}^{(O_g)}). \quad (\text{B.9})$$

**Proof of Proposition B.2.3** The equivalence between (i) and (ii) falls down from definitions. The implication (ii)  $\implies$  (iii) is trivial. We show now (iii)  $\implies$  (ii) Notice that for every Borel set  $B$  of  $\Theta$ , we have

$$\Phi(B) = \sum_{g=1}^G \Phi(B \cap \Theta^{(O_g)}) + \Phi\left(B \cap (\Theta \setminus \cup_{g=1}^G \Theta^{(O_g)})\right) \geq \sum_{g=1}^G \Phi(B \cap \Theta^{(O_g)}) = \Phi_{\Pi}(B).$$

The identity in Equation (B.9) can be rewritten as

$$\int_{\Theta} \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_{\Pi})(d\mathbf{w}) = 0.$$

From above, we know that  $(\Phi - \Phi_\Pi)$  defined a positive measure. For every Borel set  $B$  of  $\Theta$ , we have

$$\int_B \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) \leq \int_\Theta \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) = 0.$$

Since the function  $\mathbf{w} \mapsto \bigvee_{j=1}^d w^{(j)} v^{(j)}$  is strictly positive, continuous and defined on a compact set, we have that  $\bigvee_{j=1}^d w^{(j)} v^{(j)} \geq c$  for a certain constant  $c$  strictly positive and we obtain

$$c(\Phi - \Phi_\Pi)(B) \leq \int_B \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) = 0.$$

The following identity is obtained

$$\Phi(B) = \Phi_\Pi(B),$$

since  $B$  is taken arbitrary from the Borelian of  $\Theta$ , we conclude.  $\square$

One can notice that the integrals defined in (B.9) can be rewritten with the help of stable tail dependence function, that is

$$L(v^{(1)}, \dots, v^{(d)}) = \sum_{g=1}^G L^{(O_g)}(\mathbf{v}^{(O_g)}), \quad \mathbf{v} \in [0, \infty)^d,$$

since for every  $\mathbf{v} \in [0, \infty)^d$

$$L(\mathbf{v}) = \int_\Theta \bigvee_{j=1}^d w^{(j)} v^{(j)} \Phi(d\mathbf{w}).$$

### B.2.2 Additional results of Section 3.3

To establish the strong consistency of the estimator  $\hat{\nu}_{n,m}$  in (3.7), certain conditions on the mixing coefficients must be satisfied.

**Condition C.** Let  $m_n = o(n)$ . The series  $\sum_{n \geq 1} \beta(m_n)$  is convergent, where  $\beta$  is defined in Section 1.1.4.

For the sake of notational simplicity, we will write  $m = m_n$ ,  $k = k_n$ . The convergence of the series of  $\beta$ -mixing coefficients in Condition C is necessary to obtain the strong consistency of  $\hat{\nu}_{n,m}$ , and it can be achieved through the sufficiency condition of the Glivenko-Cantelli lemma for almost sure convergence.

**Proposition B.2.4.** Let  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a stationary multivariate random process. Under Conditions A and C, the madogram estimator in (3.7) is strongly consistent, i.e.,

$$|\hat{\nu}_{n,m} - \nu| \xrightarrow[n \rightarrow \infty]{a.s.} 0,$$

with  $\nu$  the theoretical madogram of the random vector  $\mathbf{X}$  with copula  $C_\infty$  given in (3.5).

Let  $C_{n,m}^o$  be the empirical estimator of the copula  $C_m$  based on the (unobservable) sample  $(U_{m,1}^{(j)}, \dots, U_{m,k}^{(j)})$  for  $j \in \{1, \dots, d\}$ . The proof of Proposition B.2.4 will use twice Lemma B.3.1,

which shows that  $\|C_{n,m}^o - C\|_\infty$  converges almost surely to 0. The proof of this lemma is postponed to B.3.1 of supplementary results.

**Proof of Proposition B.2.4** We aim to show the following convergence

$$|\hat{\nu}_{n,m} - \nu| \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 0.$$

Following Lemma A.1 of Marcon et al. (2017), we can show that

$$\hat{\nu}_{n,m}^o - \nu = \phi(C_{n,m}^o - C_\infty),$$

where  $\hat{\nu}_{n,m}^o$  given in (B.1) and  $\phi : \ell^\infty([0, 1]^d) \rightarrow \ell^\infty(\Delta_{d-1})$ ,  $f \mapsto \phi(f)$  defined by

$$\phi(f) = \frac{1}{d} \sum_{j=1}^d \int_{[0,1]} f(1, \dots, 1, \underbrace{u}_{j\text{-th component}}, 1, \dots, 1) du - \int_{[0,1]} f(u, \dots, u) du.$$

Using Conditions  $\mathcal{A}$  and  $\mathcal{C}$ , by Lemma B.3.1 in B.3.1, as  $\|C_{n,m}^o - C_\infty\|_\infty$  converges almost surely to 0, we obtain that

$$|\hat{\nu}_{n,m}^o - \nu| \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 0. \quad (\text{B.10})$$

Furthermore, using the chain of inequalities and again Lemma B.3.1 in B.3.1,

$$\begin{aligned} |\hat{\nu}_{n,m} - \hat{\nu}_{n,m}^o| &\leq 2 \sup_{j \in \{1, \dots, d\}} \sup_{x \in \mathbb{R}} |\hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x)| \\ &\leq 2 \sup_{j \in \{1, \dots, d\}} \sup_{u \in [0,1]} \left| \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{\{U_{m,i}^{(j)} \leq u\}} - u \right|. \end{aligned}$$

Then we obtain that

$$|\hat{\nu}_{n,m} - \hat{\nu}_{n,m}^o| \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 0. \quad (\text{B.11})$$

Now, write

$$|\hat{\nu}_{n,m} - \nu| \leq |\hat{\nu}_{n,m} - \hat{\nu}_{n,m}^o| + |\hat{\nu}_{n,m}^o - \nu|,$$

and use Equations (B.10) and (B.11) to obtain the statement.  $\square$

The strong consistency of the madogram in Proposition B.2.4 could be extended to the  $\alpha$ -mixing case. We present here the strong consistency of our procedure when the dimension  $d$  is fixed the sample size  $n$  grows at infinity. The main technicality of the proof has already been tackled in Proposition B.2.4 and we state the precise formulation of this theorem below.

**Theorem B.2.1.** *Consider the AI-block model as defined in Definition 3.2.1 under Condition  $\mathcal{B}$  and  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a stationary multivariate random process. For a given  $\mathcal{X}$  and its corresponding estimator  $\hat{\mathcal{X}}$ , if Conditions  $\mathcal{A}$ ,  $\mathcal{C}$  holds, then taking  $\tau = 0$*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \hat{\mathcal{O}} = \bar{\mathcal{O}} \right\} = 1.$$

**Proof of Theorem B.2.1** If  $a$  and  $b$  are not in the same cluster according to  $\bar{\mathcal{O}}$ , i.e.  $a \not\sim b$ , then  $\chi(a, b) = 0$ . Therefore, using Proposition B.2.4 along with Conditions  $\mathcal{A}$  and  $\mathcal{C}$ , we can

conclude that almost surely

$$\lim_{n \rightarrow \infty} \hat{\chi}_{n,m}(a, b) = 0 \leq \tau.$$

Now, if  $a \stackrel{\bar{O}}{\sim} b$ , then  $\chi(a, b) > 0$  and again by Propositions B.2.4 and Conditions  $\mathcal{A}$ ,  $\mathcal{C}$ , we obtain

$$\lim_{n \rightarrow \infty} \hat{\chi}_{n,m}(a, b) = \chi(a, b) > 0,$$

where the strict positiveness is obtain through Condition  $\mathcal{B}$ , hence

$$a \stackrel{\bar{O}}{\sim} b \iff \lim_{n \rightarrow \infty} \hat{\chi}_{n,m}(a, b) > \tau.$$

Let us prove Theorem B.2.1 by induction on the algorithm step  $l$ . We consider the algorithm at some step  $l - 1$  and assume that the algorithm was consistent up to this step, i.e.  $\hat{O}_j = \bar{O}_j$  for  $j = 1, \dots, l - 1$ .

If  $\lim_{n \rightarrow \infty} \hat{\chi}_{n,m}(a_l, b_l) = 0$ , then no  $b \in S$  is in the same group of  $a_l$ . Since the algorithm has been consistent up to this step  $l$ , it means that  $a_l$  is a singleton and  $\hat{O}_l = \{a_l\}$ .

If  $\lim_{n \rightarrow \infty} \hat{\chi}_{n,m}(a_l, b_l) > \tau$ , then  $a_l \stackrel{\bar{O}}{\sim} b$ . The equivalence above implies that  $\hat{O}_l = S \cap \bar{O}_l$ . Since the algorithm has been consistent up until this step, we know that  $\hat{O}_l = \bar{O}_l$ . Therefore, the algorithm remains consistent at step  $l$  with probability tending to one as  $n \rightarrow \infty$ , and Theorem B.2.1 follows by induction.  $\square$

## B.3 Further results

### B.3.1 A usefull Glivenko-Cantelli result for the copula with known margins in a weakly dependent setting

In this section, we will prove an important auxiliary result: the empirical copula estimator  $\hat{C}_{n,m}^o$  based on the weakly dependent sample  $\mathbf{U}_{m,1}, \dots, \mathbf{U}_{m,k}$  is uniformly strongly consistent towards the extreme value copula  $C$ . This result is a main tool to obtain important results in the paper such as Proposition B.2.4, Theorem B.2.1. For that purpose, the Berbee's coupling lemma is of prime interest (see, e.g., (Rio, 2017, Chapter 5)) which gives an approximation of the original process by conveniently defined independent random variables.

**Lemma B.3.1.** *Under conditions of Proposition B.2.4, we have*

$$\|C_{n,m}^o - C\|_\infty \xrightarrow[n \rightarrow \infty]{a.s.} 0.$$

**Lemma B.3.1** Using triangle inequality, one obtain the following bound

$$\|C_{n,m}^o - C\|_\infty \leq \|C_{n,m}^o - C_m\|_\infty + \|C_m - C\|_\infty. \quad (\text{B.12})$$

As  $\{C_m, n \in \mathbb{N}\}$  is an equicontinuous class of functions (for every  $m$ ,  $C_m$  is a copula hence a 1-Lipschitz function), defined on the compact set  $[0, 1]^d$  (by Tychonov's theorem) which converges pointwise to  $C$  by Condition  $\mathcal{A}$ . Then the convergence is uniform over  $[0, 1]^d$ . Thus the second term of the RHS of Equation (B.12) converges to 0 almost surely.

Now, let us prove that  $\|C_{n,m}^o - C_m\|_\infty$  converges almost surely to 0. By Berbee's coupling lemma (see (Rio, 2017, Theorem 6.1) or (Bücher and Segers, 2014, Theorem 3.1) for similar applications), we can construct inductively a sequence  $(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m})_{i \geq 0}$  such that the following three properties hold:

- (i)  $(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m}) \stackrel{d}{=} (\mathbf{Z}_{im+1}, \dots, \mathbf{Z}_{im+m})$  for any  $i \geq 0$ ;
- (ii) both  $(\bar{\mathbf{Z}}_{2im+1}, \dots, \bar{\mathbf{Z}}_{2im+m})_{i \geq 0}$  and  $(\bar{\mathbf{Z}}_{(2i+1)m+1}, \dots, \bar{\mathbf{Z}}_{(2i+1)m+m})_{i \geq 0}$  sequences are independent and identically distributed;
- (iii)  $\mathbb{P}\{(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m}) \neq (\mathbf{Z}_{im+1}, \dots, \mathbf{Z}_{im+m})\} \leq \beta(m)$ .

Let  $\bar{C}_{n,m}^o$  and  $\bar{\mathbf{U}}_{m,i}$  be defined analogously to  $C_{n,m}^o$  and  $\mathbf{U}_{m,i}$  respectively but with  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$  replaced with  $\bar{\mathbf{Z}}_1, \dots, \bar{\mathbf{Z}}_n$ . Now write

$$C_{n,m}^o(\mathbf{u}) = \bar{C}_{n,m}^o(\mathbf{u}) + \left\{ C_{n,m}^o(\mathbf{u}) - \bar{C}_{n,m}^o(\mathbf{u}) \right\}. \quad (\text{B.13})$$

We will show below that the term under brackets converges uniformly to 0 almost surely. Write  $C_{n,m}^o(\mathbf{u}) = \bar{C}_{n,m}^{o,\text{odd}}(\mathbf{u}) + \bar{C}_{n,m}^{o,\text{even}}(\mathbf{u})$  where  $\bar{C}_{n,m}^{o,\text{odd}}(\mathbf{u})$  and  $\bar{C}_{n,m}^{o,\text{even}}(\mathbf{u})$  are defined as sums over the odd and even summands of  $\bar{C}_{n,m}^o(\mathbf{u})$ , respectively. Since both of these sums are based on i.i.d. summands by properties (i) and (ii), we have  $\|C_{n,m}^o - C_m\|_\infty \xrightarrow[n \rightarrow \infty]{\text{a.s.}} 0$  using Glivenko-Cantelli (see (van der Vaart and Wellner, 1996, Chapter 2.5)).

It remains to control the term under brackets on the right hand side of Equation (B.13), we have that

$$\begin{aligned} \left| C_{n,m}^o(\mathbf{u}) - \bar{C}_{n,m}^o(\mathbf{u}) \right| &\leq \frac{1}{k} \sum_{i=1}^k \left| \mathbb{1}_{\{\bar{\mathbf{U}}_{m,i} \leq \mathbf{u}\}} - \mathbb{1}_{\{\mathbf{U}_{m,i} \leq \mathbf{u}\}} \right| \\ &\leq \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{\{(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m}) \neq (\mathbf{Z}_{im+1}, \dots, \mathbf{Z}_{im+m})\}}. \end{aligned}$$

Hence, using Markov's inequality and property (iii), we have

$$\mathbb{P} \left\{ \sup_{\mathbf{u} \in [0,1]^d} \left| \bar{C}_{n,m}^o(\mathbf{u}) - C_{n,m}^o(\mathbf{u}) \right| > \epsilon \right\} \leq \frac{\beta(m)}{\epsilon}.$$

Thus by Condition  $\mathcal{C}$ ,

$$\sum_{n \geq 1} \mathbb{P} \left\{ \sup_{\mathbf{u} \in [0,1]^d} \left| \bar{C}_{n,m}^o(\mathbf{u}) - C_{n,m}^o(\mathbf{u}) \right| > \epsilon \right\} < \infty.$$

Applying Borel-Cantelli gives the desired convergence to 0 almost surely of the term under bracket in Equation (B.13). Gathering all results gives that the term  $\|C_{n,m}^o - C_m\|_\infty$  converges almost surely to 0. Hence the statement using Equation (B.12).  $\square$

### B.3.2 Weak convergence of an estimator of $\mathcal{A}^{(O)} - \mathcal{A}$

We now state conditions on the block size  $m$  and the number of blocks  $k$ , as in Bücher and Segers (2014), to demonstrate the weak convergence of the empirical copula process based on

the (unobservable) sample  $(U_{n,m,1}^{(j)}, \dots, U_{n,m,k}^{(j)})$  for every  $j \in \{1, \dots, d\}$  under mixing conditions. An additional condition will be required within the theorem to establish the weak convergence of the rank-based copula estimator under the same mixing conditions.

**Condition  $\mathcal{F}$ .** There exists a positive integer sequence  $\ell_n$  such that the following statement holds:

- (i)  $m_n \rightarrow \infty$  and  $m_n = o(n)$
- (ii)  $\ell_n \rightarrow \infty$  and  $\ell_n = o(m_n)$
- (iii)  $k_n \alpha(\ell_n) = o(1)$  and  $(m_n/\ell_n)\alpha(\ell_n) = o(1)$
- (iv)  $\sqrt{k_n}\beta(m_n) = o(1)$

We recall that both  $m$  and  $k$  depends on  $n$ . Also, for notational convenience, we will write in the following  $\ell_n = \ell$ . Note that Condition  $\mathcal{F}$  (iii) guarantees that the limit  $C$  is an extreme value copula by (Hsing, 1989, Theorem 4.2). As usual, the weak convergence of the empirical copula process stems down from the finite dimensional convergence and the asymptotic tightness of the process which then hold from Condition  $\mathcal{F}$  (iii) and (iv) respectively. In order to apply Hadamard's differentiability to obtain the weak convergence of the empirical copula based on the sample's scaled ranks, we need a classical condition over the derivatives of the limit copula stated as follows.

**Condition  $\mathcal{G}$ .** For any  $j \in \{1, \dots, d\}$ , the  $j$ th first order partial derivative  $\dot{C}^{(j)} = \partial C / \partial u^{(j)}$  exists and is continuous on  $\{\mathbf{u} \in [0, 1]^d, u^{(j)} \in (0, 1)\}$ .

The estimator of the Pickands dependence function that we present is based on the madogram concept (Cooley et al. (2006); Marcon et al. (2017)), a notion borrowed from geostatistics in order to capture the spatial dependence structure. Our estimator is defined as

$$\hat{\mathcal{A}}_{n,m}(\mathbf{t}) = \frac{\hat{\nu}_{n,m}(\mathbf{t}) + c(\mathbf{t})}{1 - \hat{\nu}_{n,m}(\mathbf{t}) - c(\mathbf{t})},$$

where

$$\hat{\nu}_{n,m}(\mathbf{t}) = \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d \left\{ \hat{U}_{n,m,j}^{(j)} \right\}^{1/t^{(j)}} - \frac{1}{d} \sum_{j=1}^d \left\{ \hat{U}_{n,m,i}^{(j)} \right\}^{1/t^{(j)}} \right], \quad c(\mathbf{t}) = \frac{1}{d} \sum_{j=1}^d \frac{t^{(j)}}{1 + t^{(j)}},$$

and  $\hat{U}_{n,m,i}^{(j)} = \hat{F}_{n,m}^{(j)}(M_{m,i}^{(j)})$  corresponds to ranks scaled by  $k^{-1}$ . By convention, here  $u^{1/0} = 0$  for  $u \in (0, 1)$ . Let  $g \in \{1, \dots, G\}$  and define

$$\hat{\mathcal{A}}_{n,m}^{(O_g)}(\mathbf{t}^{(O_g)}) = \hat{\mathcal{A}}_{n,m}(\mathbf{0}, \mathbf{t}^{(O_g)}, \mathbf{0})$$

the empirical Pickands dependence function associated to the  $k$ -th subvector of  $\mathbf{X}_p$ . We consider the empirical process of the difference between estimates of the Pickands dependence functions of subvectors  $\mathbf{X}^{(O_g)}, g \in \{1, \dots, G\}$ , and the estimator of the Pickands dependence function of  $\mathbf{X}$ :

$$\mathcal{E}_{nG}(\mathbf{t}) = \sqrt{k} \left( \hat{\mathcal{A}}_{n,m}^{(O)}(\mathbf{t}) - \hat{\mathcal{A}}_{n,m}(\mathbf{t}) \right),$$

where  $\hat{\mathcal{A}}_{n,m}^{(O)}(\mathbf{t}) = \sum_{g=1}^G w^{(O_g)}(\mathbf{t}) \hat{\mathcal{A}}_{n,m}^{(O_g)}(\mathbf{t}^{(O_g)})$ . Noticing that multiplying the above process by  $d$  and taking  $\mathbf{t} = (d^{-1}, \dots, d^{-1})$  gives



$$\sqrt{k}\widehat{SECO}(O) = \sqrt{k} \left( \sum_{g=1}^G \hat{\theta}_{n,m}^{(O_g)} - \hat{\theta}_{n,m} \right).$$

Hence, the weak convergence of the above empirical process will immediately comes down from the one of the empirical process in  $\mathcal{E}_{nG}$ , as stated in the theorem below.

**Theorem B.3.1.** *Consider the AI-block model in Definition 3.2.1 with a given partition  $O$ , i.e.,  $\mathcal{A} = \mathcal{A}^{(O)}$  where the latter is defined in Equation (B.6). Under Conditions  $\mathcal{A}$ ,  $\mathcal{F}$ ,  $\mathcal{G}$  and  $\sqrt{k}(C_m - C) \rightsquigarrow \Gamma$ , the empirical process  $\mathcal{E}_{nG}$  converges weakly in  $\ell^\infty(\Delta_{d-1})$  to a tight Gaussian process having representation*

$$\begin{aligned} \mathcal{E}_G(\mathbf{t}) &= (1 + \mathcal{A}(\mathbf{t}))^2 \int_{[0,1]} (N_{C_\infty} + \Gamma)(u^{t^{(1)}}, \dots, u^{t^{(d)}}) du \\ &\quad - \sum_{g=1}^G w^{(O_g)}(\mathbf{t}) \left( 1 + \mathcal{A}^{(O_g)}(\mathbf{t}^{(O_g)}) \right)^2 \int_{[0,1]} (N_{C_\infty} + \Gamma)(\mathbf{1}, u^{t^{(i_g,1)}}, \dots, u^{t^{(i_g,d_g)}}) du, \end{aligned}$$

where  $N_{C_\infty}$  is a continuous tight Gaussian process with representation

$$N_{C_\infty}(u^{(1)}, \dots, u^{(d)}) = B_{C_\infty}(u^{(1)}, \dots, u^{(d)}) - \sum_{j=1}^d \dot{C}_\infty^{(j)}(u^{(1)}, \dots, u^{(d)}) B_{C_\infty}(\mathbf{1}, u^{(j)}, \mathbf{1}),$$

and  $B_{C_\infty}$  is a continuous tight Gaussian process with covariance function

$$\text{Cov}(B_{C_\infty}(\mathbf{u}), B_{C_\infty}(\mathbf{v})) = C_\infty(\mathbf{u} \wedge \mathbf{v}) - C_\infty(\mathbf{u})C_\infty(\mathbf{v}) = C_\Pi(\mathbf{u} \wedge \mathbf{v}) - C_\Pi(\mathbf{u})C_\Pi(\mathbf{v}),$$

where  $C_\Pi(\mathbf{u}^{(O_g)}) = \Pi_{g=1}^G C_\infty^{(O_g)}(\mathbf{u}^{(O_g)})$ .

**Theorem B.3.1** The proof is straightforward, notice that by the triangle diagram in Figure B.1

$$\mathcal{E}_{nG} = \psi \circ \phi \left( \sqrt{k}(\hat{\mathcal{A}}_{n,m} - \mathcal{A}) \right),$$

where  $\phi$  is detailed as

$$\begin{aligned} \phi &: \ell^\infty(\Delta_{d-1}) \rightarrow \ell^\infty(\Delta_{d-1}) \otimes (\ell^\infty(\Delta_{d-1}), \dots, \ell^\infty(\Delta_{d-1})) \\ x &\mapsto (x, \phi_1(x), \dots, \phi_G(x)), \end{aligned}$$

with for every  $g \in \{1, \dots, G\}$

$$\begin{aligned} \phi_g &: \ell^\infty(\Delta_{d-1}) \rightarrow \ell^\infty(S_d) \\ x &\mapsto w^{(O_g)}(t^{(1)}, \dots, t^{(G)})x(\mathbf{0}, t^{(i_g,1)}, \dots, t^{(i_g,d_g)}, \mathbf{0}), \end{aligned}$$

and also

$$\begin{aligned} \psi &: \ell^\infty(\Delta_{d-1}) \otimes (\ell^\infty(\Delta_{d-1}), \dots, \ell^\infty(\Delta_{d-1})) \rightarrow \ell^\infty(\Delta_{d-1}) \\ &\quad (x, \phi_1(x), \dots, \phi_G(x)) \mapsto \sum_{g=1}^G \phi_g(x) - x. \end{aligned}$$

$$\begin{array}{ccc}
 \sqrt{k} \left( \hat{\mathcal{A}}_{n,m} - \mathcal{A} \right) & \longrightarrow & \mathcal{E}_{nG} \\
 & \searrow \phi & \uparrow \psi \\
 \left( \sqrt{k} \left( \hat{\mathcal{A}}_{n,m} - \mathcal{A} \right); w^{(O_1)} \sqrt{k} \left( \hat{\mathcal{A}}_{n,m}^{(O_1)} - \mathcal{A}^{(O_1)} \right), \dots, w^{(O_G)} \sqrt{k} \left( \hat{\mathcal{A}}_{n,m}^{(O_G)} - \mathcal{A}^{(O_G)} \right) \right) & & 
 \end{array}$$

Fig. B.1 Commutative diagram of composition of function.

The function  $\phi_g$  is a linear and bounded function hence continuous for every  $g$ , it follows that  $\phi$  is continuous since each coordinate functions is continuous. As a linear and bounded function,  $\psi$  is also a continuous function. Noticing that,

$$(C_m - C_\infty)(\mathbf{1}, u, \mathbf{1}) = 0, \quad \forall n \in \mathbb{N},$$

where  $m$  is the block length for a sample size  $n$ . We thus have

$$\sqrt{k}(C_m - C_\infty)(\mathbf{1}, u, \mathbf{1}) \xrightarrow[n \rightarrow \infty]{} 0.$$

Therefore  $\Gamma(\mathbf{1}, u, \mathbf{1}) = 0$ . Combining this equality with Corollary 3.6 of [Bücher and Segers \(2014\)](#) and the same techniques as in the proof of Theorem 2.4 in [Marcon et al. \(2017\)](#), we obtain along with Conditions  $\mathcal{A}$ ,  $\mathcal{F}$ ,  $\mathcal{G}$

$$\sqrt{k}(\hat{\mathcal{A}}_{n,m}(\mathbf{t}) - \mathcal{A}(\mathbf{t})) \rightsquigarrow - \left( 1 + \hat{\mathcal{A}}_{n,m}(\mathbf{t}) \right)^2 \int_{[0,1]} (N_{C_\infty} + \Gamma)(u^{t^{(1)}}, \dots, u^{t^{(d)}}) du.$$


Applying the continuous mapping theorem for the weak convergence in  $\ell^\infty(\Delta_{d-1})$  (Theorem 1.3.6 of [van der Vaart and Wellner \(1996\)](#)) leads the result.  $\square$



## CHAPTER 4

# IDENTIFYING REGIONS OF CONCOMITANT COMPOUND PRECIPITATION AND WIND SPEED EXTREMES OVER EUROPE

This chapter is based on work currently under revision for publication in an international peer-reviewed journal.

 Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde (2023), Identifying regions of concomitant compound precipitation and wind speed extremes over Europe.

### **Abstract.**

The task of simplifying the complex spatio-temporal variables associated with climate modeling is of utmost importance and comes with significant challenges. In this research, our primary objective is to tailor clustering techniques to handle compound extreme events within gridded climate data across Europe. Specifically, we intend to identify subregions that display asymptotic independence concerning compound precipitation and wind speed extremes. To achieve this, we utilise daily precipitation sums and daily maximum wind speed data derived from the ERA5 reanalysis dataset spanning from 1979 to 2022. Our approach hinges on a tuning parameter and the application of a divergence measure to spotlight disparities in extremal dependence structures without relying on specific parametric assumptions. We propose a data-driven approach to determine the tuning parameter. This enables us to generate clusters that are spatially concentrated, which can provide more insightful information about the regional distribution of compound precipitation and wind speed extremes. In the process, we aim to elucidate the respective roles of extreme precipitation and wind speed in the resulting clusters. The proposed method is able to extract valuable information about extreme compound events while also significantly reducing the size of the dataset within reasonable computational timeframes.

## 4.1 Introduction

The occurrence of extreme weather events is often exacerbated by the convergence of distinct geographic factors and concurrent weather patterns, resulting in profound disruptions and extensive damage to society. Catastrophic climate phenomena such as floods, wildfires, and heatwaves frequently manifest due to the simultaneous intensification of multiple interacting processes. When these various processes coalesce to yield a substantial impact, it is referred to as a compound event. Among the primary manifestations of extreme weather, heavy precipitation and robust surface winds hold central positions, exerting adverse effects on both the natural world and human society. Extratropical cyclones, along with their associated wind patterns

## Identifying Regions of Concomitant Precipitation and Wind Speed Extremes

and storm surges, contribute significantly to economic and insured losses resulting from natural calamities in Europe. Furthermore, they disrupt transportation, trade, and energy supply systems, often leading to human casualties (refer to, for instance, [Pinto et al. \(2012\)](#); [Schwierz et al. \(2010\)](#)). To mitigate these impacts, it is important to better understand the dependence structure of extreme weather events. However, modeling such complex scenarios, where multiple rare events occur simultaneously, can be incredibly challenging, especially with high-dimensional climate datasets that exhibit heavy tails.

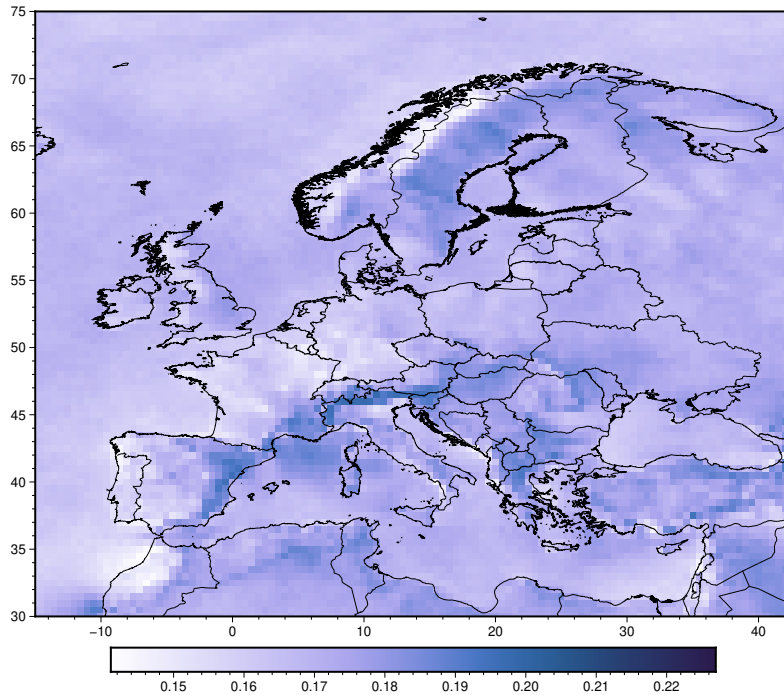


Fig. 4.1 Proportion of the total precipitation or wind speed in the ERA5 dataset that exceed their respective 0.9th quantiles.

**Considered ERA5 dataset** Large ensemble simulations present a unique opportunity for gaining deeper insights into the spatial regionalisation of compound precipitation and wind extremes. This is primarily because these simulations provide a more accurate representation of local-scale processes compared to global ensembles and do so without being hindered by data limitations. However, such simulations need to be interpreted with care as it is often largely unknown how well the employed models represent observed compound events ([Zscheischler et al. \(2021\)](#)), and differences might be large between models. In our endeavor to regionalize compound precipitation and wind speed, we turn to the ERA5 dataset ([Hersbach et al. \(2018\)](#)). This dataset allows us to investigate the correlation between daily cumulative precipitation and daily peak wind speeds throughout the extended winter between season, spanning from November to March, across Europe, for the period between 1979 and 2022. ERA5 offers a comprehensive record of atmospheric conditions, land surface characteristics, and ocean wave patterns, spanning from 1950 to the present day. It is worth noting that ERA5 supersedes the previous ERA-interim reanalysis, which began in 1979 and was initiated in 2006.

ERA5 has benefited from significant advancements in model physics, core dynamics, and data assimilation techniques developed over the past decade. Produced using 4D-Var data assimilation technology within model cycle 41r2 (Cy41r2), ERA5 incorporates improved parameterization schemes (Hersbach et al. (2020)). The dataset is available at a spatial resolution of  $0.25^\circ$  (approximately 27-28 km) on a regular grid. Our specific focus lies within the region defined by  $[-15^\circ E, 42.5^\circ E] \times [30^\circ N, 75^\circ N]$ , which encompasses Europe. We remap the original hourly data to a regularly spaced grid with a  $0.5^\circ$  spatial resolution, allowing us to compute daily precipitation totals and daily maximum wind speeds. We selected the  $0.5^\circ$  spatial resolution due to its ability to facilitate calculations within a reasonable timeframe while maintaining manageable storage requirements. The need for remapping can be circumvented with more extensive computing resources. The resulting dataset comprises 6655 daily precipitation totals and maximum wind speed measurements, covering  $91 \times 116$  grid cells with the chosen spatial resolution, totaling 10556 grid cells for clustering. To illustrate, Fig. 4.1 provides a visualization of the proportion of grid cells where either wind speed or precipitation exceed a significant threshold. As observed in both panels, there is a noticeable spatial variation in these proportions. In the following, we introduce some notations to describe the spatio-temporal process under consideration.

Consider a spatio-temporal random field denoted as  $(\mathbf{Z}_n^{(s)}, s \in D \subset \mathbb{R}^2, n \in \mathbb{N})$ . Here,  $\mathbf{Z}_n^{(s)} = (Z_n^{(s,1)}, Z_n^{(s,2)})$  represents the vector of daily total precipitation and wind speed maxima at location  $s$  on day  $n$ . We assume that  $\mathbf{Z}_n^{(s)}$  is identically distributed over  $n$  for each location  $s$  in the domain  $D$ . Now, let's suppose that we have observations available at  $d$  spatial locations for each time  $n$ . We can represent these observations as  $\mathbf{Z}_n = (\mathbf{Z}_n^{(1)}, \dots, \mathbf{Z}_n^{(d)})$ , where  $\mathbf{Z}_n$  is random vector with stationary law  $\mathbf{Z} = (\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)})$ . Each  $\mathbf{Z}^{(j)}$  is a random vector with two dimensions, corresponding to precipitation and wind speed. In the given dataset, each variable represents a location on a grid pixel scale and is characterised by multiple features that can exhibit extreme behavior when considered together. This implies that clustering methods designed to analyze the dependence structure should consider spatial extremal dependence across different locations. In the following discussion, we will review some studies that aim to understand the dependence structure of potentially high-dimensional random vectors.

**Related literature** In the field of high-dimensional extremes, researchers have made significant contributions to address the challenges associated with vectors that have univariate random margins, with a focus on unsupervised techniques such as support identification (see, e.g., Goix et al. (2017); Meyer and Wintenberger (2021)), Principal Component Analysis of the angular measure of extremes (Cooley and Thibaud (2019); Drees and Sabourin (2021)), graphical models for extremes (Asenova et al. (2021); Engelke and Hitz (2020); Hitz and Evans (2016)) and clustering methods (Janßen and Wan (2020) or Chapter 3). These methods can help identify hidden spatial patterns and sub-regions where variables are dependent on their extremes, which is crucial for regionalisation tasks.

Over the years, a number of clustering approaches have been suggested, with a focus on extremes, based on a comparison between univariate distributions. For instance, Bernard et al. (2013) analysed weekly maxima of precipitation in France and developed a clustering algorithm on a proper distance, the madogram, justified by Extreme Value Theory (EVT). The same approach was used in Bador et al. (2015) to evaluate the bias of climate model simulations of temperature maxima over Europe. In Durante et al. (2015), a four-step clustering procedure was

presented that considered a pairwise conditional Spearman’s correlation coefficient, extracted from daily-log-returns of the adjusted stock price, as a measure of tail dependence. Pappadà et al. (2018) investigated spatial sub-regions (clusters) of flood risk behavior in the Po river basin in Italy using a copula-based Agglomerative Hierarchical Clustering. In the paper by Maume-Deschamps et al. (2023), they introduce a modified spectral clustering algorithm designed for analyzing spatial extreme events. This algorithm combines spectral clustering with the concept of extremal concurrence probability, as proposed by Dombry et al. (2018). The goal of this approach is to determine whether a max-stable process exhibits a stationary dependence structure or not. Chapter 3 proposed a class of models, the Asymptotic Independent block (AI block) models, for variable clustering, which defines population-level clusters based on the independence of extremes between clusters. They exhibited an algorithm that compares the extremal dependence of univariate distributions at different locations and showed that it recovers the thinnest partition such that extremes between groups of random variables are mutually independent with high probability.

While regional analysis of univariate climate extreme events is a well-studied area of research, multivariate compound extreme events at larger scales have received less attention. Although the widely known Kullback-Leibler divergence has been adapted for use in the context of compound extreme events, it has been primarily employed to cluster data based on their bivariate extreme behavior (as demonstrated in Vignotto et al. (2021)) and to analyze compound weather and climate events (as discussed in Zscheischler et al. (2021)). However, this metric primarily summarises the differences in distribution between two sets of random variables when at least one of their components is extreme, and it does not quantify deviations from asymptotic independence. recognising sub-regions characterised by concurrent extreme precipitation and wind speed events is essential for improving extreme event modeling. This is particularly evident in works like Chatelain et al. (2020) and Engelke and Hitz (2020), which rely on the assumption of asymptotic dependence in the data. Such insights are crucial for the development of strategies to mitigate the impacts of these extreme events.

**Proposed methodology.** In this paper, our objective is to expand upon the AI block model as introduced in Chapter 3 to tackle the challenges posed by this environmental dataset. We depart from the assumption that clusters of pixels are mutually independent univariate time series, with regard to their extremes. Instead, we shift to a framework where a collection of univariate time series is recorded for each pixels, with a particular emphasis on their extreme behavior. To tackle this intrinsic problem, we introduce the concept of *constrained* AI block model, compelling pixels represent a collection of univariate time series. This concept comes into play in our environmental problematic which concerns phenomena like precipitation and wind speed extremes recorded at a specific geographic locations represented as pixels within a large ensemble dataset.

Our objective is the following: cluster a number of  $d = 10556$  pixels across Europe based on their asymptotic independence on compound precipitation and wind speed extremes where data are relatively scarce, i.e., the sample size  $n = 6655$ . To efficiently implement a fast algorithm designed for this model-based approach in such a high-dimensional setting, we employ a divergence measure that highlights the differences in extremal dependence structures for asymptotically dependent and independent random vectors. Noticeably, this divergence measure adheres to several axioms that makes it a valid measure of dependence (De Keyser and Gijbels

(2023a,b)) and also a coherent measure (Scarsini (1984)). Furthermore, this divergence is linked to a well-known quantity in Extreme Value Theory and can be consistently estimated without the need of parametric assumptions. This consistent estimation is possible under the condition of weak mixing conditions to stay within the scope of our application where departures from the independence assumption are strongly suspected.

The algorithm requires the specification of a tuning parameter, and we suggest an approach based on data to determine its value. When applied to our environmental dataset, this clustering procedure is efficient and produces clusters that are spatially concentrated, which is a pattern commonly observed in spatial processes. Furthermore, we leverage the interpretability of classical AI block models to gain insights into the role of precipitation and wind speed extremes in the compound partition. In other words, we use this approach to understand how the clustering is influenced by both wind speed and precipitation. To further analyse the results, we make use of a straightforward modification of our dissimilarity measure. This modification allows us to comment on the different clusterings obtained through various algorithms and provide valuable insights into our proposed methodology.

## 4.2 A clustering algorithm for compound extreme events

### 4.2.1 A measure for evaluating dependence between compound extremes

We consider a high-dimensional random vector  $\mathbf{Z} = (\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)})$  with law  $F$  having  $d$  marginal random vectors  $\mathbf{Z}^{(j)} = (Z^{(j,1)}, \dots, Z^{(j,p_j)})$  for  $j = 1, \dots, d$ . To accommodate vectors of different sizes which will be useful later, we introduce a different notation from the one given in the introduction. Each  $\mathbf{Z}^{(j)}$  contains  $p_j$  marginal univariate random variables  $Z^{(j,\ell)}$  for  $\ell = 1, \dots, p_j$ . In this framework  $\mathbf{Z}$  has  $q = p_1 + \dots + p_d$  marginal univariate random variables.

We call for convenience a function  $u$  on  $\mathbb{R}$  a normalising function if  $u$  is non-decreasing, right continuous, and  $u(x) \rightarrow \pm\infty$  as  $x \rightarrow \pm\infty$ . For a stationary sequence  $(\mathbf{Z}_n, n \in \mathbb{N})$  of  $\mathbf{Z}$ , we say that the distribution  $F$  belongs to the max-domain of attraction of the Extreme Value Distribution (EVD)  $H$  if the following convergence result holds for properly normalised maxima:

$$\mathbb{P} \left\{ \bigvee_{i=1}^n \mathbf{Z}_i \leq \mathbf{u}_n(\mathbf{x}) \right\} \xrightarrow{n \rightarrow \infty} H(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^q, \quad (4.1)$$

where  $\mathbf{u}_n(\mathbf{x}) = (u_n^{(1,1)}(x^{(1,1)}), \dots, u_n^{(1,p_1)}(x^{(1,1)}), \dots, u_n^{(d,p_d)}(x^{(d,p_d)}))$  is a  $q$ -dimensional vector of normalising functions. The margins  $H^{(1,1)}, \dots, H^{(d,p_d)}$  of  $H$  must be univariate extreme value distributions and the dependence structure of  $H$  is determined by the relation

$$-\ln H(\mathbf{x}) = L \left( -\ln H^{(1,1)}(x^{(1,1)}), \dots, -\ln H^{(d,p_d)}(x^{(d,p_d)}) \right), \quad (4.2)$$

for all points  $\mathbf{x}$  such that  $H^{(j,\ell)}(x^{(j,\ell)}) > 0$  for all  $j = 1, \dots, d$ ,  $\ell = 1, \dots, p_j$ . The convergence result in (4.1) with the relation in (4.2) holds under mild assumptions on the dependence between  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ , making the study of time series relevant within the framework of EVT. The stable tail dependence function  $L : [0, \infty)^q \rightarrow [0, \infty)$  can be retrieved from the distribution



function  $F$  via

$$L(\mathbf{x}) = \lim_{t \rightarrow 0} t^{-1} \mathbb{P} \left\{ F^{(1,1)}(Z^{(1,1)}) > 1 - tx^{(1,1)} \text{ or } \dots \text{ or } F^{(d,p_d)}(Z^{(d,p_d)}) > 1 - tx^{(d,p_d)} \right\}. \quad (4.3)$$

We assume that the random vector  $\mathbf{Z}$  is in the max-domain of attraction of an EVD, denoted as  $H$ . Moreover, we aim that the extremal dependence can be modelled using an AI block model (see Chapter 3) where the definition is recalled below.

**Definition 4.2.1 (Asymptotic Independent block model).** Let  $(\mathbf{Z}_n, n \in \mathbb{N})$  be a  $q$ -variate stationary random sequence with law  $F$  in the max domain of attraction of  $H$ . The random sequence  $(\mathbf{Z}_n, n \in \mathbb{N})$  is said to follow an AI block model if there exists a partition  $O = \{O_g\}_{g=1}^G$  of  $\{1, \dots, q\}$  with  $|O_g| = d_g$  and marginal extreme value distributions  $H^{(O_g)} : \mathbb{R}^{d_g} \rightarrow [0, 1]$  such that  $H = \Pi_{g=1}^G H^{(O_g)}$ .

The constrained AI block model requires improvements to the methods proposed in Chapter 3, which uses extremal correlation to detect asymptotic independence between random variables, to correctly identify the hidden partition. Asymptotic independence is a concept that describes the relationship between extremes of two random variables, denoted  $Z^{(a)}$  and  $Z^{(b)}$ . Each variable has its own cumulative distribution function, denoted  $F^{(a)}$  and  $F^{(b)}$ , respectively. The extremal correlation, denoted as  $\chi(a, b)$ , between these two random variables is formally stated by

$$\chi(a, b) = \lim_{t \rightarrow 0} \mathbb{P} \left\{ F^{(a)}(Z^{(a)}) > 1 - t | F^{(b)}(Z^{(b)}) > 1 - t \right\}.$$

The extremal correlation represents the probability of one variable being extreme given that the other is also extreme. If the extremal correlation coefficient  $\chi(a, b)$  is in the range of  $(0, 1]$ , then  $Z^{(a)}$  and  $Z^{(b)}$  are said to be asymptotically dependent. Otherwise, if  $\chi(a, b) = 0$ , the variables are asymptotically independent. For instance, the well-known bivariate Gaussian distribution with correlation coefficient  $\rho \in [-1, 1)$  satisfies  $\chi(a, b) = 0$ .

An extension beyond the bivariate case involves examining two groups of random variables. In Chapter 3, a new metric called Sum of Extremal COefficient (SECO) was introduced. To better understand the definition of the metric, the necessary notations for extremal coefficients of an extremal random vector with possibly different sizes are presented below:

$$\theta(1, \dots, d) = \lim_{t \rightarrow 0} t^{-1} \mathbb{P} \left\{ \max_{j=1, \dots, d} \max_{\ell=1, \dots, p_j} F^{(j,\ell)}(Z^{(j,\ell)}) > 1 - t \right\} \quad (4.4)$$

$$\theta(j) = \lim_{t \rightarrow 0} t^{-1} \mathbb{P} \left\{ \max_{\ell=1, \dots, p_j} F^{(j,\ell)}(Z^{(j,\ell)}) > 1 - t \right\}, \quad j = 1, \dots, d. \quad (4.5)$$

Then,  $\theta(j)$  corresponds to the extremal coefficient associated to the  $j$ th marginal random vector  $\mathbf{X}^{(j)}$ . The SECO metric is defined as the difference between the sum of the extremal coefficients of the  $d$  marginal random vectors  $\mathbf{Z}^{(j)}$  and the extremal coefficient of the entire vector, that is, formally stated

$$\text{SECO}(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)}) = \sum_{j=1}^d \theta(j) - \theta(1, \dots, d). \quad (4.6)$$

The SECO metric is always positive and quantifies the deviation from asymptotic independence between the  $d$  groups of variables. Indeed, Proposition B.2.3 in Chapter 3, Ferreira (2011) showed that the SECO metric is equal to zero if and only if the  $d$  groups of variables are independent extreme value random vectors. Moreover, the bivariate SECO between  $\mathbf{Z}^{(j)}$  and  $\mathbf{Z}^{(k)}$  simplifies to the extremal correlation when these are random variables. Recently, De Keyser and Gijbels (2023b) developed an axiomatic framework to quantify dependence between multiple groups of random variables of possibly different sizes. For self-consistency, we recall them in Appendix C.1 and we show that the SECO metric defined in (4.6) satisfies most of the stated axioms (see Lemma C.1.1 in Appendix C.1). Also, we are interested in determining whether SECO remains coherent (as defined in Durante and Sempi (2015); Scarsini (1984)) when comparing two random vectors with the same dimension, which occurs when  $p_a = p_b$ . In Appendix C.2, we examine the coherence of the SECO in nested extreme value copulae, which were introduced in Hofert et al. (2018). A further objective is to investigate how SECO behaves in specific nested models.

### 4.2.2 Clustering for compound extremes

In this section, we introduce a modified version of the ECO algorithm as presented in Chapter 3, which is capable of clustering compound extremes.

To introduce flexibility in AI block models and bring notations into our application, we consider  $\mathbf{Z}_i^{(j)}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, d$ , which are extracted from  $d$  pixels. We assume that each observation is distributed according to  $F$  which is in the max-domain of attraction an *constrained* AI block model as defined in Section 4.2.1. Any deviation from asymptotic independence between two pixels can then be measured by the empirical counterpart version of the SECO in (4.6). The empirical counterpart of the SECO between two pixels  $a$  and  $b$  is defined as:

$$\widehat{\text{SECO}}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) = \hat{\theta}(a) + \hat{\theta}(b) - \hat{\theta}(a, b), \quad (4.7)$$

In this case study, we consider  $d = 10556$  pixels. For each pixel  $j = 1, \dots, d$ , we define a bivariate random vector  $\mathbf{Z}^{(j)} = (Z^{(j,1)}, Z^{(j,2)})$ , where, as a convention in this paper,  $Z^{(j,1)}$  and  $Z^{(j,2)}$  represent the stationary distributions of daily total precipitation and wind speed maxima at location  $j$ , respectively. Hence, to stick in this context,  $\hat{\theta}(a, b)$  is the empirical counterpart of the extremal coefficient for the joint vector  $(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)})$ , and  $\hat{\theta}(j)$  is the empirical counterpart of the extremal coefficient for the random vector  $\mathbf{Z}^{(j)}$ ,  $j \in \{a, b\}$ , i.e.,

$$\hat{\theta}(a, b) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{R_i^{(a,1)} > n-k+0.5 \text{ or } R_i^{(a,2)} > n-k+0.5 \text{ or } R_i^{(b,1)} > n-k+0.5 \text{ or } R_i^{(b,2)} > n-k+0.5\}} \quad (4.8)$$

$$\hat{\theta}(j) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{R_i^{(j,1)} > n-k+0.5 \text{ or } R_i^{(j,2)} > n-k+0.5\}}, \quad j = a, b, \quad (4.9)$$

where  $R_i^{(j,\ell)}$  denote the rank of  $Z_i^{(j,\ell)}$  among  $Z_1^{(j,\ell)}, \dots, Z_n^{(j,\ell)}$ ,  $j = a, b$ ,  $\ell = 1, 2$ . The notation described above can be easily extended to pixels of varying sizes or to a number a pixels greater than 2. However, for the sake of clarity in notation and to maintain consistency with our application, we focus on the scenario where pixels consist of bivariate time series. The statistic in Equation (4.7) gauges the strength of dependence between two pixels by counting how often

one of the pixels is extreme in terms of either precipitation or wind speed, while adjusted for the number of times a component is extreme in at least one of the pixels.

Classical non-parametric estimators of the extremal coefficient, as demonstrated in equation (4.8), is only relevant when the number of variables is less than the number of observations, meaning  $d \leq n$ . Indeed, in Appendix C.3 (see Lemma C.3.1 and Lemma C.3.2), we exhibit bounds which show limitations that classical estimators of the extremal dependence structure face, preventing them from encompassing the full range of extremal dependence structures. In high-dimensional scenarios where the number of variables exceeds the number of observations ( $d > n$ ), classical estimators may fail to identify asymptotic independence in extremes. This is particularly relevant for a generalised version estimator in (4.9) for an arbitrary number  $d$  of variables. This phenomenon can also be observed in other estimators, such as the madogram, and other rank-based estimators of the tail dependence structure may suffer from the same issue. Therefore, caution must be exercised when dealing with high-dimensional data, particularly by taking a lower number of extremes (determined by the parameter  $k$ ), as advised by the upper bound which is equal to  $n/k$ . This approach allows for a wider range of values with the cost of a greater variance.

Nonetheless, while the dimension is arbitrary, the empirical estimator of the extremal coefficient in (4.9) is consistent and the asymptotic deviation is well-understood (see, for instance, Drees and Huang (1998); Einmahl et al. (2012)) in the standard assumption of independent observations. In Appendix C.4, Proposition C.4.1, we present arguments regarding the consistence of the proposed estimator of the SECO which goes beyond this classical setup of independent observations.

If  $\mathbf{Z}^{(a)}$  and  $\mathbf{Z}^{(b)}$  are asymptotically independent, there is no guarantee that an extreme event in one vector will be accompanied by an extreme event in the other vector, then the statistic in (4.7) will converge in probability to zero (see Appendix C.4, Proposition C.4.1). On the other hand, if  $\mathbf{Z}^{(a)}$  and  $\mathbf{Z}^{(b)}$  are asymptotically comonotone, then the SECO reduces to  $\hat{\theta}(a) = \hat{\theta}(b)$  almost surely, since an extreme event in one vector will always be accompanied by an extreme event in the other vector. Additionally, the lower and upper bounds of  $\widehat{\text{SECO}}(a, b)$  is given by:

$$0 \leq \widehat{\text{SECO}}(a, b) \leq \min\{\hat{\theta}(a), \hat{\theta}(b)\} \quad \text{a.s.}, \quad (4.10)$$

where the upper one is reached when  $\mathbf{Z}^{(a)}$  and  $\mathbf{Z}^{(b)}$  are asymptotically comonotone. The resulting matrix, denoted by  $\hat{\Theta}$ , is a  $d \times d$  matrix where each entry is given by

$$\hat{\Theta}(a, b) = \widehat{\text{SECO}}(a, b) / \min\{\hat{\theta}(a), \hat{\theta}(b)\}, \quad a, b \in \{1, \dots, d\}, \quad (4.11)$$

which is the normalised SECO metric.

The algorithm that we present in this section takes as input the matrix  $\hat{\Theta}$  in (4.11). This enables the division of  $d$  objects of interest into the thinnest partition possible such that mutual asymptotic independence holds between clusters. Algorithm (CAICE) summarises the procedure for clustering asymptotically independent compound extreme events. As detailed in Chapter 3 for the Algorithm ECO which is similar to the above algorithm (CAICE), the overall complexity of the estimation procedure is  $O(d^2(d \vee n \ln(n)))$ .

The  $\tau$  threshold is the only hyper-parameter in the (CAICE) Algorithm, and its selection is important in obtaining an effective partitioning. A useful tool for choosing an appropriate

**Algorithm (CAICE)** Clustering procedure for AI block models with compound extreme

---

```

1: procedure CAICE( $S, \tau, \hat{\Theta}$ )
2:   Initialise:  $S = \{1, \dots, d\}$ ,  $\hat{\Theta}(a, b)$  for  $a, b \in \{1, \dots, d\}$  and  $l = 0$ 
3:   while  $S \neq \emptyset$  do
4:      $l = l + 1$ 
5:     if  $|S| = 1$  then
6:        $\hat{O}_l = S$ 
7:     if  $|S| > 1$  then
8:        $(a_l, b_l) = \arg \max_{a, b \in S} \hat{\Theta}(a, b)$ 
9:       if  $\hat{\Theta}(a_l, b_l) \leq \tau$  then
10:         $\hat{O}_l = \{a_l\}$ 
11:       if  $\hat{\Theta}(a_l, b_l) > \tau$  then
12:         $\hat{O}_l = \{s \in S : \hat{\Theta}(a_l, s) \wedge \hat{\Theta}(b_l, s) \geq \tau\}$ 
13:        $S = S \setminus \hat{O}_l$ 
14:   return  $\hat{O} = (\hat{O}_l)_l$ 

```

---

threshold is the SECO value for the resulting partition, which has been recommended in Chapter 3, Section 3.3.4. This metric measures the divergence between the sum of the extremal coefficients of each cluster and the extremal coefficient of the entire vector (see (4.6) with group of different sizes). An effective partitioning is achieved when this metric is minimised, ideally for a moderate value of the threshold  $\tau$ . By identifying the threshold that results in the lowest SECO value, one can establish a partition of pixels, ensuring that their clusters are asymptotically independent from each other.

## 4.3 Detecting concomitant extremes of compound precipitation and wind

### 4.3.1 Non-serially independent

To statistically assert departures from serial independence of multivariate time series, we conducted a randomness test as proposed by Genest and Remillard (2004) and Ghoudi et al. (2001). We use the methodology presented in Kojadinovic and Holmes (2009) and extended in Kojadinovic and Yan (2011) to detect serial dependence in continuous multivariate time series. Briefly, let  $\mathbf{Z}_1, \mathbf{Z}_2, \dots$  be  $d$ -dimensional random vectors. We chose an integer  $k > 1$ , and for  $\mathbf{u} \in [0, 1]^{dk}$ , the vector  $\mathbf{u}_{\{j\}}$  is defined as follows:

$$u_{\{j\}}^{(i)} = (u^{(i)} - 1) \mathbb{1}_{\{i \in \{(j-1)q+1, \dots, jq\}\}} + 1,$$

We form the  $dk$ -dimensional random vector  $\mathfrak{Z}_i = (\mathbf{Z}_i, \dots, \mathbf{Z}_{i+k-1})$ , with  $i = 1, \dots, n$ . The serial independence empirical copula process in the multivariate setting thus write

$$\sqrt{n} \left( C_n^s(\mathbf{u}) - \prod_{j=1}^d C_n^s(\mathbf{u}_{\{j\}}) \right), \quad (4.12)$$

where  $C_n^s$  is the serial empirical copula process computed with  $\mathfrak{Z}_1, \dots, \mathfrak{Z}_n$ . Under the hypothesis of serial independence, one can establish the asymptotic behavior of (4.12) to a tight Gaussian process. To obtain potentially powerful test obtained above from the empirical process, Kojadinovic and Yan (2011) derived  $2^{k-1} - 1$  tests statistic based on a Möbius decomposition of the process in (4.12) in the continuous multivariate time series setting. Those tests are implemented in the copula R package (Yan (2007)).

Due to computational limitations, we only considered three  $3 \times 3$  pixels, at different locations, covering the initial 304 days of the study, which represented the two first years of observation. We analysed precipitation and wind separately and the resulting dependogram is depicted in Fig. 4 in Appendix C.6. It is notable that most of the “subsets of lags” exhibited serial dependence. The function `multSerialIndepTest` computed three  $p$ -values, all of which provided robust evidence against serial independence. Furthermore, the decreasing trend of the computed statistic implied that the dependence was weakening for observations further apart. These results align with the mixing conditions stated in Appendix C.4.

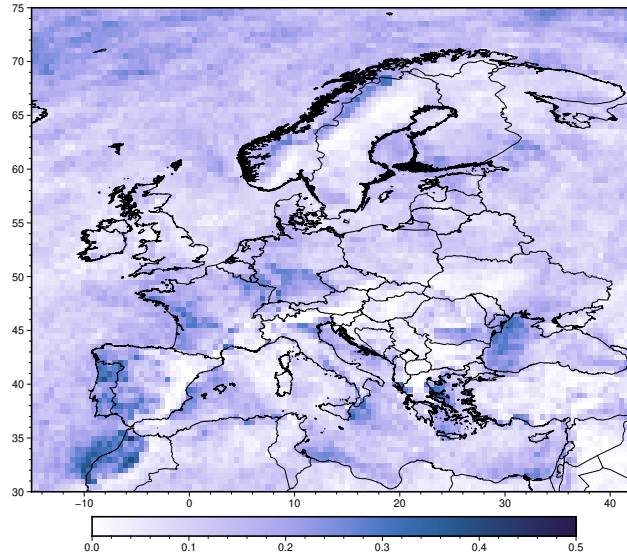


Fig. 4.2 Estimator of the extremal correlation within pixels, focusing on the daily sum of precipitation and wind speed maxima,  $\hat{\chi}$  in Equation (4.13), maps for the 100 largest values ( $k = 100$ ).

### 4.3.2 Exploratory analysis

Extreme values of daily precipitation and wind speed maxima may occur together, and we aim to understand the spatial variability of this relationship. To identify regions for which extrema are non-concomitant between them, we consider the peak over threshold approach, which are values that lie above a certain value (see, for instance, Beirlant et al. (2006); De Haan and Ferreira (2006); Resnick (2007) for an overall introduction to classical statistical methods in EVT). We conduct an exploratory analysis of the extremal dependence structure between the

two variables within pixels. The estimated  $\chi$  coefficient is defined by

$$\hat{\chi}(a) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{R_i^{(a,1)} > n-k+0.5, R_i^{(a,2)} > n-k+0.5\}}, \quad (4.13)$$

where  $R_i^{(a,\ell)}$  denotes the rank of  $Z_i^{(a,\ell)}$  among  $Z_1^{(a,\ell)}, \dots, Z_n^{(a,\ell)}$ ,  $\ell = 1, 2$ . The estimated extremal coefficient between the variables reveals the highest co-occurrences in regions along the western coasts of Portugal, Spain, France, the UK, and Norway, as well as the northeastern coast of the Mediterranean (Fig. 4.2). Conversely, the smallest co-occurrences are observed on the eastern coasts of the UK, Sweden, and Spain, over the northwestern coast of the Mediterranean, and around the Carpathian and southeastern Norwegian mountain ranges. These results are consistent with prior research (see, e.g., Martius et al. (2016)).

The low co-occurrence over eastern Norway and Spain may be attributed to the orographic enhancement of rain on the windward side of a mountain and the drying of the air as it reaches the lee (Martius et al. (2016)). This could also explain the high co-occurrences over the eastern coasts of the Mediterranean, where cyclones from the Mediterranean storm track may arrive perpendicularly to the mountains on the western coast of Italy and the eastern coasts of the Adriatic Sea (Owen et al. (2021)). Additionally, the Cierzo winds may be responsible for the low co-occurrence to the south of the Pyrenees (Martius et al. (2016)).

Nevertheless, the aforementioned analysis only considers extreme behavior within individual pixels. To dig deeper, we take a further step by using the empirical SECO outlined in Equation (4.7) to explore interactions between pixels. Noteworthy, the empirical SECO does not inherently include spatial distances between pixels. As shown in Fig. 4.3, we observe strong or moderate dependence between extremes for locations that are close to each other, while locations that are far apart have a normalised SECO near zero, indicating that extremes are weakly dependent or independent. The number of points per pixel highlights that most pixels are widely separated, with Euclidean distances between 20 to 40, and have small SECO values, indicating weak dependence or independence between compound extremes of two pixels. Therefore, even though it is not explicitly designed for this purpose, in the case of this specific dataset, the empirical SECO captures significant spatial information. This information could be highly valuable for comprehending the spatial patterns of precipitation and wind speed in Europe.

In this context, our aim is to gain a deeper understanding of the spatial patterns of extreme total precipitation and wind speed maxima across Europe. To broaden our analysis of extremal pattern within pixels, we propose employing the clustering algorithm described in (CAICE) to group pixels while considering the extremal dependence of precipitation and wind speed between pixels.

#### 4.3.3 Clustering with constrained AI block models

To delineate Europe into distinct regions that are mutually independent in compound weather extremes, we employ the clustering algorithm (CAICE). Additionally, we employ a data-driven approach for selecting a suitable threshold, as elaborated below. Our initial step involves working with the matrix  $\hat{\Theta}$  derived from Equation (4.11). The resultant clustering outcome is contingent on the value of  $\tau$ . Once Algorithm (CAICE) has been applied, we will denote the resulting partition as  $\hat{O}(\tau) = \{\hat{O}_g\}_{g=1}^G$  of the set  $\{1, \dots, d\}$ . Each cluster is characterised by a

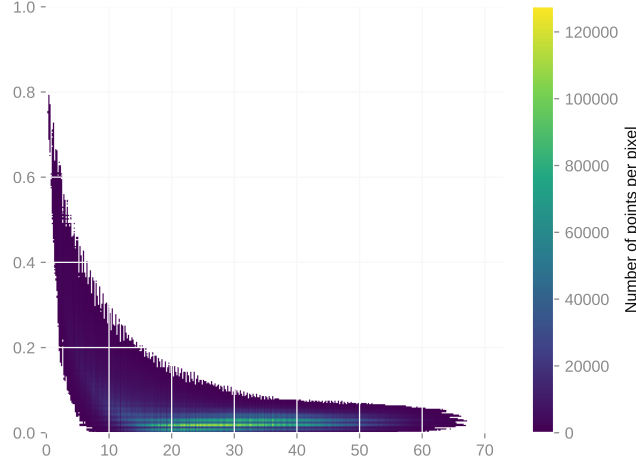


Fig. 4.3 Pairwise SECO as a function of distance between sites.

respective cardinality of  $d_g$ . As depicted in Figure 4.3, even a small  $\tau$  value within Algorithm (CAICE) can effectively partition the sub-regions of Europe. Nevertheless, a more rigorous approach to selecting the threshold value  $\tau$  need to be discussed.

Let us introduce the random vector  $\mathbf{Y}_\tau^{(g)}$ , composed of the variable index within  $\hat{O}_g$ . This means that  $\mathbf{Y}_\tau^{(g)}$  is a  $(2d_g)$ -dimensional random vector, considering two variables for each pixel in our case study. Moving forward, we can define the empirical SECO for groups of random vectors, which might have varying sizes within the given partition, using the following equation:

$$\widehat{\text{SECO}}(\mathbf{Y}_\tau^{(1)}, \dots, \mathbf{Y}_\tau^{(G)}) = \sum_{g=1}^G \hat{\theta}(g) - \hat{\theta}(1, \dots, d). \quad (4.14)$$

The estimator mentioned above varies with  $\tau$  due to its reliance on the partition  $\hat{O}(\tau)$ . As outlined in Appendix 3.3.4 of Chapter 3, the values of  $\tau$  that minimise SECO also ensure consistent recovery of our groups. For more details, we refer to Proposition 3.3.2 in Chapter 3. To address this, we construct a loss function, denoted as  $L$ , over a grid of  $\tau$  values denoted as  $\Delta$ , as follows:

$$L(\tau) = \ln \left( 1 + \left( \widehat{\text{SECO}}(\mathbf{Y}_\tau^{(1)}, \dots, \mathbf{Y}_\tau^{(G)}) - \min_{\tau \in \Delta} \widehat{\text{SECO}}(\mathbf{Y}_\tau^{(1)}, \dots, \mathbf{Y}_\tau^{(G)}) \right) \right), \quad \tau \in \Delta. \quad (4.15)$$

In our current high-dimensional context, where we have  $n = 6655$  daily observations and  $d = 10556$  pixels ( $n < d$ ), estimating the extremal coefficient for the entire vector  $\theta(1, \dots, d)$  encounters an upper bound that prevents it from reaching its theoretical maximum of  $d$ , as discussed in Appendix 4.2.2. However, the loss function  $L$  in equation (4.15) remains unaffected by this bias since it is mitigated by the subtraction operation. Nonetheless, it is essential to be mindful of this bias when computing  $L$  for partitions with larger clusters, which typically occurs for smaller values of  $\tau$ . Therefore, we recommend reducing the number of extremes to

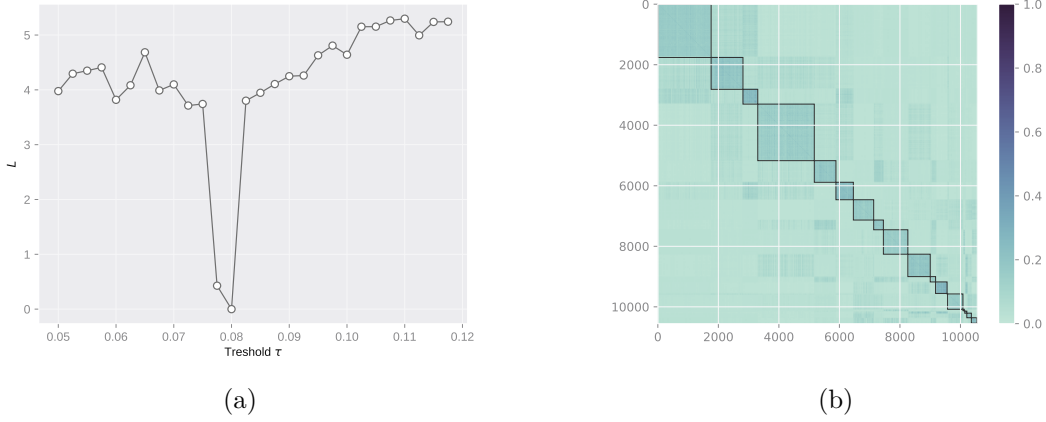


Fig. 4.4 Value of the function  $L(\tau)$  in (4.15) for different values of  $\tau \in \Delta = \{0.05, 0.0525, \dots, 0.12\}$  in Panel a. Partitions of the SECO similarity matrix with threshold  $\tau = 0.08$  in Panel b. Squares represent the clusters of variables.

$k = 30$  when evaluating  $L$ . This adjustment widens the range of values used for estimating the extremal coefficient from higher values, but it comes at the expense of increased variance in the estimation process.

The value of  $L$  with different values of  $\tau$  with  $\Delta = \{0.05, 0.0525, \dots, 0.12\}$  suggests that the best partitioning is found for  $\tau = 0.08$  (Fig. 4.4, Panel a). For this threshold, we obtain 22 clusters with 3 of them having less than 10 entities. We report the clustered SECO matrix defined in (4.11) in Fig. 4.4, Panel b.

#### 4.3.4 Results

In Fig. 4.5, we present the twelve largest clusters obtained with the partition setting  $\tau = 0.08$ . Our algorithm effectively identifies sub-regions with strong dependence within clusters, as well as *near-independence* or independence among compound extremes of daily precipitation and wind speed maxima in different clusters. The most prominent cluster, the fourth one, encompasses the North Sea and the Baltic Sea, which are connected basins. The North Sea is a marginal basin of the North Atlantic, and a shallow connection to the Baltic Sea exists through the Skagerrak and Kattegat regions. This area is particularly important for climate sciences and hydrology, and has inspired several works (see, for example, [Andrée et al. \(2022\)](#); [Gröger et al. \(2019\)](#); [Wang et al. \(2015\)](#)). Another interesting area is the ninth cluster, consisting of the southwestern Black Sea, the Levantine Basin, and the southeast of the Anatolian peninsula. The Levantine Basin is known to be one of the windiest areas of the Mediterranean Sea, and the spatial distribution of the Levantine Basin in Figure 2 of [Soukissian et al. \(2018\)](#) outlines the geometry of this cluster. Additionally, extreme storm situations typically occur in December-January in the southwestern part of the Black Sea ([Divinsky et al. \(2020\)](#)). To the best of our knowledge, there are no studies on the tail dependence structure of climate variables in the southwestern Black Sea and the Levantine Basin. Furthermore, our algorithm sometimes distinguishes between the extremal behavior of land and sea, as illustrated by the first and sixth clusters for the Norwegian Sea and the Atlantic Ocean, respectively.



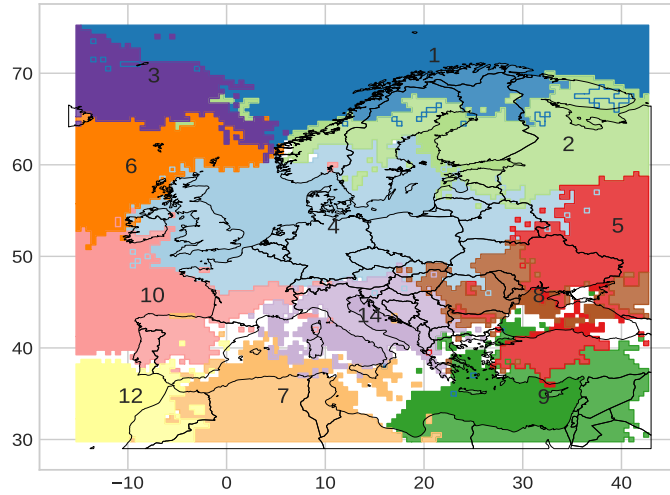


Fig. 4.5 Representation of the 12 largest clusters of the partition of the SECO matrix between pixels with Algorithm (CAICE) and threshold  $\tau = 0.08$ . Number of each cluster is depicted inside the cluster.

Identifying regions with simultaneous extreme events can be valuable for statistical modeling, and the method proposed here can address this question effectively. However, it is important to understand whether wind or precipitation is more important for the resulting clustering. To investigate this, a similar clustering process was applied separately to daily total precipitation and wind speed maxima using the extremal correlation as a dissimilarity. The resulting partitions are denoted by,  $\hat{O}^P$ , and  $\hat{O}^W$  with an adapted threshold  $\tau$ . We will denote by  $\hat{O}^{PW}$  the resulting partition for compound daily total precipitation and wind speed maxima. The figures depicting the results for the individual variables can be found in Fig. 5 in Appendix C.6.

The adapted threshold for the partition  $\hat{O}^P$  is  $\tau = 0.09$ , resulting in 70 medium-sized clusters. For daily wind speed maxima, the optimal clustering  $\hat{O}^W$  is obtained by setting  $\tau = 0.07$ , resulting in 24 clusters. Upon visual inspection, our algorithmic partitioning reveals mosaic block patterns along the diagonal, while no clear patterns could be discerned from the off-diagonal. Additionally, the off-diagonal showcases moderate asymptotic dependence between groups or asymptotic independence, indicating that the resulting clustering aligns with the purpose of AI block models. The clusters for daily wind speed maxima are larger than those for daily total precipitation, which supports previous studies indicating that heavy gusts have a larger spatial impact than precipitation events (see, for example, Pfahl and Wernli (2012); Raveh-Rubin and Wernli (2015)). With regard to spatial precipitation, several studies have shown that dependence tends to weaken for the largest observations (Lalancette et al. (2021); Le et al. (2018)). This knowledge could explain why there are a large number of clusters with only a few entities for the clustering of daily total precipitation.

To compare different clustering methods, we use the Adjusted Rand Index (ARI), a popular measure used in clustering analysis (see, for instance, Hubert and Arabie (1985); Rand (1971)). To summarise, the ARI gives a concordance score between two different partitions. It takes

values between 0 and 1 and the closer to 1, the more similar the partitions. For more details about its computation, we refer the reader to Appendix C.5 of the supplementary materials. Computing the ARI between  $\hat{O}^{PW}$  and  $\hat{O}^W$  (resp,  $\hat{O}^{PW}$  and  $\hat{O}^P$ ), we obtain

$$\text{ARI}(\hat{O}^{PW}, \hat{O}^W) = 0.5, \quad \text{ARI}(\hat{O}^{PW}, \hat{O}^P) = 0.3.$$

An ARI value of 0 indicates that the two sets are completely random and have nothing in common, while a value of 1 indicates a perfect match between the two partitions. In this case, the ARI value of 0.5 between the clustering of compound daily total precipitation and wind speed maxima and the clustering of the sole wind speed suggests that there is moderate similarity between the two sets, implying that there are fewer matching data points or clusters between them. In light of these results, we can conclude that the clustering of compound extreme is induced by both variables with a little more emphasis driven by wind speed maxima.

#### 4.3.5 Alternative clustering method using SECO

In equation (4.11),  $\hat{\Theta}$  can be envisioned as a similarity matrix, and  $1 - \hat{\Theta}$  as a dissimilarity matrix. This dissimilarity metric is smaller (or larger) for compound extremes that are dependent (or independent) between two pixels. The upper bound of the dissimilarity metric is 1, which is reached when the compound extremes are independent. Thus, it is possible to perform classical method of clustering using this dissimilarity matrix. In particular, we have chosen to explore two different methods: a quantization-based approach and a Hierarchical clustering. These two approach requiring a specification of the number of clusters (unknown in practice), we use the ‘‘silhouette coefficient’’ developed by, [Rousseeuw \(1987\)](#) and which compares the tightness of clusters with their dissociation. In practice, the number of clusters, denoted as  $K$ , is determined by selecting the maximum average silhouette coefficient.

Quantization, also known as lossy data compression in information theory, involves the task of substituting data with an efficient and compact representation that allows for the reconstruction of the original observations with a certain degree of accuracy. A clustering problem can be viewed as an optimal quantization process aimed at minimising a specific loss function (for example, see [Banerjee et al. \(2005\)](#) or ([Linder, 2002](#), p. 15)). To generate clusters through the optimal quantization process, we utilised the algorithm described in references [Laloë \(2010\)](#) and [Laloë \(2021\)](#), setting arbitrary the number of clusters to  $K = 10$  as the silhouette coefficient did not provide a suitable number of clusters.

The process of hierarchical clustering begins with a basic partition, initially comprising of  $d$  individual data points where each pixel stands alone as its own cluster. Subsequently, the data is grouped together incrementally, with clusters gradually merging until a single cluster encompassing all variables is achieved. At each step, the algorithm combines the two closest cluster centers according to a specific definition (for more details, refer to ([Giraud, 2021](#), Chapter 12)), while keeping other clusters unchanged. For this approach, the silhouette coefficient leads us to set  $K = 25$ .

The resulting partitions for this two methods are depicted in Figure 4.6, and the clustered matrices can be found in Fig. 6 and Fig. 7 for further visualizations, in Appendix C.6. Notably, the clusters obtained through the quantization-based approach in the northern regions bear a resemblance to those obtained through Algorithm (CAICE). For instance, clusters 3, 5, and

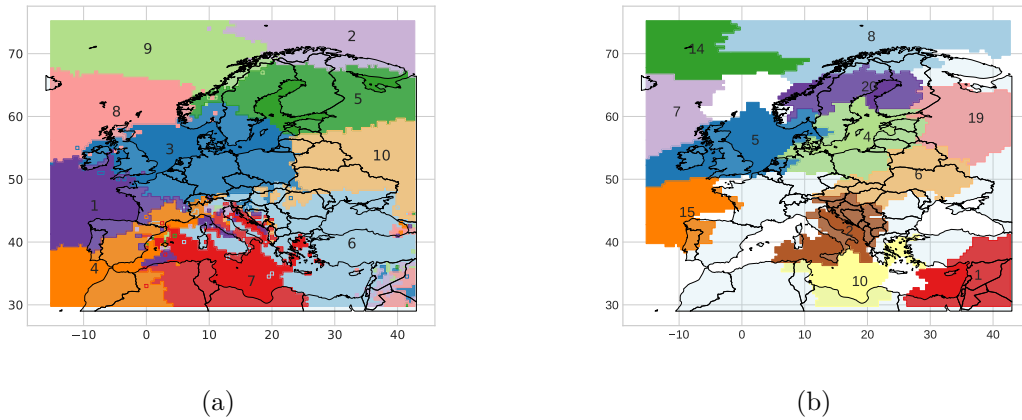


Fig. 4.6 In Panel **a** is depicted the representation of the partition obtained through quantization-based approach with  $K = 10$ . In Panel **b** a similar representation for the 12 largest clusters is proposed through hierarchical clustering approach with  $K = 25$ . These clusters were generated based on data related to daily total precipitation and wind speed maxima extremes from 10556 pixels. Number of each cluster is depicted inside the cluster.

8 in Figure 4.6, Panel **a**, are closely similar to clusters 2, 4, and 6 in Figure 4.5 as identified through the (CAICE) Algorithm. Furthermore, as observed in the (CAICE) Algorithm, the quantization-based approach effectively distinguishes extreme behaviors between land and sea. This distinction is apparent in clusters 2 and 8 in Figure 4.6, Panel **a**. These findings suggest the uniqueness of the SECO similarity measure in extracting valuable spatial information. One notable difference between the hierarchical clustering and Algorithm (CAICE) is that the clusters obtained through hierarchical clustering are smaller and depict stronger cross-dependencies. This phenomenon can be explained by the fact that some clusters which are separate in the hierarchical clustering partition are combined in the output of Algorithm (CAICE). For instance, in Fig. 4.6, Panel **b**, most of pixels of clusters 4 and 5 in the hierarchical clustering are combined into a single cluster in Algorithm (CAICE), the fourth one in Fig. 4.5. This suggests that compound extremes of those clusters are dependent, but smaller groups in the northern and Baltic Sea areas observe more concomitant extremes. Below, we investigate a hierarchical clustering of the fourth cluster given by Algorithm (CAICE) to inquiry whether or not we obtain a similar partition given by cluster 4 and 5 obtained in the hierarchical clustering.

Fig. 4.7 displays the results of hierarchical clustering analysis on the most prominent cluster, i.e., cluster 4 from the output of Algorithm (CAICE). The analysis is performed on a reduced dataset, and the number of clusters is calibrated using the silhouette coefficient Rousseeuw (1987), with the maximum value obtained for  $K = 3$  (see Fig. 8). The resulting matrix,  $\hat{\Theta}$ , reveals a strong dependence among compound extremes, indicating the presence of a whole asymptotic dependent vector with three distinct blocks with more concomitant compound extremes.

These results reveal intriguing patterns of co-occurring wind and precipitation extremes. The third cluster in Fig. 4.7, which represents central-eastern Europe, is spatially coherent and having no access to the sea. The first cluster is connected to the North Sea and consists of

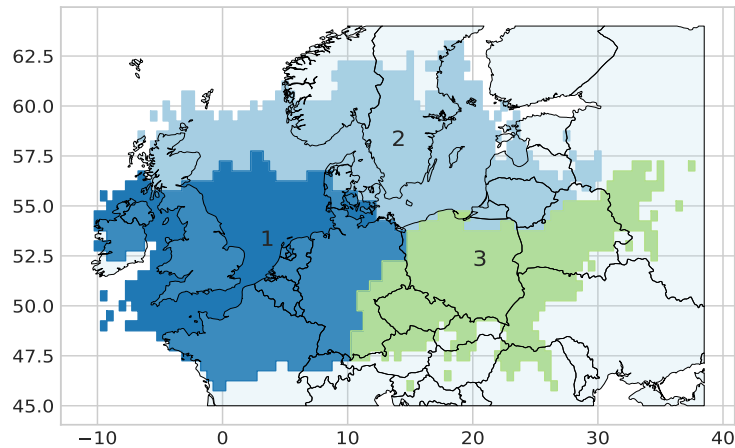


Fig. 4.7 Representation of the 3 clusters of the partition of the 1868 pixels of the fourth cluster of the partition given by Algorithm (CAICE) using extremes of daily total precipitation and wind speed maxima. Number of each cluster is depicted inside the cluster.

western European countries (except Scotland). The second cluster comprises Scandinavian countries and the Baltic Sea, with a shallow connection to the North Sea. Notably, the North Sea's section belonging to this area corresponds to the Fladen Ground, the deepest part of the North Sea in the Scottish sector.

We also observe a similar partition given by the hierarchical clustering among all locations, where the North Sea is separated from the Baltic Sea. However, there is a difference in the Fladen Ground's classification, which is in different clusters for the hierarchical clustering among all pixels belonging to the North Sea cluster, see cluster 5 in Fig. 4.6b, and to the Baltic Sea cluster for the hierarchical clustering applied to the reduced dataset, see cluster 2 in Fig. 4.7.

## 4.4 Conclusion and perspectives

Clustering spatial pixels is highly significant as it allows for a better comprehension of the inherent spatial pattern of a relevant physical phenomenon, enhances the accuracy of statistical procedures in situations where data is limited, and helps to identify regions where joint preventive measures can be taken to mitigate the impact of weather-related risks.

Traditionally, climate science research has focused on analysing single drivers or univariate dangers, which simplifies the complex dynamics of climate and its consequences. However, in reality, climate hazards often interact, leading to compound extremes. Therefore, statistical analyses that consider multiple hazards simultaneously are needed. In this paper, our goal is to identify subregions in Europe that demonstrate asymptotic independence concerning compound precipitation and wind speed extremes. In simpler terms, we want to find areas where two distinct subregions cannot experience concurrent compound extremes. To achieve this, we introduce a multivariate extreme value measure known as the SECO metric. This

metric helps us quantify the extent to which random vectors of varying sizes deviate from asymptotic independence. We not only introduce this metric but also demonstrate the reliability of a non-parametric estimator for it, even in scenarios that go beyond the typical setup of independent observations. Building on this, we propose an algorithm specifically tailored for *constrained* AI block model. This model ensures that pixels represent collections of univariate time series. Our algorithm allows us to pinpoint the largest partition where compound extremes exhibit independence over Europe. Interestingly, we uncover specific geographical patterns without relying on positional information, and it is worth noting that we do not need to pre-determine the number of clusters.

The proposed methodology can be extended to the case where the dataset has more than two variables, denoted as  $p > 2$ . Furthermore, the methodology can also be extended to pixels with varying lengths of recorded time series, all the while preserving the concept of asymptotic independence. However, caution should be taken when interpreting the results in terms of coherence, or it may not be appropriate at all.

A major issue highlighted in this paper is how to estimate the dependence structure in high-dimensional datasets, where the number of variables is greater than the number of observations. As explained in Appendix C.3, traditional estimators are not able to accurately recover the dependence structure of an extreme value random vector if its margins are not sufficiently dependent, quantified by the condition  $L(\mathbf{x}) \geq n/k$  for  $\mathbf{x} \in \mathbb{R}^d$  where  $n$  is the number of observations and  $k$  the number of considered extremes. However, extreme value data are often scarce, making the high-dimensional setting common, except in certain cases.

# APPENDIX C

## SUPPLEMENTARY MATERIALS OF CHAPTER 4

### C.1 Axioms for a valid dependence measure

In this section, we recall the axiomatic framework to quantify dependence between multiple groups of random variables of possibly different sizes [De Keyser and Gijbels \(2023b\)](#). We recall that we consider a random vector  $\mathbf{Z}$  with distribution  $F$  that is in the max domain of attraction of an multivariate extreme value distribution  $H$ . Plausible axioms, for a valid dependence measure of a random vector  $\mathbf{X}$  with cumulat, denoted as  $\mathcal{D}(\mathbf{X})$ , are as follows

- (A1) For every permutation  $\pi$  of  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}$ :  $\mathcal{D}(\mathbf{X}) = \mathcal{D}(\pi(\mathbf{X}))$ ; and for every permutation  $\pi_j$  of  $X^{(j,1)}, \dots, X^{(j,p_j)}$ , for  $j \in \{1, \dots, d\}$ , it holds:

$$\mathcal{D}(\mathbf{X}) = \mathcal{D}(\mathbf{X}^{(1)}, \dots, \pi^{(j)}(\mathbf{X}^{(j)}), \dots, \mathbf{X}^{(d)}).$$

- (A2)  $0 \leq \mathcal{D}(\mathbf{X}) \leq 1$ .  
(A3)  $\mathcal{D}(\mathbf{X}) = 0$  if and only if  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}$  are mutually independent.  
(A4)  $\mathcal{D}(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}, \mathbf{X}^{(d+1)})$  with equality if and only if  $\mathbf{X}^{(d+1)}$  is independent of  $(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)})$ .  
(A5)  $\mathcal{D}(\mathbf{X})$  is well defined for any  $q$ -dimensional random vector  $\mathbf{X}$  and is a functional of solely the copula  $C$  of  $\mathbf{X}$ .  
(A6) Let  $T^{(j,\ell)}$  for  $j = 1 \dots, d$  and  $\ell = 1, \dots, p_j$  be strictly increasing, continuous transformations. Then

$$\mathcal{D}(\mathbf{T}^{(1)}(\mathbf{X}^{(1)}), \dots, \mathbf{T}^{(d)}(\mathbf{X}^{(d)})) = \mathcal{D}(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}),$$

where  $\mathbf{T}^{(j)} = (T^{(j,1)}(X^{(j,1)}), \dots, T^{(j,p_j)}(X^{(j,p_j)}))$  for  $j = 1, \dots, d$ .

- (A7) Let  $T^{(j,\ell)}$  be a strictly decreasing, continuous transformation for a fixed  $j \in \{1, \dots, d\}$  and a fixed  $\ell \in \{1, \dots, p_j\}$ . Then

$$\mathcal{D}(\mathbf{X}^{(1)}, \dots, T^{(j,\ell)}(\mathbf{X}^{(j)}), \dots, \mathbf{X}^{(d)}) = \mathcal{D}(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}),$$

where  $T^{(j,\ell)}(\mathbf{X}^{(j)}) = (X^{(j,1)}, \dots, T^{(j,\ell)}(X^{(j,\ell)}), \dots, X^{(j,p_i)})$ .

- (A8) Let  $(\mathbf{X}_n)_{n \in \mathbb{N}}$  be a sequence of  $q$ -dimensional reduction random vectors having copulas  $(C_n)_{n \in \mathbb{N}}$ , then

$$\lim_{n \rightarrow \infty} \mathcal{D}(\mathbf{X}_n) = \mathcal{D}(\mathbf{X})$$

if  $C_n \rightarrow C$  uniformly, where  $C$  denotes the copula of  $\mathbf{X}$ .

Having those necessary materials, we now detail below which axioms hold for the SECO metric to measure the dependence among extreme of random vectors.

**Proposition C.1.1.** *Let SECO be the metric defined in (4.6). Then, it satisfies the system of axioms (A1), (A2) with the following bounds*

$$0 \leq \text{SECO}(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)}) \leq \min_{j=1, \dots, d} \left\{ \sum_{k \neq j} \theta(k) \right\},$$

(A3), (A4), (A5), (A6) and (A8) stated in *De Keyser and Gijbels (2023b)*.

**Proof** The reason why Property (A1) holds is due to the fact that the set union and addition of numbers are commutative. Proposition B.2.2 and Proposition B.2.3 from Chapter 3 imply the results stated about (A2) for the lower bound and (A3), where the latter is related to asymptotic independence. To obtain the upper bound, it is observed that

$$0 \leq \theta(1, \dots, d) \leq \min\{\theta(1), \dots, \theta(d)\}.$$

For property ((A4)), the inequality

$$\text{SECO}(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)}, \mathbf{Z}^{(d+1)}) \geq \text{SECO}(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)})$$

can be rewritten as

$$\theta(1, \dots, d) + \theta(d+1) \geq \theta(1, \dots, d, d+1),$$

which holds true by Proposition B.2.2 of Chapter 3. Moreover, this inequality holds as an equality if and only if  $\mathbf{Z}^{(d+1)}$  is asymptotically independent of  $(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)})$ , according to Proposition B.2.3 of Chapter 3.

We have for every  $u \in (0, 1)$ ,

$$\theta(1, \dots, d) = \ln(C_\infty(u, \dots, u)) / \ln(u),$$

where  $C_\infty$  is the extreme value of copula of  $H$  which is the max-domain attractor of  $F$ , the distribution of  $\mathbf{Z}$ , i.e.,

$$C_\infty(u^{(1,1)}, \dots, u^{(d,pa)}) = \exp \left\{ -L \left( -\ln u^{(1,1)}, \dots, -\ln u^{(d,pa)} \right) \right\}.$$

Hence (A5) and (A6) are fulfilled. Let  $(\mathbf{Z}_n)_{n \in \mathbf{N}}$  be a sequence of random vector, for each  $n$ , suppose that  $\mathbf{Z}_n$  is in the max-domain of attraction of a random vector  $H_n$  an extreme value distribution with extreme value copula  $C_n$ . Suppose that the sequence of extreme value copulae  $(C_n)_{n \in \mathbf{N}}$  converges uniformly to  $C_\infty$ , where  $C_\infty$  is the extreme value copula of  $H$ , an extreme value distribution. Then, for every  $\mathbf{u} \in (0, 1)$ , we have the pointwise convergence:

$$C_n(\mathbf{u}) \rightarrow C_\infty(\mathbf{u}), \quad n \rightarrow \infty.$$

Thus, as  $\ln : (0, 1) \rightarrow \mathbb{R}_-$  is continuous, we have

$$\ln(C_n(\mathbf{u})) \rightarrow \ln(C_\infty(\mathbf{u})), \quad n \rightarrow \infty.$$

Then

$$\text{SECO}(\mathbf{Z}_n^{(1)}, \dots, \mathbf{Z}_n^{(d)}) \rightarrow \text{SECO}(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)}), \quad n \rightarrow \infty.$$

□

## C.2 A coherent measure for extreme value random vectors

In this section, we borrow the notation previously introduced in the specific case where  $d = 2$  and  $p_a = p_b = p$ . Let  $\mathbf{Z}$  and  $\mathbf{W}$  be two random vectors with marginal random vectors  $\mathbf{Z}^{(a)}$  and  $\mathbf{Z}^{(b)}$  (resp.  $\mathbf{W}^{(a)}$  and  $\mathbf{W}^{(b)}$ ), each having  $p$  components. We set  $q = 2p$ . Assume that  $\mathbf{Z}$  and  $\mathbf{W}$  are in the max-domain of attraction of two extreme value distributions having stable tail dependence function  $L_{\mathbf{Z}}$  and  $L_{\mathbf{W}}$ . Here,  $L_{\mathbf{Z}}$  and  $L_{\mathbf{W}}$  are nested, and for  $\mathbf{x} \in [0, \infty)^q$ ,  $\mathbf{x}^{(a)} = (x^{(a,1)}, \dots, x^{(a,p)})$ ,  $\mathbf{x}^{(b)} = (x^{(b,1)}, \dots, x^{(b,p)})$ , we have

$$L_{\mathbf{Z}}(\mathbf{x}) = L_{\mathbf{Z}}^{(0)} \left( L_{\mathbf{Z}}^{(a)}(\mathbf{x}^{(a)}), L_{\mathbf{Z}}^{(b)}(\mathbf{x}^{(b)}) \right), \quad L_{\mathbf{W}}(\mathbf{x}) = L_{\mathbf{W}}^{(0)} \left( L_{\mathbf{W}}^{(a)}(\mathbf{x}^{(a)}), L_{\mathbf{W}}^{(b)}(\mathbf{x}^{(b)}) \right),$$

where  $L_{\mathbf{Z}}^{(0)}$  and  $L_{\mathbf{Z}}^{(j)}$ ,  $j = a, b$ , are the ‘‘mother’’ and ‘‘childrens’’ stable tail dependence function, respectively. The stable tail dependence functions  $L_{\mathbf{W}}^{(0)}$  and  $L_{\mathbf{W}}^{(j)}$ ,  $j = a, b$ , are defined similarly. The SECO for these two models thus reduces to

$$\begin{aligned} \text{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) &= L_{\mathbf{Z}}^{(a)}(\mathbf{1}^{(a)}) + L_{\mathbf{Z}}^{(b)}(\mathbf{1}^{(b)}) - L_{\mathbf{Z}}^{(0)} \left( L_{\mathbf{Z}}^{(a)}(\mathbf{1}^{(a)}), L_{\mathbf{Z}}^{(b)}(\mathbf{1}^{(b)}) \right), \\ \text{SECO}(\mathbf{W}^{(a)}, \mathbf{W}^{(b)}) &= L_{\mathbf{W}}^{(a)}(\mathbf{1}^{(a)}) + L_{\mathbf{W}}^{(b)}(\mathbf{1}^{(b)}) - L_{\mathbf{W}}^{(0)} \left( L_{\mathbf{W}}^{(a)}(\mathbf{1}^{(a)}), L_{\mathbf{W}}^{(b)}(\mathbf{1}^{(b)}) \right) \end{aligned}$$

Taking advantage of the definition introduced by [Scarsini \(1984\)](#), we say that  $\mathbf{Z}$  is more concordant in levels of extremes than  $\mathbf{W}$  if for every possible value of  $\mathbf{x} \in \mathbb{R}^q$ ,  $L_{\mathbf{Z}}^{(0)}(\mathbf{x}) \leq L_{\mathbf{W}}^{(0)}(\mathbf{x})$ . Now, let’s suppose that  $\mathbf{Z}$  and  $\mathbf{W}$  have the same marginal extremal dependence structure, which means that their stable tail dependence functions are the same for both components, i.e.,

$$L_{\mathbf{Z}}^{(a)}(\mathbf{x}^{(a)}) = L_{\mathbf{W}}^{(a)}(\mathbf{x}^{(a)}), \quad L_{\mathbf{Z}}^{(b)}(\mathbf{x}^{(b)}) = L_{\mathbf{W}}^{(b)}(\mathbf{x}^{(b)}), \quad \mathbf{x} \in \mathbb{R}^q.$$

Thus, we have

$$\begin{aligned} \text{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) - \text{SECO}(\mathbf{W}^{(a)}, \mathbf{W}^{(b)}) &= \\ L_{\mathbf{W}}^{(0)} \left( L_{\mathbf{W}}^{(a)}(\mathbf{1}^{(a)}), L_{\mathbf{W}}^{(b)}(\mathbf{1}^{(b)}) \right) - L_{\mathbf{Z}}^{(0)} \left( L_{\mathbf{W}}^{(a)}(\mathbf{1}^{(a)}), L_{\mathbf{W}}^{(b)}(\mathbf{1}^{(b)}) \right) \end{aligned}$$

which is positive. Thus the SECO is a coherent measure given the marginal random vector’s dependence structure is the same.

To better understand the behaviour of the SECO, we will consider two specific nested models: the nested Gumbel and Hüsler-Reiss models, for which  $p = 2$  and  $d = 2$ . For these models, we will analyse the corresponding stable tail dependence functions.

$$L^{\text{Gu}}(x_1, x_2) = \left( x_1^{1/\alpha} + x_2^{1/\alpha} \right)^\alpha, \quad L^{\text{HR}}(x_1, x_2) = \Phi \left( \frac{\lambda}{2} + \frac{x_2 - x_1}{\lambda} \right) x_1 + \Phi \left( \frac{\lambda}{2} + \frac{x_1 - x_2}{\lambda} \right) x_2,$$



with  $\alpha \in (0, 1]$ ,  $\lambda \in (0, \infty)$  and  $\Phi$  denotes the cumulative distribution function of a standard normal random variable. The first stable tail dependence function is known as the Gumbel (or Logistic) distribution introduced by Gumbel (1960b), the parameter  $\alpha$  represent the strength of dependence: if  $\alpha \rightarrow 0$ , then the two random variables are comonotone while if  $\alpha = 1$  it reduces to independence. The second stable tail dependence function is the Hüsler Reiss distribution (Hüsler and Reiss 1989)), both asymptotic comonotony and independence are depicted by the respective limits  $\lambda \rightarrow 0$  and  $\lambda \rightarrow \infty$ .

For the nested Gumbel model, the SECO is equal to

$$2^{\alpha_a} + 2^{\alpha_b} - \left(2^{\alpha_a/\alpha_0} + 2^{\alpha_b/\alpha_0}\right)^{\alpha_0}. \quad (\text{C.1})$$

This equation involves the parameters  $\alpha_0, \alpha_a$ , and  $\alpha_b$ , which correspond to the mother, the first, and the second random vector margins, respectively. In Fig. 1 are depicted level sets of (C.1) and the normalised version, i.e., (C.1) divided by  $\min\{2^{\alpha_a}, 2^{\alpha_b}\}$ , according to  $\alpha_0$  for different dependence structure between the marginal random vectors. Additionally, we can also calculate the SECO for the Hüsler-Reiss nested model which equals to

$$\theta^{(a)} + \theta^{(b)} - \left[ \Phi \left( \frac{\lambda_0}{2} + \frac{\theta^{(b)} - \theta^{(a)}}{\lambda_0} \right) \theta^{(a)} + \Phi \left( \frac{\lambda_0}{2} + \frac{\theta^{(a)} - \theta^{(b)}}{\lambda_0} \right) \theta^{(b)} \right], \quad (\text{C.2})$$

where  $\theta^{(j)} = 2\Phi(\lambda_j/2)$ ,  $j = a, b$  are the extremal coefficients and includes several parameters such as  $\lambda_a, \lambda_b$ , and  $\lambda_0$  which correspond to the first, the second random vector margins and the mother respectively. In Fig. 2, we represent level sets of (C.2) and the normalised version of (C.2), that is divided by  $\min\{\theta^{(a)}, \theta^{(b)}\}$ , the bound found in Proposition C.1.1.

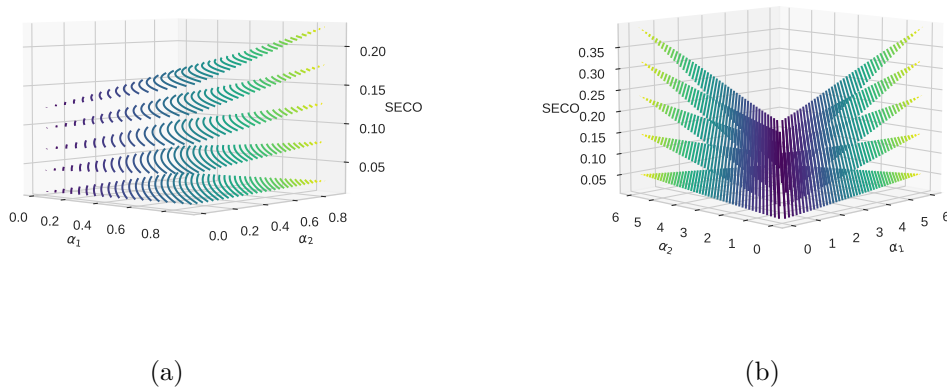


Fig. 1 Level sets of the SECO for the nested Logistic model (see (C.1)) in Panel 1a and the normalised SECO in Panel 1b for  $\alpha_0 \in \{0.91, 0.93, 0.95, 0.97, 0.99\}$  and  $\alpha_1, \alpha_2$  range in  $\{0.01, 0.02, \dots, 0.9\}$ . The coherence property ensures that the level sets are arranged in ascending order based on the values of  $\alpha_0$ .

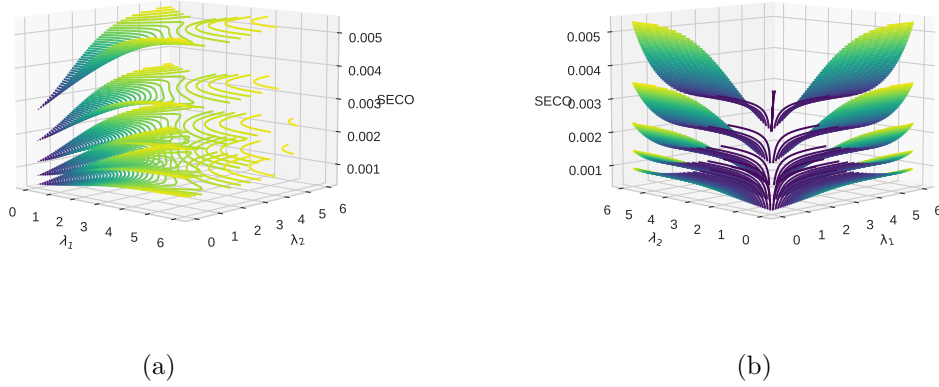


Fig. 2 Level sets of the SECO for the nested Hüsler-Reiss model (see (C.2)) in Panel 2a and the normalised SECO in Panel 2b for  $\lambda_0 \in \{6.0, 6.25, 6.5, 6.75, 7.0\}$  and  $\lambda_1, \lambda_2$  range in  $\{0.01, 0.02, \dots, 6.0\}$ . The coherence property ensures that the level sets are arranged in ascending order based on the values of  $\lambda_0$ .

In both Fig. 1 and Fig. 2, the behaviour of SECO exhibits similarities, particularly in relation to the level sets. These level sets are characterised by their monotonicity, whereby the level set's height increases with the degree of dependence among the marginal random vectors. In other words, the greater the dependence between the marginal random vectors, the higher the level set will be. Additionally, the SECO demonstrates monotonicity through  $\alpha_1, \alpha_2$  (or alternatively,  $\lambda_1, \lambda_2$ ), whereby the value of the SECO increases as the random variables in the marginal random vectors become more concordant. In simpler terms, the more closely related the random variables are, the higher the SECO will be. Moving on to the normalised SECO, we observe that the highest values are attained when the random variables within the marginal random vectors display asymmetric behaviour, meaning that one variable is comonotonic while the other is independent. Conversely, the lowest values are obtained when both variables share the same dependence structure.

### C.3 Incompleteness tail dependence structure estimation in high dimension

Throughout the section, assume that we have  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$  independent and identically distributed observations of the  $d$ -dimensional random vector  $\mathbf{Z}$ , which is in the max-domain of attraction of  $H$ , an EVD where each of the components of  $\mathbf{Z}$  are asymptotically independent. Let  $R_{n,i}^{(j)}$  denote the rank of  $Z_i^{(j)}$  among  $Z_1^{(j)}, \dots, Z_n^{(j)}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, d$ . For  $k \in \{1, \dots, n\}$ , define a nonparametric estimator of  $\theta$ , the extremal coefficient by

$$\hat{\theta}_{n,k}^{\text{EKS}} := \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{R_{n,i}^{(1)} > n+0.5-k \text{ or } \dots \text{ or } R_{n,i}^{(d)} > n+0.5-k\}}$$

see Einmahl et al. (2012).

**Lemma C.3.1.** For  $k \in \{1, \dots, n\}$  we have

$$\hat{\theta}_{n,k}^{\text{EKS}} \leq \frac{n}{k}$$

**Proof** The upper bound is trivial since

$$\sum_{i=1}^n \mathbb{1}_{\{R_{n,i}^{(1)} > n+0.5-k \text{ or } \dots \text{ or } R_{n,i}^{(d)} > n+0.5-k\}} \leq n.$$

□

Now, let us divide the sample of size  $n$  of  $\mathbf{Z}$  into  $k$  blocks of length  $m$ , so that  $k = n/m$  (where we suppose, without loss of generality that  $m$  divide  $n$ ). For the  $i$ th block, the maximum value in the  $j$ -component is denoted by

$$M_{m,i}^{(j)} = \max\{Z_t^{(j)} : t \in (im - m, im]\}.$$

Let  $R_{n,m,i}^{(j)}$  denote the rank of  $M_{m,i}^{(j)}$  among  $M_{m,1}^{(j)}, \dots, M_{m,j}^{(j)}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, d$ . Define a non parametric estimator of the multivariate madogram

$$\hat{\nu}_{n,m} := \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d \frac{R_{n,m,i}^{(j)}}{k+1} - \frac{1}{d} \sum_{j=1}^d \frac{R_{n,m,i}^{(j)}}{(k+1)} \right]$$

**Lemma C.3.2.** For  $m \in \{1, \dots, n\}$ , we have

$$\hat{\nu}_{n,m} \leq \frac{k}{k+1} - \frac{1}{2}.$$

Consequently,

$$\hat{\theta}_{n,m}^{\text{MAD}} \leq \frac{n}{m}.$$

**Proof** One can easily deduce the following upper bound

$$\bigvee_{j=1}^d \frac{R_{n,m,i}^{(j)}}{k+1} \leq \frac{k}{k+1}.$$

Thus we obtain that

$$\hat{\nu}_{n,m} \leq \frac{k}{k+1} - \frac{1}{k} \frac{1}{d} \sum_{j=1}^d \sum_{i=1}^k \frac{R_{n,m,i}^{(j)}}{k+1}.$$

The right hand side of the equation is equal to

$$\frac{1}{k} \frac{1}{d} \sum_{j=1}^d \sum_{i=1}^k \frac{R_{n,m,i}^{(j)}}{k+1} = \frac{1}{k} \frac{1}{d} \sum_{j=1}^d \frac{k(k+1)}{2(k+1)} = \frac{1}{2}.$$

Let us consider the following function

$$f: \left[0, \frac{k}{k+1} - \frac{1}{2}\right] \rightarrow \mathbb{R}$$

$$x \mapsto \frac{0.5 + x}{0.5 - x}.$$

Since it is an nondecreasing function, we must have

$$\hat{\theta}_{n,m}^{\text{MAD}} \leq f\left(\frac{k}{k+1} - \frac{1}{2}\right) = \frac{n}{m}$$

□

The consequence of both lemmas is that the nonparametric and the madogram-based estimator of the extremal coefficient can only recover values in the following range:

$$1 \leq \hat{\theta}_{n,k}^{\text{EKS}} \leq \frac{n}{k}, \quad 1 \leq \hat{\theta}_{n,m}^{\text{MAD}} \leq \frac{n}{m}.$$

If we suppose that  $d > n/k$  or  $d > n/m$ , then both estimators cannot expect to retrieve dependencies above the thresholds stated by our two lemmas. In particular, in high dimension, i.e., when  $d > n$ , these estimators are unable to detect asymptotic independence. Indeed, in cases where the  $d$  variables are asymptotically independent, extremes occur in one variable without influencing extremes in the others. Thus, when  $d > n$ , it is highly probable to observe a rank that is greater than  $n - k + 0.5$  for at least one variable and this occurs for every observation  $i = 1, \dots, n$ . Since this happens for every observation  $i = 1, \dots, n$ , the characteristic function is (with high probability) always equal to one, resulting in an overall extremal coefficient equal to  $n/k$  when taking the sum. However, in high dimensions, this cannot be equal to  $d$ , the value taken in asymptotic independence.

To illustrate these findings, we consider the following numerical setup. Consider as the sample size  $n \in \{100, 150, \dots, 1000\}$  and the high dimensional setting given by  $d = n^{1.25}$  where we want to estimate the extremal coefficient of the random vector of  $\mathbf{Z}$  where its components are asymptotically independent. In this setup, we know that the theoretical value of  $\theta$  is given by  $d$ . When studying the dependence structure of extreme events, we face the ‘‘curse of dimensionality’’ in two ways. Firstly, traditional estimators do not cover the full spectrum of possible values. Secondly, in order to expand the range of values, one may need to reduce the number of extremes considered, denoted by  $k$ , or decrease the size of block maxima, denoted by  $m$ . However, this may lead to an increase in variance or a decrease in bias, depending on the estimator used.

In Fig. 3, the bias of the estimator is clearly depicted. For each value of  $n$ , both estimators reach their upper bounds, that is  $\hat{\theta}_{n,k}^{\text{EKS}} = n/k$  and  $\hat{\theta}_{n,m}^{\text{MAD}} = n/m$ .

## C.4 Consistent estimation of SECO

Consider a  $q$ -dimensional random vector  $\mathbf{Z} = (Z^{(1,1)}, \dots, Z^{(d,p_d)})$ , where  $p_1 + \dots + p_d = q$ . This vector has a joint cumulative distribution function (c.d.f.) denoted as  $F$ , and each of its components has continuous marginal c.d.f.s  $F^{(1,1)}, \dots, F^{(d,p_d)}$ . The copula, denoted as  $C$ ,

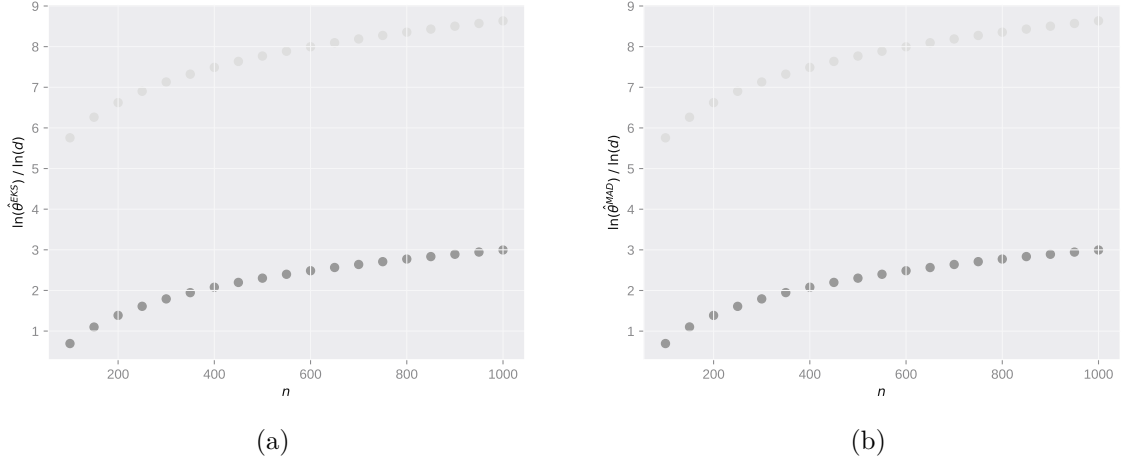


Fig. 3 Estimator  $\hat{\theta}_{n,k}^{\text{EKS}}$  (Panel 3a, black dots) and  $\hat{\theta}_{n,m}^{\text{MAD}}$  (Panel 3b, black dots) according to different values of  $n \in \{100, 150, \dots, 1000\}$  and  $d = n^{1.25}$ . The theoretical value of the extremal coefficient is depicted in grey dots. For both estimators, we took  $k = m = 50$ .

associated with  $F$  (or equivalently, with  $\mathbf{Z}$ ) is defined as the c.d.f. of another random vector  $\mathbf{U} = (U^{(1,1)}, \dots, U^{(d,p_d)})$ . These random variables  $\mathbf{U}$  are obtained through the marginal application of the probability integral transform, meaning that  $U^{(j,\ell)} = F^{(j,\ell)}(Z^{(j,\ell)})$  for  $j = 1, \dots, d$  and  $\ell = p_1, \dots, p_d$ . Notably, the marginal c.d.f.s of the copula  $C$  are uniformly distributed on the interval  $[0, 1]$ . According to Sklar's theorem, the copula  $C$  is a unique function that satisfies the following relationship for all  $\mathbf{x} = (x^{(1,1)}, \dots, x^{(d,p_d)}) \in \mathbb{R}^q$ :

$$F(x^{(1,1)}, \dots, x^{(d,p_d)}) = C \left\{ F^{(1,1)}(x^{(1,1)}), \dots, F^{(d,p_d)}(x^{(d,p_d)}) \right\}.$$

In simpler terms, this equation describes how the joint distribution  $F$  can be represented in terms of the copula  $C$  and the marginal distributions  $F^{(j,\ell)}$ .

Now, let us consider a sequence of observed data,  $\mathbf{Z}_i$  for  $i = 1, \dots, n$ , which represents a stationary time series. Importantly, each  $\mathbf{Z}_i$  follows the same distribution as  $\mathbf{Z}$ . Set  $\mathbf{U}_i = (U_i^{(1,1)}, \dots, U_i^{(d,p_d)}) \sim C$  with  $U_i^{(j,\ell)} = F^{(j,\ell)}(Z_i^{(j,\ell)})$ . Define

$$\alpha_n^{(j)}(\mathbf{u}) = \sqrt{n} \left( G_n^{(j)}(u) - u \right), \quad G_n^{(j)}(u) = n^{-1} \sum_{i=1}^n \mathbf{1}_{\{U_i^{(j)} \leq u\}},$$

denote the (unobservable) empirical processes based on  $\mathbf{U}_1, \dots, \mathbf{U}_n$ .

For any sequence  $(\mathbf{Z}_n, n \in \mathbb{N})$ , let

$$\mathcal{F}_k = \sigma(\mathbf{Z}_n, n \leq k), \quad \text{and} \quad \mathcal{G}_k = \sigma(\mathbf{Z}_n, n \geq k),$$

be the natural filtration and "reverse" filtration of the sequence  $(\mathbf{Z}_n, n \in \mathbb{N})$ . Define

$$\beta(\mathcal{A}_1, \mathcal{A}_2) = \sup \frac{1}{2} \sum_{i,j \in I \times J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|,$$

where the sup is taken over all finite  $(A_i)_{i \in I}$  and  $(B_j)_{j \in J}$  of  $\Omega$  with the sets  $A_i$  (resp.  $B_j$ ) in the sigma field  $\mathcal{A}_1$  (resp.  $\mathcal{A}_2$ ). The  $\beta$ -mixing (or completely regular) coefficient is defined as

$$\beta(\ell) = \sup_{n \in \mathbb{N}} \beta(\mathcal{F}_n, \mathcal{G}_{n+\ell}). \quad (\text{C.3})$$

For the formulation of the consistency result for our estimator of the SECO, we need a couple of conditions over the regularity of the sequence  $(\mathbf{Z}_n, n \in \mathbb{N})$  which are the following:

**Condition A.** There exists an intermediary sequence  $m = m_n$  such that  $m_n = o(n)$  and  $\beta(m_n) \rightarrow 0$  as  $n \rightarrow \infty$ , where  $\beta$  is defined in (C.3).

**Condition B.** There exists some  $\theta_1 \in (0, 1/2]$  such that, for all  $\mu \in (0, \theta_1]$  and all sequences  $\delta_n \rightarrow 0$ , we have

$$M_n(\delta_n, \mu) := \sup_{|u-v| \leq \delta_n} \frac{|\alpha_n^{(j)}(u) - \alpha_n^{(j)}(v)|}{\max\{|u-v|^\mu, n^{-\mu}\}} = o_{\mathbb{P}}(1), \quad j = 1, \dots, d$$

Condition B can, for instance, be verified in the i.i.d. case with  $\theta_1 = 1/2$  or for  $\beta$ -mixing sequence with  $\beta(n) = o(a^n)$  as  $n \rightarrow \infty$  for some  $a \in (0, 1)$ , see Proposition 4.4 of [Berghaus et al. \(2017\)](#). Here we state the proposition that the actual appendix is devoted to prove.

**Proposition C.4.1.** *Let  $k = k_n$  be an intermediary sequence. Provided that  $k \rightarrow \infty$ ,  $k/n \rightarrow 0$  as  $n \rightarrow \infty$ , under the regularity condition over the sequence  $(\mathbf{Z}_n, n \in \mathbb{N})$  stated by Condition A and Condition B and  $\mathbf{Z}$  is in the max-domain of attraction of  $H$ , then*

$$\widehat{\text{SECO}}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \text{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}).$$

**Proof** Since convergence in probability remains stable through addition and subtraction, our task is to establish the consistency of each component:  $\hat{\theta}(a)$ ,  $\hat{\theta}(b)$ , and  $\hat{\theta}(a, b)$ . We address this concern in Lemma C.4.1 below, where, without loss of generality, we focus on demonstrating the consistency of  $\hat{\theta}(a)$ . Also, without loss of generality, we study the following estimator

$$\hat{\theta}(a) = \frac{1}{k} \sum_{i=1}^n \mathbf{1}_{\{\hat{F}_n^{(a,1)}(Z_i^{(a,1)}) > 1-k/n \text{ or } \dots \text{ or } \hat{F}_n^{(a,p)}(Z_i^{(a,p)}) > 1-k/n\}}, \quad (\text{C.4})$$

where  $\hat{F}_n^{(a,\ell)}$  denotes the empirical distribution function of  $Z_1^{(a,\ell)}, \dots, Z_n^{(a,\ell)}$  for  $\ell = 1, \dots, p_a$ . This estimator mirrors the one presented in (4.8), with the exception that we employ uniform margins instead of ranks, and we omit the constant factor of  $1/2$ , which, crucially, does not alter the estimator's asymptotic behavior.

**Lemma C.4.1.** *Under the conditions of Proposition C.4.1, we have*

$$\hat{\theta}(a) = \theta_q(a) + o_{\mathbb{P}}(1),$$

with

$$\theta_t(a) = \mathbb{P} \left\{ F^{(a,1)}(Z^{(a,1)}) > 1 - t \text{ or } \dots \text{ or } F^{(a,p)}(Z^{(a,p_a)}) > 1 - t \right\} / t,$$

and  $\lim_{t \rightarrow 0} \theta_t(a) = \theta(a)$ .

**Proof** Without confusions, we set in this proof  $p_a = p$  and that  $\mathbf{Z} = (Z^{(1)}, \dots, Z^{(p)}) := (Z^{(a,1)}, \dots, Z^{(a,p)})$  is a  $p$ -dimensional random vector. We begin by introducing some useful notations. In the same spirit, define the random variable  $U_i^{(j)} = F^{(j)}(Z_i^{(j)})$  (here  $Z_i^{(j)}$  denote the  $j$ th entry of the vector  $\mathbf{Z}_i$ ) and the vectors  $\mathbf{U}_i := (U_i^{(1)}, \dots, U_i^{(p)})$  with stationary distribution  $C$ . Denote by  $\hat{G}_n^{(j)}$  the empirical distribution of  $U_1^{(j)}, \dots, U_n^{(j)}$ . Define the vector  $\hat{G}_n^{\leftarrow}(\mathbf{x}) = ((\hat{G}_n^{(1)})^{\leftarrow}(x^{(1)}), \dots, (\hat{G}_n^{(p)})^{\leftarrow}(x^{(p)}))$ , the function

$$C_n^o(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^{(1)} \leq x^{(1)}, \dots, U_i^{(p)} \leq x^{(p)}\}}$$

and

$$\hat{C}_n \left( 1 - \frac{k\mathbf{x}}{n} \right) := C_n^o \left( \hat{G}_n^{\leftarrow} \left( 1 - \frac{k\mathbf{x}}{n} \right) \right)$$

Note that the estimator  $\hat{\theta}$  depends only on the marginals ranks of  $Z_i^{(j)}$  with  $j = 1, \dots, p$ ; thus we have almost surely

$$|\hat{\theta}(a) - \theta_t(a)| = \left| \frac{n}{k} \hat{C}_n(1 - k/n, \dots, 1 - k/n) - t^{-1} C(1 - t, \dots, 1 - t) \right|$$

Standard arguments gives that under  $k \rightarrow \infty$  and  $k/n \rightarrow t \in (0, 1)$  and  $n \rightarrow \infty$ , the right hand side of the latter equation is equal to

$$\left| \frac{n}{k} \left( \hat{C}_n \left( 1 - \frac{k}{n}, \dots, 1 - \frac{k}{n} \right) - C \left( 1 - \frac{k}{n}, \dots, 1 - \frac{k}{n} \right) \right) \right| + o_{\mathbb{P}}(1).$$

Now, we can bound the first term by

$$\frac{n}{k} \left| C_n^o(\hat{G}_n^{\leftarrow}(1 - k\mathbf{1}/n)) - C(\hat{G}_n^{\leftarrow}(1 - k\mathbf{1}/n)) \right| + \frac{n}{k} \left| C(\hat{G}_n^{\leftarrow}(1 - k\mathbf{1}/n)) - C(1 - k\mathbf{1}/n) \right|.$$

Using Lipschitz continuity of  $C$ , we obtain the following upper bound:

$$\frac{n}{k} \|C_n^o - C\|_{\infty} + \frac{n}{k} \sum_{j=1}^d \|u_n^{(j)}\|_{\infty} \quad (\text{C.5})$$

where  $\|\cdot\|_{\infty}$  is the uniform norm and  $u_n^{(j)}(u) = (\hat{G}_n^{(j)})^{\leftarrow}(u) - u$ ,  $u \in [0, 1]$ .

By Berbee's coupling Lemma (Bücher and Segers (2014); Doukhan et al. (1995b)), one can construct inductively a sequence  $(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m})_{i \geq 0}$  such that the following three properties hold:

- (i)  $(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m}) \stackrel{d}{=} (\mathbf{Z}_{im+1}, \dots, \mathbf{Z}_{im+m})$  for any  $i \geq 0$ ;
- (ii) both  $(\bar{\mathbf{Z}}_{2im+1}, \dots, \bar{\mathbf{Z}}_{2im+m})_{i \geq 0}$  and  $(\bar{\mathbf{Z}}_{(2i+1)m+1}, \dots, \bar{\mathbf{Z}}_{(2i+1)m+m})_{i \geq 0}$  sequences are independent and identically distributed;

$$(iii) \mathbb{P}\{(\bar{\mathbf{Z}}_{im+1}, \dots, \bar{\mathbf{Z}}_{im+m}) \neq (\mathbf{Z}_{im+1}, \dots, \mathbf{Z}_{im+m})\} \leq \beta(m).$$

Let  $\bar{C}_n^o$  be defined analogously to  $C_n^o$  but with  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$  replaced by  $\bar{\mathbf{Z}}_1, \dots, \bar{\mathbf{Z}}_n$ . Now write

$$C_n^o(\mathbf{u}) - C(\mathbf{u}) = \left\{ C_n^o(\mathbf{u}) - \bar{C}_n^o(\mathbf{u}) \right\} + o_{\mathbb{P}}(1). \quad (C.6)$$

The term in brackets in the right hand side is  $o_{\mathbb{P}}(1)$  uniformly in  $\mathbf{u}$ , since

$$|C_n^o(\mathbf{u}) - \bar{C}_n^o(\mathbf{u})| \leq \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{\bar{\mathbf{Z}}_i \neq \mathbf{Z}_i\}}.$$

Hence by Markov's inequality, for any  $\epsilon > 0$

$$\mathbb{P} \left\{ \sup_{\mathbf{u} \in [0,1]^q} |C_n^o(\mathbf{u}) - \bar{C}_n^o(\mathbf{u})| > \epsilon \right\} \leq \frac{\beta(m)}{\epsilon}.$$

By Condition [A](#), we obtain that the first summand in brackets in (C.6) is  $o_{\mathbb{P}}(1)$  as  $n \rightarrow \infty$ , uniformly in  $\mathbf{u} \in [0, 1]^q$ . We obtain that

$$\|C_n^o - C\|_{\infty} = o_{\mathbb{P}}(1).$$

We also have

$$\sup_{u \in [0,1]} |u_n^{(j)}(u)| \leq \sup_{u \in [0,1]} |\alpha_n^{(j)}(u)| + \frac{n \sup_{u \in [0,1]} |\hat{G}_n^{(j)}(u) - \hat{G}_n^{(j)}(u-)| - 1}{n}.$$

To understand this, we start by recognizing that the maximum of either  $\alpha_n^{(j)}(\cdot)$  or  $-\alpha_n^{(j)}(\cdot)$ , and consequently,  $|\alpha_n^{(j)}(\cdot)|$ , must occur at one of the discontinuities in  $\hat{G}_n^{(j)}$ . These discontinuities correspond to the values  $\{U_{i:n}^{(j)}, 1 \leq i \leq n\}$ , where  $U_{1:n}^{(j)} \leq \dots \leq U_{n:n}^{(j)}$  represent the order statistics. Hence, the quantity

$$n \times (u_n^{(j)}(i/n) + \alpha_n^{(j)}(U_{i:n}^{(j)})) \quad (C.7)$$

is equal to the highest count of  $U_i^{(j)}$  that are equal to  $U_{i:n}^{(j)}$  minus 1. Assuming there are no ties among  $U_1^{(j)}, \dots, U_n^{(j)}$  (which, for example, happens in the i.i.d. case), this expression equals 1. Consequently, we derive the classical identity for the uniform quantile process, as outlined in (Csörgő, 1983, Section 1.4) or (Shorack and Wellner, 2009, Chapter 3).

$$\sup_{0 \leq u \leq 1} |\alpha_n^{(j)}(u)| = \sup_{0 \leq u \leq 1} |u_n^{(j)}(u)|.$$

In the general case, Equation (C.7) is limited above by the maximum count of  $U_i^{(j)}$  which are equal minus 1. It is worth noting that this maximum count can be expressed as

$$n \times \sup_{u \in [0,1]} |\hat{G}_n^{(j)}(u) - \hat{G}_n^{(j)}(u-)|.$$



We have, following the proof of lemma 4.6 in Berghaus et al. (2017)

$$\begin{aligned} \sup_{u \in [0,1]} |\hat{G}_n^{(j)}(u) - \hat{G}_n^{(j)}(u-)| &\leq \sup_{u,v \in [0,1], |u-v| \leq 1/n} |\hat{G}_n^{(j)}(u) - \hat{G}_n^{(j)}(v)| \\ &\leq \sup_{u,v \in [0,1], |u-v| \leq 1/n} |\hat{G}_n^{(j)}(u) - \hat{G}_n^{(j)}(v) - (u-v)| + \frac{1}{n} \\ &\leq \frac{1}{\sqrt{n}} \sup_{u,v \in [0,1], |u-v| \leq 1/n} |\alpha_n^{(j)}(u) - \alpha_n^{(j)}(v)| + \frac{1}{n}. \end{aligned}$$

Using Condition  $\mathcal{B}$ , the above term is  $o_{\mathbb{P}}(n^{-1/2-\mu})$  for  $\mu \in (0, \theta_1)$  and  $\theta_1 \in [0, 1/2]$ . Additionally, using  $\|C_n^o - C\|_{\infty} = o_{\mathbb{P}}(1)$ , we obtain

$$\|u_n^{(j)}\|_{\infty} = o_{\mathbb{P}}(1).$$

Thus,  $\forall t \in (0, 1)$

$$\hat{\theta}_n(a) = \theta_t(a) + o_{\mathbb{P}}(1).$$

Since  $F$  is in the max-domain of attraction of  $H$ , we have

$$\theta(a) = \lim_{t \rightarrow 0} \theta_t(a).$$

Hence the result of the lemma. □

Since convergence in probability is reliably preserved under continuous transformations, we attain the outcomes outlined in Proposition C.4.1. □

## C.5 Definition of the Adjusted Rand Index (ARI)

The ARI is computed as follows: Let  $O = \{O_g\}_{g=1, \dots, G}$  and  $S = \{S_h\}_{h=1, \dots, H}$  be two partitions with  $d$  entities, and let  $d_{gh}$  be the number of entities in cluster  $O_g$  in partition  $O$  and in cluster  $S_h$  in partition  $S$ . Denote by  $d_{g\cdot}$  (resp.  $d_{\cdot h}$ ) the number of entities in cluster  $O_g$  (resp.  $S_h$ ) in partition  $O$  (resp.  $S$ ). The ARI is evaluated using the following expressions:

$$\begin{aligned} r_0 &= \sum_{g=1}^G \sum_{h=1}^H \binom{d_{gh}}{2}, & r_1 &= \sum_{g=1}^G \binom{d_{g\cdot}}{2}, & r_2 &= \sum_{h=1}^H \binom{d_{\cdot h}}{2}, & r_3 &= \frac{2r_1 r_2}{d(d-1)}, \\ \text{ARI}(O, S) &= \frac{r_0 - r_3}{0.5(r_1 + r_2) - r_3}, \end{aligned}$$

where  $\binom{n}{k}$  is the binomial coefficient.

## C.6 Supplementary Figures

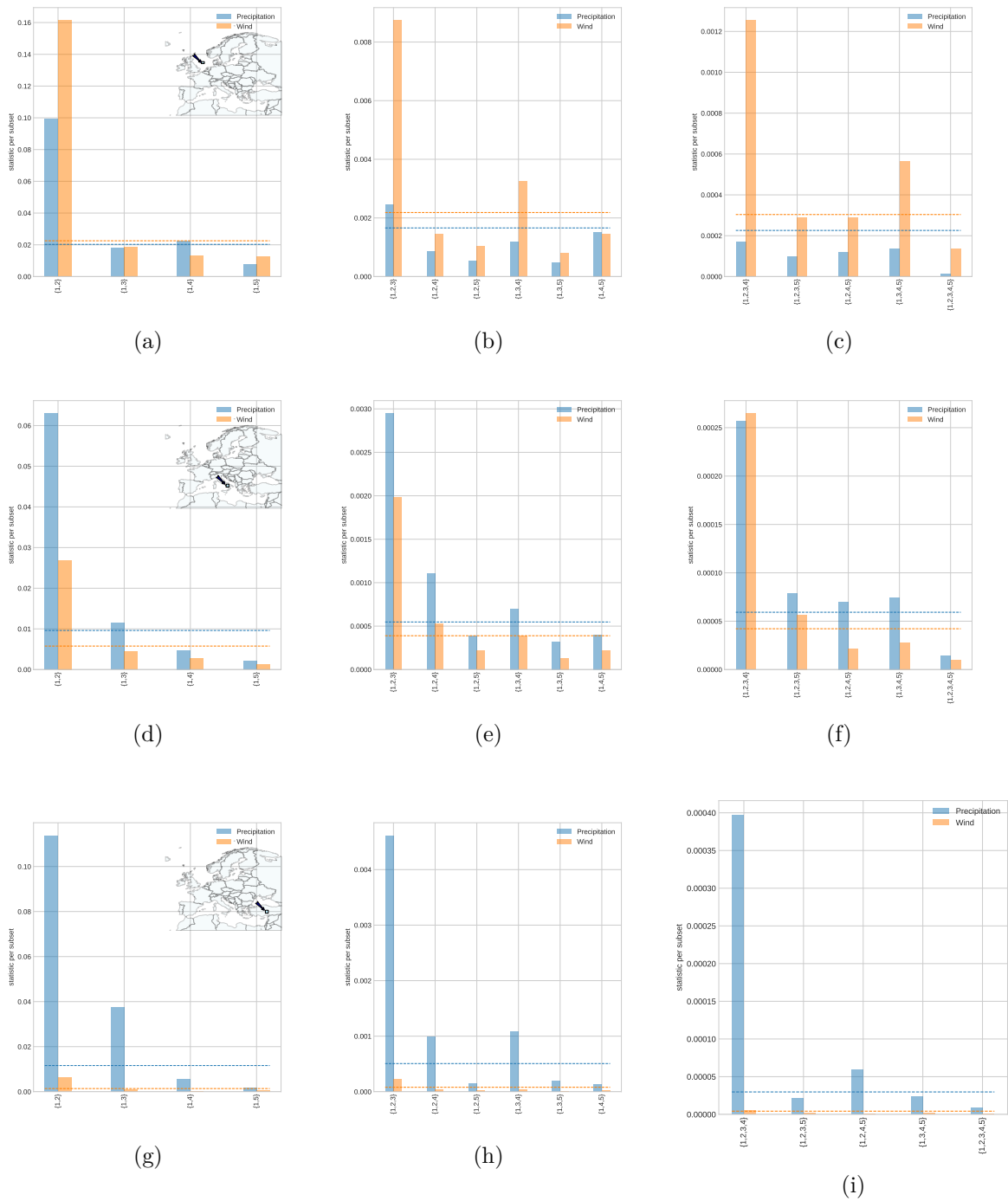


Fig. 4 The dependogram provides a concise summary of randomness test results conducted on daily total precipitation and wind speed maxima in the ERA5 dataset. This study covers three European regions, examining 304 days of observations across nine distinct pixels. The first column displays a map with each row representing an area and the nine pixels marked by red squares. Test statistics are represented by bars, and critical values are depicted by dotted horizontal lines. The dependogram columns focus on pairwise, three-wise, and four-wise randomness tests with a lag of 4.

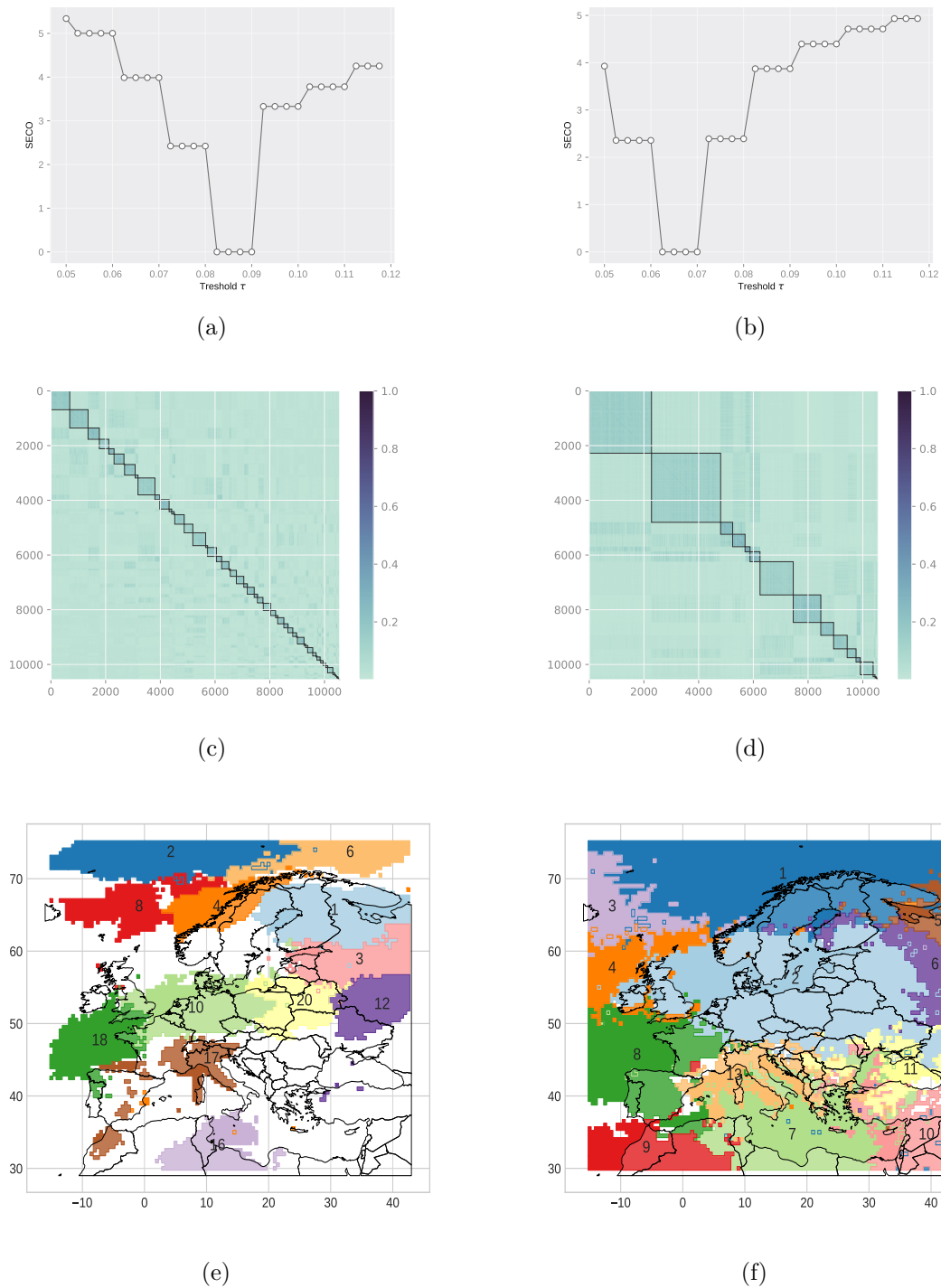


Fig. 5 Value of the function  $L$  for different values of  $\tau \in \Delta = \{0.05, 0.0525, \dots, 0.12\}$  in Panels 5a (Precipitation) and 5b (Wind). Partitions of the extremal correlation similarity matrix with threshold  $\tau = 0.09$  for Panel 5c (Precipitation) and  $\tau = 0.07$  5d (Wind). Squares represent the clusters of variables. Representation of the 12 largest clusters (in decreasing order) of the partition of the extremal correlation matrix of total precipitation and wind speed maxima with threshold  $\tau = 0.09$  and  $\tau = 0.07$ , respectively in Panels 5e and 5f. Number of each cluster is depicted at the top-left corner of the corresponding panel.

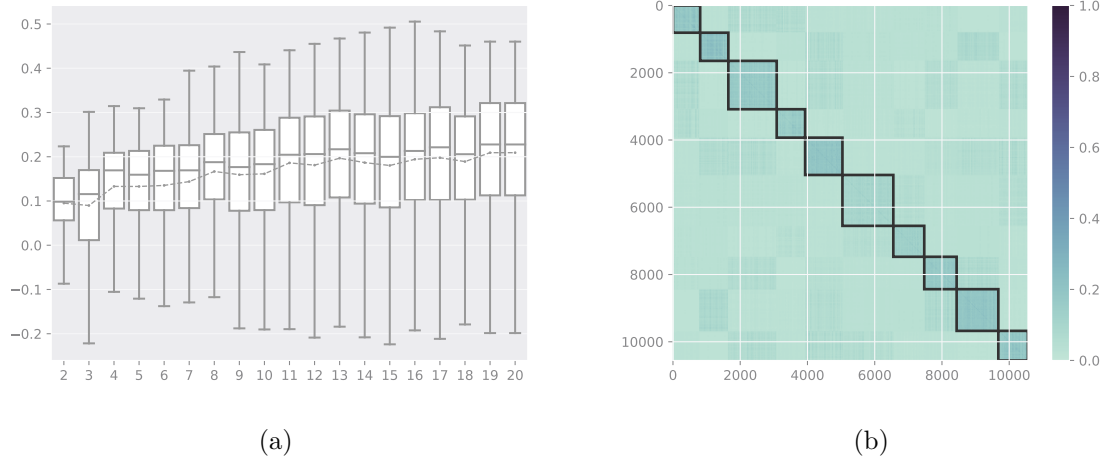


Fig. 6 Boxplots of the silhouette coefficients for different values of  $K$  using the quantization-based algorithm. Thick lines indicate the median, boxes the interquartile range and whiskers the full range of the distribution. Partitions of the SECO similarity matrix with  $K = 10$  for the quantization-based approach. Squares represent the clusters of variables.

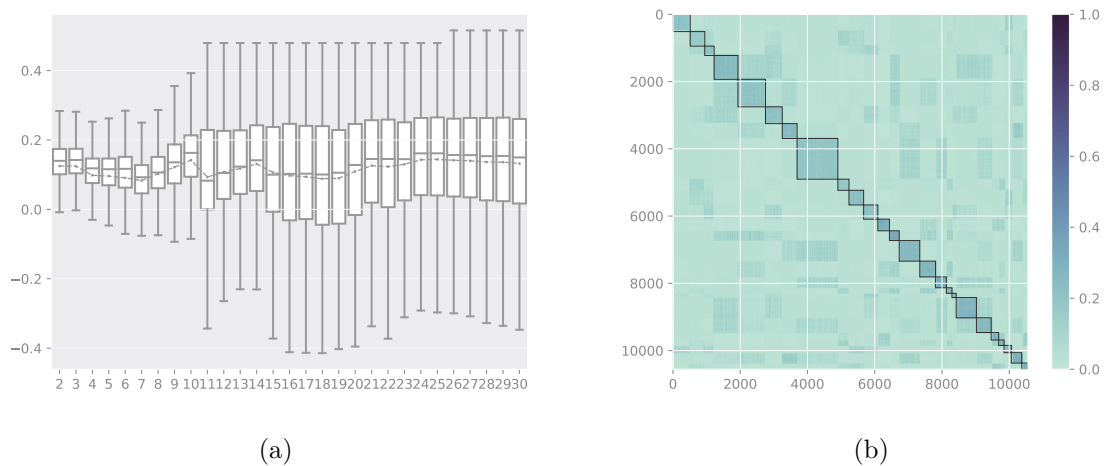


Fig. 7 Boxplots of the silhouette coefficients for different values of  $K$  using the hierarchical clustering algorithm. Thick lines indicate the median, boxes the interquartile range and whiskers the full range of the distribution. The average silhouette is depicted by the dotted line. Partitions of the SECO similarity matrix with  $K = 25$ . Squares represent the clusters of variables.

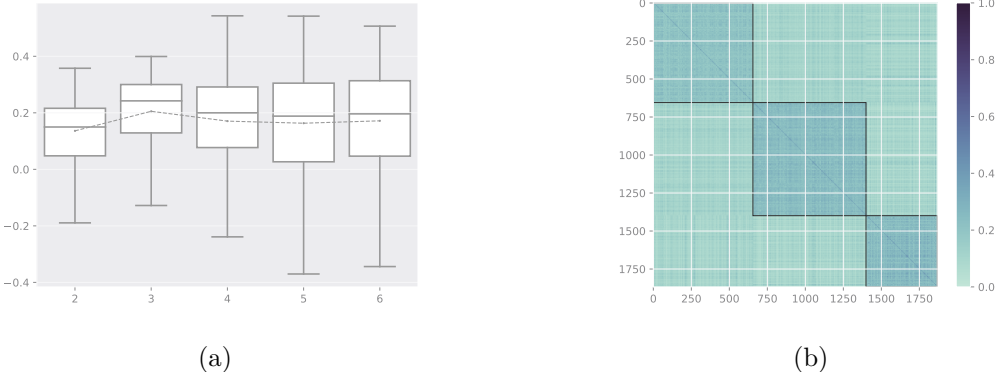


Fig. 8 Boxplots of the silhouette coefficients for different values of K performed by a hierarchical clustering algorithm on the fourth partition given by Algorithm (ECO). Thick lines indicate the median, boxes the interquartile range and whiskers the full range of the distribution. The average silhouette is depicted by the dotted line. The clustered matrix SECO is represent in Panel 8b where squares represent the clusters.

## CHAPTER 5

# ESTIMATING REGULARLY VARYING RANDOM VECTORS WITH DISCRETE SPECTRAL MEASURE USING MODEL-BASED CLUSTERING

This chapter is based on the following work, which will soon be submitted:

 Alexis Boulin (2024), Estimating Max-Stable Random Vectors with Discrete Spectral Measure using Model-Based Clustering.

### Abstract.

This study introduces a novel estimation method for the entries and structure of a matrix  $A$  in the linear factor model  $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$ . This is applied to an observable vector  $\mathbf{X} \in \mathbb{R}^d$  with  $\mathbf{Z} \in \mathbb{R}^K$ , a vector composed of independently regularly varying random variables, and lighter tail noise  $\mathbf{E} \in \mathbb{R}^d$ . This leads to max-linear models treated in classical multivariate extreme value theory. The spectral measure of the regularly varying random vector  $\mathbf{X}$  is subsequently discrete and completely characterised by the matrix  $A$ . It follows that the behaviour of its maxima can be modelled by a max-stable random vector with discrete spectral measure. Every max-stable random vector with discrete spectral measure can be written as a max-linear model. Each row of the matrix  $A$  is supposed to be both scaled and sparse. Additionally, the value of  $K$  is not known a priori. The problem of identifying the matrix  $A$  from its matrix of pairwise extremal correlation is addressed. In the presence of pure variables, which are elements of  $\mathbf{X}$  linked, through  $A$ , to a single latent factor, the matrix  $A$  can be reconstructed from the extremal correlation matrix. Our proofs of identifiability are constructive and pave the way for our innovative estimation for determining the number of factors  $K$  and the matrix  $A$  from  $n$  weakly dependent observations on  $\mathbf{X}$ . We apply the suggested method to weekly maxima rainfall and wildfires to illustrate its applicability.

## 5.1 Introduction

In this current study, our aim is to estimate the  $d \times K$  loading matrix  $A$ , which might exhibit sparsity, and serves as the parameter for the decomposition of an observable random vector  $\mathbf{X}$ . This can be expressed as

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}. \quad (5.1)$$

In this equation,  $\mathbf{Z}$  represents an unobservable,  $K$ -dimensional random vector, serving as an underlying latent factor,  $\mathbf{E} \in \mathbb{R}^d$  as a unobservable random noise. The precise count of factors,

$K$ , remains undisclosed and both  $d$  and  $K$  are permitted to increase and be larger than  $n$ , the number of observations. To establish the foundation of our framework inside extreme value theory, we assume that  $\mathbf{Z}$  comprises of asymptotic independent random variables characterised by a tail index  $\alpha$ , for the purposes of our study, we will set this tail index to a fixed value of  $\alpha$  equal to unity. As per the construction, the vector  $\mathbf{Z}$  is regularly varying with the subsequent exponent measure

$$\Lambda_{\mathbf{Z}} = \sum_{k=1}^K \delta_0 \otimes \cdots \otimes \Lambda_{Z^{(k)}} \otimes \cdots \otimes \delta_0, \quad \Lambda_{Z^{(k)}}(dy) = y^{-2} dy.$$

The random noise vector  $E \in \mathbb{R}^d$  is postulated to possess a distribution with a tail that is lighter than that of the associated factors. Hence,  $\mathbf{X}$  is also regularly varying which can be equivalently described by the existence of an angular measure  $\Phi$  where the following weak convergence holds true on the positive unit sphere for an arbitrary norm  $\|\cdot\|$  on  $\mathbb{R}^d$ ,

$$\lim_{x \rightarrow \infty} \mathbb{P} \left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in \cdot \mid \|\mathbf{X}\| > x \right\} = \Phi(\cdot),$$

where  $\Phi$  has the discrete representation

$$\Phi(\cdot) = \sum_{k=1}^K \|A_{\cdot,k}\| \delta_{\frac{A_{\cdot,k}}{\|A_{\cdot,k}\|}}(\cdot), \quad (5.2)$$

with  $\delta_x(\cdot)$  the Dirac measure that puts unit mass at  $x$  and  $A_{\cdot,k}$  is the  $k$ th column of the matrix  $A$ . Taking  $\mathbf{X}_1, \dots, \mathbf{X}_m$ ,  $m$  i.i.d. replications of  $\mathbf{X}$  in (5.1), it follows that, see for example (Kulik and Soulier, 2020, Theorem 2.1.6) for details,

$$\lim_{m \rightarrow \infty} \mathbb{P} \left\{ \bigvee_{i=1}^m c_m^{-1} X_i^{(j)} \leq x^{(j)}, j = 1, \dots, d \right\} = e^{-\sum_{a=1}^K V_{j=1}^d \frac{A_{ja}}{x^{(j)}}}, \quad \mathbf{x} \geq 0,$$

where  $c_m$  is a scaling sequence. The limiting distribution on the right is max-stable, which means there exist  $\mathbf{a}_m > 0$  and  $\mathbf{b}_m$  such that  $H^m(\mathbf{a}_m \mathbf{x} + \mathbf{b}_m) = H(\mathbf{x})$  for any  $m \in \mathbb{N}$  with  $H$  a cumulative distribution function. Furthermore, under our model in (5.1) it has the property of having a discrete angular measure. Throughout this chapter, we will refer to a random vector with a max-stable distribution as a max-stable random vector.

The expression (5.1) can also be considered as a linear adaptation of the max-linear models, sharing the same angular measure  $\Phi$ . This essentially follows from the fact that the ratio of the probabilities of the sum and the maximum of the  $A_{ij}Z^{(j)}$  exceeding  $x$  tends to 1 as  $x \rightarrow \infty$  (see (Embrechts et al., 2013, page 38) or (Kulik and Soulier, 2020, Example 2.2.8, Example 2.2.9)). Each max-stable distribution with a discrete spectral measure is inherently max-linear, see (Fougères et al., 2013, Section 3.1). The max-linear model, in turn, is dense in the class of  $d$ -dimensional multivariate extreme value distribution (Fougères et al. (2013)). Consequently, any multivariate max-stable vector can be finely approximated through a max-linear model provided that  $K$  is large. Additionally, Cooley and Thibaud (2019) established the existence of a finite natural number, denoted as  $q \in \mathbb{N}$ , such that the tail pairwise dependence matrix  $\Sigma_{\mathbf{X}}$  for any multivariate regularly varying random vector  $\mathbf{X}$  with a tail index  $\alpha = 2$  is equivalent to that

of a max-linear model with  $q$  factors. This equivalence is expressed through the relationship  $\Sigma_{\mathbf{X}} = AA^\top$ .

Corresponding max-linear models, have also been explored in the field of time series for extremes [Davis and Resnick \(1989\)](#); [Hall et al. \(2002\)](#). More recently, they have found applications in the domain of structural equation models [Gissibl and Klüppelberg \(2018\)](#); [Klüppelberg and Lauritzen \(2019\)](#), as well as in the context of clustering extremes [Avella-Medina et al. \(2021, 2022\)](#); [Janßen and Wan \(2020\)](#). Factor models of this kind find widespread use across diverse applications. For instance, they are often employed to represent underlying factors that influence financial returns ([Cui and Zhang \(2018\)](#)) as well in environmental sciences ([Kiriliouk and Zhou \(2022\)](#)).

**Outline of the literature.** Estimating parameters in linear factor models poses a difficult task, primarily because there is no spectral density that rules out standard maximum likelihood procedures. Instead, [Einmahl et al. \(2012\)](#) and [Einmahl et al. \(2018\)](#) opt for a least square estimator based on the stable tail dependence function to tackle this task. [Avella-Medina et al. \(2021\)](#); [Janßen and Wan \(2020\)](#) propose spectral clustering algorithms designed for extremes employing its output to estimate  $A_{.1}/\|A_{.1}\|, \dots, A_{.K}/\|A_{.K}\|$  and  $\|A_{.1}\|/w, \dots, \|A_{.K}\|/w$ . These parameters characterise the angular measure of the linear factor model. However, this approach falls short in estimating the matrix  $A$  which can be crucial for practical interpretation and computing failure sets (see Section 5.6.1). Additionally, these methods face limitations in higher dimensions, grappling running time difficulties or curse of dimension. Moreover these methods also assume that the number  $K$  is known a priori, a requirement that is often scarcely fulfilled in practical scenario. Addressing this hurdle, additional methods, as proposed by [Avella-Medina et al. \(2021, 2022\)](#), introduce a procedure coupled with the so-called screeplot to aid in the selection of the elusive number  $K$ . Despite the practical utility of such an approach, the theoretical underpinnings supporting these findings are still in their early stage of development. To our current understanding, methods for estimating  $A$  in higher dimensions have emerged specifically under the condition of a squared matrix  $A \in \mathbb{R}^{d \times d}$ . Notably, these methods have found fruitful application in contexts characterised by moderate dimensions. For instance, in Directed Acyclic Graph, [Klüppelberg and Krali \(2021\)](#) have made noteworthy contributions, while [Kiriliouk and Zhou \(2022\)](#) have demonstrated successful applications of their estimator in environmental and financial dataset. However, it is crucial to acknowledge that these achievements are contingent upon the specific conditions and dimensions involved. A noteworthy recent paper worth emphasising is [Zhang et al. \(2023\)](#). The paper investigates minimax risk bounds for estimators of the spectral measure in multivariate linear factor models, particularly when the number of latent factors  $K$  exceeds the dimension  $d$  and the latter is fixed. Foremost, a critical lens on the theoretical foundations reveals a reliance on a i.i.d. sample and the asymptotic framework in the mentioned literature. The assumption of serial independence may face scrutiny when these methods are extended to environmental datasets, where deviations from serial independence are legitimately suspected. Moreover, the asymptotic framework, with a fixed arbitrary dimension  $d$  while the sample size  $n \rightarrow \infty$ , may offer limited insights into the performance of estimation processes in high-dimensional setting, i.e.,  $d$  vary with  $n$  and might even surpass the sample size.

**Our contribution.** Drawing inspiration of [Bing et al. \(2020\)](#), we propose a model-based clustering via  $A$  with the crucial distinction that the covariance matrix of  $\mathbf{X}$  does not exist in



our model. Within the framework of model (5.1), we consider two components, namely  $X^{(i)}$  and  $X^{(j)}$  belonging to the vector  $\mathbf{X}$ , as akin if they share a non-zero association. This association is established through the intermediary of the matrix  $A$ , connecting them to a common latent factor  $Z^{(a)}$ . Variables exhibiting this similarity are grouped together within the cluster denoted as  $G_a$ :

$$G_a = \{j \in \{1, \dots, d\} : A_{ja} \neq 0\}, \quad \text{for each } a \in \{1, \dots, K\}. \quad (5.3)$$

Given that  $X^{(j)}$  can potentially be linked to multiple latent factors, the resulting clusters are characterised by overlap. In terms of terminology, groups that may become large without the others are called extreme directions. More precisely, if  $J \subset \{1, \dots, d\}$  is such that the components  $(X^{(j)})_{j \in J}$  can be large simultaneously while the other components  $(X^{(j)})_{j \in [d] \setminus J}$  are small, then  $J$  defines an extreme direction (see Simpson et al. (2020) for a precise definition). By examining (Mourahib et al., 2023, Example 3.7), one can observe that soft clusters in Equation (5.3) also represent the extreme directions.

In this endeavor, our focus centers on presenting a model-based clustering approach through the utilisation of  $A$ . It is worth noting, however, that the definition of  $A$  within model (5.1) lacks uniqueness without imposing additional constraints. To address this, we contemplate a variant of model (5.1) where in every rows of  $A$  undergoes scaling. To be specific, we posit the following assumption:

**Condition (i).**  $\sum_{a=1}^K A_{ja} = 1$ .

The weights  $A_{j1}, \dots, A_{jK}$  indicate the degree to which component align with each cluster. This condition dives our model into both hard and soft clustering.

Condition (i), if not explicitly specified, fails to guarantee the identifiability of  $A$  in model (5.1) solely based on the extremal correlation matrix of  $\mathbf{X}$ . For a more comprehensive understanding, we refer to the insights provided in Remark 5.2.2. Additionally, to employ the model effectively for clustering purposes, it is important to circumvent the trivial scenario where each component  $X^{(j)}$  is associated with all latent factors. To adress this concern, we permit the row of  $A_j = (A_{j1}, \dots, A_{jK})$  to exhibit sparsity for  $j$  in the range of  $\{1, \dots, d\}$ . However, it is noteworthy that this property is not required for establishing the identifiability of  $A$ .

Furthermore, relying solely on Condition (i) is insufficient to guarantee the uniqueness of  $A$  within model (5.1), see Remark 5.2.3 hereafter.

We term the following condition, denoted as (ii), the ‘‘pure variable assumption’’. In simple terms, this assumption posits the presence of at least one pure variables  $X^{(j)}$ , among the components of  $\mathbf{X}$ . These pure variables are uniquely associated with a single latent factor and no other.

**Condition (ii).** For every  $a \in \{1, \dots, K\}$ , there exists at least one indice  $j \in \{1, \dots, d\}$  such that  $A_{ja} = 1$  and  $A_{jb} = 0, \forall b \neq a$ .

Cluster denoted as  $G_a$ , established in accordance with Equation (5.3), derive their definition from the unobservable factor  $Z^{(a)}$ . In this context, a pure variable  $X^{(j)}$  serves as an observable representation of  $Z^{(a)}$ , contributing to the elucidation of the ambiguous nature of cluster  $G_a$ . Moreover, a more stringent version of Condition (ii) has a rich history, specifically

**Condition (ii’)**. For every  $a \in \{1, \dots, K\}$ , there exist at least one known indice  $j \in \{1, \dots, d\}$  such that  $A_{ja} = 1$  and  $A_{jb} = 0, \forall b \neq a$ .

This condition stands out as one of the few interpretable parametrisations of  $A$ , effectively eliminating the ambiguity associated with latent factors. In psychology, the variables generated purely through the parametrisation as referred to as factorial simple items (McDonald (1999)). A comparable condition find its roots in the topic modeling literature, where the identifiability of topics is assured under the assumption that anchor words exists, i.e., words that exclusively appear in one topic. In hydrology, the concept of pure variables was utilised to pinpoint catchments stations as representatives of pollution sources in (Tolosana-Delgado et al., 2005, page 700). In section 5.2, we demonstrate that, under Conditions (i) and (ii), the matrix  $A$  can be recovered solely through the use of bivariate measures, namely extremal correlations. In Section 5.3, we develop (SCRAM), a new soft clustering algorithm specifically for linear factor model that estimate the loading matrix  $A$  and the overlapping groups. We provide a sparse estimator  $\hat{A}$  of  $A$  that is tailored to our model specifications. Our approach follows the constructive techniques used in our identifiability proofs. We first construct an estimator  $\hat{I}$ , an estimator of the pure variable set  $I$ , and  $\hat{K}$ , an estimator of the number of clusters  $K$ . These are used to estimate the rows in  $A$  corresponding to pure variables. The remaining rows of  $A$  are estimated via an easily implementable program that is tailored to this problem. We base our theoretical study on mixing conditions over the studied process. These conditions make explicit the independence between “past” and “future”; meaning that the “past” is progressively forgotten. Mixing conditions are consequently more adapted to work in areas like finance or climate sciences where history is of considerable importance. To make this more precise, we consider processes with exponentially decaying strong mixing coefficients, as introduced in Section 5.4. The algorithm (SCRAM) recovers the number of latent variables with high probability under a strong signal condition in Section 5.4.1. We establish an upper bound on the  $L_2$  norm ( $L_2(\hat{A}, A)$ ), as defined in Section 5.4.2) for the matrix  $A$  specified by model (5.1) and subject to Conditions (i)-(ii), as discussed in Section 5.4.2. A control of cluster or, equivalently stated, extreme direction recovery is also given in Section 5.4.2.

**Notations** All bold letters  $\mathbf{x}$  correspond to vector in  $\mathbb{R}^d$ . The notation  $\delta_x$  corresponds to the Dirac measure at  $x$ . Throughout this paper, we are concerned with a simple undirected graph  $G = (V, E)$  with a finite set  $V$  of vertices and a finite set  $E$  of unordered pairs  $(v, w)$  of distinct vertices, called edges, we denote by  $\bar{E}$  its complementary adjacent matrix. A pair of vertices  $v$  and  $w$  are said to be adjacent if  $(v, w) \in E$ . For the subset  $W \subseteq V$  of vertices,  $G(W) = (W, E(W))$  with  $E(W) = \{(v, w) \in W \times W \mid (v, w) \in E\}$  is called a subgraph of  $G = (V, E)$  induces by  $W$ . Given the subset  $Q \subseteq V$  of vertices, the induced subgraph  $G(Q)$  is said to be complete if  $(v, w) \in E$  for all  $v, w \in Q$  with  $v \neq w$ . In this case, we may simply state that  $Q$  is complete subgraph. A complete subgraph is also called a clique. A clique is maximum if its cardinality is the largest among all the cliques of the graph.

## 5.2 Identifiability

Within this section, we present a demonstration that the allocation matrix  $A$ , as defined by model (5.1) and subject to conditions (i)-(ii), is identifiable, within the exception of multiplication by a permutation matrix.

As per the construction, the vector  $\mathbf{Z}$  is regularly varying, it possesses an extremal correlation matrix represented by  $I_K$ , the identity matrix. Consequently, we deduce that the vector  $\mathbf{X}$  also follows a pattern of regular variation, leading to the presence of an extremal correlation matrix denoted as  $\mathcal{X} = [\chi(i, j)]_{i=1, \dots, d; j=1, \dots, d}$ , where

$$\chi(i, j) = \lim_{x \rightarrow \infty} \frac{\mathbb{P}\{X^{(i)} > x, X^{(j)} > x\}}{\mathbb{P}\{X^{(j)} > x\}}.$$

The subsequent theorem is poised to demonstrate that the extremal correlation matrix can be elegantly formulated using exclusively the loading matrix  $A$ . However, before going further, we introduce a novel matrix operation defined over matrices  $A \in \mathcal{M}_{p,K}(\mathbb{R})$  and  $B \in \mathcal{M}_{K,q}(\mathbb{R})$ . Here, the notation  $\mathcal{M}_{p,q}(\mathbb{R})$  refers to the collection of matrices encompassing  $p$  rows and  $q$  columns, with coefficients in the real number domain.

**Definition 5.2.1.** We call  $\odot$  the application:

$$\begin{aligned} \odot: \mathcal{M}_{p,K}(\mathbb{R}) \times \mathcal{M}_{K,q}(\mathbb{R}) &\longrightarrow \mathcal{M}_{p,q}(\mathbb{R}) \\ (a_{ik}, b_{mj}) &\mapsto c_{ij}, \end{aligned}$$

where

$$c_{ij} = a_{i1} \wedge b_{1j} + \dots + a_{iK} \wedge b_{Kj}.$$

With all the essential tools at our disposal, we are ready to present the ensuing fundamental theorem.

**Theorem 5.2.1.** *Let  $\mathbf{X}$  defined in (5.1) and  $A$  satisfies Condition (i). Then  $\mathbf{X}$  is regularly varying and its extremal correlation  $\mathcal{X}$  can be written as*

$$\mathcal{X} = A \odot A^\top,$$

with

$$\chi(i, j) = \sum_{k=1}^K A_{ik} \wedge A_{jk}.$$

For any loading matrix  $A$  that adheres to model (5.1), we can subdivide the set  $[d] = \{1, \dots, d\}$  into two distinct non-overlapping segments:  $I$  and its complement, designated as  $J$ . Within each row  $A_i$  of  $A_I$ , there exists precisely at least one value  $a \in [K]$  for which  $A_{ia} = 1$ . We assign the term “pure variable set” to  $I$ , while  $J$  corresponds to the “non-pure variable set”. To be more specific, for any given matrix  $A$ , the pure variable set is outlined as follows

$$I(A) = \cup_{a=1}^K I_a, \quad I_a := \{i \in [d] : A_{ia} = 1, A_{ib} = 0 \forall b \neq a\}. \quad (5.4)$$

In Equation (5.4), we use the notation  $I(A)$  to underscore that the pure variables set finds its definition in relation to matrix  $A$ . Moving forward, we will omit this explicit statement whenever there is no ambiguity. Additionally, it is worth mentioning that the sets  $\mathcal{I} := \{I_a\}_{1 \leq a \leq K}$  constitute a partition of the pure variable set  $I$ .

To establish the identifiability of matrix  $A$ , our task is simplified by focusing on distinct identifiability of  $A_I$  and  $A_J$ , each with allowance for a transformation by a permutation matrix.

With respect to the definition of  $A_I$ , its identifiability is assured as long as the partition of the pure variable set  $I$  remains identifiable. The heart of the challenge lies in the identifiability of set  $I$  and the inherent issue of distinguishing between  $I$  and  $J$ , based solely on the distribution of the vector  $\mathbf{X}$ . This stands as the central hurdle of the problem. Theorem 5.2.2 holds a central position in our discussion. In the first part (a), it offers both a necessary and sufficient description of the set  $[K]$  by examining the extremal correlation matrix  $\mathcal{X}$ . In the second part (b), it provides a necessary and sufficient characterisation of the set  $I$  when the cardinality of  $I_a$  is greater than one. Finally, in the third part (c), it illustrates that both the set  $I$  and its partition into subsets  $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$  can be successfully identified. Let

$$M_i = \max_{j \in [d] \setminus \{i\}} \chi(i, j) \tag{5.5}$$

denote the greatest value among the entries of row  $i$  of matrix  $\mathcal{X}$  excluding  $\chi(i, i) = 1$ . Additionally, let  $S_i$  represent the index set for which  $M_i$  reaches its maximum

$$S_i = \{j \in [d] \setminus \{i\}, \chi(i, j) = M_i\}. \tag{5.6}$$

**Theorem 5.2.2.** *Assume that model (5.1) and conditions (i)-(ii) hold. Then:*

- (a) *The set  $[K]$  is any maximum clique of the undirected graph  $G = (V, E)$  where  $V = [d]$  and  $(i, j) \in E$  if  $\chi(i, j) = 0$ .*
- (b) *Let  $i \in I_a$ ,  $a \in [K]$  and  $|I_a| \geq 2$ , then*

$$j \in I_a \iff \chi(i, j) = 1 \text{ for any } j \in S_i.$$

- (c) *The pure variable set  $I$  can be determined uniquely from  $\mathcal{X}$ . Moreover its partition  $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$  is unique and can be determined from  $\mathcal{X}$  up to label permutations.*

**Remark 5.2.1.** It is crucial to emphasise that the identification of recursive max-linear models on a Directed Acyclic Graph involves examining the extremal correlation matrix and the initial nodes  $V_0$ , specifically, the nodes without parent connections (see Section 4.3 of Gissibl et al. (2018)). Furthermore the set  $V_0$  can be determined from the tail dependence matrix since it is a maximum clique of the undirected graph  $G = (V, E)$  where  $V = [d]$ , and  $(i, j) \in E$  if  $\chi(i, j) = 0$  (refer to (Gissibl et al., 2018, Theorem 2.7)), akin to the characterisation provided in Theorem 5.2.2 (a) to identify the set  $[K]$ .

The decision problem of the maximum clique problem is one of the first 21 NP-complete problems, as introduced by Karp in his influential paper on computational complexity (Karp (1972)). This problem is known for its exponential complexity as the number of vertices increases in the worst cases. However, in the real world, many graphs tend to be sparse, meaning they have low degrees of connectivity (as noted in works by Buchanan et al. (2014); Eppstein et al. (2010)). This sparsity property allows for more efficient algorithms to solve the maximum clique problem in sparse graphs compared to general graphs.

In our framework, we have made the implicit practical assumption that the rows of the matrix  $A$  are sparse. Consequently, the complement of the adjacency matrix  $E$ , denoted as  $\bar{E}$ , in the graph defined in Theorem 5.2.2 (a), is also sparse. In this context, a faster method to find a

maximum clique is presented through the following binary problem:

$$\begin{aligned} \max_{x^{(i)}} \quad & \sum_{i=1}^d x^{(i)} \\ \text{s.t.} \quad & x^{(i)} + x^{(j)} \leq 1, \quad \forall (i, j) \in \bar{E} \\ & x^{(i)} \in \{0, 1\}, \quad i = 1, \dots, d. \end{aligned}$$

In this edge-based formulation, any valid solution defines a clique  $C$  in the graph  $G$  as follows: a vertex  $i$  belongs to the clique if  $x_i = 1$ , and otherwise  $x_i = 0$ . In our numerical studies, the use of this problem accelerated the estimation process, reducing computation time from minutes to seconds for large dimensions. In situations where the sparsity of matrix  $A$  is less emphasised, attention shifts to the adjacency matrix  $E$  which now becomes sparse. In such instances, well-known algorithms efficiently operates on the graph  $E$ . An excellent demonstration of this efficiency is found in the classical algorithm authored by [Bron and Kerbosch \(1973\)](#). For a more in-depth understanding of these intricacies, a comprehensive exploration awaits in [Appendix D.1](#).

The identifiability of the allocation matrix  $A$  and that of the collection of clusters  $\mathcal{G} = \{G_1, \dots, G_K\}$  in (5.4) use the results of [Theorem 5.2.2](#) in crucial ways. We state the result in [Theorem 5.2.3](#) below.

**Theorem 5.2.3.** *Assume that model (5.1) and conditions (i)-(ii) hold,  $A$  can be uniquely recovered from  $\mathcal{X} = A \odot A^\top$ , up to column permutations. This implies that the associated soft clusters  $G_a$ , for  $1 \leq a \leq K$ , are identifiable, up to label switching.*

**Remark 5.2.2.** If Condition (i) is replaced with

$$\sum_{a=1}^K A_{ja} \leq 1,$$

then the loading matrix is no longer identifiable from  $\mathcal{X}$ . Indeed, consider  $\mathbf{X} = A\mathbf{Z}$  and  $\tilde{\mathbf{X}} = \tilde{A}\mathbf{Z}$  with  $\tilde{A} = \lambda A$  for some  $\lambda \in (0, 1)$  and  $A$  verifies Condition (i). By [Theorem 5.2.1](#), we have

$$\mathcal{X} = A \odot A^\top,$$

and using the same tools involved in the proof of [Theorem 5.2.1](#), we have

$$\tilde{\chi}(i, j) = \lim_{x \rightarrow \infty} \frac{\mathbb{P}\{\tilde{X}_j > x, \tilde{X}_i > x\}}{\mathbb{P}\{\tilde{X}_i > x\}} = \frac{\sum_{k=1}^K (\lambda A_{ik}) \wedge (\lambda A_{jk})}{\sum_{k=1}^K (\lambda A_{ik})} = A_{ik} \wedge A_{jk}.$$

Thus  $\mathcal{X} = \tilde{\mathcal{X}}$ , and we cannot recover  $A$  from the extremal correlation matrix.

**Remark 5.2.3.** We show that the pure variable assumption stated in Condition (ii) is needed for the identifiability of  $A$  with the extremal correlation, up to a permutation. Consider the specific example where  $d = 3$  and  $K = 2$

$$A^{(1)} = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix}.$$

Then with some computations using Theorem 5.2.1, we obtain that the extremal correlation of  $\mathbf{X}$  is equal to

$$\mathcal{X} = \begin{pmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}.$$

Now taking

$$A^{(2)} = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix},$$

leads the same extremal correlation matrix. Thus, if  $A$  does not satisfy (ii),  $A$  is generally not identifiable with the extremal correlation matrix.

### 5.3 Estimation

Suppose that  $(\mathbf{X}_t, t \in \mathbb{Z}) = (X_t^{(1)}, \dots, X_t^{(d)}, t \in \mathbb{Z})$  is a multivariate strictly stationary process, and that  $(\mathbf{X}_t, t = 1, \dots, n)$  is observable data. Let  $m \in \{1, \dots, n\}$  be a block size parameter and, for  $i = 1, \dots, k$  and  $j = 1, \dots, d$ , let  $M_{m,i}^{(j)} = \max\{X_t^{(j)} : t \in [(i-1)m, \dots, im]\}$  be the maximum of the  $i$ th block observations in the  $j$ th coordinate. For  $\mathbf{x} = (x^{(1)}, \dots, x^{(d)})$ , let

$$\begin{aligned} \mathbf{M}_{m,i} &= (M_{m,i}^{(1)}, \dots, M_{m,i}^{(d)}), \\ F_m^{(j)}(x) &= \mathbb{P}\{M_{m,1}^{(j)} \leq x\}, \\ \mathbf{F}_m(\mathbf{x}) &= (F_m^{(1)}(x^{(1)}), \dots, F_m^{(d)}(x^{(d)})), \\ U_{m,i}^{(j)} &= F_m^{(j)}(M_{m,i}^{(j)}), \\ \mathbf{U}_{m,i} &= (U_{m,i}^{(1)}, \dots, U_{m,i}^{(d)}). \end{aligned}$$

Subsequently, we assume that the marginals of  $X_1^{(1)}, \dots, X_1^{(d)}$  are continuous. In that case, the marginals of  $\mathbf{M}_{m,1}$  are continuous as well and

$$C_m(\mathbf{u}) = \mathbb{P}\{U_{m,1}^{(1)} \leq \mathbf{u}\}, \quad \mathbf{u} \in [0, 1]^d,$$

is the unique copula associated with  $\mathbf{M}_{m,1}$ . Let us consider the following set  $\Delta_{d-1} = \{(w^{(1)}, \dots, w^{(d)}) \in [0, \infty)^d : \sum_{j=1}^d w^{(j)} = 1\}$  which is the unit simplex in  $\mathbb{R}^d$ . Throughout, we shall work under the following data generative process.

**Definition 5.3.1 (Data generative process).** Let  $(\mathbf{X}_t, t \in \mathbb{Z})$  be a multivariate strictly stationary random process and  $(\mathbf{X}_t, t = 1, \dots, n)$  its observable data. Let  $m \in \{1, \dots, n\}$  and  $C_m$  the copula of  $\mathbf{M}_{m,1}$  such that  $C_m$  is positive quadrant dependent, meaning that

$$C_m(\mathbf{u}) \geq \prod_{j=1}^d u^{(j)}, \quad \mathbf{u} \in [0, 1]^d. \quad (5.7)$$

There exist a copula  $C_\infty$ , a finite Borel measure  $\Phi$  on the unit positive sphere as defined in equation (5.2) such that

$$\lim_{m \rightarrow \infty} C_m(\mathbf{u}) = C_\infty(\mathbf{u}), \quad \mathbf{u} \in [0, 1]^d, \quad (5.8)$$

where

$$C_\infty(\mathbf{u}) = \exp \left\{ -L \left( -\ln(u^{(1)}), \dots, -\ln(u^{(d)}) \right) \right\}$$

and the stable tail dependence function  $L : [0, \infty)^d \rightarrow [0, \infty)$  is described by

$$L(z^{(1)}, \dots, z^{(d)}) = \sum_{a=1}^K \bigvee_{j=1}^d A_{ja} z^{(j)}.$$

Since extreme value copulae are positive quadrant dependent (see, e.g., (Resnick, 2008, Section 5.8)), it is expected that in practice  $C_m$ , i.e., a proxy to  $C_\infty$  will also verify this property for a sufficiently large  $m$ . The max-domain of attraction in (5.8) and the definition of the stable tail dependence function indicate that our observations are in the max-domain of attraction of a max-stable distribution with discrete angular measure. Then, this discrete max-stable distribution is characterised by a vector of latent factors  $\mathbf{Z} \in \mathbb{R}^K$  and a matrix  $A \in \mathbb{R}^{d \times K}$  that we want to estimate. As we will see, we will provide in Section 5.4 a non-asymptotic analysis of our estimator, which is valid for any  $d, n, m$  and  $k$ . Such an analysis avoids the need to have the max-domain of attraction condition given in Equation (5.8) to derive the main bound of the estimator's risk. However, an implicit link between  $C_m$  and  $C_\infty$  will be given through a bias term, which is expected to be smaller with respect to the block's length if (5.8) is satisfied in the data. Nonetheless, for presentation purposes, our statistical findings will be presented with this condition satisfied.

Our estimation procedure is inspired from Bing et al. (2020) and consists of the following four steps:

- (a) Estimate the number of clusters  $K$ , the pure variable set  $I$  and the partition  $\mathcal{I}$ ;
- (b) Estimate  $A_I$ , the submatrix of  $A$  with rows  $A_i$ . that correspond to  $i \in I$ ;
- (c) Estimate  $A_J$ , the submatrix of  $A$  with rows  $A_j$ . that correspond to  $j \in J$ ;
- (d) Estimate the overlapping clusters  $\mathcal{G} = \{G_1, \dots, G_K\}$ .

### 5.3.1 Estimation of $I$ and $\mathcal{I}$

In the context of our analysis, we need to estimate the submatrices, denoted as  $A_I$  and  $A_J$ , separately. To begin with  $A_I$ , we initiate the estimation process by determining  $[K]$ , which subsequently allows us to identify  $I$  and its partition, denoted as  $\mathcal{I} = \{I_1, \dots, I_K\}$ . This partition can be uniquely constructed from the extremal correlation matrix  $\mathcal{X}$ , as demonstrated in Theorem 5.2.2. To perform this step, we employ the constructive proof provided by Theorem 5.2.2, substituting the unknown  $\mathcal{X}$  with its sampled counterpart, referred to as the extremal correlation matrix:

$$\hat{\mathcal{X}} = [\hat{\chi}_{n,m}(i, j)]_{i=1, \dots, d, j=1, \dots, d}.$$

The quantity  $\hat{\chi}_{n,m}(i, j)$  is the sampling version of the *pre-asymptotic* extremal correlation,  $\chi_m(i, j)$  between  $M_{m,1}^{(i)}$  and  $M_{m,1}^{(j)}$  using block-maxima approach, i.e.,

$$\chi_m(i, j) = 2 - \frac{0.5 + \nu_m(i, j)}{0.5 - \nu_m(i, j)}, \quad \nu_m(i, j) = \frac{1}{2} \mathbb{E} \left[ |U_{m,1}^{(i)} - U_{m,1}^{(j)}| \right].$$

The quantity  $\nu_m(i, j)$  is the bivariate madogram (Cooley et al. (2006)) between  $M_{m,1}^{(i)}$  and  $M_{m,1}^{(j)}$ . Since  $\nu_m(i, j) \geq 0$ , it stems down that  $\chi_m(i, j) \leq 1$  and given that  $C_m$  satisfies (5.7), we have  $\nu_m(i, j) \leq 1/6$ , implying that  $\chi_m(i, j) \geq 0$ . These last quantities can be approached with the empirical madogram using the following relationship:

$$\hat{\chi}_{n,m}(i, j) = 2 - \frac{0.5 + \hat{\nu}_{n,m}(i, j)}{0.5 - \hat{\nu}_{n,m}(i, j)},$$

where  $\hat{\nu}_{n,m}(i, j)$  is the bivariate empirical estimator of the madogram which can be easily generalised by applying the equality  $|a - b|/2 = \max(a, b) - (a + b)/2$  with its multivariate counterpart

$$\hat{\nu}_{n,m}(\{1, \dots, d\}) = \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d \hat{U}_{n,m,i}^{(j)} - \frac{1}{d} \sum_{j=1}^d \hat{U}_{n,m,i}^{(j)} \right].$$

Here, we have standardised the marginals by ranking the observed block maxima componentwise. For a given value  $x \in \mathbb{R}$ ,  $j = 1, \dots, d$ , and block size  $m$ , we define:

$$\hat{F}_{n,m}^{(j)}(x) = \frac{1}{k+1} \sum_{i=1}^k \mathbb{1}_{\{M_{m,i}^{(j)} \leq x\}},$$

and we consider observable pseudo-observations of  $U_{m,i}^{(j)}$  as follows:

$$\hat{U}_{n,m,i}^{(j)} = \hat{F}_{n,m}^{(j)}(M_{m,i}^{(j)}).$$

To elaborate on our approach, we construct a graph denoted as  $G = (V, E)$ , where the vertex set is represented as  $V = [d]$ . We utilise the sample version of part (a) of Theorem 5.2.2 to identify the largest vector that is asymptotically independent. Using this set of indices, we then employ the sample version of part (b) of Theorem 5.2.2 to determine whether a specific index  $j$  qualifies as a pure variable. If an index is not categorised as pure, we include it in the set that estimates  $J$ . However, if it is deemed pure, we retain the estimated set  $\hat{S}_i$ , as defined in (5.6), of indices  $j \neq i$  that are strongly associated, through their extremal correlations, with  $i$ . Subsequently, we utilise the constructive proof of part (c) of Theorem 5.2.2 to declare that  $\hat{S}_i \cup i = \hat{I}^{(i)}$  serves as an estimator for one of the partition sets within  $\mathcal{I}$ .

For a comprehensive understanding of the algorithm, including the specification of a tuning parameter denoted as  $\delta$ , please refer to Algorithm (PureVar) in Appendix D.2. The discussion pertaining to the tuning parameter  $\delta$  will be presented in detail in Section 5.4.1.

### 5.3.2 Estimation of the allocation matrix $A$ and soft clusters.

Based on the estimators  $\hat{I}$ ,  $\hat{K}$ , and  $\hat{\mathcal{I}} = \{\hat{I}_1, \dots, \hat{I}_{\hat{K}}\}$  obtained from Algorithm (PureVar), we estimate the matrix  $A_I$ . This estimation takes the form of a matrix with dimensions  $|\hat{I}| \times \hat{K}$ , where each row corresponding to an index  $i$  in  $\hat{I}$  contains  $\hat{K} - 1$  zeros and one entry equal to 1. This procedure induces the following estimator of  $A_I$

$$\hat{A}_{ka} = \hat{A}_{la} = 1, \quad \text{for } k, l \in \hat{I}_a, \quad a \in [\hat{K}]. \quad (5.9)$$



We continue by estimating the matrix  $A_J$ , row by row. To explain our approach, we first outline the structure of each row, denoted as  $A_{j\cdot}$ , within the matrix  $A_J$ , for  $j \in J$ . We should note that each  $A_{j\cdot}$  satisfies sparsity conditions and  $\sum_{a=1}^K A_{ja} = 1$ , as stipulated by Condition (i). To simplify the exposition, we rearrange  $\mathcal{X}$  and  $A$  as follows

$$\mathcal{X} = \begin{bmatrix} \mathcal{X}_{II} & \mathcal{X}_{IJ} \\ \mathcal{X}_{JI} & \mathcal{X}_{JJ} \end{bmatrix}, \quad A = \begin{bmatrix} A_I \\ A_J \end{bmatrix}.$$

Model (5.1) and Theorem 5.2.1 imply the following decomposition of the extremal correlation matrix of  $\mathbf{X}$

$$\mathcal{X} = \begin{bmatrix} \mathcal{X}_{II} & \mathcal{X}_{IJ} \\ \mathcal{X}_{JI} & \mathcal{X}_{JJ} \end{bmatrix} = \begin{bmatrix} A_I \odot A_I^\top & A_I \odot A_J^\top \\ A_J^\top \odot A_I & A_J \odot A_J^\top \end{bmatrix}.$$

In particular,  $\mathcal{X}_{IJ} = A_I \odot A_J^\top$ . Thus for each  $i \in I_a$  with some  $a \in [K]$  and  $j \in J$ , we have

$$\chi(i, j) = A_{ja}.$$

Averaging the above display over all  $i \in I_a$  yields

$$\frac{1}{|I_a|} \sum_{i \in I_a} \chi(i, j) = A_{ja}.$$

Hence

$$\beta^{(j)} := A_j = \left( \frac{1}{|I_1|} \sum_{i \in I_1} \chi(i, j), \dots, \frac{1}{|I_K|} \sum_{i \in I_K} \chi(i, j) \right),$$

which can be estimated from the data as follows. For each  $j \in \hat{J}$ , we denote an estimator for the  $a$ -th element of  $\beta^{(j)}$  using a simple approach. This estimator is represented as follows

$$\bar{\chi}^{(j)} = \left( \frac{1}{|\hat{I}_1|} \sum_{i \in \hat{I}_1} \hat{\chi}_{n,m}(i, j), \dots, \frac{1}{|\hat{I}_{\hat{K}}|} \sum_{i \in \hat{I}_{\hat{K}}} \hat{\chi}_{n,m}(i, j) \right).$$

It is important to note that this estimator is neither sparse nor an element of the unit simplex. Given the value  $\bar{\chi}^{(j)}$ , our objective is to determine a Euclidean projection of  $\bar{\chi}^{(j)}$  that lies within the space  $\mathbb{B}_0(s) = \{\beta \in \mathbb{R}^{\hat{K}}, \sum_{j=1}^{\hat{K}} \mathbf{1}_{\{\beta^{(j)} \neq 0\}} \leq s\}$ , i.e., vectors with at most  $s$  non-zero entries, and the unit simplex  $\Delta_{\hat{K}-1} = \{\beta \in \mathbb{R}^{\hat{K}}, \beta^{(j)} \geq 0, \sum_{j=1}^{\hat{K}} \beta^{(j)} = 1\}$  :

$$\mathcal{P}(\hat{\beta}^{(j)}) \in \underset{\beta: \beta \in \mathbb{B}_0(s) \cap \Delta_{\hat{K}-1}}{\operatorname{argmin}} \quad \|\beta - \bar{\chi}^{(j)}\|_2. \quad (5.10)$$

Kyriallidis et al. (2013) have demonstrated the feasibility of computing such a projection efficiently, using a simple greedy algorithm. This algorithm is outlined in (HTSP) (Hard Thresholding and Simplex Projector), and it involves two main steps: first, identifying the support of  $\hat{\beta}^{(j)}$ , and then determining these values associated with this support. Consequently, in order to construct an estimator for the support, we select only the coordinates indexed by  $a$  where  $\bar{\chi}_a^{(j)}$  exceeds a

threshold  $\delta$ . This selection results in a sparse estimator for  $\beta_a^{(j)}$  as follows

$$\bar{\beta}_a^{(j)} = \bar{\chi}_a^{(j)} \mathbf{1}_{\{\bar{\chi}_a^{(j)} > \delta\}}, \quad a = 1, \dots, \hat{K}.$$

This estimator  $\bar{\beta}^{(j)}$  is often referred to as the hard thresholding estimator, where  $\delta > 0$  represents the threshold value. However, it is essential to note that the estimator  $\bar{\beta}^{(j)}$  does not inherently belong to the unit simplex. To address this, we can obtain an alternative estimator, denoted as  $\hat{\beta}^{(j)}$ , by projecting  $\bar{\beta}^{(j)}$  onto the unit simplex within the  $\hat{K}$ -dimensional space. The projection operation onto the unit simplex is achieved by utilising a specific mathematical operator, and this operator is defined as

$$(\mathcal{P}_{\Delta_{\hat{K}-1}}(\beta))_j = [\beta^{(j)} - \tau]_+, \quad \tau := \frac{1}{\rho} \left( \sum_{i=1}^{\rho} \beta^{(i)} - 1 \right),$$

for  $\rho := \max\{k, \beta^{(j)} > \frac{1}{k}(\sum_{j=1}^k w^{(j)} - 1)\}$ . Hence, by denoting  $\hat{\mathcal{S}} = \text{supp}(\bar{\beta}^{(j)})$ , we obtain

$$\hat{\beta}^{(j)} \Big|_{\hat{\mathcal{S}}} = \mathcal{P}_{\Delta_{\hat{K}-1}} \left( \bar{\beta}^{(j)} \Big|_{\hat{\mathcal{S}}} \right), \quad \hat{\beta}^{(j)} \Big|_{\hat{\mathcal{S}}^c} = 0 \quad (5.11)$$

Next, we construct the matrix  $\hat{A}_{\hat{J}}$  with rows corresponding to the estimators  $\hat{\beta}^{(j)}$  for each  $j \in \hat{J}$ . Our final estimator, denoted  $\hat{A}$ , for the matrix  $A$ , is obtained by concatenating  $\hat{A}_{\hat{J}}$  and  $\hat{A}_{\hat{J}}$ . The statistical properties of the final estimator are examined in Section 5.4, where we also provide detailed specifications of the tuning parameter necessary for its development.

Recalling the definition of groups in (5.3), the soft clusters are estimated by

$$\hat{\mathcal{G}} = \{\hat{G}_1, \dots, \hat{G}_{\hat{K}}\}, \quad \hat{G}_a = \{i \in [d] : \hat{A}_{ia} \neq 0\}, \quad \text{for each } a \in [\hat{K}]. \quad (5.12)$$

Variables  $X^{(j)}$  that are associated (via  $\hat{A}$ ) with the same latent factor  $Z^{(a)}$  are therefore placed in the same group  $\hat{G}_a$ . We demonstrate in Section 5.4.2 that the overlapping clusters or extreme directions can be controlled with high probability.

## 5.4 Statistical guarantees

This section serves a dual purpose. Firstly, we derive a bound akin to Bernstein's inequality for the uniform norm of the sampled version of the extremal correlation. This is achieved by utilising a sequence of dependent and strictly stationary random variables denoted as  $(\mathbf{X}_t, t \in \mathbb{Z})$ . Secondly, we employ the developed techniques to investigate statistical guarantees for the following aspects:

- (a) The estimated number of clusters, denoted as  $\hat{K}$ ;
- (b) The estimated pure variable set  $\hat{I}$  and its corresponding estimated partition  $\hat{\mathcal{L}}$ ;
- (c) The estimated allocation matrix  $\hat{A}$  and its adjustment to account for the unknown sparsity of rows in matrix  $A$ ;
- (d) Guarantees to recover overlapping clusters / extreme directions in terms of Total False Positive Proportion (TFPP) and Total False Negative Proportion (TFNP).

We recall the definition of strongly mixing sequences, introduced by Rosenblatt (1956a): For any two  $\sigma$ -algebras  $\mathcal{A}$  and  $\mathcal{B}$ , we define the  $\alpha$ -mixing coefficient by

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

Let  $(\mathbf{X}_t, t \in \mathbb{Z})$  be a sequence of real-valued random variables defined on  $(\Omega, \mathcal{A}, \mathbb{P})$ . This sequence will be called strongly mixing if

$$\alpha(n) := \sup_{k \geq 1} \alpha(\mathcal{F}_k, \mathcal{G}_{k+n}) \rightarrow 0 \text{ as } n \rightarrow \infty, \quad (5.13)$$

where  $\mathcal{F}_j := \sigma(\mathbf{X}_i, i \leq j)$  and  $\mathcal{G}_j := \sigma(\mathbf{X}_i, i \geq j)$  for  $j \geq 1$ . Throughout, we assume the sequence  $(\mathbf{X}_t, t \in \mathbb{Z})$  has exponentially decaying strong mixing coefficients, that is

$$\alpha(n) \leq \exp\{-2cn\}, \quad (5.14)$$

for a certain  $c > 0$ . Our first result is the following exponential inequality.

**Theorem 5.4.1.** *Let  $(\mathbf{X}_t, t \in \mathbb{Z})$  be a sequence of stationary random variables with exponential decaying strong mixing coefficients given in (5.14). Then, there are constants  $c_0 > 0$  and  $c_1 > 0$  which depend only on  $c$  such that*

$$\mathbb{P} \left\{ \sup_{1 \leq i < j \leq d} |\hat{\chi}_{n,m}(i, j) - \chi_m(i, j)| \geq c_1 \left( \sqrt{\frac{\ln(kd)}{k}} + \frac{\ln(k) \ln \ln(k) \ln(kd)}{k} \right) \right\} \leq d^{-c_0},$$

where  $n, m \in \{1, \dots, n\}$  denote respectively the sample and the block size,  $k = \lfloor n/m \rfloor \geq 4$  and  $\chi_m(i, j) \in [0, 1]$  is the pre-asymptotic extremal correlation between  $M_{m,1}^{(i)}$  and  $M_{m,1}^{(j)}$ .

In the context of our section, let us introduce some mathematical notations and concepts. Consider the quantity  $\chi(i, j)$ , which represents the extremal correlation between  $X^{(i)}$  and  $X^{(j)}$ , within the max-domain of attraction as specified by Equation (5.8). We also define a crucial parameter denoted as:

$$d_m = \sup_{1 \leq i < j \leq d} |\chi_m(i, j) - \chi(i, j)|,$$

where  $\chi_m(i, j)$  is the pre-asymptotic extremal correlation between  $M_{m,1}^{(i)}$  and  $M_{m,1}^{(j)}$ . This parameter characterises the explicit bias between the subsymptotic framework and the max-domain of attraction. It essentially quantifies the rate at which the system converges to its asymptotic extreme behavior. Additionally, we introduce the following new event:

$$\mathcal{E} = \mathcal{E}(\delta) := \left\{ \sup_{1 \leq i < j \leq d} |\hat{\chi}_{n,m}(i, j) - \chi(i, j)| \leq \delta \right\}. \quad (5.15)$$

Now, we state that

$$\delta = d_m + c_1 \left( \sqrt{\frac{\ln(kd)}{k}} + \frac{\ln(k) \ln \ln(k) \ln(kd)}{k} \right), \quad (5.16)$$

for some absolute constant  $c_1 > 0$ . Furthermore, we require that  $\ln(d) = o(k)$ . This condition ensures, with the additional max-domain of attraction in (5.8), that  $\delta = o(1)$  provided that  $m = o(n)$ . Taking  $c_1 > 0$  large enough, Theorem 5.4.1 guarantees that  $\mathcal{E}$  holds with high probability:

$$\mathbb{P}(\mathcal{E}) \geq 1 - d^{-c_0},$$

for some positive constant  $c_0 > 0$ . The order of the threshold  $\delta$  involves known quantity such as  $d$ ,  $k$  and a unknown parameter  $d_m$ . For the latter, there is no simple manner to choose optimally this parameter, as there is no simple way to determine how fast is the convergence to the asymptotic extreme behavior, or how far into the tail the asymptotic dependence structure appears. In particular, Equation (5.8) does not contain any information about the rate of convergence of  $C_m$  to  $C_\infty$ . More precise statements about this rate can be made with second order conditions. Let a regularly varying function  $\Psi : \mathbb{N} \rightarrow (0, \infty)$  with coefficient of regular variation  $\rho_\Psi < 0$  and a continuous non-zero function  $S$  on  $[0, 1]^d$  such that

$$C_m(\mathbf{u}) - C_\infty(\mathbf{u}) = \Psi(m)S(\mathbf{u}) + o(\Psi(m)), \quad \text{for } m \rightarrow \infty, \quad (5.17)$$

uniformly in  $\mathbf{u} \in [0, 1]^d$  (see, e.g., Bücher et al. (2019) for a proper introduction to this condition). In this case, we can show that  $d_m = O(\Psi(m))$ .

#### 5.4.1 Statistical guarantees for $\hat{K}$ , $\hat{I}$ and $\hat{\mathcal{I}}$ .

We now move forward with the analysis of the statistical performance of our estimator, denoted as  $\hat{I}$ , which aims to estimate  $I$ . Alongside this estimation, we also consider its associated partition. This particular problem falls within the broader category of pattern recovery problems. It is well established that, given sufficiently strong signal conditions, we can reasonably expect that  $\hat{I} = I$  with a high level of confidence. These conditions are stated below

**Condition (SSC).** Let  $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$ .

(SSC1)  $\forall k \notin \mathcal{I}, A_{ka} < 1 - 2\delta, \forall a \in [K]$ ;

(SSC2)  $\forall k \notin \mathcal{I}, \exists a, b \in [K]$  such that  $A_{ka} > 2\delta$  and  $A_{kb} > 2\delta$ .

In Theorem 5.4.2, we establish a critical result that provides a high-confidence guarantees for recovery of  $K$ ,  $I$  and  $\mathcal{I}$  under the condition (SSC). This theorem has several key implications. In the first aspect of these implications, our theorem demonstrate that, with high probability, the estimated set  $\hat{I}$  is equal to the set of pure variables  $I$ . This implies that our procedure correctly identifies these fundamental variables. Ambiguous variables  $X^{(j)}$  with  $j \notin I$  that exhibit associations with multiple latent factors not exceeding the  $1 - 2\delta$  threshold, as dictated by Condition (SSC1), are deliberately excluded from  $\hat{I}$ . This exclusion is crucial to maintain the accuracy of our method. In the construction of the graph  $G$  in part (a) of Theorem 5.2.2, Condition (SSC2) guarantees that only ambiguous variables are excluded from a maximum clique of  $G$ . This is of utmost importance because it ensures that the number of cluster  $K$  is determined accurately.

**Theorem 5.4.2.** *Let  $(\mathbf{X}_t, t \in \mathbb{Z})$  be a sequence of stationary random variables with exponential decaying strong mixing coefficients given in (5.14), satisfying the data generative process given in Definition 5.3.1. Under Conditions (i)-(ii) and Condition (SSC), then*

(a)  $\hat{K} = K$ ;

(b)  $I = \hat{I}$ .

Moreover, there exists a label permutation  $\pi$  of the set  $\{1, \dots, K\}$  such that the output  $\hat{\mathcal{I}} = \{\hat{I}_a\}_{1 \leq a \leq K}$  from Algorithm (PureVar) satisfies

(c)  $I_{\pi(a)} = \hat{I}_{\pi(a)}$ .

All results hold with probability larger than  $1 - d^{-c_0}$  for a positive constant  $c_0$ .

### 5.4.2 Statistical guarantees for $\hat{A}$

In this section, we present and discuss the statistical properties of the estimator  $\hat{A}$  and its control over the relationship between the support of  $A$  and the support of  $\hat{A}$ . It is worth noting that  $\delta$  is defined in Equation (5.15) and the estimator of  $A$  relies solely on this tuning parameter. Theorem 5.4.3 established an adaptative finite sample upper bound for exponentially decaying  $\alpha$ -mixing observations. Both  $d$  and  $K$  are allowed to grow with  $n$ . We consider the loss function for two  $d \times K$  matrices  $A, A'$  as

$$L_2(A, A') := \min_{P \in S_K} \|AP - A'\|_{\infty, 2} \quad (5.18)$$

where  $S_K$  is the group of all  $K \times K$  permutation matrices and

$$\|A\|_{\infty, 2} := \max_{1 \leq j \leq d} \|A_{j \cdot}\|_2 = \max_{1 \leq j \leq d} \left( \sum_{i=1}^K |A_{ij}|^2 \right)^{1/2};$$

for a generic matrix  $A \in \mathbb{R}^{d \times K}$ . Finally given  $\delta$  in (5.16), we define

$$J_1 = \{j \in J : \text{for any } a \in [K] \text{ with } A_{ja} \neq 0, A_{ja} > 2\delta\}. \quad (5.19)$$

**Theorem 5.4.3.** *Assume the conditions in Theorem 5.4.2 hold. Set  $s = \max_{j \in [d]} \|A_{j \cdot}\|_0$ ,  $s(j) = \sum_{a=1}^K \mathbb{1}_{\{A_{ja} > 0\}}$  and  $t(j) = \sum_{a=1}^K \mathbb{1}_{\{A_{ja} \leq 2\delta\}}$  for each  $j \in J$ . Then for the estimator  $\hat{A}$  the following holds.*

(a) *An upper bound:*

$$L_2(\hat{A}, A) \leq 4\sqrt{s}\delta,$$

(b) *A guarantee for support recovery:*

$$\text{supp}(A_{J_1}) \subseteq \text{supp}(\hat{A}) \subseteq \text{supp}(A),$$

(c) *Cluster recovery:*

$$\begin{aligned} TFPP(\hat{\mathcal{G}}) &= \frac{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}=0, \hat{A}_{ja}>0\}}}{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}=0\}}} = 0, \\ TFNP(\hat{\mathcal{G}}) &= \frac{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0, \hat{A}_{ja}=0\}}}{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0\}}} \leq \frac{\sum_{j \in J \setminus J_1} t(j)}{|I| + \sum_{j \in J} s(j)}, \end{aligned}$$

with probability larger than  $1 - d^{-c_0}$  for a positive constant  $c_0$ .

Some comments on the above results are in order. On a high level, larger dimensions  $d$ , larger values of  $d_m$  lead to a larger bound. The effects of dimension  $d$  and bias  $d_m$  are intuitive: larger dimensions or more bias make the matrix recovery problem more difficult. The dependence on the number of latent factors  $K$  is implicitly conveyed through the sparsity index  $s$ , and its impact on the bound is also straightforward: the sparser the matrix, the better the bound.

It is important to note that the dimension  $d$  is permitted to grow as  $\ln(d) = o(k)$ , while the procedure still preserving an accurate estimation of  $A$  under additional assumptions. Specifically, it is required that  $m = o(n)$  and  $C_m$  verifies Equation (5.17) for which there exists a constant  $K_\Psi$  independent of  $d$  such that  $d_m \leq K_\Psi \Psi(m)$  for  $m$  large enough. In particular, if the dimension  $\ln(d)$  is varying exponentially with  $k$ , the bound in Theorem 5.4.3 (a) using this procedure is not meaningful in the worst cases, meaning that the distance  $L_2(\hat{A}, A)$  does not decrease when the sample size  $n$  grows. Additionally, we have demonstrate that  $\hat{A}$  possesses another desirable property, namely variable selection, or exact recovery sparsity pattern for row  $j \in J_1$ .

To the best of our knowledge, the estimation of identifiable sparse loading matrices  $A$  and overlapping clusters (or termed equivalently as extreme directions) in the model (5.1), meeting Conditions (i)-(ii), when both  $I$  and  $K$  are unknown, has not been explored in existing literature. Our results fill this gap in the research landscape. However, there is an extensive body of literature on related problems, as outlined in the introduction.

## 5.5 Numerical results

### 5.5.1 A data-driven selection method for the tuning parameter

Theorems 5.4.2-5.4.3 outline the theoretical rate of  $\delta$ , but only up to constants. Below, we propose a method for selecting  $\delta$  based only on data. We opt for a finely tuned grid of values  $\delta_\ell = c_\ell(d_m + \sqrt{\ln(d)/k})$  with  $1 \leq \ell \leq M$  as suggested by (5.16) for  $k$  sufficiently large and omitting log factors in  $k$ . This selection process involves varying the proportionality constants  $c_\ell$ . For each  $\delta_\ell$  chosen, we determine the number of clusters  $\hat{K}(\ell)$  using Algorithm (SCRAM), and the corresponding matrix,  $\hat{A}(\ell)$ . We denote the associated overlapping clusters as  $\hat{G}(\ell) = \{\hat{G}_1(\ell), \dots, \hat{G}_{\hat{K}(\ell)}(\ell)\}$ . Define

$$\mathcal{L}(\hat{G}(\ell)) = \sum_{j=1}^d \sum_{a \in [\hat{K}(\ell)]} \left( \hat{A}_{ja}(\ell) - \hat{\chi}_{n,m}(j, a) \mathbb{1}_{\{\hat{A}_{ja}(\ell) > 0\}} \right)^2. \quad (5.20)$$

We then proceed to select  $\delta^*$  as the value of  $\delta_\ell$  that minimises the criteria in (5.20). This leads to data-driven selection of  $\delta^* = c^*(d_m + \sqrt{\ln(d)/k})$ . While our choice may not be optimal in certain challenging scenarios, it provides effective data-based guidelines for our comparative analysis in Section 5.5.3 and real-world data evaluation in Section 5.6.

### 5.5.2 Performance of the proposed methodology in finite sample setting

In this section, we investigate the finite-sample performance of our algorithm to estimate the matrix  $A$  in a linear factor model described in (5.1) by means of a simulation study.

**The setup.** As a time series model, we consider the discrete-time,  $d$ -variate moving maxima process  $(\mathbf{Z}_t, t \in \mathbb{Z})$  of order  $p \in \mathbb{N}$  given by

$$\mathbf{Z}_t^{(a)} = \bigvee_{\ell=0}^p \rho^\ell \epsilon_{t+\ell}^{(a)}, \quad (t \in \mathbb{Z}, a = 1, \dots, K), \quad \rho \in (0, 1). \quad (5.21)$$

Here  $(\epsilon_t, t \in \mathbb{Z})$  is an i.i.d. sequence of  $K$ -dimensional random vectors having the following distribution

$$\mathbb{P}\{\epsilon_1 \leq \mathbf{x}\} = \varphi^{\leftarrow} \left( \varphi(P(x^{(1)})) + \dots + \varphi(P(x^{(K)})) \right), \quad \mathbf{x} \in \mathbb{R}^K,$$

where  $\varphi$  is the Archimedean generator  $\varphi : [0, 1] \rightarrow [0, \infty]$ ,  $\varphi^{\leftarrow}$  its generalised inverse and  $P(x) = 1 - 1/x$  for  $x \geq 1$  is the cumulative distribution function of a standard Pareto random variable. We assume the existence of the limit of

$$\lim_{s \rightarrow 0} \frac{\varphi(1 - st)}{\varphi(1 - s)} = t, \quad t \in (0, \infty), \quad (5.22)$$

i.e., the upper tail exhibits asymptotic independence. In Appendix D.6 (as seen in Proposition D.6.1), we demonstrate that the maxima of  $(\mathbf{Z}_t, t \in \mathbb{Z})$  belong to the maximum domain of attraction of independent Fréchet distributions. Consequently, the related process verifies the max-domain of attraction (5.8)

$$\mathbf{X}_t = A\mathbf{Z}_t + \mathbf{E}_t, \quad (5.23)$$

where  $\mathbf{E}_t$  represents independent and identically distributed replications of a lighter tail distributions. Noteworthy, since  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is  $p$ -dependent, the overall process  $(\mathbf{X}_t, t \in \mathbb{Z})$  is  $\alpha$ -mixing.

We generate the data in the following way. We set the number of clusters  $K$  to be 20 and simulate the latent variables  $Z = (Z^{(1)}, \dots, Z^{(K)})$  from (5.21) and  $\varphi = (t^{-1} - 1)$ . The error terms  $E_i^{(1)}, \dots, E_i^{(d)}$  are independently generate from a standard normal distribution. To speed up the simulation process, we set the first 20 rows of  $A$  as pure variables. This means these rows are predetermined and do not change during the process so that the best permutation matrix that achieves the best estimation error is the identity. To generate  $A_j$ , for any  $j \in J$ , we randomly assign the cardinality  $s_j$  of the support  $A_j$  to a number in  $\{2, 3, 4\}$ , with equal probability. Then, we randomly select the support from the set  $\{1, 2, \dots, K\}$  with a cardinality equal to  $s_j$ . We then sample  $U_1, \dots, U_{s_j}$  uniformly over the segment  $[0.35, 0.65]$ . Finally, we assign the value of variables in the support as the corresponding sampled value divided by the sum of all sampled values for variables in the support. Thus, we can generate  $\mathbf{X}$  according to the model in (5.23) and setting  $\rho = 0.8$  and  $p = 2$ .

**The target values.** Our simulation study aims at investigating the performance of our algorithm to recover the number of latent variables  $K$  and the performance of  $\hat{A}$  as estimators of  $A$ . When the number of clusters is correctly identified, we compute the norm  $L_2(\hat{A}, A)$  in (5.18).

**Calibrating parameters.** In practice, based on Equation (5.16), we recommend the following choice, for sufficiently large  $k$ , and omitting logarithmic terms in  $k$  with  $d_m = c_2/m$  as a rule of

thumb (see Appendix D.6 for technical details of this heuristic):

$$\delta = \frac{c_2}{m} + c_1 \sqrt{\frac{\ln(d)}{k}},$$

and set  $c_1 = 1.2$  and  $c_2 = 1.0$  in Algorithm (SCRAM). We have found that these choices for  $c_1$  and  $c_2$  not only yield good overall performance but are also robust with respect to the dimension  $d$ , the block's length  $m$  and the number of blocks  $k$ .

**Results and discussion.** Figure 5.1, Panel a present simulation results on exact recovery rate of number of clusters  $K$  and estimation error  $L_2(\hat{A}, A)$  in the case of a fixed  $m = 15$  and with varying  $k \in \{300, 500, 700, 1000\}$ . The shape of the functions are as expected; the estimation error decreases when the number of blocks  $k$  increases from 300 to 1000, which is in line with our theoretical results. The simulations are conducted on an Ubuntu system version 22.04 with 2.5 GHz Intel Core i7 CPU and 32 GB memory. Even with  $d = 1000$  and  $k = 1000$ , the computing time of our method for each simulation is around 5 seconds. In Figure 5.1, Panel b, we present simulation results on exact recovery rate of the number of cluster  $\hat{K}$  and estimation error  $L_2(\hat{A}, A)$  with a fixed sample size  $n = 5000$ . We plot these two quantities against the number of blocks  $k$ . From the picture, we see that, as expected, the performance of the estimator is first decreasing in  $k$  while it increases after a certain threshold. This phenomenon is an illustration of the bias-variance tradeoff that runs is the choice of the corresponding block length. Regarding the exact recovery rate and the performance of the estimator, we observe a rather good performance for a value of  $k = 700$ , corresponding to block length  $m = 7$ .

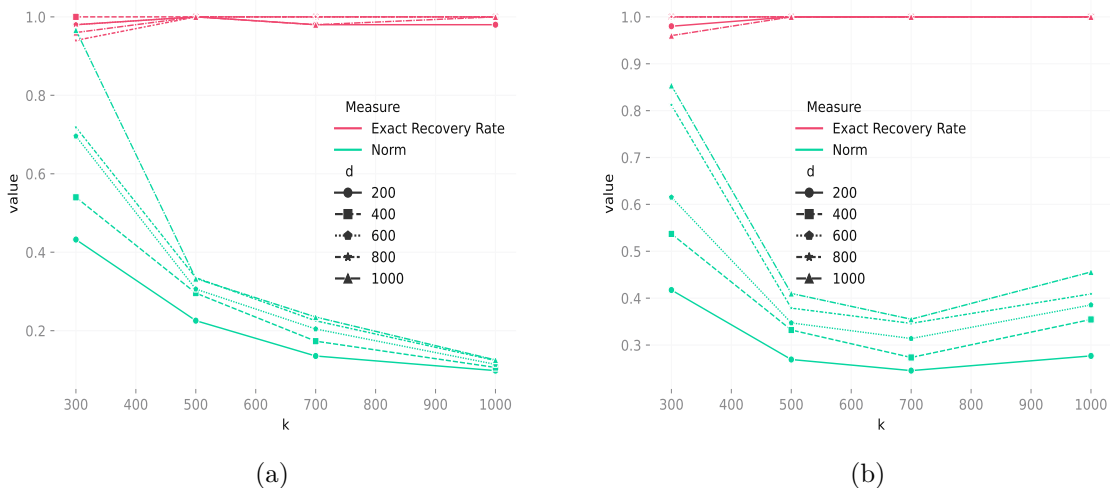


Fig. 5.1 In Panel a, simulation results on exact recovery rate of number of clusters  $K$  (in red) of  $\mathcal{I}$  and  $L_2(\hat{A}, A)$  (in green) with fixed  $m = 20$  and varying  $d \in \{200, 400, 600, 800, 1000\}$ . In Panel b, simulations results on exact recovery rate (in blue) of  $\mathcal{I}$  and  $L_2(\hat{A}, A)$  (in red) for fixed  $n = 5000$  and varying number of blocks and  $d \in \{200, 400, 600, 800, 1000\}$ .

Finally, within Table 5.1, we display the parameter  $\delta^*$  which minimises the average of the criteria in (5.20) using the data-driven selection method proposed in Section 5.5.1 over 50 runs



with different numbers of block maxima  $k$  and dimensions  $d$ . Additionally, we give the average ERR over these 50 runs. These results provide support for this criterion in selecting  $\delta$ , yielding favorable outcomes in terms of ERR. However, it is also evident that this criterion may not be optimal in certain difficult scenarios, particularly for large values of  $d$  and small values of  $k$ .

$k / d$	$c^* / \text{ERR}$				
	200	400	600	800	1000
300	(1.2, 1)	(1.13, 1)	(1.05, 0.98)	(1, 0.92)	(0.985, 0.86)
500	(1.41, 1)	(1.38, 1)	(1.3, 1)	(1.28, 1)	(1.29, 1)
700	(1.63, 1)	(1.6, 1)	(1.5, 1)	(1.47, 1)	(1.33, 1)
1000	(1.83, 1)	(1.64, 1)	(1.74, 1)	(1.69, 1)	(1.35, 1)

Table 5.1 Data-driven selection of  $c^*$  using the average criterion (5.20) and Exact Recovery Rate (ERR) of latent factors  $K = \hat{K}$  over 50 runs with varying  $k \in \{300, 500, 700, 1000\}$  and  $d \in \{200, 400, 600, 800, 1000\}$ .

### 5.5.3 Numerical comparisons

In this section, we analyse how well our method performs in recovering extreme directions in contrast to DAMEX (Goix et al. (2017)), and its efficiency for estimating normalised columns  $A_{\cdot k} / \|A_{\cdot k}\| =: \mathbf{a}_k$ ,  $k \in [K]$  compare to sKmeans (Janßen and Wan (2020)). Other algorithms (namely, Chiapino et al. (2019); Fomichov and Ivanovs (2022); Meyer and Wintenberger (2023)) in the literature were also considered; however, since they suffer from computational weaknesses or yield poor performance where  $\mathbf{X}$  is decomposed as a linear factor model described in (5.1) under Conditions (i)-(ii), they are omitted from the presentation of the results. We borrow the identical setup of a moving-maxima process as in Section 5.5.2, we hence moved to an elucidation of the target values.

**Target values.** Our simulation study aims to investigate the performance of our algorithm in determining the number of extreme directions and assessing its performance, when  $\hat{K} = K$ , using the TFPP and TFNP metrics compared to the DAMEX Algorithm. Additionally, we assess the disparity between the true centroids  $\mathbf{a}_1, \dots, \mathbf{a}_K$  and the estimated centroids  $\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_K$  as produced by both our procedure and sKmeans when  $\hat{K} = K$  by:

$$D(\{\mathbf{a}_1, \dots, \mathbf{a}_K\}, \{\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_K\}) = \min_{\pi} \sqrt{\sum_{k=1}^K \|\hat{\mathbf{a}}_{\pi(k)} - \mathbf{a}_k\|_2^2}, \quad (5.24)$$

where the min is taken over all permutation  $\pi$  of  $\{1, \dots, K\}$ . By the definition of (5.24), the number of factors in the experiment is reduced to  $K = 6$  due to memory limitations.

**Results and Discussion.** Figure 5.2, Panels a-c depict results on exact recovery rate of the number of clusters  $K$ , TFNP and TFPP for both Algorithms (SCRAM) and DAMEX over 50 simulations. The exact recovery rate of Algorithm of (SCRAM) is always better than the one of DAMEX for any configurations of  $d$  and  $n$ . Moreover, our procedure appears to be more resilient than DAMEX to a decrease in the sample size  $n$  and an increase in the dimension  $d$ .

However, when the number of latent factors is correctly retrieved by DAMEX Algorithm, it exhibits a better performance than (SCRAM) in terms of TFNP and TFPP. The average value of  $D(\{\mathbf{a}_1, \dots, \mathbf{a}_K\}, \{\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_K\})$  over 50 realisations for different values of  $d$  and  $n$  are show in Figure 5.2, Panel d. It can be seen that the locations of the points of mass of the spectral measure are most precisely estimated by (SCRAM) Algorithm.

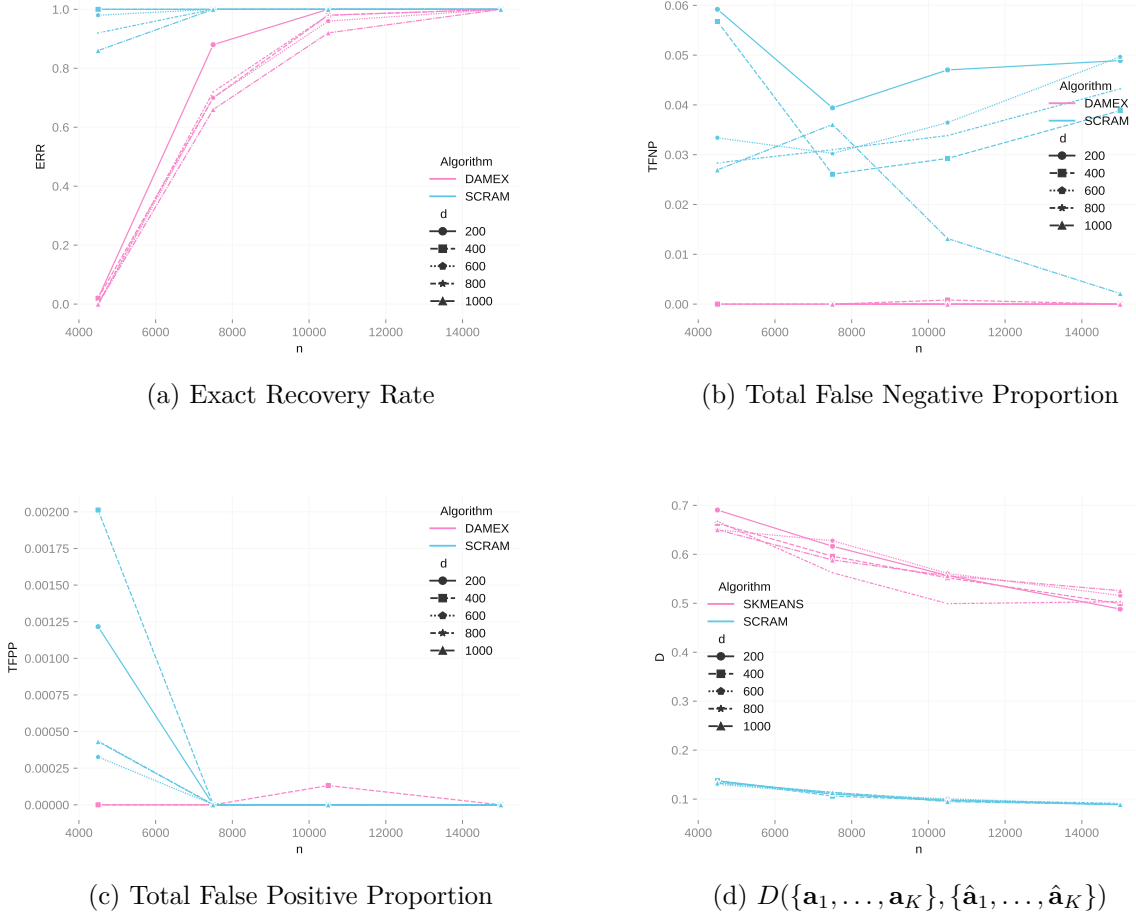


Fig. 5.2 Results of numerical comparisons between (SCRAM) Algorithm (in blue) and DAMEX Algorithm (in pink) in Panels a-c. Comparison of  $D(\{\mathbf{a}_1, \dots, \mathbf{a}_K\}, \{\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_K\})$  between (SCRAM) Algorithm (in blue) and sKmeans (in pink) are given in Panel d.

**Calibrating parameters.** In the (SCRAM), the parameter  $\delta$  is selected using the method proposed in Section 5.5.1. To fasten computations, we utilise the simulation outcomes detailed in Section 5.5.2, selecting  $\delta^*$  as the threshold that minimises the average value of the criteria across 50 iterations given in Table 5.1. In the DAMEX Algorithm, we adopt the approach recommended by the authors, selecting the  $\lfloor \sqrt{n} \rfloor$  largest values. Due to the propensity of the DAMEX Algorithm to return numerous extreme directions for many  $\epsilon$  values (tuning parameter of the DAMEX Algorithm), we opt to merge overlapping directions. This adjustment is crucial for enabling the DAMEX Algorithm to accurately recover the true number of latent factors. Furthermore, we determine  $\epsilon$  through trial and error using the exact recovery rate of

$K$  (post-merging step) as a benchmark, and settle on  $\epsilon = 0.3$ . Calibrating parameters of the sKmeans algorithm is relatively straightforward. We simply select the  $\lfloor \sqrt{n} \rfloor$  largest values and we designate the true number of latent factors  $K = 6$  (which is typically unknown in practice) as the number of clusters.

## 5.6 Applications

### 5.6.1 Extreme precipitations in France

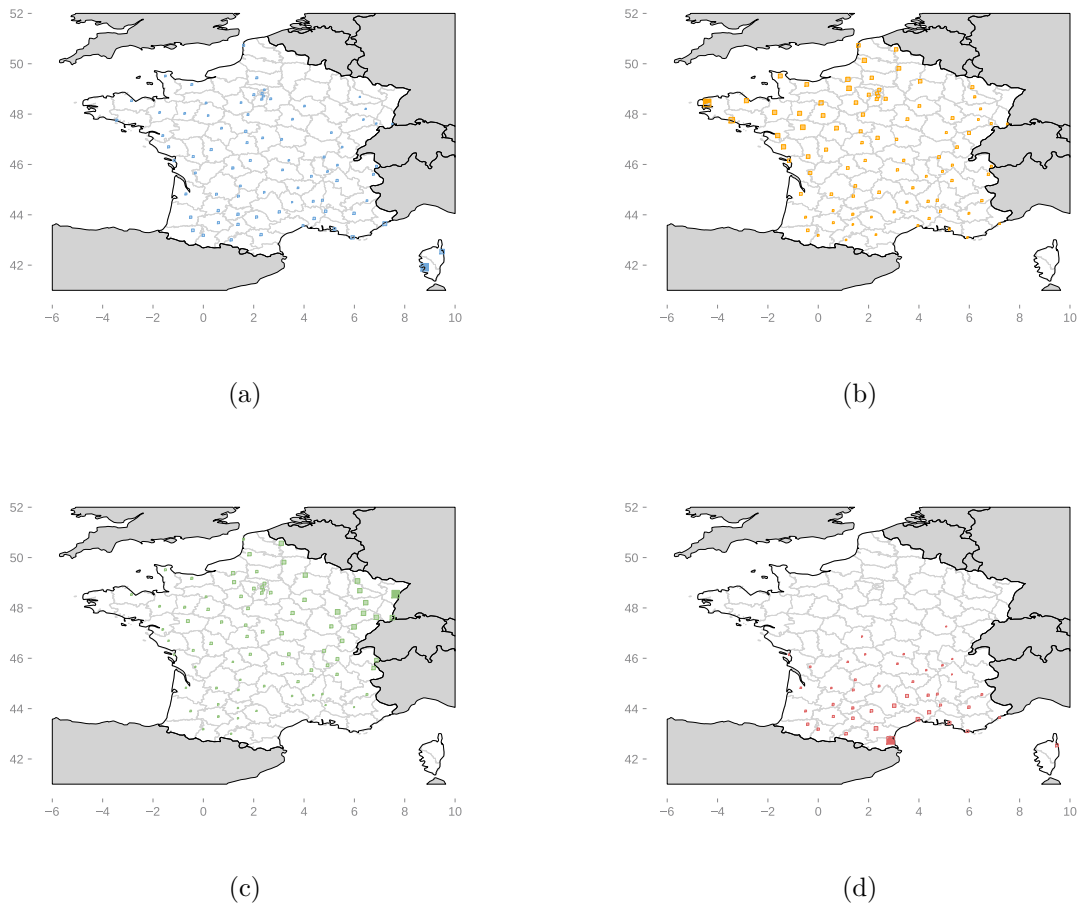


Fig. 5.3 In Panel a, we show the clusters for the Corsica latent variable. Panel b displays clusters for the western latent variable. Panel c shows clusters for the eastern latent variable. Finally, Panel d depicts clusters for the southern latent variable. The strength of association is indicated by the size and color intensity of each square.

In our analysis, we focus on weekly maxima of hourly precipitation recorded at 92 weather stations in France during the fall season, spanning from September to November, for the years 1993 to 2011, resulting in 228 block maxima. This dataset was provided by Météo-France and has been previously used in [Bernard et al. \(2013\)](#). The selection of stations was based on their data quality and ensuring a relatively uniform coverage of France.

We use the process described in Section 5.5.1 to choose the tuning parameter  $\delta$ . The entire process recommends using  $c_\ell \approx 0.82$  as the most suitable threshold value for our analysis. Employing the designated threshold, we unveil four latent variables situated in the western, eastern, southern regions of metropolitan France and Corsica. It is crucial to highlight that our process operates solely based on rainfall records, devoid of any geographical information. Consequently, discerning consistent spatial structures from just rainfall measurements is not a straightforward outcome. Spatial representation of clusters are depicted in Figure 5.3. The Corsican cluster highlighted in Figure 5.3, Panel a, where the pure variable is located at Ajaccio, exhibits a strong association within the island, while other associations rapidly decline on the mainland of France. The western area above Bordeaux, indicated in Figure 5.3 Panel b, exhibits robust dependencies with the central region around Paris. However, beyond these regions, the associations with the latent variable rapidly decrease. Symmetrically, the eastern region, spanning from Lyon and covering the Vosges mountains, Alsace, the Franche-Comté and regions in northeastern France, depicted in Figure 5.3 Panel c, displays dependencies with the central regions while diminishing rapidly outside this area. In contrast, the western cluster shows a broader distribution spanning across the entire country. The southern cluster, in Figure 5.3 Panel d, showcases spatial dependencies over Corsica and Mediterranean cities. These associations rapidly fell-off, resulting in the formation of a less spread-out cluster. The clustering results for locations align quite close with [Bernard et al. \(2013\)](#); [Maume-Deschamps et al. \(2023\)](#) dividing France into north and south regions. The key distinction lies in our clusters being overlapping, providing a more nuanced understanding of the variability of each location's affiliation to a cluster. It is noteworthy that the farther a location is from the pure variable, the lesser the corresponding affiliation.

Except for the corsican cluster, the interpretation of the resulting clusters seems straightforward. The extreme rainfall in northern France can be attributed to disturbances originating from the Atlantic, impacting regions like Brittany, Paris, and other northern areas. In the southern regions of France, particularly during the fall, intense rainfall events typically arise from southern winds compelling warm and moist air to interact with the mountainous terrain of the Pyrénées, Cévennes, and Alps. This interaction often leads to the development of severe thunderstorms. While these events can be quite localised, they frequently impact a substantial portion of the Mediterranean coastal area. In the eastern regions, despite the presence of various microclimates, the Vosges mountains serve a delineation between the temperate oceanic climate in the western part and the continental climate in the eastern part, particularly in the Upper Rhine Basin.

Since the exponent measure of the linear factor model  $\mathbf{X}$  in equation (5.1) is discrete, calculating probabilities of extreme events, denoted as  $\mathbb{P}\{\mathbf{X} \in C\}$  for a set of interest  $C$ , becomes a straightforward task. In our environmental dataset, determining regions and probabilities as

$$C_a(\mathbf{x}) = \cup_{j \in G_a} \{y^{(j)} > x^{(j)}\}, p_a(\mathbf{x}) = \mathbb{P}\{\mathbf{X} \in C_a(\mathbf{x})\}, a \in \{1, 2, 3\}$$

is a common approach, especially when an extreme event at any location could potentially result in a climatological catastrophe. Letting  $\hat{A}_{j\ell}$  be the element of the estimated  $\hat{A}$ , one can show that

$$\hat{p}_a(\mathbf{x}) = \sum_{\ell=1}^3 \max_{j \in G_a} \frac{\hat{A}_{j\ell}}{x^{(j)}}, a \in \{1, 2, 3\}.$$

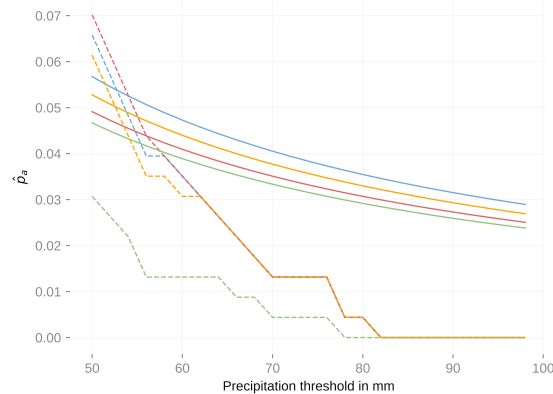


Fig. 5.4 Approximations of  $\hat{p}_a$  concerning precipitation in millimeters are illustrated through straight lines in blue, green, red and yellow representing the corsican, western, eastern and southern clusters, respectively. Empirical estimations are portrayed with dashed lines, mirroring the color code for their respective clusters.

In Figure 5.4, we illustrate  $\hat{p}_a(\mathbf{x})$  where  $\mathbf{x}$  is selected within the range of 50 to 100 mm for precipitation. The obtained estimation are of the same magnitude of those of (Kiriliouk and Zhou, 2022, Appendix B) for Switzerland. A comparison with empirical estimates reveals that the latter tends to underestimate the probability of heavy rainfall events. Moreover, the linear factor model exhibits the capability to extrapolate, maintaining informative values even as the empirical estimates plummet to zero, losing their informativeness.

### 5.6.2 Wildfires in French Mediterranean

Our case study focuses on the southeastern part of France, covering an area of 80500 km<sup>2</sup>. This region, prone to wildfires, encompasses a broad range of bioclimatic, environmental, and anthropogenic gradients. Approximately 60% of study area consists of easily ignitable forested areas or vegetation types, such as shrubland and other natural herbaceous vegetation. Wildfires face challenges in spreading through the various available cover types. The observation period for this study is 1995-2018, specifically during the extended summer months (June-October). Gridded weather reanalysis data from the SAFRAN model of Météo-France, with an 8km resolution, is utilised for analysis. This dataset has also undergone extensive examination in Koh et al. (2023), from which we obtain the data.

Understanding the joint impact of variables like temperature, precipitation, and wind speed on fire activity patterns is highly intricate. Various meteorological indices on fire activity patterns have been developed, including the widely used unitless Fire Weather Index (FWI), originally designed for Canadian forests. Typically, FWI values are directly interpreted and used for fire danger mapping. However, our approach involves studying its spatial variability through a linear factor model. In our methodology, we extract monthly maxima of FWI during the extended summer months over the 1143 pixels, resulting in 100 observations. Through a data-driven approach to select the threshold, as explained in Section 5.5.1, we choose  $\delta^* \approx 0.1765$  to obtain two latent factors (see Figure 5.5. These factors are directly interpretable in terms of elevation

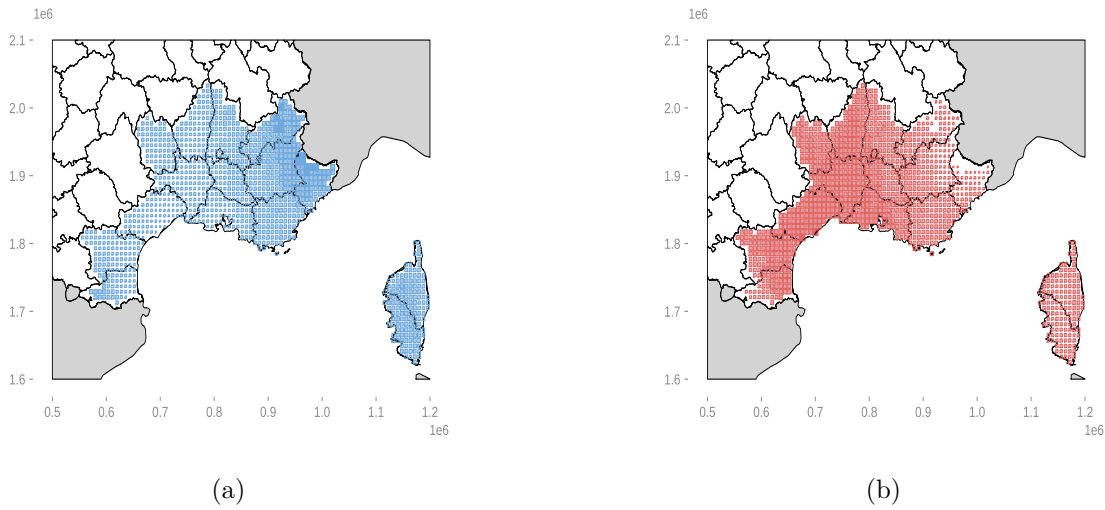


Fig. 5.5 In Panel **a**, we show the spatial cluster for the first latent variable. Panel **b** displays the spatial cluster for the second latent variable. The strength of association is indicated by the size and color intensity of each square.

(refer to Figure 5.5 Panel **a** and Figure 5.5 Panel **b**). Indeed, the first cluster demonstrates a strong association within mountainous areas (Western Alps, Corsica, and Pyrénées), while the second exhibits associations within lowlands prone to fire activity (Fréjaville and Curt (2015)) and heatwaves (Ruffault et al. (2016)), mid-elevation hinterlands, and foothills.

## 5.7 Discussion

We have introduced a comprehensive methodology for estimating the parameters of a discrete spectral measure of a max-stable distribution. Our approach lies into model-based clustering and proves to be both rapid and convenient, particularly suited for moderate dimensions. Additionally, we have provided statistical assurances for our method, ensuring favorable outcomes even in high dimensions where the relationship between the dimensionality, represented by  $d$ , and the sample size, denoted as  $n$ , may vary and potentially result in larger  $d$  values. These results are robust, grounded in general conditions that span a diverse array of applications. Our methodology however does not accommodate multivariate regularly varying distributions with an unknown tail index  $\alpha$ . Instead of the considered framework, let  $\mathbf{Z}$  consists of independent regularly varying random variables with a known tail index  $\alpha$ , then Theorem 5.2.1 can be reformulated as

$$\mathcal{X} = \bar{A} \odot \bar{A}^\top,$$

where  $\bar{A}$  represents the standardised loading matrix of  $\mathbf{X}$ , that is

$$\bar{A} = (\bar{A}_{ja})_{d \times K} = \left( \frac{A_{ja}^\alpha}{\sum_{a=1}^K A_{ja}^\alpha} \right)_{j=1, \dots, d, a=1, \dots, K}.$$

Then, the entire proposed procure applies to  $\bar{A}$  as long as the pure variable condition is satisfied, namely Condition (ii). Setting the scaling condition  $\sum_{a=1}^K A_{ja}^\alpha = 1$  for any  $j = 1, \dots, d$ , we can, however estimate the stable-tail dependence function of  $\mathbf{X}$ , as it is expressed as

$$L(z^{(1)}, \dots, z^{(d)}) = \sum_{a=1}^K \bigvee_{j=1, \dots, d} \left( A_{ja} z^{(j)} \right)^\alpha, \quad z^{(1)}, \dots, z^{(d)} > 0.$$

However, estimating the matrix  $A$  now requires the estimation of the tail index  $\alpha$ , demanding more intricate technical details within our non-asymptotic framework with weakly dependent observations. This aspect remain of significant interest for applications. Indeed, since marginal standardisation does not impact dependence modeling, it is a common practice to separate marginal and dependence modeling. This involves standardising each marginal to a common distribution and then focusing solely on modeling the dependence structure. However, in certain applications where the goal is to estimate failure regions of the form  $\{\sum_{j=1}^d v^{(j)} X^{(j)} > x\}$  with  $x$  being large and  $\sum_{j=1}^d v^{(j)} = 1$ ,  $v^{(j)} > 0$ , such an approach may be suboptimal. In these cases, it is necessary to directly model the original vector  $\mathbf{X}$ . These failure regions are particularly relevant in climate applications, as demonstrated by Kiriliouk and Naveau (2020) and Kiriliouk and Zhou (2022).

One can also consider the contaminated linear factor model

$$\mathbf{X} = \mathbf{AZ} + \sigma\eta,$$

where  $A \in \mathbb{R}^{d \times K}$  satisfies Condition (i),  $\mathbf{Z} = (Z^{(1)}, \dots, Z^{(K)})$  is a  $K$ -dimensional vector with i.i.d. standard Fréchet distributed components,  $\sigma > 0$  regulates the signal-to-noise ratio and  $\eta$  is a common factor noise distributed as a standard Fréchet. The standardised loading matrix  $\bar{A}$  of  $\mathbf{X}$  is expressed as:

$$\bar{A} = \left( \frac{A_{ja} + \sigma}{1 + \sigma} \right)_{j=1, \dots, d, a=1, \dots, K}.$$

The current challenge is that Theorem 5.2.2 (a) no longer applies since

$$\chi(i, j) = \frac{\sigma}{1 + \sigma} \neq 0, \quad i \in I_a, j \in I_b, b \neq a.$$

So, we can no longer recover latent factors using pairwise asymptotic independence obtained from the proposed model in (5.1). However, taking  $a \in [K]$  with  $|I_a| \geq 2$ , it is readily verified that

$$\chi(i, k) = \frac{A_{ka} + \sigma}{1 + \sigma} < 1 = \chi(i, j)$$

for any  $k \notin I_a$  and  $i, j \in I_a$ . Thus, a procedure to identify  $[K]$  is possible using the more stringent condition

**Condition (ii’’).** For any  $a \in \{1, \dots, K\}$ , there exist at least two indices  $j \in \{1, \dots, d\}$  such that  $A_{ja} = 1$  and  $A_{jb} = 0$ ,  $\forall b \neq a$ .

Without knowledge of  $\sigma$ , the matrix  $A$  can be recovered up to a multiplication constant (and by multiplication by a permutation matrix). It is also crucial to emphasise that such Condition (ii’’) paves the way to reduce the computational complexity of our procedure method.

A possible extension of the methodology is matrix-valued data, which has become increasingly prevalent in many applications. Most existing clustering methods for this type of data are tailored to the mean model and do not account for the (extremal) dependence structure of the variables. To extract information from the extremal dependence structure for clustering, we can propose a new latent variable model for the variables arranged in matrix form, with some unknown loading matrices representing the clusters for rows and columns.

Drawing an analogy with the linear factor model studied in this paper, assume the variables are stacked as a random matrix  $X \in \mathbb{R}^{p \times q}$  which follows the decomposition

$$X = AZB^\top + E,$$

where  $Z \in \mathbb{R}^{K_1 \times K_2}$  is a latent variable matrix which is regularly varying, meaning that there exist a scaling sequence  $\{c_n\}$  and a measure  $\Lambda_Z$  on  $\mathcal{M}_{K_1, K_2}(\mathbb{R}_+)$  such that the following vague convergence holds:

$$n\mathbb{P} \left\{ c_n^{-1} Z \in \cdot \right\} \xrightarrow[n \rightarrow \infty]{v} \Lambda_Z(\cdot).$$

$A \in \mathbb{R}^{p \times K_1}$  and  $B \in \mathbb{R}^{q \times K_2}$  are the unknown loading matrices for the rows and the columns, respectively.  $E \in \mathbb{R}^{p \times q}$  represents the random noise matrix with entries having lighter tails.





# APPENDIX D

## SUPPLEMENTARY MATERIALS OF CHAPTER 5

### D.1 Investigation into the computation time of clique algorithm

In this section, we explore the computation time required to identify a clique using the extremal correlation matrix, irrespective of whether the matrix is sparse or not. To achieve this, we examine the approach outlined in the main paper, which involves the following binary problem,

$$\begin{aligned} \max_{x^{(i)}} \quad & \sum_{i=1}^d x^{(i)} \\ \text{s.t.} \quad & x^{(i)} + x^{(j)} \leq 1, \quad \forall (i, j) \in \bar{E} \\ & x^{(i)} \in \{0, 1\}, \quad i = 1, \dots, d, \end{aligned}$$

and the Bron-Kerbosh algorithm (Bron and Kerbosch (1973)). The matrix  $A$  is constructed as follows: we designate the initial rows to comprise the first 20 pure variables. For generating  $A_j$ , where  $j \in J$ , we randomly choose the support from the set  $1, 2, \dots, 20$  with a sparsity  $s \in \{2, 3, \dots, 15\}$ . Subsequently, we form the extremal correlation matrix  $\mathcal{X} = A \odot A^\top$  and investigate a clique using the two aforementioned methods in 20 replications. We examine three scenarios with varying dimensions, denoted as  $d \in \{100, 200, 300\}$ . We denote the time spent recovering the clique through the adjacency matrix  $E$  computed with the extremal correlation matrix  $\mathcal{X}$  as  $T_{BK}$  for the Bron-Kerbosh algorithm and  $T_{MILP}$  for the binary problem. The results are illustrated in Figure D.1. For a concise interpretation of the numerical results, when  $d = 300$  and  $s = 3$ , the binary problem is 768 times faster than the Bron-Kerbosh algorithm, and conversely, the Bron-Kerbosh algorithm is 100 times faster when  $s = 15$ .

As anticipated, when the sparsity  $s$  is low ( $s < 4$ ), the binary problem proves to be the most effective in recovering the maximum clique, whereas the Bron-Kerbosh algorithm exhibits superior performance for a sparsity index  $s \geq 4$ . Regardless of the dimension, the (log) average ratio is decreasing and shows a rapid deceleration when  $s \geq 4$ . The contrast between the two methods becomes more pronounced with increasing considered dimensions.

### D.2 Algorithm

We give below the specifics of Algorithm (PureVar) motivated in Section (5.3), the Algorithm (HTSP) and summarize our final algorithm in Algorithm (SCRAM) (Soft Clustering lineAR fActor Model).

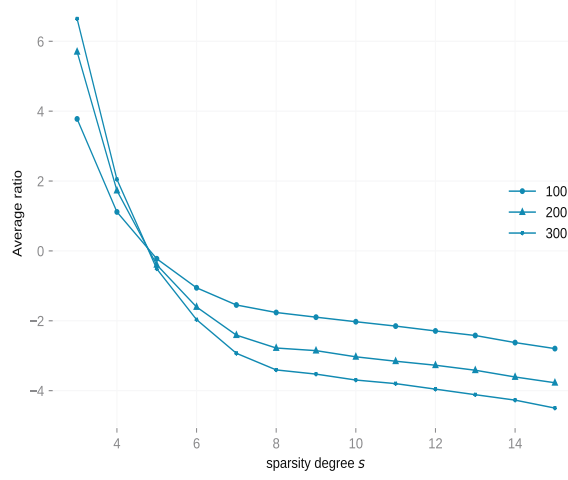


Fig. D.1 Average ratio  $T_{BK}/T_{MILP}$  in seconds set to the log scale with respect to the sparsity degree  $s \in \{2, 3, \dots, 15\}$  and  $d \in \{100, 200, 300\}$ .

---

**Algorithm (PureVar)**

---

- 1: **procedure** PUREVAR( $\hat{\mathcal{X}}, \delta$ )
  - 2:     Initialize:  $\mathcal{I} = \emptyset$
  - 3:     Construct the graph  $G = (V, E)$  where  $V = [d]$  and  $(i, j) \in E$  if  $\hat{\chi}_{n,m}(i, j) \leq \delta$
  - 4:     Find a maximum clique,  $\hat{\mathcal{G}}$ , of  $G$
  - 5:     **for**  $i \in \hat{\mathcal{G}}$  **do**
  - 6:          $\hat{I}^{(i)} = \{j \in [d] \setminus \{i\} : 1 - \hat{\chi}_{n,m}(i, j) \leq \delta\}$
  - 7:          $\hat{I}^{(i)} = \hat{I}^{(i)} \cup \{i\}$
  - 8:          $\hat{\mathcal{I}} = \text{MERGE}(\hat{I}^{(i)}, \hat{\mathcal{I}})$
  - 9:     Return  $\hat{\mathcal{I}}$  and  $\hat{K}$  as the number of sets in  $\hat{\mathcal{I}}$
- 

---

**Algorithm (MERGE)**

---

- 1: **procedure** MERGE( $\hat{I}^{(i)}, \hat{\mathcal{I}}$ )
  - 2:     **for**  $G \in \hat{\mathcal{I}}$  **do**
  - 3:         **if**  $G \cap \hat{I}^{(i)} \neq \emptyset$  **then**
  - 4:              $G = G \cap \hat{I}^{(i)}$
  - 5:         Return  $\hat{\mathcal{I}}$
  - 6:      $\hat{I}^{(i)} \in \hat{\mathcal{I}}$
  - 7:     Return  $\hat{\mathcal{I}}$
-

---

**Algorithm (HTSP)**


---

- 1: **procedure** HTSP( $\hat{\mathcal{X}}, \delta, \hat{I}$ )
  - 2:   **for**  $j \in [d] \setminus \hat{I}$  **do**
  - 3:      $\bar{\chi}^{(j)} = \left( \frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} \hat{\chi}_{n,m}(i, j) \right)_{a=1, \dots, \hat{K}}$
  - 4:      $\bar{\beta}^{(j)} = \left( \bar{\chi}_a^{(j)} \mathbb{1}_{\{\bar{\chi}_a^{(j)} > \delta\}} \right)_{a=1, \dots, \hat{K}}$
  - 5:      $\hat{\mathcal{S}} = \text{supp}(\bar{\beta}^{(j)})$
  - 6:      $\hat{\beta}^{(j)}|_{\hat{\mathcal{S}}} = \mathcal{P}_{\Delta_{\hat{K}-1}}(\bar{\beta}^{(j)}|_{\hat{\mathcal{S}}}), \hat{\beta}^{(j)}|_{\hat{\mathcal{S}}^c} = 0$
- 

---

**Algorithm (SCRAM)**


---

- 1: **procedure** SCRAM( $\hat{\mathcal{X}}$ , the tuning parameter  $\delta$ )
  - 2:   Apply Algorithm (PureVar) to obtain the number of clusters  $\hat{K}$ , the estimated set of pure variables  $\hat{I}$  and its partition of  $\hat{\mathcal{I}}$ .
  - 3:   Estimate  $A_I$  by  $\hat{A}_{\hat{I}}$  from (5.9).
  - 4:   Estimate  $A_J$  by  $\hat{A}_j$  applying Algorithm (HTSP). Combine  $\hat{A}_{\hat{I}}$  with  $\hat{A}_j$  to obtain  $\hat{A}$ .
  - 5:   Estimate fuzzy clusters  $\hat{\mathcal{G}} = \{\hat{G}_1, \dots, \hat{G}_K\}$  from (5.12) by using  $\hat{A}$ .
  - 6:   Output  $\hat{A}$  and  $\hat{\mathcal{G}}$ .
- 

## D.3 Proofs of Section 5.2

**Proof of Theorem 5.2.1** Let  $i, j \in \{1, \dots, d\}$  be arbitrary with  $i \neq j$ . Define  $Y^{(i)} = \sum_{a=1}^K A_{ia} Z^{(a)}$  and  $Y^{(j)} = \sum_{a=1}^K A_{ja} Z^{(a)}$ . Note that  $\mathbf{Y}$  and  $\mathbf{X}$  have the same exponent measure since they differ only by a sum of a random variable with a lighter tail (see, e.g., (Kulik and Soulier, 2020, Lemma 1.3.2)). So we only have to compute bivariate extremal correlations for  $\mathbf{Y}$  to obtain those of  $\mathbf{X}$ . In order to obtain bivariate regular variation of  $\mathbf{Y}^{(i,j)} = (Y^{(i)}, Y^{(j)})$ , consider the map  $\psi$  from  $\mathbb{R}_+^K \rightarrow [0, \infty)^2$  defined by

$$\psi(z^{(1)}, \dots, z^{(K)}) = \left( \sum_{a=1}^K A_{ia} z^{(a)}, \sum_{a=1}^K A_{ja} z^{(a)} \right).$$

For a measurable subset  $A$  of  $\mathbb{R}^2$ , separated from 0, we obtain by corollary 2.1.14 of Kulik and Soulier (2020):

$$\begin{aligned} \Lambda_{\mathbf{Y}^{(i,j)}}(A) &= \Lambda_{\mathbf{Z}} \circ \psi^{-1}(A) = \sum_{a=1}^K \delta_0 \otimes \dots \otimes \Lambda_{Z^{(a)}} \otimes \dots \otimes \delta_0 \circ \psi^{-1}(A) \\ &= \sum_{a=1}^K \int_0^\infty \mathbb{1}_A(A_{ia}s, A_{ja}s) s^{-2} ds. \end{aligned}$$

Applying to  $A = (1, \infty) \times (1, \infty)$  and  $A = (1, \infty) \times \mathbb{R}_+$ , we get respectively

$$\begin{aligned} \Lambda_{\mathbf{Y}^{(i,j)}}((1, \infty) \times (1, \infty)) &= \Lambda_{\mathbf{Z}} \circ \psi^{-1}(A) = \sum_{a=1}^K \int_0^\infty \mathbb{1}_{(1, \infty) \times (1, \infty)}(A_{ia}s, A_{ja}s) s^{-2} ds \\ &= \sum_{a=1}^K \int_0^\infty \mathbb{1}_{\{s > 1/A_{ia}, s > 1/A_{ja}\}} s^{-2} ds \\ &= \sum_{a=1}^K \left( \frac{1}{A_{ia}} \vee \frac{1}{A_{ja}} \right)^{-1} = \sum_{a=1}^K (A_{ia} \wedge A_{ja}), \end{aligned}$$

and

$$\Lambda_{\mathbf{Y}^{(i,j)}}((1, \infty) \times \mathbb{R}_+) = \sum_{a=1}^K A_{ia} = 1.$$

Thus

$$\chi(i, j) = \lim_{x \rightarrow \infty} \frac{\mathbb{P}\{Y^{(i)} > x, Y^{(j)} > x\}}{\mathbb{P}\{Y^{(i)} > x\}} = \frac{\Lambda_{\mathbf{Y}^{(i,j)}}((1, \infty) \times (1, \infty))}{\Lambda_{\mathbf{Y}^{(i,j)}}((1, \infty) \times \mathbb{R}_+)} = \sum_{a=1}^K A_{ia} \wedge A_{ja}.$$

□

We know state and prove two lemmata that are crucial for the main results of this section. All results are proved under the condition that model (5.1) and Conditions (i)-(ii) hold.

**Lemma D.3.1.** *For any  $a \in [K]$ ,  $i \in I_a$  and  $|I_a| \geq 2$  we have*

1.  $\chi(i, j) = 1$  for all  $j \in I_a$ ,
2.  $\chi(i, j) < 1$  for all  $j \notin I_a$ .

**Proof of Lemma D.3.1** For any given  $i \in \{1, \dots, d\}$ , we define the set  $s(i) := \{1 \leq a \leq K : A_{ia} \neq 0\}$ . For any  $i \in I_a$  and  $j \neq i$ , we have

$$\chi(i, j) = \sum_{a \in s(i)} A_{ia} \wedge A_{ja} = A_{ja} \leq 1,$$

we observe that we have equality in the above display for  $j \in I_a$  and strict inequality for  $j \notin I_a$  which proves the lemma. □

**Lemma D.3.2.** *We have  $S_i \cup \{i\} = I_a$  and  $M_i = 1$  for any  $i \in I_a$ , with  $|I_a| \geq 2$  and  $a \in [K]$ .*

**Proof of Lemma D.3.2** Lemma D.3.1 implies that, for any  $i \in I_a$ ,  $M_i = 1$  and  $S_i = I_a \setminus \{i\}$  which proves the lemma. □

### Proof of Theorem 5.2.2

**Proof of (a)** By condition (ii), for any  $a \in [K]$ , there exists  $i_a \in [d]$  such that  $X^{(i_a)} = Z^{(a)} + E^{(i_a)}$ . By its very nature under the model (5.1), the vector  $(X^{(i_1)}, \dots, X^{(i_K)})$  is the largest vector being asymptotically independent, i.e.,

$$\chi(i, j) = 0, \quad \forall i, j \in \{i_1, \dots, i_K\}, \tag{D.1}$$

see (Resnick, 2008, Proposition 5.24). Let us construct the simple undirected graph  $G = (V, E)$  with a finite set of vertices  $V = [d]$  and a finite set of ordered pairs  $(i, j)$  of edges such that  $(i, j) \in E$  if  $\chi(i, j) = 0$ . Through this construction of  $G$ , the search for a maximum clique in  $G$  is equivalent to searching for a set of indices, denoted as  $\{i_1, \dots, i_K\}$ , satisfying equation (D.1). Consequently, we established (a).

**Proof of (b)** Consider any  $j \in [d]$  with  $M_i = 1$  for  $i \in I_a$ . Since  $|I_a| \geq 2$ , by Lemma D.3.1, the maximum is achieved for any pairs  $j, k \in I_a$ . However, if  $j \notin I_a$ , we have  $\chi(j, k) < 1$  for all  $k \neq j$ . Hence  $j \in I_a$  and this concludes the proof of the sufficiency part. It remains to prove the necessity part.

Let  $i \in I_a$  for some  $a \in [K]$  and  $j \in I_a \cap S_i$ . Since  $j \in S_i$  and  $|I_a| \geq 2$ , we have  $\chi(i, j) = M_i = 1$ , as a result of Lemma D.3.2, which proves (b).

**Proof of (c)** We start with the following construction approach. Let  $N = [d]$  be the set of all variables indices and  $O = \emptyset$ . Let  $M_i$  and  $S_i$  be defined in (5.5) and (5.6), respectively.

- (1) Construct the undirected graph  $G = (V, E)$  where  $(i, j) \in E$  if  $\chi(i, j) = 0$ .
- (2) Find a maximum clique of  $G$  denoted as  $\bar{Q}$ .
- (3) Choose  $i \in \bar{Q}$  and calculate  $M_i$  and  $S_i$ .
  - (a) If  $M_i = 1$ , set  $I^{(i)} = S_i \cup \{i\}$ ,  $O = O \cup \{i\}$  and  $\bar{Q} \setminus \{i\}$ .
  - (b) Otherwise, replace  $\bar{Q}$  by  $\bar{Q} \setminus \{i\}$ .
- (4) Repeat Step (3) until  $\bar{Q} = \emptyset$ .

We show that  $\{I^{(i)} : i \in O\} = \mathcal{I}$ . Let  $i \in O$  be arbitrary fixed. By (a) and (b) of Theorem 5.2.2, we have  $i \in I$ . Thus, there exists  $a \in [K]$  such that  $i \in I_a$ . If  $|I_a| \geq 2$ , by Lemma D.3.2,  $i \in I_a$  implies  $I_a = S_i \cup \{i\} = I^{(i)}$ . On the other hand, let  $a \in [K]$  be arbitrary fixed. By condition (ii), there exists, at least one  $i \in I_a$ . If  $|I_a| = 1$ , then by (a), we have  $I^{(i)} = I_a$ . If  $|I_a| \geq 2$  and  $j \in I_a$ , then  $\chi(i, j) = 1$  and  $j \in S_i$ , once again, by Lemma D.3.2,  $S_i \cup \{i\} = I_a$ , that is  $I^{(i)} = I_a$ .  $\square$

### Proof of Theorem 5.2.3

Theorem 5.2.2 establishes that the set  $\mathcal{X}$  uniquely determines both  $I$  and its partition  $\mathcal{I}$ , with the exception of potential permutations of labels. When we have  $I$  and its partition  $\mathcal{I}$  available, represented as  $\{I_1, \dots, I_K\}$ , for any index  $i$  belonging to  $I$ , there exists a single integer  $1 \leq a \leq K$  such that  $i \in I_a$ . We then construct a row vector  $A_i$  of dimension  $K$ , akin to the canonical basis  $\mathbf{e}_a$  in  $\mathbb{R}^K$ , where the element at position  $a$  equals 1, and all other elements are 0. Consequently, the matrix  $A_I$ , which has dimensions  $|I| \times K$  and is composed of rows  $A_i$ , is uniquely determined, except for possible multiplications by permutation matrices.

We show below that  $A_J$  is also identifiable up to a signed permutation matrix. We begin by observing that, for each  $i \in I_k$  for some  $k \in [K]$  and any  $j \in J$ , Model (5.1) implies

$$\chi(i, j) = \sum_{a \in s(i)} A_{ia} \wedge A_{ja} = A_{jk}$$

and, after averaging over all  $i \in I_k$ ,

$$A_{jk} = \frac{1}{|I_k|} \sum_{i \in I_k} \chi(i, j).$$

Repeating this for every  $k \in [K]$ , we obtain the formula

$$A_j = \left( \frac{1}{|I_1|} \sum_{i \in I_1} \chi(i, j), \dots, \frac{1}{|I_K|} \sum_{i \in I_K} \chi(i, j) \right),$$

for each  $j \in J$ , which shows that  $A_j$  can be determined uniquely from  $\mathcal{X}$  up to a permutation. Therefore,  $A_j$  is identifiable which concludes the proof.  $\square$

## D.4 Proof of Section 5.3

For the sake of notations, we set  $\hat{\nu}_{n,m}(\{1, \dots, d\}) := \hat{\nu}_{n,m}$ .

**Lemma D.4.1.** *Let  $(\mathbf{X}_t, t \in \mathbb{Z})$  satisfies the conditions in Theorem 5.4.1. Choose a  $c_2 \in (0, \infty)$ , and let  $z = y - k^{-c_2}$  for any  $y \geq k^{-c_2}$ . Then for  $k \geq 4$ , there is a constant  $c_1 > 0$  such that*

$$\mathbb{P} \left\{ |\hat{\nu}_{n,m} - \nu_m| \geq y + \frac{1}{k+1} \right\} \leq (2d+1) \exp \left\{ -\frac{c_1 k z^2}{1 + z \ln k (\ln \ln k)} + c_2 \ln k \right\}.$$

**Proof** Since for every  $j \in \{1, \dots, d\}$ ,  $U_{m,1}^{(j)}$  is uniformly distributed under the unit segment, we directly obtain

$$\mathbb{E} \left[ \frac{1}{d} \sum_{j=1}^d U_{m,1}^{(j)} \right] = \frac{1}{2}.$$

Furthermore, by simple computations, and using the very nature of scaled ranks, we have

$$\frac{1}{k} \sum_{i=1}^k \frac{1}{d} \sum_{j=1}^d \hat{U}_{n,m,i}^{(j)} = \frac{1}{k} \sum_{i=1}^k \frac{1}{d} \sum_{j=1}^d \frac{i}{k+1} = \frac{1}{2}.$$

Hence, the desired quantity that we want to control can be simply rewritten as:

$$|\hat{\nu}_{n,m} - \nu_m| = \left| \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d \hat{U}_{n,m,i}^{(j)} - \mathbb{E} \bigvee_{j=1}^d U_{m,i}^{(j)} \right] \right|.$$

Introduce by  $\tilde{U}_{n,m,i}^{(j)}$  the scaled ranks using the factor  $k$  (instead of  $k+1$  for  $\hat{U}_{n,m,i}^{(j)}$ ),  $j = 1, \dots, d$ ,  $i = 1, \dots, k$ . Direct computation gives

$$|\hat{\nu}_{n,m} - \nu_m| \leq \frac{1}{k+1} + \left| \frac{1}{k} \sum_{i=1}^k \left[ \bigvee_{j=1}^d \tilde{U}_{n,m,i}^{(j)} - \mathbb{E} \bigvee_{j=1}^d U_{m,i}^{(j)} \right] \right|.$$

The remaining term can be upper bounded by

$$E_1 + E_2 := \left| \frac{1}{k} \sum_{i=1}^k \left[ \prod_{j=1}^d \tilde{U}_{n,m,i}^{(j)} - \prod_{j=1}^d U_{m,i}^{(j)} \right] \right| + \left| \frac{1}{k} \sum_{i=1}^k \left[ \prod_{j=1}^d U_{m,i}^{(j)} - \mathbb{E} \prod_{j=1}^d U_{m,i}^{(j)} \right] \right|.$$

Furthermore,  $E_1$  is upper bounded by the following well-known quantity

$$E_1 \leq \sup_{j \in \{1, \dots, d\}} \sup_{i \in \{1, \dots, k\}} \left| \tilde{U}_{n,m,i}^{(j)} - U_{m,i}^{(j)} \right| \leq \sup_{j \in \{1, \dots, d\}} \sup_{x \in \mathbb{R}} \left| \hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x) \right|.$$

Thus, applying union bound and Lemma D.7.3 to  $E_1$  and Lemma D.7.2 to  $E_2$  (taking  $n = k$  in the statement of both lemmas), we deduce the result.  $\square$

In the proof of Theorem 5.4.1, for the sake of simplicity in presentation, we make the blanket assumption that  $\hat{\nu}_{n,m}(i, j) \leq 1/6$  (since, by construction,  $\hat{\nu}_{n,m}(i, j) \geq 0$ ), so that, using Lemma B.1.1 in Chapter 3,

$$|\hat{\chi}_{n,m}(i, j) - \chi_m(i, j)| \leq 9|\hat{\nu}_{n,m}(i, j) - \nu_m(i, j)|.$$

If the assumption does not hold, one can introduce the estimator  $\tilde{\chi}_{n,m}(i, j) = \max(\hat{\chi}_{n,m}(i, j), 0)$  and  $\tilde{\nu}_{n,m}(i, j) = \min(\hat{\nu}_{n,m}(i, j), 1/6)$ . Notice that these estimators are, by construction, projection estimators since:

$$\tilde{\chi}_{n,m}(i, j) = \arg \min_{x \in [0, 1]} |\hat{\chi}_{n,m}(i, j) - x|, \quad \tilde{\nu}_{n,m}(i, j) = \arg \min_{x \in [0, 1/6]} |\hat{\nu}_{n,m}(i, j) - x|.$$

Consequently, we obtain the same bound given in Theorem 5.4.1 for  $|\tilde{\chi}_{n,m}(i, j) - \chi_m(i, j)|$  using the following arguments:

$$\tilde{\chi}_{n,m}(i, j) = \max \left( 2 - \frac{0.5 + \hat{\nu}_{n,m}(i, j)}{0.5 - \hat{\nu}_{n,m}(i, j)}, 0 \right) = \max (f(\hat{\nu}_{n,m}(i, j)), f(1/6)) = f(\tilde{\nu}_{n,m}(i, j)),$$

where

$$f: [0, 1/6] \rightarrow [0, 1] \\ x \mapsto 2 - \frac{0.5 + x}{0.5 - x},$$

which is nonincreasing and 9-Lipschitz by Lemma B.1.1 in Chapter 3. By definition  $\chi_m(i, j) = \max(\chi_m(i, j), 0)$ . We obtain

$$\begin{aligned} |\tilde{\chi}_{n,m}(i, j) - \chi_m(i, j)| &= |f(\min(\hat{\nu}_{n,m}(i, j), 1/6)) - f(\min(\nu_m(i, j), 1/6))| \\ &\leq 9|\tilde{\nu}_{n,m}(i, j) - \nu_m(i, j)|, \end{aligned}$$

and using that

$$|\tilde{\nu}_{n,m}(i, j) - \nu_m(i, j)| = |\min(\hat{\nu}_{n,m}(i, j), 1/6) - \min(\nu_m(i, j), 1/6)| \leq |\hat{\nu}_{n,m}(i, j) - \nu_m(i, j)|,$$



using either  $\min(x, y) = \frac{x+y-|x-y|}{2}$  or the projection onto a convex set is 1-Lipschitz to obtain the last inequality. we manage to obtain the same bound for  $|\tilde{\chi}_{n,m}(i, j) - \chi_m(i, j)|$  as  $|\hat{\chi}_{n,m}(i, j) - \chi_m(i, j)|$  as stated.

**Proof of Theorem 5.4.1** Taking Lemma D.4.1 with  $d = 2$  and  $c_2 = 1$ , we obtain

$$\mathbb{P} \left\{ |\hat{\nu}_{n,m}(i, j) - \nu_m(i, j)| \geq y + \frac{1}{k+1} \right\} \leq 5 \exp \left\{ -\frac{c_1 k z^2}{1 + z \ln k (\ln \ln k)} + \ln k \right\},$$

where  $z = y - k^{-1}$ . Now taking

$$z = c_1 \left( \sqrt{\frac{\ln \left( \frac{k}{\delta} \right)}{k}} + \frac{\ln k \ln \ln(k) \ln \left( \frac{k}{\delta} \right)}{k} \right), \quad (\text{D.2})$$

it implies that with probability at least  $1 - \delta$

$$|\hat{\nu}_{n,m}(i, j) - \nu_m(i, j)| \leq c_1 \left( \sqrt{\frac{\ln \left( \frac{k}{\delta} \right)}{k}} + \frac{\ln k \ln \ln(k) \ln \left( \frac{k}{\delta} \right)}{k} \right) + \frac{1}{k+1} + \frac{1}{k}.$$

Using now Lemma B.1.1 in Chapter 3, stating that,

$$|\hat{\chi}_{n,m}(i, j) - \chi_m(i, j)| \leq 9 |\hat{\nu}_{n,m}(i, j) - \nu_m(i, j)|,$$

and we obtain that with probability at least  $1 - d^{-c_0}$  through a union bound

$$\sup_{1 \leq i < j \leq d} |\hat{\chi}_{n,m}(i, j) - \chi_m(i, j)| \leq c_1 \left( \sqrt{\frac{\ln(kd)}{k}} + \frac{\ln k \ln \ln(k) \ln(kd)}{k} \right),$$

for a sufficiently large constant  $c_1$ , thus the desired result.  $\square$

## D.5 Proof of Section 5.4

### D.5.1 Proof of Section 5.4.1

**Lemma D.5.1.** *Under Condition (SSC), for any  $i \in I_a$  with some  $a \in [K]$ , the following inequalities hold on the event*

- (A1)  $\hat{\chi}_{n,m}(i, j) \leq \delta$  for all  $j \in I_b$  for some  $b \in [K]$  with  $b \neq a$ ;
- (A2)  $1 - \hat{\chi}_{n,m}(i, j) \leq \delta$  for all  $j \in I_b$ ;
- (A3)  $1 - \hat{\chi}_{n,m}(i, k) > \delta$  for all  $k \notin I_a$ .

**Proof** For the entire proof, we work under the event  $\mathcal{E}$  defined in (5.15). To prove (A1), we observe that for any  $j \in I_b$ , with  $b \in [K]$  and  $b \neq a$ ,  $\chi(i, j) = 0$ , whence

$$\hat{\chi}_{n,m}(i, j) \leq \chi(i, j) + \delta = \delta.$$

So (A1) holds.

To prove (A2), taking  $j \in I_a$  gives  $\chi(i, j) = 1$ , then

$$1 - \hat{\chi}_{n,m}(i, j) \leq 1 - \chi(i, j) + \delta = \delta,$$

and hence (A2) holds.

To obtain (A3), for  $k \notin I_a$ , Condition (SSC1) implies

$$\chi(i, j) = A_{ka} < 1 - 2\delta.$$

Next, we obtain

$$1 - \hat{\chi}_{n,m}(i, j) \geq 1 - \delta - \chi(i, j) = 1 - \delta - A_{ka} > \delta.$$

□

**Proof of Theorem 5.4.2** We work on the event  $\mathcal{E}$  throughout the proof. Without loss of generality, we assume that the label permutation  $\pi$  is the identity. Following Bing et al. (2020), we start by point out that the three following claims are sufficient to prove (a)-(c). Let  $[\hat{K}]$  and  $\hat{I}^{(i)}$  be respectively the set of integers in the maximum clique  $\bar{\mathcal{G}}$  of  $G$  given in Step 4 of Algorithm (PureVar) and the set defined in Step 6 of Algorithm (PureVar).

- (1) For any  $i \in J$ , we have  $\text{Pure}(i) = \text{False}$ ;
- (2) For any  $i \in I_a$  and  $a \in [K]$ , we have  $\text{Pure}(i) = \text{True}$ ,  $I_a = \hat{I}^{(i)}$ .

To prove (1), let  $i \in J$  be fixed. We first prove that  $\text{Pure}(i) = \text{False}$  under  $\hat{I}^{(i)} \cap I \neq \emptyset$ . Under this hypothesis, we have  $[\hat{K}] \subseteq [K]$  and no variables  $i \in J$  belongs to  $[\hat{K}]$  by Step 4 of Algorithm (PureVar). Now, if  $i$  was taken at Step 6 of Algorithm (PureVar), then by  $\hat{I}^{(i)} \cap I \neq \emptyset$ , there exists  $b \in [K]$  and  $j \in I_b$  such that

$$1 - \hat{\chi}_{n,m}(i, j) \leq \delta,$$

which is prevented from (A3) of Lemma D.5.1. This shows that for any  $i \in J$ , if  $\hat{I}^{(i)} \cap I \neq \emptyset$ , then  $\text{Pure}(i) = \text{False}$ .

Therefore to complete the proof of (1), we show  $\hat{I}^{(i)} \cap I \neq \emptyset$  is impossible when  $i \in J$ , under our assumptions. By construction of the algorithm, we have to verify that no  $i \in J$  belongs to  $i \in [\hat{K}]$  in Step 4 of Algorithm (PureVar). We have, using (A1) of Lemma D.5.1, for every  $k \in I_a$  and  $j \in I_b$  with  $a, b \in [K]$

$$\hat{\chi}_{n,m}(k, j) \leq \delta.$$

Hence  $[K]$  is a clique and  $[K] \subseteq [\hat{K}]$ . Now suppose  $i \in [\hat{K}]$  while  $i \in J$ , then we have

$$\hat{\chi}_{n,m}(i, j) \leq \delta \text{ for any } j \in [\hat{K}], j \neq i. \tag{D.3}$$

Take  $k \in I_{a^*}$  and  $j \in I_{b^*}$  for  $a^*, b^* \in [K]$  such that  $A_{ia^*} > 2\delta$  and  $A_{ib^*} > 2\delta$  where the existence of such indices in  $[K]$  is guaranteed by Condition (SSC2). We hence obtain

$$\chi(i, k) = \sum_{a \in s(i)} A_{ia} \wedge A_{ka} \geq A_{ia^*} \wedge A_{ka^*} = A_{ia^*}$$

where the last inequality follows from  $k \in I_{a^*}$ . Hence,

$$\chi(i, k) > A_{ia^*} > 2\delta$$

where the last inequality stems down from Condition **(SSC2)**. Then, under  $\mathcal{E}$ ,

$$\hat{\chi}_{n,m}(i, k) \geq \chi(i, k) - \delta > \delta.$$

The same arguments hold for  $j \in I_{b^*}$  and hence

$$\hat{\chi}_{n,m}(i, k) > \delta \text{ and } \hat{\chi}_{n,m}(i, j) > \delta,$$

which contradicts **(D.3)** and guarantees that  $\hat{I}^{(i)} \cap I = \emptyset$  is impossible when  $i \in J$ . Indeed, the maximum clique that we can obtain from Step 4 of Algorithm **(PureVar)** is  $[\hat{K}] \setminus \{i\}$  by the above inequality.

To prove **(2)**, since  $\hat{I}^{(i)} \cap I \neq \emptyset$  with  $i \in I_a$  under  $\mathcal{E}$  from the discussion of **(1)**, then the statement of **(2)** should only be verified at step 6 of the algorithm since only pure variables are gathered at step 4 of the algorithm. Now, from step 6 of Algorithm **(PureVar)**, we have to show that

$$1 - \hat{\chi}_{n,m}(i, j) \leq \delta,$$

for any  $j \in \hat{I}^{(i)}$  and  $j \in I_a$ . Since  $i \in I_a$ , **(A2)** in Lemma **D.5.1** states that the above inequality stands. Thus we have shown that for any  $i \in I_a$ , **Pure(i) = True**. We conclude the proof.  $\square$

## D.5.2 Proof of Section 5.4.2

### Proof of Theorem 5.4.3

**Proof of (a)** The proof of Theorem 5.4.3, item **(a)** implies two steps:

- (S1) We write  $\bar{A} = AP$ , and prove the first error bound for  $\hat{A}_{\hat{I}} - \bar{A}_{\hat{I}}$ ;
- (S2) We prove the error bounds  $\hat{A}_j - \bar{A}_j$ .

For ease of the notation and without loss of generality, we make the blanket assumption that the permutation matrix  $P$  is the identity so that  $\bar{A} = A$  for the remainder of the proof. Let  $s(j) = \|A_j\|_0$  for  $j = 1, \dots, d$ . For the first step **(S1)**, from the construction of  $\hat{A}_{\hat{I}}$  and parts **(a)-(c)** in Theorem 5.4.2, we have for any  $i \in \hat{I}_a$  and the definition of  $I$  implies  $A_{ia} = 1$ . Then

$$\|\hat{A}_{\hat{I}} - A_{\hat{I}}\|_{\infty} = \max_{j \in \hat{I}} \|\hat{A}_j - A_j\|_{\infty} = 0$$

Then for any  $j \in \hat{I}$ , we have

$$\|\hat{A}_j - A_j\|_2 = 0.$$

For the second step of the proof **(S2)**, we will make use of the results of the Lemma stated here first and proved at the end of this section.

**Lemma D.5.2.** *Under the conditions of Theorem 5.4.3, on the event  $\mathcal{E}$ , we have  $\beta_a^{(j)} = 0$  implies  $\bar{\beta}_a^{(j)} = 0$ , for any  $j \in \hat{J}$  and  $a \in [\hat{K}]$ .*

Let us head into the proof of (S2). For each  $j \in \hat{J}$ , we have by the very nature of our estimator:

$$\|\hat{A}_j - A_j\|_2 = \|\hat{\beta}^{(j)} - \beta^{(j)}\|_2 \leq \|\bar{\beta}^{(j)} - \beta^{(j)}\|_2 + \|\hat{\beta}^{(j)} - \bar{\beta}^{(j)}\|_2.$$

Because this is a projection

$$\|\hat{\beta}^{(j)}\big|_{\hat{\mathcal{S}}} - \bar{\beta}^{(j)}\big|_{\hat{\mathcal{S}}}\|_2 \leq \|\beta^{(j)}\big|_{\hat{\mathcal{S}}} - \bar{\beta}^{(j)}\big|_{\hat{\mathcal{S}}}\|_2,$$

hence

$$\|\hat{\beta}^{(j)} - \bar{\beta}^{(j)}\|_2 \leq \|\beta^{(j)} - \bar{\beta}^{(j)}\|_2.$$

Then

$$\|\hat{A}_j - A_j\|_2 = \|\hat{\beta}^{(j)} - \beta^{(j)}\|_2 \leq 2\|\bar{\beta}^{(j)} - \beta^{(j)}\|_2.$$

Furthermore, we can show that

$$\|\bar{\beta}^{(j)} - \beta^{(j)}\|_\infty \leq 2\delta,$$

indeed, for any  $a \in \{1, \dots, K\}$  with  $\bar{\chi}_a^{(j)} > \delta$  we have

$$|\bar{\beta}_a^{(j)} - \beta_a^{(j)}| = \left| \frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} \hat{\chi}_{n,m}(i, j) - A_{ja} \right| \leq \frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} |\hat{\chi}_{n,m}(i, j) - A_{ja}|.$$

By Theorem 5.4.2,  $\hat{I} = I$ , then if  $i \in I_a$ , then  $A_{ja} = \chi(i, j)$  and as we are on the event  $\mathcal{E}$ , we obtain that

$$|\hat{\chi}_{n,m}(i, j) - \chi(i, j)| \leq \delta.$$

Thus,

$$|\bar{\beta}_a^{(j)} - \beta_a^{(j)}| \leq 2\delta,$$

whenever  $a \in \{1, \dots, K\}$  with  $\bar{\chi}_a^{(j)} > \delta$ . Now take  $a \in \{1, \dots, K\}$  such that  $\bar{\chi}_a^{(j)} \leq \delta$ , we obtain

$$|\bar{\beta}_a^{(j)} - \beta_a^{(j)}| = A_{ja}.$$

If  $i \in I_a$ , then  $\chi(i, j) = A_{ja}$  and under the event  $\mathcal{E}$ , we obtain

$$A_{ja} \leq \hat{\chi}_{n,m}(i, j) + \delta.$$

Then for any  $i \in I_a$

$$A_{ja} \leq \frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} \hat{\chi}_{n,m}(i, j) + \delta = \bar{\chi}_a^{(j)} + \delta \leq 2\delta.$$

Hence, we have, as stated

$$\|\bar{\beta}^{(j)} - \beta^{(j)}\|_\infty \leq 2\delta.$$

Then following Lemma D.5.2 and using  $\hat{K} = K$  on the event  $\mathcal{E}$  gives

$$\|\bar{\beta}^{(j)} - \beta^{(j)}\|_2 = \left( \sum_{a=1}^K |\bar{\beta}_a^{(j)} - \beta_a^{(j)}|^2 \right)^{1/2} = \left( \sum_{a \in s(j)} |\bar{\beta}_a^{(j)} - \beta_a^{(j)}|^2 \right)^{1/2} \leq 2\sqrt{s(j)}\delta.$$

This completes the proof of the last step and of Theorem 5.4.3 (a).

**Proof of (b).** In the initial stage of the proof, let us demonstrate that, under the event  $\mathcal{E}$

$$\text{supp}(\hat{\beta}^{(j)}) = \text{supp}(\bar{\beta}^{(j)}), \forall j \in \{1, \dots, d\}. \quad (\text{D.4})$$

Through our initial construction, we immediately obtain that  $\hat{\beta}_a^{(j)} = 0$  whenever  $\bar{\beta}_a^{(j)} = 0$  for any  $a \in [\hat{K}]$ . Now, let us consider any  $a \in [\hat{K}]$  with  $\bar{\beta}_a^{(j)} > 0$ , a condition equivalent to  $\bar{\beta}_a^{(j)} > \delta$ . Our task is to establish that  $\bar{\beta}_a^{(j)} > \tau$  where

$$\tau := \frac{1}{\rho} \left( \sum_{a=1}^{\rho} \bar{\beta}_a^{(j)} - 1 \right) \text{ and } \rho = \max \left\{ p \in \text{supp}(\bar{\beta}^{(j)}) : \bar{\beta}_p^{(j)} > \frac{1}{p} \left( \sum_{a=1}^p \bar{\beta}_a^{(j)} - 1 \right) \right\}.$$

Let us show that  $p = \text{supp}(\bar{\beta}^{(j)})$ . Indeed for any  $a \in \text{supp}(\bar{\beta}^{(j)})$

$$\bar{\beta}_a^{(j)} = \frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} \hat{\chi}_{n,m}(i, j).$$

If  $i \in I_a$ , then  $\chi(i, j) = A_{ja}$  and we obtain the following inequality under the event  $\mathcal{E}$

$$\hat{\chi}_{n,m}(i, j) \leq A_{ja} + \delta.$$

We obtain simultaneously

$$\hat{\chi}_{n,m}(i, j) \leq A_{ja} + \delta, \forall i \in \hat{I}_a, \quad \bar{\beta}_a^{(j)} \leq A_{ja} + \delta, \quad a \in \text{supp}(\bar{\beta}^{(j)}).$$

Summing across all instances of  $a \in \text{supp}(\bar{\beta}^{(j)})$  and employing Condition (i),

$$\sum_{a=1}^p \bar{\beta}_a^{(j)} \leq 1 + p\delta,$$

this leads us to achieve

$$\frac{1}{p} \left( \sum_{a=1}^p \bar{\beta}_a^{(j)} - 1 \right) \leq \delta.$$

Building upon our initial assumption, we can express

$$\bar{\beta}_a^{(j)} > \delta \geq \frac{1}{p} \left( \sum_{a=1}^p \bar{\beta}_a^{(j)} - 1 \right).$$

From this, we can infer  $\rho = \text{supp}(\bar{\beta}^{(j)})$  but also  $\bar{\beta}_a^{(j)} > \tau$ , hence  $\hat{\beta}_a^{(j)} > 0$ . We obtain the result of the initial stage stated in (D.4).

Let us remember that Lemma D.5.2 suggests  $\text{supp}(\bar{\beta}^{(j)}) \subseteq \text{supp}(\beta^{(j)})$ , and by the initial stage of the proof (see Equation (D.4)), we infer  $\text{supp}(\hat{\beta}^{(j)}) \subseteq \text{supp}(\beta^{(j)})$  for any  $j \in \hat{J}$ . Hence  $\text{supp}(\hat{A}_{\hat{J}}) \subseteq \text{supp}(A_{\hat{J}})$ . Furthermore Theorem 5.4.2 provides the result that  $\hat{I}_a = I_a$  for all  $a \in [\hat{K}]$ . From the way we construct  $\hat{A}_{\hat{J}}$ , we have  $\text{supp}(\hat{A}_{\hat{J}}) \subseteq \text{supp}(A_{\hat{J}})$ . Therefore, we have proved  $\text{supp}(\hat{A}) \subseteq \text{supp}(A)$ .

On the contrary, considering any  $j \in J_1$ , we have the knowledge that  $\beta_a^{(j)} > 2\delta$  for every  $a \in \text{supp}(\beta^{(j)})$ . Exploiting this insight and the additional observation that  $|\bar{\chi}_a^{(j)} - \beta_a^{(j)}| \leq \delta$ , we

deduce

$$|\bar{\chi}_a^{(j)}| \geq |\beta_a^{(j)}| - |\bar{\chi}_a^{(j)} - \beta_a^{(j)}| > \delta,$$

which implies  $\bar{\beta}_a^{(j)} > 0$ , hence  $\text{supp}(\beta^{(j)}) \subseteq \text{supp}(\bar{\beta}^{(j)})$  with  $j \in J_1$ . Using the initial stage of the proof, we have  $\text{supp}(A_{J_1}) \subseteq \text{supp}(\hat{A}_{J_1})$ .

**Proof of (c).** The proof is in the same line of (Bing et al., 2020, Theorem 7), we recall it for consistency. We first show that  $TFPP(\hat{\mathcal{G}}) = 0$ . From the result of part (b), we know that  $\text{supp}(\hat{A}) \subseteq \text{supp}(A)$ . Thus,

$$\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}=0, \hat{A}_{ja}>0\}},$$

which implies  $TFPP(\hat{\mathcal{G}}) = 0$ . In order to prove the result of  $TFNP(\hat{\mathcal{G}})$ , observe

$$\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0\}} = |I| + \sum_{j \in J} s(j),$$

with  $s(j) = \|A_j\|_0$  for each  $j \in J$ . For a given  $I$ , we partition  $[d] = I \cup J_1 \cup (J \setminus J_1)$ . Theorem 5.4.2 implies that  $\hat{I} = I$  and from the way we construct  $\hat{A}_{\hat{I}}$ , we have

$$\sum_{j \in I} \mathbb{1}_{\{\hat{A}_{ja}>0, A_{ja}=0\}} = 0.$$

Next, we consider the set  $J_1$ . Part (b) gives  $\text{supp}(A_{J_1}) = \text{supp}(\hat{A}_{J_1})$  which yields

$$\sum_{j \in J_1, a \in [K]} \mathbb{1}_{\{\hat{A}_{ja}>0, A_{ja}=0\}} = 0.$$

Finally, we consider the set  $J \setminus J_1$ . By examining the proof of part (b), we have necessarily  $\hat{A}_{ja} > 0$  if  $A_{ja} > 2\delta$  for any  $j \in J_1$ , and  $a \in [K]$ . Thus,

$$\sum_{j \in J \setminus J_1, a \in [K]} \mathbb{1}_{\{A_{ja}>0, \hat{A}_{ja}=0\}} \leq \sum_{j \in J \setminus J_1} t(j).$$

We hence obtain by combining the above inequalities

$$TFNP(\hat{\mathcal{G}}) = (\hat{\mathcal{G}}) = \frac{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0, \hat{A}_{ja}=0\}}}{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0\}}} \leq \frac{\sum_{j \in J \setminus J_1} t(j)}{|I| + \sum_{j \in J} s(j)} \leq \frac{\sum_{j \in J_1} t(j)}{|I| + \sum_{j \in J} s(j)}.$$

□

To conclude this section, we give below the proof of the intermediary result used in the proof.

**Proof of Lemma D.5.2** Suppose that  $\beta_a^{(j)} = 0$ , which implies, by definition,  $A_{ja} = 0$ . Now take  $i \in I_a$ , we thus have under  $\mathcal{E}$

$$\hat{\chi}_{n,m}(i, j) \leq \chi(i, j) + \delta = A_{ja} + \delta \leq 2\delta.$$

We thus obtain, under the event  $\mathcal{E}$  and  $\beta_a^{(j)} = 0$ , that

$$\bar{\chi}_a^{(j)} = \frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} \hat{\chi}_{n,m}(i, j) \leq 2\delta$$

and we obtain

$$\bar{\beta}_a^{(j)} = 0.$$

□

## D.6 Proofs of Section 5.5

Denote in this section by  $M_n^{(a)}$  the maximum of  $Z_i^{(a)}$ ,  $1 \leq i \leq n$  and  $1 \leq a \leq K$ . Also, for convenience, we call a function  $u$  on  $\mathbb{R}$  a normalizing function if  $u$  is non-decreasing, right continuous, and  $u(x) \rightarrow \pm\infty$ , as  $x \rightarrow \pm\infty$ .

**Proposition D.6.1.** *Suppose  $(\mathbf{Z}_t, t \in \mathbb{Z})$  is a moving maxima process of order  $p$  as described in (5.21) where margins of  $\epsilon_1$  are standard Pareto with an Archimedean copula function with upper tail equal to 1, see (5.22), then there exist sequences  $u_n^{(a)}$ ,  $1 \leq a \leq K$ , such that*

$$\mathbb{P} \left\{ M_n^{(a)} \leq u_n^{(a)}(x^{(a)}), 1 \leq a \leq K \right\} \rightarrow \prod_{a=1}^K e^{-\frac{1}{x^{(a)}}}, \quad \mathbf{x} \in (0, \infty)^K.$$

**Proof** This result is a direct application of (Hsing, 1989, Theorem 5.2) where most prominent arguments are taking from Examples in (Hsing, 1989, Section 6). For definitions of conditions  $D(u_n(x^{(1)}), \dots, u_n(x^{(K)}))$  and  $D''(u_n(x^{(1)}), \dots, u_n(x^{(K)}))$ , we also refer to Hsing (1989). For  $u_n^{(a)}(x) = \frac{nx(1-\rho^{p+1})}{1-\rho}$ ,  $n \geq 1$ ,  $x \in \mathbb{R}$ , we obtain

$$\mathbb{P} \left\{ M_n^{(a)} \leq u_n^{(a)}(x) \right\} = \left( \prod_{\ell=0}^p \mathbb{P} \left\{ \epsilon_1^{(a)} \leq \rho^{-\ell} u_n(x) \right\} \right)^n = \left( \prod_{\ell=0}^p \left( 1 - \frac{\rho^\ell(1-\rho)}{nx(1-\rho^{p+1})} \right) \right)^n.$$

Noticing that

$$\prod_{\ell=0}^p \left( 1 - \frac{\rho^\ell(1-\rho)}{nx(1-\rho^{p+1})} \right) = \exp \left\{ \sum_{\ell=0}^p \ln \left( 1 - \frac{\rho^\ell(1-\rho)}{nx(1-\rho^{p+1})} \right) \right\},$$

and using  $\exp\{x\} = 1 + x + O(x^2)$  and  $\ln(1-x) = -x + O(x^2)$  as  $x \rightarrow 0$ , we obtain that

$$\prod_{\ell=0}^p \left( 1 - \frac{\rho^\ell(1-\rho)}{nx(1-\rho^{p+1})} \right) = 1 - \frac{1}{nx} + O\left(\frac{1}{n^2}\right),$$

hence

$$\mathbb{P} \left\{ M_n^{(a)} \leq u_n^{(a)}(x) \right\} \xrightarrow{n \rightarrow \infty} e^{-\frac{1}{x}} \mathbf{1}_{\{x \geq 0\}}.$$

Furthermore, since  $\sigma(\mathbf{Z}_t, t \leq 0)$  and  $\sigma(\mathbf{Z}_t, t \geq p+1)$  are two independent  $\sigma$ -fields, the condition  $D(u_n(x^{(1)}), \dots, u_n(x^{(K)}))$  holds immediately for  $(\mathbf{Z}_t, t \in \mathbb{Z})$  for each  $\mathbf{x} \in \mathbb{R}^d$ . Thus, it suffices to show that the condition  $D''(u_n(x^{(1)}), \dots, u_n(x^{(K)}))$  holds for each  $\mathbf{x} \in (0, \infty)^K$ . For any fixed

$\mathbf{x} \in (0, \infty)^K$ , one obtains simply the estimate

$$\begin{aligned} 1 - \Pi_{\ell=0}^{i-1} \mathbb{P} \left\{ \epsilon_{1,1} \leq \rho^{-\ell} u_n(x^{(1)}) \right\} &= \frac{1 - \rho^i}{1 - \rho^{p+1} n x_1} + O\left(\frac{1}{n^2}\right), \\ 1 - \Pi_{\ell=p-i+1}^p \mathbb{P} \left\{ \epsilon_{1,2} \leq \rho^{-\ell} u_n(x^{(2)}) \right\} &= \frac{\rho^{p-i+1}(1 - \rho^i)}{1 - \rho^{p+1} n x_2} + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Now, by (5.22), one can obtain the following estimate:

$$1 - \Pi_{\ell=p-i+1}^p \mathbb{P} \left\{ \epsilon_{1,1} \leq \rho^{-\ell-i} u_n(x^{(1)}), \epsilon_{1,2} \leq \rho^{-\ell} u_n(x^{(2)}) \right\} = \frac{\rho^i(1 - \rho^{p-i+1})}{(1 - \rho^{p+1}) n x_1} + \frac{(1 - \rho^{p-i+1})}{(1 - \rho^{p+1}) n x_2} + O\left(\frac{1}{n^2}\right), \quad (\text{D.5})$$

for all  $n$ . Indeed

$$\begin{aligned} &1 - \mathbb{P} \left\{ \epsilon_{1,1} \leq \rho^{-\ell-i} u_n(x^{(1)}), \epsilon_{1,2} \leq \rho^{-\ell} u_n(x^{(2)}) \right\} = \\ &1 - \varphi^{\leftarrow} \left( \varphi \left( 1 - \frac{\rho^{\ell+i}(1 - \rho)}{(1 - \rho^{p+1}) n x_1} \right) + \varphi \left( 1 - \frac{\rho^{\ell}(1 - \rho)}{(1 - \rho^{p+1}) n x_2} \right) \right) = \\ &\frac{n}{n} \times \left( 1 - \varphi^{\leftarrow} \left( \varphi \left( 1 - \frac{1}{n} \right) \left\{ \frac{\varphi \left( 1 - \frac{\rho^{\ell+i}(1 - \rho)}{(1 - \rho^{p+1}) n x_1} \right)}{\varphi \left( 1 - \frac{1}{n} \right)} + \frac{\varphi \left( 1 - \frac{\rho^{\ell}(1 - \rho)}{(1 - \rho^{p+1}) n x_2} \right)}{\varphi \left( 1 - \frac{1}{n} \right)} \right\} \right) \right). \end{aligned}$$

The function  $x \mapsto 1/\varphi(1 - 1/x)$  is regularly varying at infinity with index 1. Therefore, its inverse function, the function  $t \mapsto 1/(1 - \varphi^{\leftarrow}(1/t))$  is regularly varying at infinity with index 1 ((Bingham et al., 1989, Theorem 1.5.12)), and thus the function  $1 - \varphi^{\leftarrow}$  is regularly varying at zero with index 1. We also have

$$\frac{\varphi \left( 1 - \frac{\rho^{\ell+i}(1 - \rho)}{(1 - \rho^{p+1}) n x_1} \right)}{\varphi \left( 1 - \frac{1}{n} \right)} \xrightarrow{n \rightarrow \infty} \frac{\rho^{\ell+i}(1 - \rho)}{x_1}, \quad \frac{\varphi \left( 1 - \frac{\rho^{\ell}(1 - \rho)}{(1 - \rho^{p+1}) n x_2} \right)}{\varphi \left( 1 - \frac{1}{n} \right)} \xrightarrow{n \rightarrow \infty} \frac{\rho^{\ell}(1 - \rho)}{x_2}.$$

By the uniform convergence theorem ((Bingham et al., 1989, Theorem 1.5.2)), the below term

$$n \times \left( 1 - \varphi^{\leftarrow} \left( \varphi \left( 1 - \frac{1}{n} \right) \left\{ \frac{\varphi \left( 1 - \frac{\rho^{\ell+i}(1 - \rho)}{(1 - \rho^{p+1}) n x_1} \right)}{\varphi \left( 1 - \frac{1}{n} \right)} + \frac{\varphi \left( 1 - \frac{\rho^{\ell}(1 - \rho)}{(1 - \rho^{p+1}) n x_2} \right)}{\varphi \left( 1 - \frac{1}{n} \right)} \right\} \right) \right)$$

converges to

$$\frac{\rho^{\ell+i}(1 - \rho)}{x_1} + \frac{\rho^{\ell}(1 - \rho)}{x_2}.$$

And then, after elementary estimation, we obtain (D.5). Hence for  $1 \leq i \leq p$ , we have

$$\begin{aligned} &\mathbb{P} \left\{ Z_1^{(1)} > u_n(x^{(1)}), Z_i^{(2)} > u_n(x^{(2)}) \right\} = \\ &1 - \mathbb{P} \left\{ Z_1^{(1)} \leq u_n(x^{(1)}) \right\} - \mathbb{P} \left\{ Z_1^{(2)} \leq u_n(x^{(2)}) \right\} + \mathbb{P} \left\{ Z_1^{(1)} \leq u_n(x^{(1)}), Z_i^{(2)} \leq u_n(x^{(2)}) \right\} = \\ &1 - \Pi_{\ell=0}^p \mathbb{P} \left\{ \epsilon_1^{(1)} \leq \rho^{-\ell} u_n(x^{(1)}) \right\} - \Pi_{\ell=0}^p \mathbb{P} \left\{ \epsilon_1^{(2)} \leq \rho^{-\ell} u_n(x^{(2)}) \right\} + \Pi_{\ell=0}^{i-1} \mathbb{P} \left\{ \epsilon_1^{(1)} \leq \rho^{-\ell} u_n(x^{(1)}) \right\} \times \\ &\Pi_{\ell=0}^{p-i} \mathbb{P} \left\{ \epsilon_1^{(1)} \leq \rho^{-\ell-i} u_n(x^{(1)}), \epsilon_1^{(2)} \leq \rho^{-\ell} u_n(x^{(2)}) \right\} \Pi_{\ell=p-i+1}^p \mathbb{P} \left\{ \epsilon_1^{(2)} \leq \rho^{-\ell} u_n(x^{(2)}) \right\} = O\left(\frac{1}{n^2}\right). \end{aligned}$$



Since for  $i > p$ , we trivially obtain that

$$\mathbb{P} \left\{ Z_1^{(1)} > u_n(x^{(1)}), Z_i^{(2)} > u_n(x^{(2)}) \right\} = O \left( \frac{1}{n^2} \right),$$

and noticing that  $\epsilon_i^{(1)}$  and  $\epsilon_i^{(2)}$  are playing symmetric role to  $\epsilon_i^{(j)}$  and  $\epsilon_i^{(k)}$  for  $1 \leq j < k \leq K$ , one concludes from this that the condition  $D''(u_n(x^{(1)}), \dots, u_n(x^{(K)}))$  holds for each  $\mathbf{x} \in (0, \infty)^K$ . Hence, applying (Hsing, 1989, Theorem 5.2), we obtain the result.  $\square$

We give below technical details on the heuristics  $d_m = O(1/m)$  made in Section 5.5. Let us consider the model in (5.1) without noise, i.e.,  $\mathbf{X} = \mathbf{AZ}$ . For  $\theta > 0$ , the Clayton copula is defined as

$$C_\theta(u, v) = \left[ 1 + \{(u^{-\theta} - 1) + (v^{-\theta} - 1)\} \right]^{-1/\theta}, \quad (u, v) \in [0, 1]^2.$$

The copula of the pair of componentwise maxima of an i.i.d. sample of size  $m$  from continuous distribution with copula  $C_\theta$  is equal to

$$\{C_\theta(u^{1/m}, v^{1/m})\}^m = C_{\theta/m}(u, v).$$

For establishing the heuristic, consider  $a, b \in [K]$  with  $a \neq b$  and  $i \in I_a, j \in I_b$ . Let  $C_m^{(i,j)}$  denote the copula between the pair  $M_m^{(i)}$  and  $M_m^{(j)}$ . Remember that these maxima are drawn from Pareto distributions, appropriately scaled in the independent setting by  $m$ . Denote  $C_\infty^{(i,j)}$  the extreme value copula between two independent standard Fréchet. The pre-asymptotic madogram is hence defined by

$$\begin{aligned} \nu_m(i, j) &= \frac{1}{2} - \int_0^1 C_m^{(i,j)}(u, u) du = \frac{1}{2} - \int_0^1 C_\theta^{(i,j)}(u^{1/m}, u^{1/m})^m du \\ &= \frac{1}{2} - \int_0^1 C_{\theta/m}^{(i,j)}(u, u) du. \end{aligned}$$

Using the same computations, we obtain for the madogram

$$\nu(i, j) = \frac{1}{2} - \int_0^1 C_\infty^{(i,j)}(v, v) dv.$$

Since  $C_\theta$  is positive lower orthant dependent for any  $\theta > 0$ , it follows that  $C_{\theta/m}$  is positive lower orthant dependent for any  $m \geq 1$  and  $\theta > 0$ . We hence obtain that  $\nu_m(i, j) \in [0, 1/6]$  and using that the function  $f(x) = (0.5 + x)/(0.5 - x)$  is Lipschitz for  $x \in [0, 1/6]$ , we obtain

$$|\chi_m(i, j) - \chi(i, j)| = \left| \frac{0.5 + \nu_m(i, j)}{0.5 - \nu_m(i, j)} - \frac{0.5 + \nu(i, j)}{0.5 - \nu(i, j)} \right| \leq 9|\nu_m(i, j) - \nu(i, j)|.$$

Now

$$|\nu_m(i, j) - \nu(i, j)| = \left| \int_0^1 C_{\theta/m}^{(i,j)}(u, u) du - \int_0^1 C_\infty^{(i,j)}(u, u) du \right|.$$

Now, applying (Bücher and Segers, 2014, Proposition 4.3), we obtain

$$\int_0^1 \left| C_{\theta/m}^{(i,j)}(u, u) - C_\infty^{(i,j)}(u, u) \right| dv \leq \sup_{u,v \in [0,1]^2} \left| C_{\theta/m}^{(i,j)}(u, v) - C_\infty^{(i,j)}(u, v) \right| = O\left(\frac{1}{m}\right).$$

## D.7 Supplementary Lemmata

**Lemma D.7.1.** *Let  $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$  where both  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are regularly varying with respective exponent measures  $\Lambda_{\mathbf{X}_1}$ ,  $\Lambda_{\mathbf{X}_2}$  and the following tail balance condition holds:*

$$\lim_{x \rightarrow \infty} \frac{\mathbb{P}\{\|\mathbf{X}_2\| > x\}}{\mathbb{P}\{\|\mathbf{X}_1\| > x\}} = 0. \quad (\text{D.6})$$

Then  $\mathbf{X}$  is regularly varying with exponent measure  $\Lambda_{\mathbf{X}}$  given by

$$\Lambda_{\mathbf{X}} = \Lambda_{\mathbf{X}_1}.$$

Under Condition (D.6) one may expect that the tail behavior of  $\mathbf{X}$  is mainly influenced by that of  $\mathbf{X}_1$ .

**Proof** Without loss of generality, we suppose that  $\mathbf{X}_1$  is regularly varying with tail index  $\alpha$  equal to unity. We must prove that the sequence of measure  $\{\Lambda_x\}$  defined by

$$\Lambda_x(\cdot) = \mathbb{P}\left\{x^{-1}(\mathbf{X}_1, \mathbf{X}_2) \in \cdot\right\} / \mathbb{P}\{\|\mathbf{X}_1\| > x\},$$

is the only possible limit along a subsequence by applying (Kulik and Soulier, 2020, Lemma B.1.29) and that the measure  $\Lambda_{(\mathbf{X}_1, \mathbf{X}_2)}$  defined by  $\Lambda_{(\mathbf{X}_1, \mathbf{X}_2)} = \Lambda_{\mathbf{X}_1} \otimes \delta_0$  is the only possible limit along a subsequence by applying (Kulik and Soulier, 2020, Lemma B.1.31). Let  $f$  be a bounded uniformly continuous function with support in  $A_1 \times \mathbb{R}^d$  with  $A_1$  separated from zero. Fix  $\epsilon > 0$ . Then there exists  $\|\mathbf{x}_2\| \leq \eta$  which implies  $\|f(\mathbf{x}_1, \mathbf{x}_2) - f(\mathbf{x}_1, 0)\| \leq \epsilon$ . By Condition (D.6) and since  $A_1$  is separated from zero, we have

$$\begin{aligned} \lim_{x \rightarrow \infty} \Lambda_x(f) &= \lim_{x \rightarrow \infty} \mathbb{E}\left[f(x^{-1}(\mathbf{X}_1, \mathbf{X}_2)) \mathbf{1}_{\{\|\mathbf{X}_2\| > \eta x\}}\right] / \mathbb{P}\{\|\mathbf{X}_1\| > x\} \\ &= \lim_{x \rightarrow \infty} \frac{\mathbb{E}\left[f(x^{-1}(\mathbf{X}_1, \mathbf{X}_2)) \mathbf{1}_{\{\|\mathbf{X}_2\| > \eta x\}}\right]}{\mathbb{P}\{\|\mathbf{X}_2\| > x\}} \lim_{x \rightarrow \infty} \frac{\mathbb{P}\{\|\mathbf{X}_2\| > x\}}{\mathbb{P}\{\|\mathbf{X}_1\| > x\}} = 0. \end{aligned}$$

For every  $\eta > 0$ , since  $\|\mathbf{X}_1\|$  and  $\|\mathbf{X}_2\|$  are both regularly varying conserving the same tail index, the assumption implies

$$\lim_{x \rightarrow \infty} \frac{\mathbb{P}\{\|\mathbf{X}_2\| > \eta x\}}{\mathbb{P}\{\|\mathbf{X}_1\| > x\}} = \lim_{x \rightarrow \infty} \frac{\mathbb{P}\{\|\mathbf{X}_2\| > \eta x\}}{\mathbb{P}\{\|\mathbf{X}_1\| > \eta x\}} \lim_{x \rightarrow \infty} \frac{\mathbb{P}\{\|\mathbf{X}_1\| > \eta x\}}{\mathbb{P}\{\|\mathbf{X}_1\| > x\}} = 0 \times \eta^{-1} = 0.$$

Then,

$$\begin{aligned}
 & \limsup_{x \rightarrow \infty} \mathbb{E} \left[ \left| f(x^{-1}(\mathbf{X}_1, \mathbf{X}_2)) - f(x^{-1}(\mathbf{X}_1, 0)) \right| \right] / \mathbb{P} \{ \|\mathbf{X}_1\| > x \} \\
 & \leq \limsup_{x \rightarrow \infty} \mathbb{E} \left[ \left| f(x^{-1}(\mathbf{X}_1, \mathbf{X}_2)) - f(x^{-1}(\mathbf{X}_1, 0)) \right| \mathbb{1}_{\{\|\mathbf{X}_2\| \leq \eta x\}} \right] / \mathbb{P} \{ \|\mathbf{X}_1\| > x \} \\
 & + \limsup_{x \rightarrow \infty} \mathbb{E} \left[ \left| f(x^{-1}(\mathbf{X}_1, \mathbf{X}_2)) - f(x^{-1}(\mathbf{X}_1, 0)) \right| \mathbb{1}_{\{\|\mathbf{X}_2\| > \eta x\}} \right] / \mathbb{P} \{ \|\mathbf{X}_1\| > x \} \leq \epsilon.
 \end{aligned}$$

Since  $\epsilon$  is arbitrary and  $\delta_0 \otimes \Lambda_{\mathbf{X}_2}(f) = 0$ , this proves that

$$\begin{aligned}
 \lim_{x \rightarrow \infty} \Lambda_{(\mathbf{X}_1, \mathbf{X}_2)}(f) &= \lim_{x \rightarrow \infty} \mathbb{E} \left[ f(x^{-1}(\mathbf{X}_1, 0)) \right] / \mathbb{P} \{ \|\mathbf{X}_1\| > x \} \\
 &= \Lambda_{\mathbf{X}_1} \otimes \delta_0(f) = \Lambda_{(\mathbf{X}_1, \mathbf{X}_2)}(f).
 \end{aligned}$$

If now  $f(\mathbf{x}_2, \mathbf{x}_2) = g(\mathbf{x}_2)$  with  $g$  a continuous function with support separated from zero, then  $\Lambda_{\mathbf{X}_1} \otimes \delta_0 = 0$  and

$$\begin{aligned}
 \lim_{x \rightarrow \infty} \Lambda_x(f) &= \lim_{x \rightarrow \infty} \mathbb{E} \left[ f(x^{-1}(\mathbf{X}_2, \mathbf{X}_2)) \right] / \mathbb{P} \{ \|\mathbf{X}_1\| > x \} = \lim_{x \rightarrow \infty} \mathbb{E} \left[ g(x^{-1} \mathbf{X}_2) \right] / \mathbb{P} \{ \|\mathbf{X}_1\| > x \} \\
 &= \lim_{x \rightarrow \infty} \frac{\mathbb{P} \{ \|\mathbf{X}_2\| > x \}}{\mathbb{P} \{ \|\mathbf{X}_1\| > x \}} \lim_{x \rightarrow \infty} \frac{\mathbb{E} [g(x^{-1} \mathbf{X}_2)]}{\mathbb{P} \{ \|\mathbf{X}_2\| > x \}} \\
 &= 0 \times \Lambda_{\mathbf{X}_2}(f) = \Lambda_{(\mathbf{X}_1, \mathbf{X}_2)}(f).
 \end{aligned}$$

This proves that  $\Lambda_{\mathbf{X}} = \Lambda_{\mathbf{X}_1} \otimes \delta_0$  is the only possible limit for the sequence  $\{\Lambda_x\}$  along any subsequence. We must now prove that  $\{\Lambda_x\}$  is relatively compact. Define  $U_n = \{(\mathbf{x}_1, \mathbf{x}_2) : \|\mathbf{x}_1\| + \|\mathbf{x}_2\| > e^n\}$ . The sets  $U_n$ ,  $n \in \mathbb{Z}$  satisfy the assumptions of (Kulik and Soulier, 2020, Lemma B.1.29). This proves that the sequence  $\{\Lambda_x\}$  is relatively compact and we conclude that  $\Lambda_x \xrightarrow{v\#} \Lambda_{(\mathbf{X}_1, \mathbf{X}_2)}$ . We directly obtain that for any  $A \in \mathcal{B}(\mathbb{R}^d)$  separated from 0

$$\Lambda_{\mathbf{X}_1 + \mathbf{X}_2}(A) = \Lambda_{\mathbf{X}_1}(A),$$

hence the result.  $\square$

We recall the following Bernstein inequality from Merlevède et al. (2009) and Lemma S.2 in Cordoni and Sancetta (2023) where we recall the proof of the second for consistency purposes.

**Lemma D.7.2.** *Let  $(Y_t)_{t \geq 1}$ , be a sequence of mean zero, stationary random variables whose absolute values is uniformly bounded by  $\bar{y} < \infty$ , and with exponential decaying strong mixing coefficients. Then for  $n \geq 4$  and  $z \geq 0$ , there is a constant  $c_1 > 0$ , depending only on the mixing coefficient and such that*

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n Y_i \right| \geq z \right\} \leq \exp \left\{ - \frac{c_1 n z^2}{\bar{y}^2 + z \bar{y} \ln n (\ln \ln n)} \right\}$$

**Lemma D.7.3.** *Under the assumptions of Lemma D.7.2, choose a  $c_2 \in (0, \infty)$ , and let  $z := y - n^{-c_2}$  for any  $y \geq n^{-c_2}$ . Then there is a constant  $c_1 \geq 0$  such that*

$$\mathbb{P} \left\{ \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \geq y \right\} \leq 2 \exp \left\{ - \frac{c_1 n z^2}{1 + z \ln n (\ln \ln n)} + c_2 \ln(n) \right\}$$

**Proof** Using standard techniques, we replace the supremum by the maximum over a finite number of elements. We then apply Lemma D.7.2.

To do so, for fixed but arbitrary  $\epsilon > 0$ , we construct intervals  $[x_\ell^L, x_\ell^U]$  with  $\ell = 1, 2, \dots, N(\epsilon)$ , such that  $|F(x) - F(z)| \leq \epsilon$  for  $x, z \in [x_\ell^L, x_\ell^U]$ . Fix an arbitrary  $\epsilon > 0$  and divide interval  $[0, 1]$  into  $N(\epsilon) \geq \epsilon^{-1}$ , some positive integer to be chosen later, intervals  $[t_{\ell-1}, t_\ell[$  where  $0 = t_0 < t_1 < \dots < t_{N(\epsilon)} = 1$  such that  $t_\ell - t_{\ell-1} \leq \epsilon$ . Now, take  $N(\epsilon)$  the smallest integer greater than or equal to  $\epsilon^{-1}$ . Define variables  $-\infty \leq x_1^L \leq x_2^L \leq \dots \leq x_{N(\epsilon)}^L = \infty$  as  $x_\ell^L := \inf \{x \in \mathbb{R}, F(x) \geq t_{\ell-1}\}$ . Similarly define variables  $-\infty \leq x_1^U \leq x_2^U \leq \dots \leq x_{N(\epsilon)}^U = \infty$  as  $x_\ell^U := \sup \{x \in \mathbb{R} : F(x) \leq t_\ell\}$ . This construction has the aforementioned properties. Note that we can have  $[x_\ell^L, x_\ell^U]$  equal to a singleton, i.e.,  $x_\ell^L = x_\ell^U$  if there are discontinuities in  $F$  and such that discontinuities are larger than  $\epsilon$ .

From the fact that  $F(x)$  and  $\hat{F}_n(x)$  are monotonically increasing, we have that  $F(x_\ell^L) \leq F(x) \leq F(x_\ell^U)$  and  $\hat{F}_n(x_\ell^L) \leq \hat{F}_n(x) \leq \hat{F}_n(x_\ell^U)$  for  $x \in [x_\ell^L, x_\ell^U]$ . Also recall that  $\mathbb{E}\hat{F}_n(x) = F(x)$ . In consequence

$$\begin{aligned} \hat{F}_n(x) - F(x) &\leq \hat{F}_n(x_\ell^U) - F(x_\ell^L) \leq \hat{F}_n(x_\ell^U) - F(x_\ell^U) + \epsilon, \\ \hat{F}_n(x) - F(x) &\geq \hat{F}_n(x_\ell^L) - F(x_\ell^U) \geq \hat{F}_n(x_\ell^L) - F(x_\ell^L) - \epsilon. \end{aligned}$$

using monotonicity and the fact  $|F(x_\ell^U) - F(x_\ell^L)| \leq \epsilon$  by construction. And thus for any  $x$

$$\hat{F}_n(x_\ell^L) - F(x_\ell^L) - \epsilon \leq \hat{F}_n(x) - F(x) \leq \hat{F}_n(x_\ell^U) - F(x_\ell^U) + \epsilon.$$

Hence, for any  $x$

$$|\hat{F}_n(x) - F(x)| \leq \max_{\ell \in \{1, \dots, N(\epsilon)-1\}} \max \left\{ |\hat{F}_n(x_\ell^U) - F(x_\ell^U)|, |\hat{F}_n(x_\ell^L) - F(x_\ell^L)| \right\} + \epsilon$$

Since it holds for any  $x$ , we obtain

$$\sup_{x \in [0, 1]} |\hat{F}_n(x) - F(x)| \leq \max_{\ell \in \{1, \dots, N(\epsilon)-1\}} \max \left\{ |\hat{F}_n(x_\ell^U) - F(x_\ell^U)|, |\hat{F}_n(x_\ell^L) - F(x_\ell^L)| \right\} + \epsilon.$$

Set  $\epsilon = n^{-c_2}$ . Using union bound and apply Lemma D.7.2 twice with  $Y_i = (1 - \mathbb{E})\mathbb{1}_{\{X_i \leq x\}}$  for arbitrary, but fixed  $x$ , and  $z = y - \epsilon = y - n^{-c_2}$ .  $\square$



## CHAPTER 6

# CONCLUSION AND PERSPECTIVES

Dependent measurements exhibiting extremes in a high-dimensional setting are prevalent in various statistical and machine learning applications. Despite their common occurrence, many extremal multivariate analysis methods overlook this type of data and adhere to the i.i.d. and “ $n$  goes to infinity,  $d$  fixed” framework. This thesis has focused on developing, studying, and practically applied methods for variable clustering of a strictly stationary multivariate mixing process in high dimensions.

We have primarily introduced two novel methods for variable clustering. The first method is a hard clustering algorithm (referred to as Algorithm (ECO)) for a specific model-based clustering where a partition of the set  $\{1, \dots, d\}$  is desired, and clusters are defined as groups of variables that are mutually independent with respect to their extremes. The second method is a soft clustering approach (referred to as Algorithm (SCRAM)) designed to estimate the entries of linear factor models, a classical model used to study dependence in large dimensions. An added specificity here is that latent factors could exhibit extreme behavior and clusters, i.e., groups that are attached to a same latent factor are permitted to overlap.

Theoretical guarantees for these two algorithms were provided, assuming fixed values of  $d$  and  $n$ , and considering various mixing conditions. Additionally, all proposed methods rely solely on bivariate measures, specifically the extremal correlation (see Definition 1.1.5 in Chapter 1). This reliance on bivariate measures is the primary reason why our procedure has a finite sample bound dependent on the logarithm of the dimension  $d$ .

All proposed methods have been implemented and demonstrated in practice, showing promising results in climate sciences. Furthermore, I believe that these methods could be valuable in other fields as well. I encourage readers to explore potential applications beyond climate sciences and finance.

To conclude, I would like to share a non-mathematical perspective on the challenges faced by statistical sciences. In essence, there seems to be a trilemma, where it is difficult for a statistical method to simultaneously possess all three of the following qualities:

- Flexibility;
- Statistical efficiency in high dimensions;
- Interpretability.

Flexibility refers to a method’s capacity to adapt and accurately represent our observations. Methods that can precisely fit our data are considered flexible, while those that struggle to do so are seen as not flexible. A method is deemed statistically efficient in high dimensions if it can maintain meaningful outcomes even when the number of features (dimensions) exceeds the number of observations. Interpretability, on the other hand, relates to the extent to which

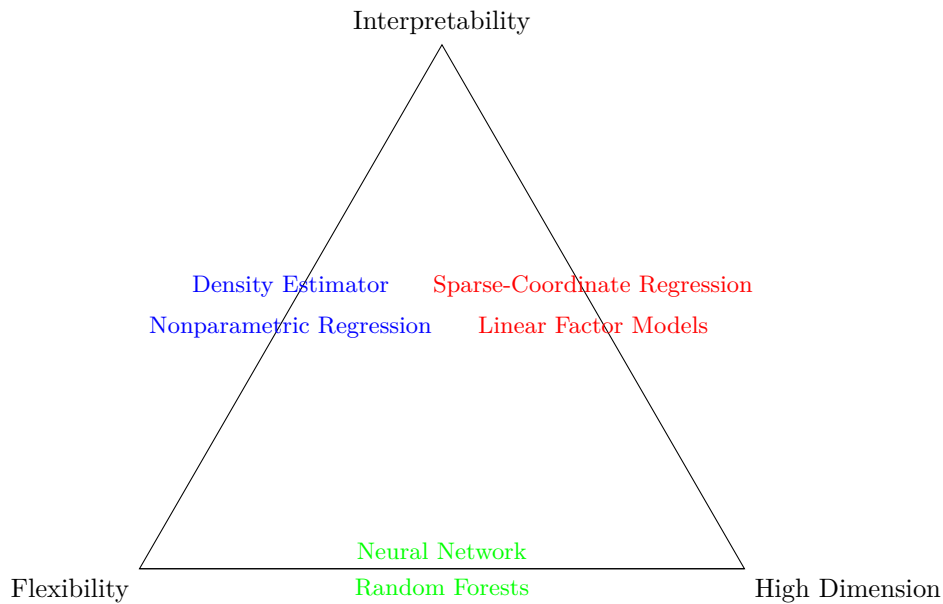


Fig. 6.1 The concept of the impossible trilemma in statistical sciences can be illustrated using a diagram with labeled edges representing different statistical methods. We have methods that prioritise flexibility and interpretability (shown in blue). Methods that excel in high-dimensional settings and are flexible (depicted in green) sacrifice interpretability. Methods that prioritise interpretability and efficiency in high dimensions (shown in green) lack the flexibility to adapt to complex data patterns.

a human can understand and consistently predict the model’s outputs. Figure 6.1 depicts this trade-off. On one side, we have some examples of methods that prioritise flexibility and interpretability (shown in blue). However, these methods are unable to achieve efficiency in high dimensions. Conversely, methods that excel in high-dimensional settings and are flexible (depicted in green) sacrifice interpretability. Finally, methods that prioritise interpretability and efficiency in high dimensions (shown in red) lack the flexibility to adapt to complex data patterns. Regarding the two methods proposed in this PhD thesis, they belong to the red side of the trilemma diagram. Both AI-block in Chapter 3 models and linear factor models in Chapter 5 are interpretable and scalable in high dimensions. However, they lack flexibility in adapting to very complex data patterns.

Now, we present some perspectives currently under study. In Section 6.1, we derive a result similar to Theorem 1.1.13 in Chapter 1 for the strong consistency of the  $K$ -means using madogram. This result has been stated at the early stages of my PhD, prompted by the recognition of this method’s prominence in applied sciences (see the references thereafter). However, when it comes to variable clustering, I find that this method will always have limitations. Theoretically speaking, as mentioned in Section 1.1.5 in Chapter 1, the statistical study of  $K$ -means is somehow linked to the problem of clustering a mixture of isotropic Gaussians, which has led to extensive literature in both statistics and machine learning. While this framework is beyond the scope of variable clustering (as it mainly concerns data clustering) and extreme value theory (since it involves Gaussians), several authors have considered  $K$ -means using the angular measure  $\Phi$  and demonstrated its connection with linear factor models. However, I

strongly doubt its efficiency in a high-dimensional setting based on our knowledge about its concentration (see Cléménçon et al. (2023)). Given the significance of this result, a brief note will be written to share this finding without extensive elaboration. The spirit of Section 6.2 and Section 6.3 is more connected to the manuscript. The Section 6.2 is mainly driven by the thirst to obtain a minimax risk for an estimator inside a class of multivariate models suited for extremes. An attempt was made to tackle the class of Linear Factor models in Chapter 5, yet certain obstacles persist, rendering it an ongoing endeavor. Nonetheless, a successful endeavor has been achieved for a different yet somewhat related area: the linear regression model with regularly varying design. A surprising result emerges as we obtain that the optimal rate is  $n^{-1}$ , mirroring that attained for classical linear regression with uniform noise, as seen in, for instance, Yi and Neykov (2024). This stands in stark contrast to the considerably slower  $n^{-1/2}$  rate for Ordinary Least Squares (OLS), which is optimal for Gaussian noise. The idea driven by Section 6.3 marks a paradigm shift within this manuscript. Throughout this manuscript, we have always assumed a strictly stationary process, which is, in practice, inaccurate. The idea presented in Section 6.3 is to develop a test to detect changepoints in the extremal dependence structure of a random vector when its angular measure is discrete, in the spirit of Chapter 5.

## 6.1 Strong Consistency of madogram-based K-means under mixing conditions

The  $K$ -means procedure is a way to identify distinct groups within a population. This procedure involves partitioning a set of data into  $G$  groups (to be consistent with our notation). To do this, we first choose cluster centers  $\psi_1, \dots, \psi_G$  for the points  $\mathbf{Z}_1, \dots, \mathbf{Z}_n \in \mathbb{R}^d$  in order to minimise

$$W_n := \frac{1}{n} \sum_{i=1}^n \min_{g \in \{1, \dots, G\}} D(\mathbf{Z}_i, \psi_g),$$

where  $D : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$  is a distance function or, more generally, a dissimilarity function in  $\mathbb{R}^d$ . The motivation is to identify cluster centers such that distances of the observations to their nearest cluster center are minimised. Accordingly, all observations which are closest to the same cluster center are viewed as belonging to the same group.

While the original version of  $K$ -means uses the Euclidean distance, several alternatives choices of distances have been suggested. As the extremal dependence structure can be described with the angular measure  $\Phi$  (see Resnick (2008), section 5 for details), a natural way to measure the distance between two points is by their angle. This corresponds to the spherical  $K$ -means clustering which is described as follow: for a given integer  $G$ , solve the following optimization problem

$$\frac{1}{n} \sum_{i=1}^n \min_{g \in \{1, \dots, G\}} D(\mathbf{Y}_i, \psi_g),$$

with  $\mathbf{Y}_i$ , i.i.d. observations from  $\mathbf{Y}$ , a random variable living on the unit sphere with law  $\Phi$ . Consistency results with i.i.d. observations and for sufficiently many large observations had been proved for this algorithm in Janßen and Wan (2020). The consistency result gives that the centroids obtained by minimising the program above are close to the true centroids of the angular distribution.



## Conclusion and perspectives

---

In the framework of [Bador et al. \(2015\)](#); [Bernard et al. \(2013\)](#); [Saunders et al. \(2021\)](#), the madogram is considered as a dissimilarity measure. This criterion can be read in the present context of block maxima method as

$$W_n = \frac{1}{k} \sum_{i=1}^k \min_{g \in \{1, \dots, G\}} \frac{1}{2} |\hat{U}_{n,m,i} - \psi_g| = \int_{[0,1]^d} \min_{g \in \{1, \dots, G\}} \frac{1}{2} |\mathbf{u} - \psi_g| d\hat{C}_{n,m}(\mathbf{u}),$$

where  $\hat{C}_{n,m}$  is the empirical copula of the pseudo-observations  $(\hat{U}_{n,m,i}^{(1)}, \dots, \hat{U}_{n,m,i}^{(d)})$  of the uniform margins of the componentwise maxima  $(M_{m,1}^{(1)}, \dots, M_{m,1}^{(d)})$  defined as

$$\hat{C}_{n,m}(\mathbf{u}) = \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{\{\hat{U}_{n,m,i} \leq \mathbf{u}\}}, \quad \mathbf{u} \in [0, 1]^d. \quad (6.1)$$

For a copula  $C_m$  of  $(M_{m,1}^{(1)}, \dots, M_{m,1}^{(d)})$  in the max-domain of attraction of an extreme value copula  $C_\infty$ , let  $\Psi = \{\psi_1, \dots, \psi_G\}$ , be a set of cluster centers with  $\psi_g \in \mathbb{R}^d$ ,  $g \in \{1, \dots, G\}$  and consider the averaged distance from any observation to the closest element of  $\Psi$  as

$$W(\Psi, C) = \int_{[0,1]^d} \min_{\psi \in \Psi} \frac{1}{2} |\mathbf{u} - \psi| dC_\infty(\mathbf{u}).$$

To the best of our knowledge, consistency results for  $K$ -means procedure using the madogram have not yet been established. The following proposition tries to bridge this gap.

**Proposition 6.1.1.** *Let  $(\mathbf{Z}_t, t \in \mathbb{Z})$  be a stationary multivariate random process with continuous univariate margins such that  $C_m(\mathbf{u}) \rightarrow C_\infty(\mathbf{u})$  for  $\mathbf{u} \in [0, 1]^d$  and Condition  $\mathcal{C}$  in Chapter 3 hold. For each  $\hat{C}_{n,m}$  in (6.1) and a given value  $G \in \mathbb{N}$ , denote by  $\Psi_G^n$  a random set which minimises*

$$W(\Psi, \hat{C}_{n,m}) = \int_{[0,1]^d} \min_{\psi \in \Psi} \frac{1}{2} |\mathbf{u} - \psi| d\hat{C}_{n,m}(\mathbf{u}),$$

*among all sets  $\Psi \subset [0, 1]^d$  with at most  $G$  elements. Accordingly, let us define  $\Psi_G$  the optimal set when we replace  $\hat{C}_{n,m}$  by  $C_\infty$  and assume that for a given value of  $G$ , the set  $\Psi_G$  is uniquely determined. Thus  $\Psi_G^n$  converges almost surely to  $\Psi_G$  as  $n \rightarrow \infty$ .*

From Proposition 6.1.1, the madogram seems to be a relevant dissimilarity to estimate the set of theoretical cluster centers with respect to the extreme value copula of  $\mathbf{X}$ .

## 6.2 Estimating Sparse Linear Regression with Randomly Varying Design

Consider a vector of positive coefficients denoted as  $A$ , represented as  $A = (A^{(1)}, \dots, A^{(d)})$ . The relationship between the observed random variables  $Y$  and  $Z^{(j)}$  is implicitly established through the linear representation

$$Y = A^\top \mathbf{Z} + \xi, \quad A \in \mathbb{R}_+^d, \quad (6.2)$$

## 6.2 Estimating Sparse Linear Regression with Randomly Varying Design

---

where  $\mathbf{Z}$  is a regularly varying random vector with tail index  $\alpha$  and  $\xi \in \mathbb{R}^n$  is a random vector with lighter tail index, independent of  $\mathbf{Z}$ .

Extreme-value theory has gained considerable traction in the field of research. However, there remains a scarcity of statistical methodologies capable of deciphering the intricate structure of complex extreme events. While low-dimensional models, along with their associated methods and theories, abound in the fields, the exploration of high-dimensional models has been notably limited. In recent years, a significant surge of interest has emerged in the theory of estimation within high-dimensional statistical models, particularly in various sparsity scenarios. The primary impetus behind sparse estimation lies in the observation that, in numerous practical applications, the number of variables far exceeds the number of available observations. A classical illustration of sparse estimation in traditional statistics is the challenge of estimating a sparse regression vector from a collection of linear measurements (see, for example, [Bickel et al. \(2009\)](#); [Bunea et al. \(2007\)](#); [Lounici \(2008\)](#)).

These notions of sparsity can be defined in terms of  $\ell_0$ -balls, defined as:

$$\mathbb{B}_0(s) := \left\{ v = (v^{(j)})_{j=1,\dots,d} \in \mathbb{R}^d : \sum_{j=1}^d \mathbb{1}_{\{v^{(j)} \neq 0\}} \leq s \right\},$$

where  $s < \infty$  is a given constant. This ball corresponds to the set of vectors  $v$  with at most  $s$  non-zero elements.

In this current discussion, we delve into the max-linear regression model with a focus on sparsity, where only a few elements, of size  $s$ , of  $A$  in (6.2) hold non-zero values. The rationale behind this model is twofold: The linear relationship enables factors to engage in competition with each other. In the domain of risk analysis, the risk of loss manifests as the sum of numerous individual risks, representing whichever risk is the most significant. Consequently, the sparsity assumption within this context implies that only a limited number of factors can contribute to the overall risk. This, coupled with the linear structure, offers a level of interpretability that surpasses many other model structures in the field.

The minimax rate of convergence characterises the fundamental limitation of the estimation accuracy. It also captures the interdependence between the different parameters in the model. There is a rich line of work of such fundamental limits (see, for example, [Tsybakov \(2008\)](#)). However, the concept of minimax rates in extreme-value theory is inherently tied to parsimony. A major focus in the present section is on derivation of lower bounds in linear regression model where the design is regularly varying, which is a key step in establishing minimax optimal rates of convergence. We quantify the estimation error by the norm  $\|\cdot\|_2$ . We establish minimax lower bounds to quantify the statistical optimality of certain estimation procedures. In this view, we are interested in the worst-case performance of estimation procedures for the model (6.2) over a family of vectors  $A$ . In particular, we define the class of models as follows

**Definition 6.2.1.** We define a class of models

$$\mathbf{Y} = A^\top \mathbf{Z} + \xi,$$

where  $A \in \mathbb{B}_0(s)$ ,  $\mathbf{Z}$  admits a density with respect to the Lebesgue measure given by  $p_{\mathbf{Z}}(\mathbf{z}) = \alpha^d \prod_{j=1}^d (1 + z^{(j)})^{-(\alpha+1)}$ ,  $\alpha > 1$ , which corresponds to a product of Pareto with the same tail index. Also, the Lebesgue-density of  $\xi$ , denoted by  $p_{\xi}$ , verifies:

**Condition (A).**

$$\int_{\mathbb{R}} p_{\xi}(u-t) \ln \left( \frac{p_{\xi}(u-t)}{p_{\xi}(u)} \right) du \leq p_* t, \forall t \in \mathbb{R}_+$$

where  $0 < p_* < \infty$ .

We start by establishing the minimax lower bounds for estimation of vector over  $\mathbb{B}_0(s)$  (Theorem 6.2.1). We denote by  $\inf_{\hat{A}}$  the infimum over all estimators  $\hat{A}$  with values in  $\mathbb{R}_+^d$ ,  $\mathbb{P}_{\mathbf{X}}^n$  the  $n$ -fold product measure of probability measure  $\mathbb{P}_{\mathbf{X}}$ , and  $\mathbb{E}_{\mathbf{X}}^n$  the expectation with respect to  $\mathbb{P}_{\mathbf{X}}^n$ .

**Theorem 6.2.1.** *Let  $d \geq 2$  and  $s \geq 1$  with  $s \leq 4d/5$ . Suppose that we observe  $\mathbf{X}_i = (Y_i, \mathbf{Z}_i)$   $n$  i.i.d. pairs in the class of models given in Definition (6.2.1). Then under Condition (A),*

(i)

$$\inf_{\hat{A}} \sup_{A \in \mathbb{B}_0(s)} \mathbb{P}_{\mathbf{X}}^n \left\{ \|\hat{A} - A\|_2 \geq C \sqrt{\frac{s}{n^2}} \ln(d/s) \right\} \geq \beta,$$

(ii)

$$\inf_{\hat{A}} \sup_{A \in \mathbb{B}_0(s)} \mathbb{E}_{\mathbf{X}}^n \left[ \|\hat{A} - A\|_2 \right] \geq \tilde{C} \sqrt{\frac{s}{n^2}} \ln(d/s),$$

where  $0 < \beta < 1$ ,  $C > 0$ , and  $\tilde{C} > 0$  are absolute constants.

### 6.3 Changepoint Detection for High-Dimensional Extremal Dependence

In contexts where data are collected over time, one of the simplest generalisations of an independent and identically distributed data stream is provided by changepoint models. In this framework, we make the hypothesis that our data may be segmented into two shorter, homogeneous series, taking the structure of a linear factor models studied in Chapter 5,

$$\mathbf{X}_t = A_t \mathbf{Z}_t + \mathbf{E}_t, \quad A_t \in \mathbb{R}^{d \times K}, \quad t = 1, \dots, n \quad (6.3)$$

Of course, the structural break, or changepoint between these series is often of interest in applications. This can be formulated as the following hypothesis testing problem

$$H_0 : A_1 = \dots = A_n \text{ v.s. } H_1 : A^{(1)} := A_1 = \dots = A_{n_1} \neq A_{n_1+1} = \dots = A_n =: A^{(2)},$$

where  $n_1$  is the possible but unknown changepoint location. In the case of a single changepoint, (6.3) reduces to:

$$\mathbf{X}_t = A^{(1)} \mathbf{Z}_t \mathbf{1}_{\{t \leq n_1\}} + A^{(2)} \mathbf{Z}_t \mathbf{1}_{\{t > n_1\}} + \mathbf{E}_t, \quad t = 1, \dots, n.$$

Denote by  $\|A\|_{(k)} = \sum_{\ell=1}^k \sigma_{\ell}(A)$  the Ky-Fan( $k$ ) norm of a matrix  $A \in \mathbb{R}^{d \times K}$  where  $k \leq K$  and  $\sigma_{\ell}(A)$  denotes the  $\ell$ -th largest singular value of  $A$ . For observations  $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$ , we

### 6.3 Changepoint Detection for High-Dimensional Extremal Dependence

---

denote the corresponding sample matrix using just the observations  $1, \dots, n_1$  and  $n_1 + 1, \dots, n$  as  $\hat{A}_{n_1}^{(1)}$  and  $\hat{A}_{n_1}^{(2)}$  respectively. A test statistic to test  $H_0$  against  $H_1$  could be of the following form:

$$T = \max_{n_1=2, \dots, n-2} \max_{s=1, \dots, K} \sqrt{n} \|\hat{A}_{n_1}^{(1)} - \hat{A}_{n_1}^{(2)}\|_{(k)}. \quad (6.4)$$

The concept involves conducting a mathematical analysis to establish asymptotic power, which would offer insights into the performance of the statistic  $T$  in (6.4) based on Ky-Fan( $k$ ) norms. Two scenarios could be considered: (1) when  $d$  is fixed as  $n$  tends to infinity, and (2) when  $d$  tends to infinity as  $n$  tends to infinity.

Designing methods to detect changepoints in the extremal dependence structure is of great interest in the context of climate change. Such methods are needed to address the challenges posed by processes that go beyond the stationarity of observed data. This is highlighted in studies such as [Gonzalez et al. \(2023\)](#); [Naveau et al. \(2014\)](#); [Naveau and Thao \(2022\)](#).



# References

- Alexander J. McNeil, Rudiger Frey, P. E. (2005). *Quantitative Risk Management - Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press.
- Ali, M. M., Mikhail, N., and Haq, M. (1978). A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8(3):405–412.
- Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A., and Fermanian, J.-D. (2020). Estimation of copulas via maximum mean discrepancy.
- Alvarez, M., Sala, C., Sun, Y., Pérez, J., Zhang, K., Montanez, A., Bonomi, G., Veeramachaneni, K., Ramírez, I., and Hofman, F. (2021). Copulas. <https://github.com/sdv-dev/Copulas>.
- Andrée, E., Drews, M., Su, J., Larsen, M. A. D., Drønen, N., and Madsen, K. S. (2022). Simulating wind-driven extreme sea levels: Sensitivity to wind speed and direction. *Weather and Climate Extremes*, 36:100422.
- Asenova, S., Mazo, G., and Segers, J. (2021). Inference on extremal dependence in the domain of attraction of a structured Hüsler-Reiss distribution motivated by a Markov tree with latent variables. *Extremes*, 24(3):461–500.
- Avella-Medina, M., Davis, R. A., and Samorodnitsky, G. (2021). Spectral learning of multivariate extremes. *arXiv preprint arXiv:2111.07799*.
- Avella-Medina, M., Davis, R. A., and Samorodnitsky, G. (2022). Kernel pca for multivariate extremes. *arXiv preprint arXiv:2211.13172*.
- Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367.
- Bador, M., Naveau, P., Gilleland, E., Castellà, M., and Arivelo, T. (2015). Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over europe. *Weather and climate extremes*, 9:17–24.
- Balkema, A. A. and Resnick, S. I. (1977). Max-infinite divisibility. *Journal of Applied Probability*, 14(2):309–319.
- Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with Bregman divergences. *Journal of machine learning research*, 6(10).
- Baudin, M., Dutfoy, A., Iooss, B., and Popelin, A.-L. (2017). Openturns: An industrial software for uncertainty quantification in simulation. In *Handbook of uncertainty quantification*, pages 2001–2038. Springer.

## References

---

- Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.
- Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). *Statistics of extremes: theory and applications*. John Wiley & Sons.
- Belzile, L. et al. (2022). *mev: Modelling Extreme Values*. R package version 1.14.
- Berbee, H. C. (1979). Random walks with stationary increments and renewal theory. *MC Tracts*.
- Berghaus, B., Bücher, A., and Volgushev, S. (2017). Weak convergence of the empirical copula process with respect to weighted metrics. *Bernoulli*, 23(1):743 – 772.
- Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in france. *Journal of climate*, 26(20):7929–7937.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732.
- Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering. *The Annals of Statistics*, 48(4):2055 – 2081.
- Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular variation*. Number 27. Cambridge university press.
- Bock, D. and Chapman, J. (2021). Copulae. <https://github.com/DanielBok/copulae>.
- Boistard, H., Chauvet, G., and Haziza, D. (2016). Doubly robust inference for the distribution function in the presence of missing survey data. *Scandinavian Journal of Statistics*, 43(3):683–699.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press.
- Bradley, P. S., Bennett, K. P., and Demiriz, A. (2000). Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0.
- Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.
- Buchanan, A., Walteros, J. L., Butenko, S., and Pardalos, P. M. (2014). Solving maximum clique in sparse graphs: an  $\mathcal{O}(nm + n^2 d/4)$  algorithm for  $d$ -degenerate graphs. *Optimization Letters*, 8:1611–1617.
- Bücher, A., Dette, H., and Volgushev, S. (2011). New estimators of the pickands dependence function and a test for extreme-value dependence. *The Annals of Statistics*, pages 1963–2006.
- Bücher, A. and Segers, J. (2014). Extreme value copula estimation based on block maxima of a multivariate stationary time series. *Extremes*, 17:495–528.
- Bücher, A., Volgushev, S., and Zou, N. (2019). On second order conditions in the multivariate block maxima and peak over threshold method. *Journal of Multivariate Analysis*, 173:604–619.

- Bunea, F., Giraud, C., Luo, X., Royer, M., and Verzelen, N. (2020). Model assisted variable clustering: Minimax-optimal recovery and algorithms. *The Annals of Statistics*, 48(1):111 – 137.
- Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1(none):169 – 194.
- Capéraà, P., Fougères, A.-L., and Genest, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84:567–577.
- Carrera, D., Santana, R., and Lozano, J. A. (2016). Vine copula classifiers for the mind reading problem. *Progress in Artificial Intelligence*, 5:289–305.
- Charpentier, A. and Segers, J. (2009). Tails of multivariate archimedean copulas. *Journal of Multivariate Analysis*, 100(7):1521–1537.
- Chatelain, S., Fougères, A.-L., and Nešlehová, J. G. (2020). Inference for Archimax copulas. *The Annals of Statistics*, 48(2):1025 – 1051.
- Chiapino, M. and Sabourin, A. (2017). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer.
- Chiapino, M., Sabourin, A., and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22:193–222.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.
- Cléménçon, S., Jalalzai, H., Lhaut, S., Sabourin, A., and Segers, J. (2023). Concentration bounds for the empirical angular measure with statistical learning applications. *Bernoulli*, 29(4):2797 – 2827.
- Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.
- Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):377–392.
- Cooley, D., Davis, R. A., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.
- Cooley, D., Naveau, P., and Poncet, P. (2006). Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*, pages 373–390. Springer.
- Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.
- Cordoni, F. and Sancetta, A. (2023). Consistent causal inference for high-dimensional time series. *arXiv preprint arXiv:2307.03074*.
- Csörgő, M. (1983). *Quantile processes with statistical applications*. SIAM.
- Cui, Q. and Zhang, Z. (2018). Max-linear competing factor models. *Journal of Business & Economic Statistics*, 36(1):62–74.



## References

---

- Davis, R. A. and Resnick, S. I. (1989). Basic properties and prediction of max-arma processes. *Advances in applied probability*, 21(4):781–803.
- Davydov, Y. A. (1970). The invariance principle for stationary processes. *Theory of Probability & Its Applications*, 15(3):487–498.
- De Haan, L. and Ferreira, A. (2006). *Extreme value theory: an introduction*, volume 3. Springer.
- De Keyser, S. and Gijbels, I. (2023a). Copula-based divergence measures for Dependence Between Random Vectors. In García-Escudero, L. A., Gordaliza, A., Mayo, A., Lubiano Gomez, M. A., Gil, M. A., Grzegorzewski, P., and Hryniewicz, O., editors, *Building Bridges between Soft and Statistical Methodologies for Data Science*, pages 104–111, Cham. Springer International Publishing.
- De Keyser, S. and Gijbels, I. (2023b). Parametric dependence between random vectors via copula-based divergence measures. *arXiv preprint arXiv:2302.13611*.
- Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique d’indépendance. *Bulletin Royal Belge de L’Académie des sciences*, 65, 274–292.
- Deheuvels, P. (1991). On the limiting behavior of the pickands estimator for bivariate extreme-value distributions. *Statistics & Probability Letters*, 12(5):429–439.
- Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review*, 73(1):111–129.
- Divinsky, B. V., Fomin, V. V., Kosyan, R. D., and Ratner, Y. D. (2020). Extreme wind waves in the Black Sea. *Oceanologia*, 62(1):23–30.
- Dombry, C., Engelke, S., and Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika*, 103(2):303–317.
- Dombry, C. and Ferreira, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25(3):1690–1723.
- Dombry, C., Ribatet, M., and Stoev, S. (2018). Probabilities of concurrent extremes. *Journal of the American Statistical Association*, 113(524):1565–1582.
- Doukhan, P., Massart, P., and Rio, E. (1995a). Invariance principles for absolutely regular empirical processes. In *Annales de l’IHP Probabilités et statistiques*, volume 31, pages 393–427.
- Doukhan, P., Massart, P., and Rio, E. (1995b). Invariance principles for absolutely regular empirical processes. *Annales de l’I.H.P. Probabilités et statistiques*, 31:393–427.
- Drees, H. (2001). Minimax risk bounds in extreme value theory. *The Annals of Statistics*, 29(1):266–294.
- Drees, H. and Huang, X. (1998). Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–46.
- Drees, H. and Sabourin, A. (2021). Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908 – 943.
- Durante, F., Pappadà, R., and Torelli, N. (2015). Clustering of time series via non-parametric tail dependence estimation. *Statistical Papers*, 56:701–721.
- Durante, F. and Sempì, C. (2015). *Principles of Copula Theory*. Taylor & Francis.

- Einmahl, J. H., Kiriliouk, A., and Segers, J. (2018). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, 21:205–233.
- Einmahl, J. H. J., Krajina, A., and Segers, J. (2012). An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764 – 1793.
- Einmahl, J. H. J. and Lin, T. (2006). Asymptotic normality of extreme value estimators on  $\mathcal{C}([0, 1])$ . *The Annals of Statistics*, 34(1):469–492.
- Einmahl, J. H. J. and Segers, J. (2021). Empirical tail copulas for functional data. *The Annals of Statistics*, 49(5):2672 – 2696.
- Embleton, J., Knight, M. I., and Ombao, H. (2020). Multiscale modelling of replicated nonstationary time series. *arXiv preprint arXiv:2005.09440*.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.
- Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.
- Engelke, S. and Volgushev, S. (2022). Structure learning for extremal tree models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2055–2087.
- Eppstein, D., Löffler, M., and Strash, D. (2010). Listing all maximal cliques in sparse graphs in near-optimal time. In *Algorithms and Computation: 21st International Symposium, ISAAC 2010, Jeju Island, Korea, December 15-17, 2010, Proceedings, Part I 21*, pages 403–414. Springer.
- Falk, M., Hüsler, J., and Reiss, R. (2010). *Laws of Small Numbers: Extremes and Rare Events*. Springer Basel.
- Fermanian, J.-D., Radulovic, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860.
- Ferreira, H. (2011). Dependence between two multivariate extremes. *Statistics & Probability Letters*, 81(5):586–591.
- Ferreira, H., Martins, A., and Temido, M. (2021). Extremal behaviour of a periodically controlled sequence with imputed values. *Statistical Papers*, 62:1–23.
- Fomichov, V. and Ivanovs, J. (2022). Spherical clustering in detection of groups of concomitant extremes. *Biometrika*.
- Fonseca, C., Pereira, L., Ferreira, H., and Martins, A. P. (2015). Generalized madogram and pairwise dependence of maxima over two regions of a random field. *Kybernetika*, 51(2):193–211.
- Fougères, A.-L., De Haan, L., and Mercadier, C. (2015). Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2):903–934.
- Fougères, A.-L., Mercadier, C., and Nolan, J. P. (2013). Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, 116:109–129.
- Frank, M. (1979). On the simultaneous associativity of  $f(x, y)$  and  $x + y - f(x, y)$ . *Aequationes mathematicae*, 19:194–226.

## References

---

- Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial journal*, 2(1):1–25.
- Fréjaville, T. and Curt, T. (2015). Spatiotemporal patterns of changes in fire regime and climate: defining the pyroclimates of south-eastern france (mediterranean basin). *Climatic Change*, 129:239–251.
- Gaetan, C. and Guyon, X. (2008). *Modélisation et statistique spatiales*. Mathématiques & applications. Springer, Berlin Heidelberg New York.
- Galambos, J. (1977). Bonferroni inequalities. *The Annals of Probability*, pages 577–581.
- Geffroy, J. (1958). Contributions à la théorie des valeurs extrêmes. *Publ. Inst. Statist. Univ. Paris*, 7:37–185.
- Geffroy, J. (1959). Contribution à la théorie des valeurs extrêmes (suite). In *Annales de l'ISUP*, volume 8, pages 123–185.
- Genest, C., Ghouli, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.
- Genest, C. and Remillard, B. (2004). Test of independence and randomness based on the empirical copula process. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 13:335–369.
- Genest, C. and Segers, J. (2009). Rank-based inference for bivariate extreme-value copulas. *The Annals of Statistics*, 37(5B):2990–3022.
- Ghouli, K., Kulperger, R. J., and Rémillard, B. (2001). A nonparametric test of serial independence for time series and residuals. *Journal of Multivariate Analysis*, 79(2):191–218.
- Gijbels, I., Omelka, M., and Veraverbeke, N. (2015). Estimation of a copula when a covariate affects only marginal distributions. *Scandinavian Journal of Statistics*, 42(4):1109–1126.
- Giraud, C. (2021). *Introduction to high-dimensional statistics*. CRC Press.
- Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693 – 2720.
- Gissibl, N., Klüppelberg, C., and Otto, M. (2018). Tail dependence of recursive max-linear models with regularly varying noise variables. *Econometrics and statistics*, 6:149–167.
- Goix, N., Sabourin, A., and Cléménçon, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on learning theory*, pages 843–860. PMLR.
- Goix, N., Sabourin, A., and Cléménçon, S. (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pages 75–83. PMLR.
- Goix, N., Sabourin, A., and Cléménçon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31.
- Goldstein, S. (1979). Maximal coupling. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 46(2):193–204.
- Gonçalves, A. R., Von Zuben, F. J., and Banerjee, A. (2016). Multi-task sparse structure learning with gaussian copula models. *J. Mach. Learn. Res.*, 17(1):1205–1234.

- Gonzalez, P., Naveau, P., Thao, S., and Worms, J. (2023). A statistical method to model non-stationarity in precipitation records changes. *Geophysical Research Letters*.
- Gröger, M., Arneborg, L., Dieterich, C., Höglund, A., and Meier, H. (2019). Summer hydrographic changes in the Baltic Sea, Kattegat and Skagerrak projected in an ensemble of climate scenarios downscaled with a coupled regional ocean–sea ice–atmosphere model. *Climate Dynamics*, 53:5945–5966.
- Gudendorf, G. and Segers, J. (2010). Extreme-value copulas. In *Copula theory and its applications*, volume 198 of *Lect. Notes Stat. Proc.*, pages 127–145. Springer, Heidelberg.
- Guillou, A., Naveau, P., and Schorgen, A. (2014). Madogram and asymptotic independence among maxima. *REVSTAT*, 12(2):119–134.
- Guillou, A., Padoan, S. A., and Rizzelli, S. (2018). Inference for asymptotically independent samples of extremes. *Journal of Multivariate Analysis*, 167:114–135.
- Gumbel, E. (1960a). Distributions de valeurs extrêmes en plusieurs dimensions. *Publications de l’institut de Statistique de l’Université de Paris*, 9:171–173.
- Gumbel, E. J. (1960b). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.
- Hall, A. and Scotto, M. (2008). On the extremes of randomly sub-sampled time series. *REVSTAT – Statistical Journal Volume*, 6:151–164.
- Hall, P., Peng, L., and Yao, Q. (2002). Moving-maximum models for extrema of time series. *Journal of statistical planning and inference*, 103(1-2):51–63.
- Hall, P. and Tajvidi, N. (2000). Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli*, 6(6):835–844.
- Hamori, S., Motegi, K., and Zhang, Z. (2019). Calibration estimation of semiparametric copula models with data missing at random. *Journal of Multivariate Analysis*, 173:85–109.
- Hamori, S., Motegi, K., and Zhang, Z. (2020). Copula-based regression models with data missing at random. *Journal of Multivariate Analysis*, 180:104654.
- Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, 585(7825):357–362.
- Hersbach, Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.
- Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2018). ERA5 hourly data on single levels from 1959 to present.

## References

---

- Hitz, A. and Evans, R. (2016). One-component regular variation and graphical modeling of extremes. *Journal of Applied Probability*, 53(3):733–746.
- Hofert, M., Huser, R., and Prasad, A. (2018). Hierarchical Archimax copulas. *Journal of Multivariate Analysis*, 167:195–211.
- Hofert, M., Mächler, M., and McNeil, A. J. (2012). Likelihood inference for archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis*, 110:133–150.
- Hsing, T. (1989). Extreme value theory for multivariate stationary sequences. *Journal of Multivariate Analysis*, 29(2):274–291.
- Huang, X. (1992). *Statistics of bivariate extreme values*. Thesis Publishers Amsterdam.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.
- Hult, H. and Lindskog, F. (2006). Regular variation for measures on metric spaces. *Publications de l'Institut Mathématique*, 80(94):121–140.
- Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286.
- Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability & Its Applications*, 7(4):349–382.
- Janßen, A. and Wan, P. (2020).  $k$ -means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233.
- Joe, H. (1990). Families of min-stable multivariate exponential and multivariate extreme value distributions. *Statistics & Probability Letters*, 9:75–81.
- Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.
- Joe, H. (2014). *Dependence modeling with copulas*. CRC press.
- Jun Yan (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21.
- Karp, R. M. (1972). *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA.
- Kimeldorf, G. and Sampson, A. R. (1989). A framework for positive dependence. *Annals of the Institute of Statistical Mathematics*, 41(1):31–45.
- Kiriliouk, A. and Naveau, P. (2020). Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. *The Annals of Applied Statistics*, 14(3):1342 – 1358.
- Kiriliouk, A. and Zhou, C. (2022). Estimating probabilities of multivariate failure sets based on pairwise tail dependence coefficients. *arXiv preprint arXiv:2210.12618*.
- Klüppelberg, C. and Krali, M. (2021). Estimating an extreme bayesian network via scalings. *Journal of Multivariate Analysis*, 181:104672.
- Klüppelberg, C. and Lauritzen, S. (2019). Bayesian networks for max-linear models. *Network Science: An Aerial View*, pages 79–97.

- Koh, J., Pimont, F., Dupuy, J.-L., and Opitz, T. (2023). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *The Annals of Applied Statistics*, 17(1):560 – 582.
- Kojadinovic, I. and Holmes, M. (2009). Tests of independence among continuous random vectors based on cramér–von mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154.
- Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula r package. *Journal of Statistical Software*, 34:1–20.
- Kojadinovic, I. and Yan, J. (2011). Tests of serial independence for continuous multivariate time series based on a möbius decomposition of the independence empirical copula process. *Annals of the Institute of Statistical Mathematics*, 63:347–373.
- Kontorovich, L. A. and Ramanan, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126 – 2158.
- Kulik, R. and Soulier, P. (2020). *Heavy-tailed time series*. Springer.
- Kyrillidis, A., Becker, S., Cevher, V., and Koch, C. (2013). Sparse projections onto the simplex. In *International Conference on Machine Learning*, pages 235–243. PMLR.
- Lalancette, M., Engelke, S., and Volgushev, S. (2021). Rank-based estimation under asymptotic dependence and independence, with applications to spatial extremes. *The Annals of Statistics*, 49(5):2552–2576.
- Laloë, T. (2010). L 1-quantization and clustering in banach spaces. *Mathematical Methods of Statistics*, 19:136–150.
- Laloë, T. (2021). Quantization based clustering: An iterative approach. *Pattern Recognition Letters*, 142:51–57.
- Le, P. D., Davison, A. C., Engelke, S., Leonard, M., and Westra, S. (2018). Dependence properties of spatial rainfall extremes and areal reduction factors. *Journal of Hydrology*, 565:711–719.
- Linder, T. (2002). Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning*, pages 163–210. Springer.
- Lindskog, F., Resnick, S. I., and Roy, J. (2014). Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probability Surveys*, 11(none):270 – 314.
- Lopez-Paz, D., Hernández-Lobato, J. M., and Zoubin, G. (2013). Gaussian process vine copulas for multivariate dependence. In *International Conference on Machine Learning*, pages 10–18. PMLR.
- Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90 – 102.
- Marcon, G., Padoan, S., Naveau, P., Muliere, P., and Segers, J. (2017). Multivariate nonparametric estimation of the pickands dependence function using bernstein polynomials. *Journal of statistical planning and inference*, 183:1–17.
- Marius Hofert and Martin Mächler (2011). Nested Archimedean copulas meet R: The nacopula package. *Journal of Statistical Software*, 39(9):1–20.

## References

---

- Marshall, A. W. and Olkin, I. (1983). Domains of Attraction of Multivariate Extreme Value Distributions. *The Annals of Probability*, 11(1):168 – 177.
- Marshall, A. W. and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association*, 83(403):834–841.
- Martius, O., Pfahl, S., and Chevalier, C. (2016). A global quantification of compound precipitation and wind extremes. *Geophysical Research Letters*, 43(14):7709–7717.
- Maume-Deschamps, V., Ribereau, P., and Zeidan, M. (2023). Detecting the stationarity of spatial dependence structure using spectral clustering.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.
- McNeil, A. J. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059 – 3097.
- Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5, pages 273–293. Institute of Mathematical Statistics.
- Meyer, N. and Wintenberger, O. (2021). Sparse regular variation. *Advances in Applied Probability*, 53(4):1115–1148.
- Meyer, N. and Wintenberger, O. (2023). Multivariate sparse clustering for extremes. *Journal of the American Statistical Association*, (just-accepted):1–23.
- Mishra, A. K. and Singh, V. P. (2011). Drought modeling – a review. *Journal of Hydrology*, 403(1):157–175.
- Mohri, M. and Rostamizadeh, A. (2010). Stability bounds for stationary  $\varphi$ -mixing and  $\beta$ -mixing processes. *Journal of Machine Learning Research*, 11(2).
- Mourahib, A., Kiriliouk, A., and Segers, J. (2023). Multivariate generalized pareto distributions along extreme directions. *arXiv preprint arXiv:2311.04618*.
- Naveau, P., Guillou, A., Cooley, D., and Diebolt, J. (2009). Modelling pairwise dependence of maxima in space. *Biometrika*, 96(1):1–17.
- Naveau, P., Guillou, A., and Rietsch, T. (2014). A non-parametric entropy-based approach to detect changes in climate extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(5):861–884.
- Naveau, P. and Thao, S. (2022). Multimodel errors and emergence times in climate attribution studies. *Journal of Climate*, 35(14):4791–4804.
- Nelsen, R. B. (2006). *An introduction to copulas*. Springer.
- Nicolas, M. L. (2022). pycop: a python package for dependence modeling with copulas. *Zenodo Software Package*, 70:7030034.
- Nolde, N. and Wadsworth, J. L. (2021). Linking representations for multivariate extremes via a limit set.
- Oliveira, J. D. T. and Galambos, J. (1977). The asymptotic theory of extreme order statistics. *International Statistical Review*, 47:230.

- Ombao, H., Von Sachs, R., and Guo, W. (2005). Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531.
- Owen, L. E., Catto, J. L., Stephenson, D. B., and Dunstone, N. J. (2021). Compound precipitation and wind extremes over europe and their relationship to extratropical cyclones. *Weather and Climate Extremes*, 33:100342.
- Palacios-Rodriguez, F., Bernardino, E. D., and Mailhot, M. (2023). Smooth copula-based generalized extreme value model and spatial interpolation for extreme rainfall in central eastern canada. *Environmetrics*, 34(3):e2795.
- Pappadà, R., Durante, F., Salvadori, G., and De Michele, C. (2018). Clustering of concurrent flood risks via hazard scenarios. *Spatial Statistics*, 23:124–142.
- Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.
- Patton, A. J. and Wiley, J. (2006). Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics*, pages 147–173.
- Peligrad, M. (1983). A note on two measures of dependence and mixing sequences. *Advances in Applied Probability*, 15(2):461–464.
- Pfahl, S. and Wernli, H. (2012). Spatial coherency of extreme weather events in germany and switzerland. *International journal of climatology*, 32(12):1863–1874.
- Pickands, J. (1981). Multivariate extreme value distribution. *Proceedings 43th, Session of International Statistical Institution, 1981*, 49:859–878.
- Pinto, J. G., Karremann, M. K., Born, K., Della-Marta, P. M., and Klawa, M. (2012). Loss potentials associated with european windstorms under future climate conditions. *Climate Research*, 54(1):1–20.
- Pollard, D. (1981). Strong Consistency of  $K$ -Means Clustering. *The Annals of Statistics*, 9(1):135 – 140.
- Portier, F. and Segers, J. (2018). On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, 166:160–181.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.
- Raveh-Rubin, S. and Wernli, H. (2015). Large-scale wind and precipitation extremes in the mediterranean: a climatological analysis for 1979–2012. *Quarterly Journal of the Royal Meteorological Society*, 141(691):2404–2417.
- Resnick, S. (2008). *Extreme Values, Regular Variation, and Point Processes*. Applied probability. Springer.
- Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.
- Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771.
- Rio, E. (1993). Covariance inequalities for strongly mixing processes. In *Annales de l’IHP Probabilités et statistiques*, volume 29, pages 587–597.



## References

---

- Rio, E. (1999). *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31. Springer Science & Business Media.
- Rio, E. (2017). *Asymptotic theory of weakly dependent random processes*, volume 80. Springer.
- Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized pareto distributions. *Bernoulli*, 12(5):917–930.
- Rosenblatt, M. (1956a). A central limit theorem and a strong mixing condition. *Proceedings of the national Academy of Sciences*, 42(1):43–47.
- Rosenblatt, M. (1956b). Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pages 832–837.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Ruffault, J., Moron, V., Trigo, R., and Curt, T. (2016). Objective identification of multiple large fire climatologies: an application to a mediterranean ecosystem. *Environmental Research Letters*, 11(7):075006.
- Saunders, K., Stephenson, A., and Karoly, D. (2021). A regionalisation approach for rainfall based on extremal dependence. *Extremes*, 24(2):215–240.
- Scarsini, M. (1984). On measures of concordance. *Stochastica*, 8(3):201–218.
- Schepsmeier, U., Stoeber, J., Brechmann, E., Graeler, B., Nagler, T., Erhardt, T., Almeida, C., Min, A., Czado, M., Hofmann, M., and Kiliches, M. (2019). Vinecopula : Statistical inference of vine copulas. *Package "VineCopula". R package, version 2.3.0.*
- Schlather, M. and Tawn, J. (2002). Inequalities for the extremal coefficients of multivariate extreme value distributions. *Extremes*, 5(1):87–102.
- Schwierz, C., Köllner-Heck, P., Zenklusen Mutter, E., Bresch, D. N., Vidale, P.-L., Wild, M., and Schär, C. (2010). Modelling european winter wind storm losses in current and future climate. *Climatic change*, 101:485–514.
- Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18(3):764 – 782.
- Segers, J. (2015). Hybrid copula estimators. *J. Statist. Plann. Inference*, 160:23–34.
- Segers, J., Zhao, Y., and Meinguet, T. (2017). Polar decomposition of regularly varying time series in star-shaped metric spaces. *Extremes*, 20:539–566.
- Shoeb, A. H. (2009). *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology.
- Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.
- Sibuya, M. et al. (1960). Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210.
- Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

- Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2021). A geometric investigation into the tail dependence of vine copulas. *Journal of Multivariate Analysis*, 184:104736.
- Sklar, A. (1959). Fonctions de répartition à  $n$  dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.
- Smith, R. L. (1990). Max-stable processes and spatial extremes. *unpublished work*.
- Soukissian, T., Karathanasi, F., Axaopoulos, P., Voukouvalas, E., and Kotroni, V. (2018). Offshore wind climate analysis and variability in the Mediterranean Sea. *International Journal of Climatology*, 38(1):384–402.
- Stephenson, A. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59.
- Stephenson, A. G. (2002). evd: Extreme Value Distributions. *R News*, 2(2).
- Strokorb, K. (2020). Extremal independence old and new.
- Strzelczyk, A., Aledo-Serrano, A., Coppola, A., Didelot, A., Bates, E., Sainz-Fuertes, R., and Lawthom, C. (2023). The impact of epilepsy on quality of life: Findings from a european survey. *Epilepsy & Behavior*, 142:109179.
- Sun, Y., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Learning vine copula models for synthetic data generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5049–5057.
- Takahashi, R. (1987). Some properties of multivariate extreme value distributions and multivariate tail equivalence. *Annals of the Institute of Statistical Mathematics*, 39:637–647.
- Takahashi, R. (1994). Asymptotic independence and perfect dependence of vector components of multivariate extreme statistics. *Statistics & Probability Letters*, 19(1):19–26.
- Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415.
- Tawn, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253.
- Tolosana-Delgado, R., Otero, N., Pawlowsky-Glahn, V., and Soler, A. (2005). Latent compositional factors in the llobregat river basin (spain) hydrogeochemistry. *Mathematical Geology*, 37:681–702.
- Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer New York, New York, NY.
- Veeramachaneni, K., Cuesta-Infante, A., and O'Reilly, U.-M. (2015). Copula graphical models for wind resource estimation. In *IJCAI*.
- Vignotto, E., Engelke, S., and Zscheischler, J. (2021). Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain and Ireland. *Weather and climate extremes*, 32:100318.

## References

---

- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.
- Volkonskii, V. and Rozanov, Y. A. (1959). Some limit theorems for random functions. i. *Theory of Probability & Its Applications*, 4(2):178–197.
- Wang, S., Dieterich, C., Döscher, R., Höglund, A., Hordoir, R., Meier, H. M., Samuelsson, P., and Schimanke, S. (2015). Development and evaluation of a new regional coupled atmosphere–ocean model in the North Sea and Baltic Sea. *Tellus A: Dynamic Meteorology and Oceanography*, 67(1):24284.
- Wintenberger, O. (2010). Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, 15:489 – 503.
- Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301.
- Xia, Y., Fabian, P., Stohl, A., and Winterhalter, M. (1999). Forest climatology: estimation of missing values for bavaria, germany. *Agricultural and Forest Meteorology*, 96(1):131–144.
- Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21:1–21.
- Yi, Y. and Neykov, M. (2024). Non-asymptotic bounds for the  $\ell_\infty$  estimator in linear regression with uniform noise. *Bernoulli*, 30(1):534 – 553.
- Zhang, X., Blanchet, J., Marzouk, Y., Nguyen, V. A., and Wang, S. (2023). Wasserstein-based minimax estimation of dependence in multivariate regularly varying extremes. *arXiv preprint arXiv:2312.09862*.
- Zou, N., Volgushev, S., and Bücher, A. (2021). Multiple block sizes and overlapping blocks for multivariate time series extremes. *The Annals of Statistics*, 49(1):295 – 320.
- Zscheischler, J., Naveau, P., Martius, O., Engelke, S., and Raible, C. C. (2021). Evaluating the dependence structure of compound precipitation and wind speed extremes. *Earth system dynamics*, 12(1):1–16.

## APPENDIX E

# A PYTHON PACKAGE FOR SAMPLING FROM COPULAE

This chapter has been recognised as an original contribution to a scientific journal:



Alexis Boulin. 2023. “A Python Package for Sampling from Copulae: Clayton.” *Computo*, January.

### Abstract.

The package `clayton` is designed to be intuitive, user-friendly, and efficient. It offers a wide range of copula models, including Archimedean, Elliptical, and Extreme. The package is implemented in pure Python, making it easy to install and use. In addition, we provide detailed documentation and examples to help users get started quickly. We also conduct a performance comparison with existing R packages, demonstrating the efficiency of our implementation. The `clayton` package is a valuable tool for researchers and practitioners working with copulas in Python.

## E.1 Introduction

Modeling dependence relations between random variables is a topic of interest in probability theory and statistics. The most popular approach is based on the second moment of the underlying random variables, namely, the covariance. It is well known that only linear dependence can be captured by the covariance and it is only characteristic for a few models, e.g., the multivariate normal distribution or binary random variables. As a beneficial alternative to dependence, the concept of copulas, going back [Sklar \(1959\)](#), has drawn a lot of attention. The copula  $C : [0, 1]^d \rightarrow [0, 1]$  of a random vector  $\mathbf{X} = (X_0, \dots, X_{d-1})$  with  $d \geq 2$  allows us to separate the effect of dependence from the effect of the marginal distribution, such that:

$$\mathbb{P}\{X_0 \leq x_0, \dots, X_{d-1} \leq x_{d-1}\} = C(\mathbb{P}\{X_0 \leq x_0\}, \dots, \mathbb{P}\{X_{d-1} \leq x_{d-1}\}),$$

where  $(x_0, \dots, x_{d-1}) \in \mathbb{R}^d$ . The main consequence of this identity is that the copula completely characterizes the stochastic dependence between the margins of  $\mathbf{X}$ .

In other words, copulae allow us to model marginal distributions and dependence structure separately. Furthermore, motivated by Sklar’s theorem, the problem of investigating stochastic dependence is reduced to the study of multivariate distribution functions under the unit hypercube  $[0, 1]^d$  with uniform margins. The theory of copulae has been of prime interest for many applied fields of science, such as quantitative finance ([Patton \(2012\)](#)) or environmental sciences ([Mishra and Singh \(2011\)](#)). This increasing number of applications has led to a demand for statistical methods. For example, semiparametric estimation ([Genest et al. \(1995\)](#)),

nonparametric estimation (Fermanian et al. (2004)) of copulae or nonparametric estimation of conditional copulae (Gijbels et al. (2015); Portier and Segers (2018)) have been investigated. These results are established for a fixed arbitrary dimension  $d \geq 2$ , but several investigations (e.g. Einmahl and Lin (2006); Einmahl and Segers (2021)) are done for functional data for the tail copula, which captures dependence in the upper tail.

Software implementation of copulas has been extensively studied in R, for example in the packages Jun Yan (2007); Schepsmeier et al. (2019); Stephenson (2002). However, methods for working with copulas in Python are still limited. As far as we know, copula-dedicated packages in Python are mainly designed for modeling, such as Alvarez et al. (2021) and Bock and Chapman (2021). These packages use maximum likelihood methods to estimate the copula parameters from observed data and generate synthetic data using the estimated copula model. Other packages provide sampling methods for copulas, but they are typically restricted to the bivariate case and the conditional simulation method (see, for example, Baudin et al. (2017)). Additionally, these packages often only consider Archimedean and elliptical copulas, and do not include the extreme value class in arbitrary dimensions  $d \geq 2$  (Nicolas (2022)). In this paper, we propose to implement a wide range of copulas, including the extreme value class, in arbitrary fixed dimension  $d \geq 2$ .

Through this paper we adopt the following notational conventions: all the indices will start at 0 as in Python. Consider  $(\Omega, \mathcal{A}, \mathbb{P})$  a probability space and let  $\mathbf{X} = (X_0, \dots, X_{d-1})$  be a  $d$ -dimensional random vector with values in  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , with  $d \geq 2$  and  $\mathcal{B}(\mathbb{R}^d)$  the Borel  $\sigma$ -algebra of  $\mathbb{R}^d$ . This random vector has a joint distribution  $F$  with copula  $C$  and its margins are denoted by  $F_j(x) = \mathbb{P}\{X_j \leq x\}$  for all  $x \in \mathbb{R}$  and  $j \in \{0, \dots, d-1\}$ . Denote by  $\mathbf{U} = (U_0, \dots, U_{d-1})$  a  $d$  random vector with copula  $C$  and uniform margins. All bold letters  $\mathbf{x}$  will denote a vector of  $\mathbb{R}^d$ .

The `clayton` package, whose Python code can be found in this [GitHub repository](#), uses object-oriented features of the Python language. The package contains classes for Archimedean, elliptical, and extreme value copulas. In section E.2, we briefly describe the classes defined in the package. Section E.3 presents methods for generating random vectors. In section E.4, we apply the `clayton` package to model pairwise dependence between maxima. Section E.5 discusses potential improvements to the package and provides concluding remarks. The appendices from E.6 to E.10 define and illustrate all the parametric copula models implemented in the package.

## E.2 Classes

The `clayton` package defines three main classes: **Multivariate**, **Archimedean**, and **Extreme**. The **Multivariate** class is designed for defining multivariate copulas (including the bivariate case) and is located at the highest level of the code architecture. It contains methods for instantiating a copula object or for sampling from a copula with desired margins using inversion methods, for example. The **Archimedean** and **Extreme** classes are children of the **Multivariate** class and represent copulas from the Archimedean and extreme value families, respectively. The **Gaussian** and **Student** classes represent elliptical copulas. This hierarchical structure is relevant theoretically, as Archimedean and extreme value copulas are studied independently (see, for example, Charpentier and Segers (2009) and Genest and Segers (2009)), and practically, as they contain effective sampling methods. However, the **Gaussian** and **Student** classes are

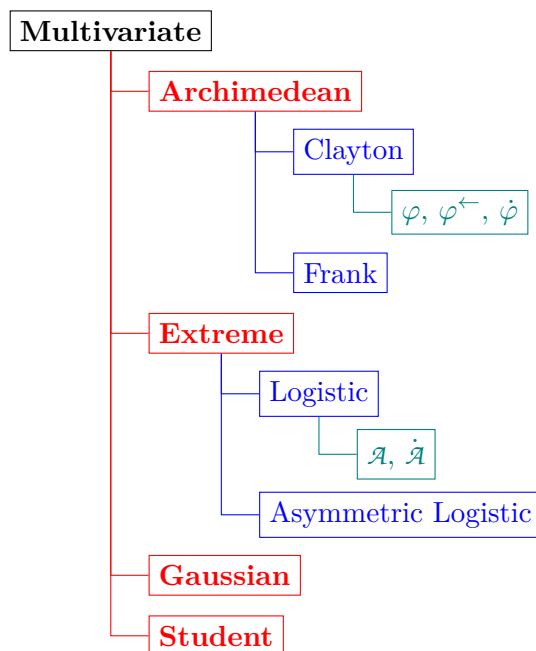


Fig. E.1 The figure shows an object diagram that structures the code. The **Multivariate** class serves as the root and is used to instantiate all its child classes **Archimedean**, **Extreme**, **Gaussian**, and **Student** in red. The blue-colored classes correspond to various parametric copula models, and the green-colored classes represent examples of methods. Symbols  $\varphi$ ,  $\varphi^{\leftarrow}$ ,  $\dot{\varphi}$  correspond to the generator function, its inverse, and its derivative, respectively, while  $\mathcal{A}$ ,  $\dot{\mathcal{A}}$  refer to the Pickands dependence function and its derivative.

split, as the most effective sampling algorithms are specific to each and cannot be generalized in a broader elliptical class.

The architecture of the code is shown in Figure E.1. At the third level of the architecture, we find important parametric models of Archimedean and extreme value copulas (depicted as blue in the figure). These parametric models contain methods such as the generator function  $\varphi$  (see Section E.2.1) for Archimedean copulas and the Pickands dependence function  $\mathcal{A}$  (see Section E.2.2) for extreme value copulas (depicted as green in the figure). We provide a brief overview of Archimedean copulas and some of their properties in high-dimensional spaces in Section E.2.1. A characterization of extreme value copulas is given in Section E.2.2. The sections from E.6 to E.10 define and illustrate all the copula models implemented in the package.

### E.2.1 The Archimedean class

Let  $\varphi$  be a generator that is a strictly decreasing, convex function from  $[0, 1]$  to  $[0, \infty]$  such that  $\varphi(1) = 0$  and  $\varphi(0) = \infty$ . We denote the generalized inverse of  $\varphi$  by  $\varphi^{\leftarrow}$ . Consider the following equation:

$$C(\mathbf{u}) = \varphi^{\leftarrow}(\varphi(u_0) + \cdots + \varphi(u_{d-1})). \quad (\text{E.1})$$

If this relation holds and  $C$  is a copula function, then  $C$  is called an Archimedean copula. A necessary condition for (E.1) to be a copula is that the generator  $\varphi$  is a  $d$ -monotonic function,

i.e., it is differentiable up to the order  $d$  and its derivatives satisfy

$$(-1)^k (\varphi)^{(k)}(x) \geq 0, \quad k \in \{1, \dots, d\} \quad (\text{E.2})$$

for  $x \in (0, \infty)$  (see Corollary 2.1 of [McNeil and Nešlehová \(2009\)](#)). Note that  $d$ -monotonic Archimedean inverse generators do not necessarily generate Archimedean copulas in dimensions higher than  $d$  (see [McNeil and Nešlehová \(2009\)](#)). As a result, some Archimedean subclasses are only implemented for the bivariate case as they do not generate an Archimedean copula in higher dimensions. In the bivariate case, (E.2) can be interpreted as  $\varphi$  being a convex function.

The `clayton` package implements common one-parameter families of Archimedean copulas, such as the Clayton ([Clayton \(1978\)](#)), Gumbel ([Gumbel \(1960a\)](#)), Joe ([Joe \(1997\)](#)), Frank ([Frank \(1979\)](#)), and AMH ([Ali et al. \(1978\)](#)) copulas for the multivariate case. It is worth noting that all Archimedean copulas are symmetric, and in dimensions 3 or higher, only positive associations are allowed. For the specific bivariate case, the package also implements other families, such as those numbered from 4.2.9 to 4.2.15 and 4.2.22 in Section 4.2 of [Nelsen \(2006\)](#). Definitions and illustrations of these parametric copula models can be found in appendices E.6 and E.8.

### E.2.2 The Extreme class

Investigating the notion of copulae within the framework of multivariate extreme value theory leads to the extreme value copulae (see [Gudendorf and Segers \(2010\)](#) for an overview) defined as

$$C(\mathbf{u}) = \exp(-L(-\ln(u_0), \dots, -\ln(u_{d-1}))), \quad \mathbf{u} \in (0, 1]^d, \quad (\text{E.3})$$

where  $L : [0, \infty)^d \rightarrow [0, \infty)$  the stable tail dependence function which is convex, homogeneous of order one, namely  $L(c\mathbf{x}) = cL(\mathbf{x})$  for  $c > 0$  and satisfies  $\max(x_0, \dots, x_{d-1}) \leq L(x_0, \dots, x_{d-1}) \leq x_0 + \dots + x_{d-1}$ ,  $\forall \mathbf{x} \in [0, \infty)^d$ . Let  $\Delta^{d-1} = \{\mathbf{w} \in [0, 1]^d : w_0 + \dots + w_{d-1} = 1\}$  be the unit simplex. The Pickands dependence function  $\mathcal{A} : \Delta^{d-1} \rightarrow [1/d, 1]$  characterizes  $L$  by its homogeneity, which is the restriction of  $L$  to the unit simplex  $\Delta^{d-1}$ :

$$L(x_0, \dots, x_{d-1}) = (x_0 + \dots + x_{d-1})\mathcal{A}(w_0, \dots, w_{d-1}), \quad w_j = \frac{x_j}{x_0 + \dots + x_{d-1}}, \quad (\text{E.4})$$

for  $j \in \{1, \dots, d-1\}$  and  $w_0 = 1 - w_1 - \dots - w_{d-1}$  with  $\mathbf{x} \in [0, \infty)^d \setminus \{\mathbf{0}\}$ . The Pickands dependence function characterizes the extremal dependence structure of an extreme value random vector and verifies  $\max\{w_0, \dots, w_{d-1}\} \leq \mathcal{A}(w_0, \dots, w_{d-1}) \leq 1$  where the lower bound corresponds to comonotonicity and the upper bound corresponds to independence. Estimating this function is an active area of research, with many compelling studies having been conducted on the topic (see, for example, [Bücher et al. \(2011\)](#); [Gudendorf and Segers \(2010\)](#)).

From a practical point of view, the family of extreme value copulae is very rich and arises naturally as the limiting distribution of properly normalised componentwise maxima. Furthermore, it contains a rich variety of parametric models and allows asymmetric dependence, that is, for the bivariate case:

$$C(u_0, u_1) \neq C(u_1, u_0).$$

In the multivariate framework, the logistic copula (or Gumbel, see [Gumbel \(1960a\)](#)), the asymmetric logistic copula ([Tawn \(1990\)](#)), the Hüsler and Reiss distribution ([Hüsler and](#)

Reiss (1989)), the t-EV copula (Demarta and McNeil (2005)), Bilogistic model (Smith (1990)) are implemented. It's worth noting that the logistic copula is the sole model that is both Archimedean and extreme value. The library includes bivariate extreme value copulae such as asymmetric negative logistic (Joe (1990)), asymmetric mixed (Tawn (1988)). The reader is again invited to read from E.7 to E.9 for precise definitions of these models.

## E.3 Random vector generator

We propose a Python-based implementation of a random vector generator that is capable of generating random vectors from a wide variety of copulas. The `clayton` package requires a few external libraries in order to function properly. These libraries are commonly used in scientific Python programming and are easy to install.

The required libraries are:

- `numpy` version 1.6.1 or newer. This is the fundamental package for scientific computing, it contains linear algebra functions and matrix / vector objects (Harris et al. (2020)).
- `scipy` version 1.7.1 or newer. A library of open-source software for mathematics, science and engineering (Virtanen et al. (2020)).

The `clayton` package provides two methods for generating random vectors: `sample_unimargin` and `sample`. The first method generates a sample where the margins are uniformly distributed on the unit interval  $[0, 1]$ , while the second method generates a sample from the chosen margins.

In Section E.3.1, we present an algorithm that uses the conditioning method to sample from a copula. This method is very general and can be used for any copula that is sufficiently smooth (see Equations (E.5) and (E.6) below). However, the practical infeasibility of the algorithm in dimensions higher than 2 and the computational intensity of numerical inversion call for more efficient ways to sample in higher dimensions. The purpose of Section E.3.2 is to present such methods and to provide details on the methods used in the `clayton` package. In each section, we provide examples of code to illustrate how to instantiate a copula and how to sample with `clayton`.

In the following sections, we will use Python code that assumes that the following packages have been loaded:

```
>>> import clayton
>>> from clayton.rng import base, evd, archimedean, monte_carlo
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> from scipy.stats import norm, expon
>>> np.random.seed(42)
```

### E.3.1 The bivariate case

In this subsection, we address the problem of generating a bivariate sample from a specified joint distribution with  $d = 2$ . Suppose that we want to sample a bivariate random vector  $\mathbf{X}$  with copula  $C$ . In the case where the components are independent, the sampling procedure is



straightforward: we can independently sample  $X_0$  and  $X_1$ . However, in the general case where the copula is not the independence copula, this approach is not applicable.

One solution to this problem is to use the conditioning method to sample from the copula. This method relies on the fact that given  $(U_0, U_1)$  with copula  $C$ , the conditional law of  $U_1$  given  $U_0$  is written as:

$$c_{u_0}(u_1) \triangleq \mathbb{P}\{U_1 \leq u_1 | U_0 = u_0\} = \frac{\partial C(u_0, u_1)}{\partial u_0}. \quad (\text{E.5})$$

This allows us to first sample  $U_0$  from a uniform distribution on the unit interval, and then to use the copula to generate  $U_1$  given  $U_0$ . Finally, we can transform the resulting sample  $(U_0, U_1)$  into the original space by applying the inverse marginal distributions  $F_0^{-1}$  and  $F_1^{-1}$  to  $U_0$  and  $U_1$  respectively. Thus, an algorithm for sampling bivariate copulas is given in Algorithm 8. Algorithm 8 presents a procedure for generating a bivariate sample from a copula. The algorithm takes as input the length of the sample  $n$ , as well as the parameters of the copula  $(\theta, \psi_1, \psi_2)$ . The output is a bivariate sample from the desired copula model, denoted  $(u_0^{(1)}, u_1^{(1)}), \dots, (u_0^{(n)}, u_1^{(n)})$ . This algorithm is applicable as long as the copula has a first partial derivative with respect to its first component.

---

**Algorithm 8** Conditional sampling from a copula

---

- 1: **Data:** sample's length  $n$ .
  - 2: Parameter of the copula  $\theta, \psi_1, \psi_2$ .
  - 3: **Result:** Bivariate sample from the desired copula model  $\{(u_0^{(1)}, u_1^{(1)}), \dots, (u_0^{(n)}, u_1^{(n)})\}$ .
  - 4: **procedure** SAMPLING( $n, \theta, \psi_1, \psi_2$ )
  - 5:     Generate two independent uniform random observations on the  $[0, 1]$  segment  $u_0$  and  $t_1$ .
  - 6:     Set  $u_1 = c_{u_0}^{\leftarrow}(t_1)$  where  $c_{u_0}^{\leftarrow}$  denotes the generalized inverse of  $c_{u_0}$ .
  - 7:     The desired pair is  $(u_0, u_1)$ .
- 

For step 6 of the algorithm, we need to find  $u_1 \in [0, 1]$  such that  $c_{u_0}(u_1) - t_1 = 0$  holds. This  $u_1$  always exists because for every  $u \in ]0, 1[$ , we have  $0 \leq c_{u_0}(u) \leq 1$ , and the function  $u \mapsto c_{u_0}(u)$  is increasing (see Theorem 2.2.7 of Nelsen (2006) for a proof). This step can be solved using the `brentq` function from the `scipy` package. A sufficient condition for a copula to have a first partial derivative with respect to its first component in the Archimedean and extreme value cases is that the generator  $\varphi$  and the Pickands dependence function  $\mathcal{A}$  are continuously differentiable on  $]0, 1[$ , respectively. In this case, the first partial derivatives of the copula are given by:

$$\begin{aligned} \frac{\partial C}{\partial u_0}(u_0, u_1) &= \frac{\varphi'(u_0)}{\varphi'(C(u_0, u_1))}, & (u_0, u_1) \in ]0, 1[^2, \\ \frac{\partial C}{\partial u_0}(u_0, u_1) &= \frac{C(u_0, u_1)}{u_0} \mu(t), & (u_0, u_1) \in ]0, 1[^2, \end{aligned}$$

where  $t = \ln(u_1)/\ln(u_0 u_1) \in (0, 1)$  and  $\mu(t) = \mathcal{A}(t) - t\mathcal{A}'(t)$ .

We now have all the necessary theoretical tools to give details on how the `clayton` package is designed. The file `base.py` contains the `Multivariate` class and the `sample` method to generate random numbers from  $\mathbf{X}$  with copula  $C$ . To do so, we use the inversion method that is to sample from  $\mathbf{U}$  using Algorithm 8 and we compose the corresponding uniform margins by  $F_j^{\leftarrow}$ . Equation (E.5) indicates that the sole knowledge of  $\mathcal{A}$  and  $\varphi$  and their respective

derivatives are needed in order to perform the sixth step of Algorithm 8. For that purpose, `cond_sim` method located inside **Archimedean** and **Extreme** classes performs Algorithm 8. Then each child of the bivariate **Archimedean** (resp. **Extreme**) class is thus defined by its generator  $\varphi$  (resp.  $\mathcal{A}$ ), its derivative  $\varphi'$  (resp.  $\mathcal{A}'$ ) and its inverse  $\varphi^{\leftarrow}$  as emphasized in greed in Figure E.1. Namely, we perform Algorithm 8 for the **Archimedean** subclasses Frank, AMH, Clayton (when  $\theta < 0$  for the previous three), Nelsen\_9, Nelsen\_10, Nelsen\_11, Nelsen\_12, Nelsen\_13, Nelsen\_14, Nelsen\_15 and Nelsen\_22. For the **Extreme** class, such algorithm is performed for the **AsyNegLog** and **AsyMix**. For other models, faster algorithms are known and thus implemented, we refer to Section E.3.2 for details.

The following code illustrates the random vector generation for a bivariate Archimedean copula. By defining the parameter of the copula and the sample's length, the constructor for this copula is available and can be called using the `Clayton` method, such as:

```
>>> n_sample, theta = 1024, -0.5
>>> copula = archimedean.Clayton(theta=theta, n_sample=n_sample)
```

To obtain a sample with uniform margins and a Clayton copula, we can use the `sample_unimargin` method, as follows:

```
>>> sample = copula.sample_unimargin()
```

Here, the `sample` object is a numpy array with 2 columns and 1024 rows, where each row contains a realization from a Clayton copula (see Figure E.2).

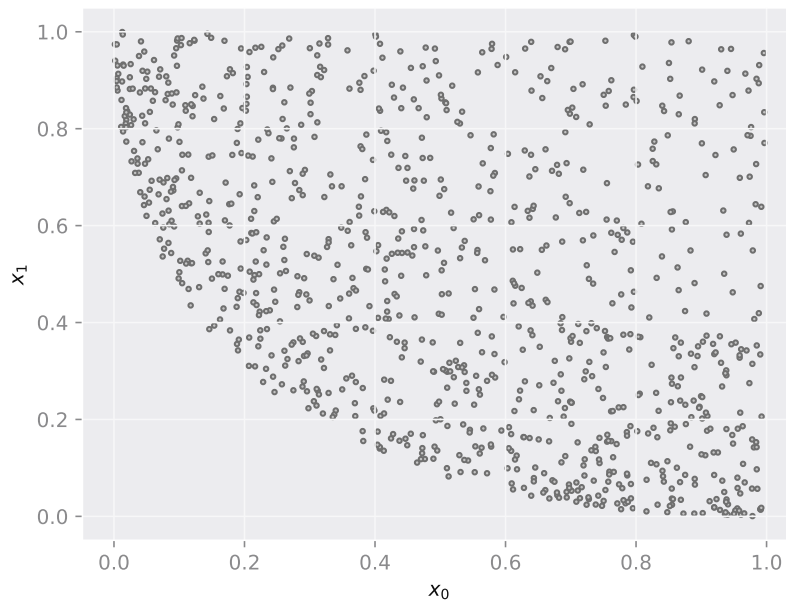


Fig. E.2 Scatterplot of a sample from a Clayton copula ( $\theta = -0.5$ ).

### E.3.2 The multivariate case

We will now address the generation of multivariate Archimedean and Extreme value copulae proposed in the Clayton package. In the multivariate case, the link between partial derivatives and the conditional law remains. Indeed, let  $(U_0, \dots, U_{d-1})$  be a  $d$ -dimensional random vector with uniform margins and copula  $C$ . The conditional distribution of  $U_k$  given the values of  $U_0, \dots, U_{k-1}$  is

$$\mathbb{P}\{U_k \leq u_k | U_0 = u_0, \dots, U_{k-1} = u_{k-1}\} = \frac{\partial^{k-1} C(u_0, \dots, u_k, 1, \dots, 1) / \partial u_0 \dots \partial u_{k-1}}{\partial^{k-1} C(u_0, \dots, u_{k-1}, 1, \dots, 1) / \partial u_0 \dots \partial u_{k-1}}, \quad (\text{E.6})$$

for  $k \in 1, \dots, d-1$ . The conditional simulation algorithm may be written as follows.

1. Generate  $d$  independent uniform random on  $[0, 1]$  variates  $v_0, \dots, v_{d-1}$ .
2. Set  $u_0 = v_0$ .
3. For  $k = 1, \dots, d-1$ , evaluate the inverse of the conditional distribution given by Equation (E.6) at  $v_k$ , to generate  $u_k$ .

Nevertheless, the evaluation of the inverse conditional distribution becomes increasingly complicated as the dimension  $d$  increases. Furthermore, it can be difficult for some models to derive a closed form of Equation (E.6) that makes it impossible to implement it in a general algorithm with only the dimension  $d$  as an input. For multivariate Archimedean copulas, [McNeil and Nešlehová \(2009\)](#) give a method to generate a random vector from the  $d$ -dimensional copula  $C$  with generator  $\varphi$  (see Section 5.2 of [McNeil and Nešlehová \(2009\)](#)). A stochastic representation for Archimedean copulas generated by a  $d$ -monotone generator is given by

$$\mathbf{U} = (\varphi^{\leftarrow}(RS_1), \dots, \varphi^{\leftarrow}(RS_d)) \sim C, \quad (\text{E.7})$$

where  $R \sim F_R$ , the radial distribution which is independent of  $S$  and  $S$  is distributed uniformly in the unit simplex  $\Delta^{d-1}$ . One challenging aspect of this algorithm is to have an accurate evaluation of the radial distribution of the Archimedean copula and thus to numerically inverse this distribution. The associated radial distribution for the Clayton copula is given in Example 3.3 [McNeil and Nešlehová \(2009\)](#) while those of the Joe, AMH, Gumbel and Frank copulas are given in [Hofert et al. \(2012\)](#). In general, one can use numerical inversion algorithms for computing the inverse of the radial distribution, however it will lead to spurious numerical errors. Other algorithms exist when the generator is known to be the Laplace-Stieltjes transform, denoted as  $\mathcal{LS}$ , of some positive random variables (see [Frees and Valdez \(1998\)](#); [Marshall and Olkin \(1988\)](#)). This positive random variable is often referenced as the frailty distribution. In this framework, Archimedean copulas allow for the stochastic representation

$$\mathbf{U} = (\varphi^{\leftarrow}(E_1/V), \dots, \varphi^{\leftarrow}(E_d/V)) \sim C, \quad (\text{E.8})$$

with  $V \sim F = \mathcal{LS}^{-1}[\varphi^{\leftarrow}]$  the frailty and  $E_1, \dots, E_d$  are distributed i.i.d. according to a standard exponential and independent of  $V$ . Algorithm 9 presents a procedure for generating a multivariate sample from an Archimedean copula where the frailty distribution is known. The algorithm takes as an input the length of the sample  $n$ , as well as the parameter of the copula  $\theta$ . The output is a  $d$ -variate sample from the desired copula model, denoted  $\{(u_0^{(1)}, \dots, u_{d-1}^{(1)}), \dots, (u_0^{(n)}, \dots, u_{d-1}^{(n)})\}$ .

## E.4 Case study: modeling pairwise dependence between spatial maximas with missing data

---

**Algorithm 9** Sampling from Archimedean copula using frailty distribution

---

- 1: **Data:** sample's length  $n$ .
  - 2: Parameter of the copula  $\theta$ .
  - 3: **Result:** multivariate sample from the desired copula model.
  - 4: **procedure** SAMPLING( $n, \theta$ )
  - 5:   Sample  $V \sim F = \mathcal{LS}^{-1}[\varphi^{\leftarrow}]$ .
  - 6:   Sample  $E_1, \dots, E_d \stackrel{i.i.d.}{\sim} \mathcal{E}(1)$ , independent of  $V$ .
  - 7:   Return  $\mathbf{U} = (\varphi^{\leftarrow}(E_1/V), \dots, \varphi^{\leftarrow}(E_d/V))$ .
- 

In this framework, we define `_frailty_sim` method defined inside the **Archimedean** class which performs Algorithm 9. Then, each Archimedean copula is defined by the generator  $\varphi$ , it's inverse  $\varphi^{\leftarrow}$  and the frailty distribution denoted as  $\mathcal{LS}^{-1}[\varphi^{\leftarrow}]$  as long as we know the frailty. This is the case for Joe, Clayton, AMH or Frank.

For the extreme value case, algorithms have been proposed, as in Stephenson (2003) (see Algorithms 2.1 and 2.2), who proposes sampling methods for the Gumbel and the asymmetric logistic model. These algorithms are implemented in the `clayton` package. Note that these algorithms are model-specific, thus the `sample_unimargin` method is exceptionally located in the corresponding child of the multivariate **Extreme** class. Another procedure designed by Dombry et al. (2016) to sample from multivariate extreme value models using extremal functions (see Algorithm 2 of the reference cited above) is also of prime interest. For the implemented models using this algorithm, namely **Hüsler-Reiss**, **tEV**, **Bilogistic** and **Dirichlet** models, a method called `_rextfunc` is located inside each classes which allows to generate an observation from the according law of the extremal function.

Samples from the Gaussian and Student copula are directly given by Algorithm 5.9 and 5.10 respectively of Alexander J. McNeil (2005). As each algorithm is model specific, the `sample_unimargin` method is located inside the **Gaussian** and **Student** classes.

We present how to construct a multivariate Archimedean copula and to generate random vectors from this model. Introducing the parameters of the copula, we appeal the following lines to construct our copula object:

```
>>> d, theta, n_sample = 3, 2.0, 1024
>>> copula = archimedean.Clayton(theta=theta, n_sample=n_sample,
>>> dim=d)
```

We now call the `sample_unimargin` method to obtain randomly generated vectors.

```
sample = copula.sample_unimargin()
```

We thus represent in three dimensions in Figure E.3.

## E.4 Case study: modeling pairwise dependence between spatial maximas with missing data

We now proceed to a case study where we use our Python package to assess, under a finite sample framework, the asymptotic properties of an estimator of the  $\lambda$ -madogram when data

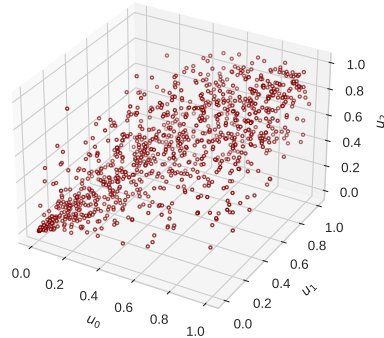


Fig. E.3 Scatterplot of a sample from a Clayton copula ( $\theta = 2.0$ ).

are completely missing at random (MCAR). This case study comes from numerical results of Chapter 2. The  $\lambda$ -madogram belongs to a family of estimators, namely the madogram, which is of prime interest in environmental sciences, as it is designed to model pairwise dependence between maxima in space. See, for example, [Bador et al. \(2015\)](#); [Bernard et al. \(2013\)](#); [Saunders et al. \(2021\)](#), where the madogram was used as a dissimilarity measure to perform clustering. In several fields, such as econometrics ([Wooldridge \(2007\)](#)) or survey theory ([Boistard et al. \(2016\)](#)), the MCAR hypothesis appears to be a strong hypothesis, but in environmental research, this hypothesis is more realistic, as the missingness of one observation is usually due to instruments, communication, and processing errors that may be reasonably supposed to be independent of the quantity of interest. In Section E.4.1, we define objects and properties of interest, while in Section E.4.2, we describe a detailed tutorial in Python and with the `clayton` package to compare the asymptotic variance with an empirical counterpart of the  $\lambda$ -madogram with  $\lambda = 0.5$ .

### E.4.1 Background

It was emphasized that the possible dependence between maxima can be described with the extreme value copula. This function is completely characterized by the Pickands dependence function (see Equation (E.4)), which is equivalent to the  $\lambda$ -madogram introduced by [Naveau et al. \(2009\)](#) and defined as

$$\nu(\lambda) = \mathbb{E} \left[ \left| F_0(X_0)^{1/\lambda} - F_1(X_1)^{1/(1-\lambda)} \right| \right], \quad (\text{E.9})$$

with  $\lambda \in (0, 1)$ , and if  $\lambda = 0$  and  $0 < u < 1$ , then  $u^{1/\lambda} = 0$  by convention. The  $\lambda$ -madogram took its inspiration from the extensively used geostatistics tool, the variogram (see Chapter 1.3 of [Gaetan and Guyon \(2008\)](#) for a definition and some classical properties). The  $\lambda$ -madogram can be interpreted as the  $L_1$ -distance between the uniform margins elevated to the inverse of the corresponding weights  $\lambda$  and  $1 - \lambda$ . This quantity describes the dependence structure between extremes by its relation with the Pickands dependence function. If we suppose that  $C$

## E.4 Case study: modeling pairwise dependence between spatial maximas with missing data

---

is an extreme value copula as in Equation (E.3), we have

$$\mathcal{A}(\lambda) = \frac{\nu(\lambda) + c(\lambda)}{1 - \nu(\lambda) - c(\lambda)}, \quad (\text{E.10})$$

with  $c(\lambda) = 2^{-1}(\lambda/(1 - \lambda) + (1 - \lambda)/\lambda)$  (see Proposition 3 of Marcon et al. (2017) for details).

We consider independent and identically distributed i.i.d. copies  $\mathbf{X}_1, \dots, \mathbf{X}_n$  of  $\mathbf{X}$ . In the presence of missing data, we do not observe a complete vector  $\mathbf{X}_i$  for  $i \in \{1, \dots, n\}$ . We introduce  $\mathbf{I}_i \in \{0, 1\}^2$  which satisfies,  $\forall j \in \{0, 1\}$ ,  $I_{i,j} = 0$  if  $X_{i,j}$  is not observed. To formalize incomplete observations, we introduce the incomplete vector  $\tilde{\mathbf{X}}_i$  with values in the product space  $\otimes_{j=1}^2 (\mathbb{R} \cup \{\text{NA}\})$  such as

$$\tilde{X}_{i,j} = X_{i,j}I_{i,j} + \text{NA}(1 - I_{i,j}), \quad i \in \{1, \dots, n\}, j \in \{0, \dots, d - 1\}.$$

We thus suppose that we observe a 4-tuple such as

$$(\mathbf{I}_i, \tilde{\mathbf{X}}_i), \quad i \in \{1, \dots, n\}, \quad (\text{E.11})$$

i.e. at each  $i \in \{1, \dots, n\}$ , several entries may be missing. We also suppose that for all  $i \in \{1, \dots, n\}$ ,  $\mathbf{I}_i$  are i.i.d copies from  $\mathbf{I} = (I_0, I_1)$  where  $I_j$  is distributed according to a Bernoulli random variable  $\mathcal{B}(p_j)$  with  $p_j = \mathbb{P}(I_j = 1)$  for  $j \in \{0, 1\}$ . We denote by  $p$  the probability of observing completely a realization from  $\mathbf{X}$ , that is  $p = \mathbb{P}(I_0 = 1, I_1 = 1)$ . In Chapter 2, hybrid and corrected estimators, respectively denoted as  $\hat{\nu}_n^{\mathcal{H}}$  and  $\hat{\nu}_n^{\mathcal{H}*}$ , are proposed to estimate nonparametrically the  $\lambda$ -madogram in presence of missing data completely at random. Furthermore, a closed expression of their asymptotic variances for  $\lambda \in ]0, 1[$  is also given. This result is summarized in the following proposition.

**Proposition E.4.1** (Proposition 2.2.1 in Chapter 2). *Let  $(\mathbf{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$  be a sample given by (E.11). For  $\lambda \in ]0, 1[$ , if  $C$  is an extreme value copula in (E.3) with Pickands dependence function  $\mathcal{A}$ , we have*

$$\begin{aligned} \mathcal{E}_n^{\mathcal{H}}(\lambda) &\triangleq \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}}(\lambda) - \nu(\lambda) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left( 0, \mathcal{S}^{\mathcal{H}}(p_0, p_1, p, \lambda) \right), \\ \mathcal{E}_n^{\mathcal{H}*}(\lambda) &\triangleq \sqrt{n} \left( \hat{\nu}_n^{\mathcal{H}*}(\lambda) - \nu(\lambda) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left( 0, \mathcal{S}^{\mathcal{H}*}(p_0, p_1, p, \lambda) \right), \end{aligned}$$

where  $\nu(\lambda)$  is defined in (E.9),  $\mathcal{S}^{\mathcal{H}}(p_0, p_1, p, \lambda)$  and  $\mathcal{S}^{\mathcal{H}*}(p_0, p_1, p, \lambda)$  are the asymptotic variances of the random variables where the closed expression is given in 2.2.1 in Chapter 2.

### E.4.2 Numerical results

Benefiting from generating data with `clayton` we are thus able, with Monte Carlo simulation, to assess theoretical results given by Proposition E.4.1 in a finite sample setting. For that purpose, we implement a `MonteCarlo` class (in `monte_carlo.py` file) which contains some methods to perform some Monte Carlo iterations for a given extreme value copula. Now, we set up parameters to sample our bivariate dataset. For this subsection, we choose the asymmetric negative logistic model (see Section E.7 for a definition) with parameters  $\theta = 10$ ,  $\psi_1 = 0.1$ ,  $\psi_2 = 1.0$ .

```
>>> n_sample = 1024
>>> theta, psi1, psi2 = 10, 0.1, 1.0
```

We choose the standard normal and exponential as margins. To simulate this sample, the following lines should be typed:

```
>>> copula = evd.AsyNegLog(theta=theta, psi1=psi1, psi2=psi2,
>>> n_sample=n_sample)
>>> sample = copula.sample(inv_cdf=[norm.ppf, expon.ppf])
```

The  $1024 \times 2$  array `sample` contains 1024 realization of the **asymmetric negative logistic** model where the first column is distributed according to a standard normal random variable and the second column as a standard exponential. This distribution is depicted in Figure E.4. To obtain it, one needs the following lines of command:

```
>>> fig, ax = plt.subplots()
>>> ax.scatter(sample[:,0], sample[:,1],
>>> edgecolors='#6F6F6F', color='#C5C5C5', s=5)
>>> ax.set_xlabel(r'$x_0$')
>>> ax.set_ylabel(r'$x_1$')
>>> plt.show()
```

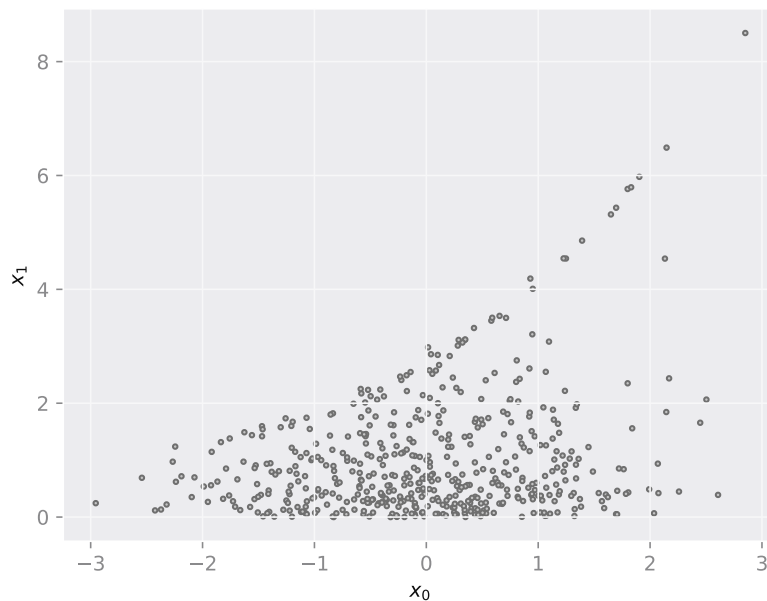


Fig. E.4 A realization from the **asymmetric negative logistic model** with Gaussian and Exponent margins and parameters  $\theta = 10, \psi_1 = 0.1, \psi_2 = 1.0$  and sample's length  $n = 1024$ .

Before going into further details, we will present the missing mechanism. Let  $V_0$  and  $V_1$  be random variables uniformly distributed under the  $]0, 1[$  segment with copula  $C_{(V_0, V_1)}$ . We set  $I_0 = \mathbb{1}_{\{V_0 \leq p_0\}}$  and  $I_1 = \mathbb{1}_{\{V_1 \leq p_1\}}$ . It is thus immediate that  $I_0 \sim \mathcal{B}(p_0)$  and  $I_1 \sim \mathcal{B}(p_1)$  and  $p \triangleq \mathbb{P}\{I_0 = 1, I_1 = 1\} = C_{(V_0, V_1)}(p_0, p_1)$ . For our illustration, we will take  $C_{(V_0, V_1)}$  as a Joe

## E.4 Case study: modeling pairwise dependence between spatial maximas with missing data

copula with parameter  $\theta = 2.0$  (we refer to Section E.6 for a definition of this copula). For this copula, it is more likely to observe a realization  $v_0 \geq 0.8$  from  $V_0$  if  $v_1 \geq 0.8$  from  $V_1$ . If we observe  $v_1 < 0.8$ , the realization  $v_0$  is close to being independent of  $v_1$ . In climate studies, extreme events could damage the recording instrument in the surrounding regions where they occur, thus the missingness of one variable may depend on others. We initialize the copula  $C_{(V_0, V_1)}$  with the following line:

```
>>> copula_miss = archimedean.Joe(theta=2.0, n_sample=n_sample)
```

For a given  $\lambda \in ]0, 1[$ , we now want to estimate a  $\lambda$ -madogram with a sample from the asymmetric negative logistic model, where some observations are missing due to the missing mechanism described above. We will repeat this step several times to compute an empirical counterpart of the asymptotic variance. The `MonteCarlo` object has been designed for this purpose: we specify the number of iterations  $n_{iter}$  (take  $n_{iter} = 1024$ ), the chosen extreme value copula (asymmetric negative logistic model), the missing mechanism (described by  $C_{(V_0, V_1)}$  and  $p_0 = p_1 = 0.9$ ), and  $\lambda$  (noted `w`). We can write the following lines of code:

```
>>> u = np.array([0.9, 0.9])
>>> n_iter, P, w = 256, [[u[0], copula_miss._c(
>>>     u)], [copula_miss._c(u), u[1]]], np.array([0.5, 0.5])
>>> monte = monte_carlo.MonteCarlo(n_iter=n_iter, n_sample=n_sample,
>>>     copula=copula, copula_miss=copula_miss, weight=w, matp=P)
```

The `MonteCarlo` object is thus initialized with all parameters needed. We may use the `simu` method to generate a `DataFrame` (a `Pandas` object) composed out 1024 rows and 3 columns. Each row contains an estimate of the  $\lambda$ -madogram,  $\hat{\nu}_n^{\mathcal{H}^*}$  in Proposition E.4.1 (FMado), the sample length  $n$  (`n`) and the normalized estimation error (scaled). We thus call the `simu` method.

```
>>> df_wmado = monte.simu(inv_cdf = [norm.ppf, expon.ppf], corr = True)
>>> print(df_wmado.head())
```

	FMado	n	scaled
0	0.147648	512.0	-0.140255
1	0.160095	512.0	-0.141402
2	0.159303	512.0	0.123480
3	0.156156	512.0	0.052269
4	0.152242	512.0	-0.036300

Where `corr=True` specifies that we compute the corrected estimator,  $\hat{\nu}_n^{\mathcal{H}^*}$  in Proposition E.4.1. Now, using the `var_mado` method defined inside in the `Extreme` class, we obtain the asymptotic variance for the given model and parameters from the missing mechanism. We obtain this quantity as follows

```
>>> var_mado = copula.var_mado(w, p=copula_miss._c(u), P=P, corr=True)
>>> print(var_mado)
0.015417245591834503
```

We propose here to check numerically the asymptotic normality with variance  $\mathcal{S}^{\mathcal{H}^*}$  of the normalized estimation error of the corrected estimator. We have all data in hand and the asymptotic variance was computed by lines above. We thus write:



```
>>> fig, ax = plt.subplots()
>>> sigma = np.sqrt(var_mado)
>>> x = np.linspace(min(df_wmado['scaled']), max(df_wmado['scaled']), 1000)
>>> gauss = gauss_function(x, 0, sigma)
>>> sns.displot(data=df_wmado, x="scaled", color='#C5C5C5', kind='hist',
>>>             stat='density', common_norm=False, alpha=0.5, fill=True,
>>>             linewidth=1.5, bins = 32)
>>> plt.plot(x,gauss, color = 'darkblue')
>>> plt.show()
```

Result of these lines might be found in Figure E.5.

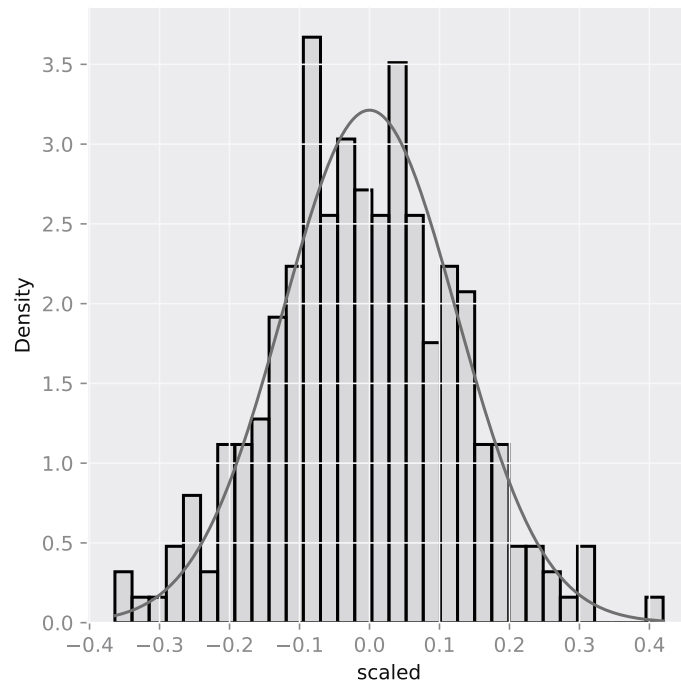


Fig. E.5 Histogram of  $\mathcal{E}_n^{\mathcal{H}^*}$  in Proposition E.4.1 where the solid line is the density of a centered Gaussian with variance  $\mathcal{S}^{\mathcal{H}^*}$ .

## E.5 Discussion

### E.5.1 Comparison of `clayton` with R packages

To compare `clayton` to existing packages in R, we consider the copula package (Kojadinovic and Yan (2010)) and `mev` (Belzile et al. (2022)) for sampling from Archimedean and multivariate extreme value distributions, respectively. To run the experiment, we use two computer clusters. The first cluster consists of five nodes, each with two 18-core Xeon Gold 3.1 GHz processors and 192 GB of memory, with 2933 MHz per socket. The second cluster has two CPU sockets, each containing a Xeon Platinum 8268 2.90 GHz processor with 24 cores. These configurations provide a significant amount of computational power and are well-suited for handling complex,

data-intensive tasks. We use the first cluster to install the `copula` package and sample from the **Clayton**, **Frank**, and **Joe** models. We consider an increasing dimension  $d \in \{50, 100, \dots, 1600\}$  for a fixed sample size of  $n = 1000$ . We use the second cluster to install the `mev` package and call some of its methods to sample from the **Husler Reiss**, **Logistic**, and **TEV** distributions. Sampling from the latter is fast, but sampling from the two others is time consuming. Therefore, we only consider dimensions  $d \in \{25, 50, \dots, 250\}$  for a fixed sample size of  $n = 1000$ .

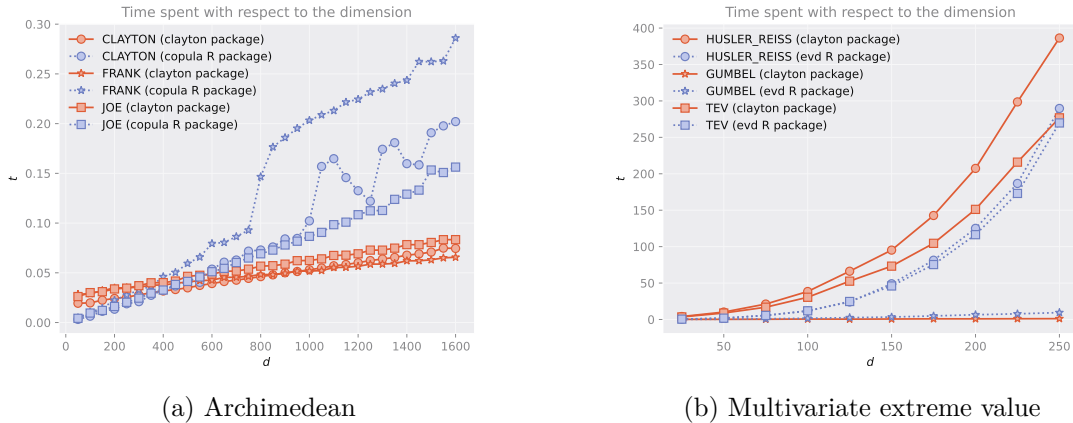


Fig. E.6 Comparison results. Time spent (in seconds) to sample from the corresponding models with respect to the dimension  $d$ . The left panel shows the results for sampling from **Clayton**, **Frank** and **Joe** using `clayton` in Python and `copula` in R. The right panel shows the results for sampling from **HuslerReiss**, **Logistic** and **TEV** by `clayton` in Python and `mev` in R. In both cases, 1000 vectors are generated for each model.

The figure shows the results of a comparison between the `clayton` and `copula` packages in R, and the `mev` package in Python. The comparison shows that the `clayton` package is more efficient at sampling from **Clayton**, **Frank** and **Joe** copulae than the `copula` package. The gap in efficiency may be due to the choice of algorithms used in the `clayton` package, which uses frailty distributions. The time required for sampling increases linearly with the dimension for the `clayton` package, but shows a more erratic behavior for the `copula` package.

When comparing the `clayton` and `mev` packages, it is clear that `mev` is more efficient. This is likely due to the fact that `mev` is written in C++, while `clayton` is written in Python. The `mev` package uses the algorithm of [Stephenson \(2003\)](#) to sample from the Logistic distribution, which is more efficient than the algorithm using frailty distributions used in `clayton`.

## E.5.2 Conclusion

This paper presents the construction and some implementations of the Python package `clayton` for random copula sampling. This is a seminal work in the field of software implementation of copula modeling in Python and there is much more potential for growth. It is hoped that the potential diffusion of the software through those who need it may bring further implementations for multivariate modeling with copulas under Python. For example, choosing a copula to fit the data is an important but difficult problem. A robust approach to estimating copulas has been investigated recently by [Alquier et al. \(2020\)](#) using Maximum Mean Discrepancy. In relation to

our example, semiparametric estimation of copulas with missing data could be of great interest, as proposed by [Hamori et al. \(2019\)](#).

Additionally, implementation of the algorithm proposed by [McNeil and Nešlehová \(2009\)](#) for generating random vectors for Archimedean copulas has been tackled, but as expected, numerical inversion gives spurious results, especially when the parameter  $\theta$  and the dimension  $d$  are high. Furthermore, as the support of the radial distribution is contained in the real line, numerical inversion leads to increased computational time. Further investigation is needed in order to generate random vectors from classical Archimedean models using the radial distribution.

A direction of improvement for the `clayton` package is dependence modeling with Vine copulas, which have recently been a tool of high interest in the machine learning community (see, e.g., [Carrera et al. \(2016\)](#); [Gonçalves et al. \(2016\)](#); [Lopez-Paz et al. \(2013\)](#); [Veeramachaneni et al. \(2015\)](#) or [Sun et al. \(2019\)](#)). This highlights the need for dependence modeling with copulas in Python, as a significant part of the machine learning community uses this language. In relation to this paper, Vine copulas may be useful for modeling dependencies between extreme events, as suggested by [Nolde and Wadsworth \(2021\)](#); [Simpson et al. \(2021\)](#). Furthermore, other copula models could be implemented to model further dependencies. These implementations will expand the scope of dependence modeling with Python and provide high-quality, usable tools for anyone who needs them.

## E.6 Bivariate Archimedean models

Table E.1 Bivariate Archimedean models in clayton module.

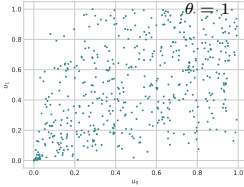
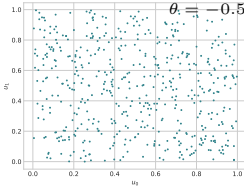
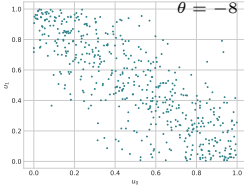
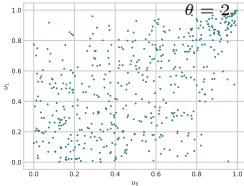
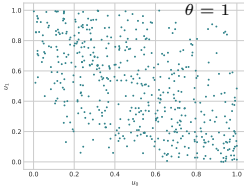
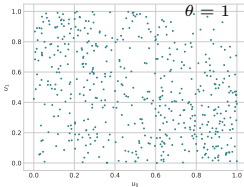
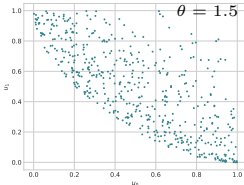
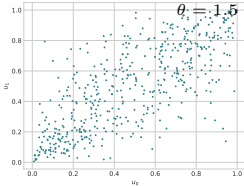
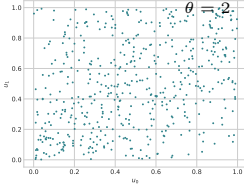
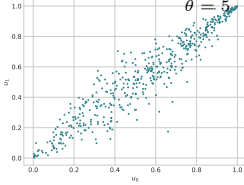
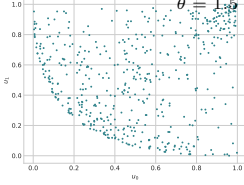
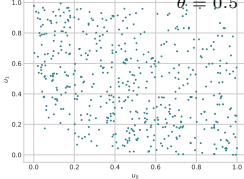
Name	$\varphi(t)$	Constraints	Figure
<b>Clayton</b>	$\frac{1}{\theta}(t^{-\theta} - 1)$	$\theta \in [-1, \infty) \setminus \{0\}$	
<b>AMH</b>	$\ln\left(\frac{1-\theta(1-t)}{t}\right)$	$\theta \in [-1, 1)$	
<b>Frank</b>	$-\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$\theta \in \mathbb{R} \setminus \{0\}$	
<b>Joe</b>	$-\ln(1 - (1-t)^\theta)$	$\theta \in [1, \infty)$	
<b>Nelsen n°9</b>	$\ln(1 - \theta \ln(t))$	$\theta \in ]0, 1]$	
<b>Nelsen n°10</b>	$\ln(2t^{-\theta} - 1)$	$\theta \in ]0, 1]$	

Table E.2 Bivariate archimedean models in **clayton** module.

Name	$\varphi(t)$	Constraints	Figure
Nelsen n°11	$\ln(2 - t^\theta)$	$\theta \in ]0, 0.5]$	
Nelsen n°12	$(\frac{1}{t} - 1)^\theta$	$\theta \in ]0, \infty) \setminus \{0\}$	
Nelsen n°13	$(1 - \ln(t))^\theta - 1$	$\theta \in ]0, \infty[$	
Nelsen n°14	$(t^{-\frac{1}{\theta}} - 1)^\theta$	$\theta \in ]1, \infty)$	
Nelsen n°15	$(1 - t^{\frac{1}{\theta}})^\theta$	$[1, \infty)$	
Nelsen n°22	$\arcsin(1 - t^\theta)$	$\theta \in [0, 1]$	

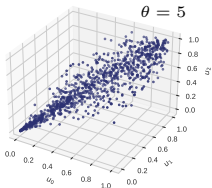
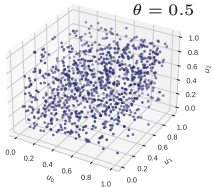
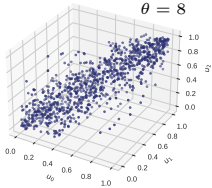
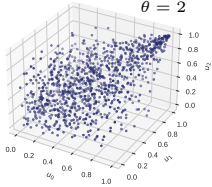
## E.7 Implemented bivariate extreme models

Table E.3 Bivariate extreme models in **clayton** module.

Name	$\mathcal{A}(w)$	Constraints	Figure
<b>Gumbel</b>	$[w^{1/\theta} + (1-w)^{1/\theta}]^\theta$	$\theta \in ]0, 1]$	
<b>Galambos</b>	$1 - [w^{-\theta} + (1-w)^{-\theta}]^{-\frac{1}{\theta}}$	$\theta \in [0, \infty)$	
<b>Asy. log.</b>	$(1 - \psi_1)w + (1 - \psi_2)(1 - w) + [(\psi_1 w)^\theta + (\psi_2(1 - w))^\theta]^{\frac{1}{\theta}}$	$\theta \in [1, \infty)$ $\psi_1, \psi_2 \in (0, 1]$	
<b>Asy. neg. log.</b>	$1 - [(\psi_1 w)^{-\theta} + (\psi_2(1 - w))^{-\theta}]^{-\frac{1}{\theta}}$	$\theta \in [0, \infty)$ $\psi_1, \psi_2 \in (0, 1]$	
<b>Asy. mixed</b>	$1 - (\theta + \psi_1)w + \theta w^2 + \psi_1 w^3$	$\theta \geq 0,$ $\theta + 3\psi_1 \geq 0,$ $\theta + \psi_1 \leq 1,$ $\theta + 2\psi_1 \leq 1.$	
<b>Husler Reiss</b>	$(1 - w)\Phi(\theta + \frac{1}{2\theta} \ln(\frac{1-w}{w})) + w\Phi(\theta + \frac{1}{2\theta} \ln(\frac{w}{1-w}))$	$\theta \in (0, \infty)$	
<b>t-EV</b>	$w t_{\psi_1+1}(z_w) + (1-w) t_{\psi_1+1}(z_{1-w})$	$\theta \in (-1, 1)$	

## E.8 Multivariate Archimedean copulae

Table E.4 Multivariate archimedean models in **clayton** module.

Name	$\varphi(t)$	Constraints	Figure
<b>Clayton</b>	$\frac{1}{\theta}(t^{-\theta} - 1)$	$\theta \in [0, \infty) \setminus \{0\}$	
<b>AMH</b>	$\ln\left(\frac{1-\theta(1-t)}{t}\right)$	$\theta \in [-1, 1)$	
<b>Frank</b>	$-\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right)$	$\theta \in \mathbb{R} \setminus \{0\}$	
<b>Joe</b>	$-\ln(1 - (1-t)^\theta)$	$\theta \in [1, \infty)$	

---

## E.9 Multivariate extreme models

Before giving the main details, we introduce some notations. Let  $B$  be the set of all nonempty subsets of  $\{1, \dots, d\}$  and  $B_1 = \{b \in B, |b| = 1\}$ , where  $|b|$  denotes the number of elements in the set  $b$ . We note by  $B_{(j)} = \{b \in B, j \in b\}$ . For  $d = 3$ , the Pickands is expressed as

$$\begin{aligned} \mathcal{A}(\mathbf{w}) = & \alpha_1 w_1 + \psi_1 w_2 + \phi_1 w_3 + \left( (\alpha_2 w_1)^{\theta_1} + (\psi_2 w_2)^{\theta_1} \right)^{1/\theta_1} + \left( (\alpha_3 w_2)^{\theta_2} + (\phi_2 w_3)^{\theta_2} \right)^{1/\theta_2} \\ & + \left( (\psi_3 w_2)^{\theta_3} + (\phi_3 w_3)^{\theta_3} \right)^{1/\theta_3} + \left( (\alpha_4 w_1)^{\theta_4} + (\psi_4 w_2)^{\theta_4} + (\phi_4 w_3)^{\theta_4} \right)^{1/\theta_4}, \end{aligned}$$

where  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_4)$ ,  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_4)$ ,  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_4)$  are all elements of  $\Delta^3$ . We take  $\boldsymbol{\alpha} = (0.4, 0.3, 0.1, 0.2)$ ,  $\boldsymbol{\psi} = (0.1, 0.2, 0.4, 0.3)$ ,  $\boldsymbol{\phi} = (0.6, 0.1, 0.1, 0.2)$  and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_4) = (0.6, 0.5, 0.8, 0.3)$  as the dependence parameter.

The Dirichlet model is a mixture of  $m$  Dirichlet densities, that is

$$h(\mathbf{w}) = \sum_{k=1}^m \theta_k \frac{\Gamma(\sum_{j=1}^d \sigma_{kj})}{\prod_{j=1}^d \Gamma(\sigma_{kj})} \prod_{j=1}^d w_j^{\sigma_{kj}-1},$$

with  $\sum_{k=1}^m \theta_k = 1$ ,  $\sigma_{kj} > 0$  for  $k \in \{1, \dots, m\}$  and  $j \in \{1, \dots, d\}$ . Let  $\mathcal{D} \in [0, \infty)^{(d-1) \times (d-1)}$  denotes the space of symmetric strictly conditionnaly negative definite matrices that is

$$\begin{aligned} \mathcal{D}_k = \left\{ \Gamma \in [0, \infty)^{k \times k} : a^\top \Gamma a < 0 \text{ for all } a \in \mathbb{R}^k \setminus \{\mathbf{0}\} \text{ with } \sum_{j=1}^{d-1} a_j = 0, \right. \\ \left. \Gamma_{ii} = 0, \Gamma_{ij} = \Gamma_{ji}, \quad 1 \leq i, j \leq k \right\}. \end{aligned}$$

For any  $2 \leq k \leq d$  consider  $m' = (m_1, \dots, m_k)$  with  $1 \leq m_1 < \dots < m_k \leq d$  define

$$\Sigma_m^{(k)} = 2 \left( \Gamma_{m_i m_k} + \Gamma_{m_j m_k} - \Gamma_{m_i m_j} \right)_{m_i m_j \neq m_k} \in [0, \infty)^{(d-1) \times (d-1)}.$$

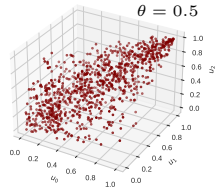
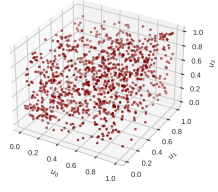
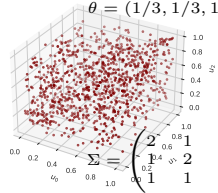
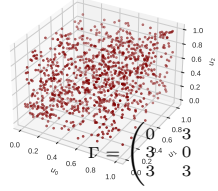
Furthermore, note  $S(\cdot | \Sigma_m^{(k)})$  denote the survival function of a normal random vector with mean vector  $\mathbf{0}$  and covariance matrix  $\Sigma_m^{(k)}$ . We now define :

$$h_{km}(\mathbf{y}) = \int_{y_k}^{\infty} S \left( (y_i - z + 2\Gamma_{m_i m_k})_{i=1}^{k-1} | \Gamma_{km} \right) e^{-z} dz$$

for  $2 \leq k \leq d$ . We denote by  $\Sigma^{(k)}$  the summation over all  $k$ -vectors  $m = (m_1, \dots, m_k)$  with  $1 \leq m_1 < \dots < m_k \leq d$ .



Table E.5 Multivariate extreme models in **clayton** module.

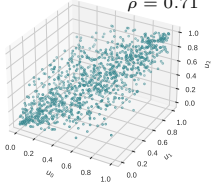
Name	$\mathcal{A}(\mathbf{w})$	Constraints	Figure
Logistic	$\left(\sum_{j=1}^d w_j^{\frac{1}{\theta}}\right)^\theta$	$\theta \in ]0, 1]$	
Asy. Log.	$\sum_{b \in B} \left(\sum_{j \in b} (\psi_{j,b} w_j)^{\frac{1}{\theta_b}}\right)^{\theta_b}$	$\theta_b \in ]0, 1] \forall b \in B \setminus B_1,$ $\psi_{j,b} \in [0, 1] \forall b \in B \forall j \in b,$ $\sum_{b \in B(j)} \psi_{j,b} = 1, j \in \llbracket d-1 \rrbracket,$ $\theta_b = 1 \forall b \in B \setminus B_1 \implies$ $\psi_{j,b} = 0 \forall j \in b.$	
Dirichlet	Not specified	$\sum_{k=1}^m \theta_k = 1,$ $\sigma_{kj} > 0, k \in \{1, \dots, m\},$ $j \in \{1, \dots, d\}$	 $\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$
Hüsler Reiss	$\sum_{k=1}^d (-1)^{k+1} \times$ $\Sigma^{(k)} h_{km}(u_{m_1}, \dots, u_{m_k})$	$\Gamma \in \mathcal{D}_d$	 $\Gamma = \begin{pmatrix} 0 & 3 & 3 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{pmatrix}$

## E.10 Multivariate elliptical dependencies

Let  $\mathbf{X} \sim \mathbf{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$  be an elliptical distributed random vector with cumulative distribution  $F$  and marginal  $F_0, \dots, F_{d-1}$ . Then, the copula  $C$  of  $F$  is called an elliptical copula. We denote by  $\phi$  the standard normal distribution function and  $\phi_\Sigma$  the joint distribution function of  $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$ , where  $\mathbf{0}$  is the  $d$ -dimensional vector composed out of 0. In the same way, we note  $t_\theta$  the distribution function of a standard univariate  $t$  distribution and by  $t_{\theta, \Sigma}$  the joint distribution function of the vector  $\mathbf{X} \sim t_d(\theta, \mathbf{0}, \Sigma)$ . A  $d$  squared matrix  $\Sigma$  is said to be positively semi definite if for all  $u \in \mathbb{R}^d$  we have :

$$u^\top \Sigma u \geq 0$$

Table E.6 Multivariate elliptical models in **clayton** module.

Name	C	Constraints	Figure
<b>Gaussian</b>	$\phi_\Sigma(\phi^\leftarrow(u_0), \dots, \phi^\leftarrow(u_{d-1}))$	$\Sigma$ PSD	
<b>Student</b>	$t_{\theta, \Sigma}(t_\theta^\leftarrow(u_0), \dots, t_\theta^\leftarrow(u_{d-1}))$	$\theta > 0, \Sigma$ PSD	