Limit Laws under Long-Range Dependence and Heavy Tails and Clustering-Based Estimation for Multivariate Extremes

by

He Tang

(Under the Direction of Shuyang Bai)

Abstract

We consider a class of stationary processes that exhibit both long-range dependence and heavy tails. While separate limit theorems for the partial sums and the maxima of such processes have recently been established—featuring novel limiting objects—this work develops the joint sum-and-max limit theorems for this class. In the finite-variance case, the limiting behavior consists of two independent components: a fractional Brownian motion (from the sum) and a long-range dependent random sup measure (from the maximum). In contrast, in the infinite-variance regime, the limit comprises two dependent components: a stable Lévy process and a random sup measure. Their dependence is characterized through the local time and range of a stable subordinator. To establish this result, we also prove a joint convergence theorem for the local time and range of subordinators, which may be of independent interest.

In parallel, we investigate the estimation of multivariate extreme value models with a discrete spectral measure using spherical clustering techniques. The primary methodological contribution is a new order selection criterion—selecting the number of spectral atoms (or clusters)—based on an augmented silhouette width index. This criterion introduces a penalty term that discourages overly small clusters and insufficient separation between cluster centers. We prove that the method consistently recovers the true number of atoms in the spectral measure, enabling consistent estimation of the order of max-linear factor models, which lack standard likelihood-based tools for model selection. Our second contribution is a large deviation analysis that quantifies the convergence quality of clustering-based estimation of spectral measures. Finally, we demonstrate how the discrete spectral measure estimation can be translated into parameter estimation for heavy-tailed factor models, supported by simulations and real-world data examples that illustrate both order selection and model inference.

Index words: Asymptotic Dependence, Infinitely Divisible Processes, Long-range Ddependence, Stable Subordinator, Weak Convergence, Multivariate Extremes, Spherical Clustering, Linear Factor Models, Penalized Silhouette Method, Order Selection.

Limit Laws under Long-Range Dependence and Heavy Tails and Clustering-Based Estimation for Multivariate Extremes

by

He Tang

B.S., Nankai University, 2018, CHINA
M.S., George Washington University, 2020

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Doctor of Philosophy

Athens, Georgia

2025

Limit Laws under Long-Range Dependence and Heavy Tails and Clustering-Based Estimation for Multivariate Extremes

by

He Tang

Major Professor:	Shuyang Bai
Committee:	Ting Zhang T.N. Sriram Yuan Ke

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2025

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Shuyang Bai, for his unwavering support, insightful guidance, and constant encouragement throughout my doctoral studies. His mentorship has been instrumental in shaping my research direction and academic development, and I am truly fortunate to have had the opportunity to learn from him.

I would also like to thank my colleague and collaborator, Dr. Shiyuan Deng, whose substantial contributions greatly enhanced this work.

I would also like to sincerely thank the members of my dissertation committee—Dr. Ting Zhang, Dr. Tharuvai N Sriram, and Dr. Yuan Ke—for their valuable feedback, constructive suggestions, and generous investment of time and expertise. Their thoughtful comments have significantly enriched the quality of this work.

I am deeply grateful for the stimulating academic environment provided by the Department of Statistics at University of Georgia, and for all the opportunities that have supported my growth as a researcher.

Contents

Acknowledgments **iv**

1 Introduction **1**

1.1 Background 1

1.2 Literature Review 23

2 The Setup **29**

2.1 A Class of Long-Range Dependent Processes Generated by Conservative Flows 29

2.2 Heavy-tailed Factor Models 36

3 Main Results **39**

3.1 Theory 39

3.2 Methods 48

3.3 Simulations 50

4 Applications **54**

4.1 Air Pollution Data 55

4.2 River Discharge Data 57

5 Conclusion and Future Work **60**

Appendices **62**

A Proofs **62**

A.1 Finite Variance Case 62

A.2 Infinite Variance Case 64

A.3 Dependence of Limits in the Infinite Variance Case 73

A.4 Joint Convergence on Subordinators 75

A.5 Criterion for Weak Convergence 77

A.6 Consistency of Order Selection via Penalized Silhouette 81

Bibliography **89**

Chapter 1

Introduction

1.1 Background

1.1.1 Extreme Value Theory and Domains of Attraction

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of stationary random variables, where $\mathbb{N} = \{1, 2, \ldots\}$. We are interested in the asymptotic behavior of the partial maximum

$$M_n := \max(X_1, \ldots, X_n),$$

as $n \to \infty$. In particular, we seek non-degenerate limit distributions for properly normalized maxima, that is, sequences of constants $\{a_n > 0\}$ and $\{b_n\} \subset \mathbb{R}$ such that

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G, \tag{1.1}$$

for some non-degenerate distribution function G. Here, $\xrightarrow{d}$ denotes convergence in distribution.

In the case where $\{X_n\}$ is an independent and identically distributed (i.i.d.) sequence, the possible limit distributions are well understood and are characterized in the classical theory of extreme value distributions, as summarized in de Haan and Ferreira, 2006. These limiting distributions, referred to as *extreme value distributions*, are characterized up to affine transformations of the form $x \mapsto ax + b$, where $a > 0$ and $b \in \mathbb{R}$. The class of all extreme value distributions can be denoted as

$$G_\gamma(ax + b), \quad a > 0,\ b \in \mathbb{R},$$

where G_γ is defined by

$$G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \quad \text{for } 1+\gamma x > 0,$$

with $\gamma \in \mathbb{R}$. For $\gamma = 0$, this expression is interpreted as the limit:

$$G_0(x) = \exp(-e^{-x}), \quad x \in \mathbb{R},$$

which is known as the *Gumbel distribution* or *double exponential distribution*.

Depending on the value of the shape parameter γ, three canonical types of extreme value distributions arise:

- **Fréchet Type ($\gamma > 0$)**: By taking $G_\gamma((x-1)/\gamma)$, and letting $\alpha = 1/\gamma > 0$, we obtain the Fréchet distribution:

$$\Phi_\alpha(x) := \begin{cases} 0, & x \le 0, \\ \exp(-x^{-\alpha}), & x > 0. \end{cases}$$

- **Gumbel Type ($\gamma = 0$)**: As already noted, the limiting distribution becomes

$$G_0(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}.$$

- **Reverse Weibull Type ($\gamma < 0$)**: By transforming $G_\gamma(-(1+x)/\gamma)$ and letting $\alpha = -1/\gamma > 0$, we obtain the reverse Weibull distribution:

$$\Psi_\alpha(x) := \begin{cases} \exp(-(-x)^\alpha), & x < 0, \\ 0, & x \ge 0. \end{cases}$$

Let F denote the common distribution function of the stationary sequence $\{X_n\}_{n\in\mathbb{N}}$. We say that F belongs to the *domain of attraction* of an extreme value distribution G_γ, written $F \in \mathcal{D}(G_\gamma)$, if the convergence (1.1) holds with G replaced by G_γ. The characterization of the domain of attraction depends on the value of the shape parameter γ, and the corresponding conditions on the distribution F are as follows:

- **Fréchet Case ($\gamma > 0$)**: In this case, the right endpoint of the support,

$$x^* := \sup\{x \in \mathbb{R} : F(x) < 1\},$$

is infinite. The distribution F belongs to the domain of attraction of the Fréchet distribution if

$$\lim_{t \to \infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}, \quad \text{for all } x > 0.$$

This condition implies that the tail $1 - F$ is *regularly varying* at infinity with index $-1/\gamma$. For a comprehensive treatment of regularly varying functions, see Bingham et al., 1989.

- **Reverse Weibull Case ($\gamma < 0$)**: Here, the right endpoint x^* is finite. The distribution F belongs to the domain of attraction of the reverse Weibull distribution if

$$\lim_{t \downarrow 0} \frac{1 - F(x^* - tx)}{1 - F(x^* - t)} = x^{-1/\gamma}, \quad \text{for all } x > 0.$$

  This condition describes the behavior of the distribution near its finite upper bound.

- **Gumbel Case ($\gamma = 0$)**: In this case, the right endpoint x^* may be either finite or infinite. The distribution F belongs to the domain of attraction of the Gumbel distribution if there exists a positive function f such that

$$\lim_{t \uparrow x^*} \frac{1 - F(t + xf(t))}{1 - F(t)} = e^{-x}, \quad \text{for all } x \in \mathbb{R}.$$

  The auxiliary function f typically satisfies certain regularity conditions, such as slowly varying behavior near x^*, and plays a role similar to a local scale.

These conditions provide the necessary and sufficient characterizations for a distribution F to belong to one of the three max-stable domains of attraction.

The convergence in distribution of the normalized partial maxima, as in equation (1.1), also holds in the presence of dependence among the random variables $\{X_n\}$. However, to ensure the validity of such a limit, certain dependence conditions must be imposed on the sequence. These conditions are formulated to control the extent of clustering of extreme values due to dependence. A comprehensive treatment of such dependence structures is given in Leadbetter et al., 1983, where conditions such as the $D(u_n)$ and $D'(u_n)$ conditions are introduced. These conditions ensure that the extremes of the sequence behave asymptotically as if they were nearly independent, thereby preserving the classical extreme value limit laws in the dependent setting.

1.1.2 Limit Theorem and Domains of Attraction

The limiting theory for the partial sums $S_n := X_1 + \cdots + X_n$ is a classical counterpart to the extreme value theory for the partial maxima M_n. Specifically, we seek sequences of normalizing constants $\{c_n\}$ with $c_n > 0$, and $\{d_n\} \subset \mathbb{R}$, such that

$$\frac{S_n - d_n}{c_n} \xrightarrow{d} U,$$

where U is a non-degenerate distribution. In the case where $\{X_n\}$ are i.i.d., the fundamental results are summarized in Feller, 1971, Chapter XVII.

A central result is that the only possible non-degenerate limit distributions for normalized sums of i.i.d. random variables are the *stable distributions*. These distributions, which generalize the classical central limit theorem, are characterized through their characteristic functions.

Let $p, q \in [0, 1]$ with $p + q = 1$, which parameterize the skewness of the distribution, where p reflects the contribution from the right tail and q from the left. Then the characteristic function of a stable random variable X, with stability index $\alpha \in (0, 2]$, scale parameter $\sigma > 0$, and location parameter $\mu \in \mathbb{R}$, is given by:

$$\varphi(t; \alpha, p) = \exp\left\{-\sigma^\alpha |t|^\alpha \left[1 - i(p-q)\omega(t,\alpha)\,\mathrm{sgn}(t)\right] + i\mu t\right\},$$

where

$$\omega(t, \alpha) = \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right), & \alpha \neq 1, \\ -\frac{2}{\pi}\log|t|, & \alpha = 1; \end{cases}$$

and $\mathrm{sgn}(t)$ is the sign function defined as follows:

$$\mathrm{sgn}(t) := \begin{cases} -1 & \text{if } t < 0, \\ 0 & \text{if } t = 0, \\ 1 & \text{if } t > 0. \end{cases}$$

This form includes the Gaussian distribution as a special case when $\alpha = 2$, where the characteristic function reduces to that of a normal distribution with finite variance.

Without loss of generality, we say that a distribution function F belongs to the *domain of attraction* of a stable law, denoted $F \in \mathcal{D}(\alpha, p)$, if the limiting distribution U has the characteristic function $\varphi(t; \alpha, p)$ as above.

Sufficient conditions for $F \in \mathcal{D}(\alpha, p)$ depend on the tail behavior of F. When $0 < \alpha < 2$, a necessary and sufficient condition is that there exists a regularly varying function $L(x)$ at infinity, such that:

$$1-F(x) \sim p \cdot \frac{2-\alpha}{\alpha} L(x), \quad \text{and} \quad F(-x) \sim q \cdot \frac{2-\alpha}{\alpha} L(x), \quad \text{as } x \to \infty.$$

This condition describes the power-law behavior of the tails, with balance given by the parameters p and q.

When $\alpha = 2$, which corresponds to the normal distribution, the requirement becomes the function

$$\mu(x) := \int_{-x}^{x} y^2 \, dF(y)$$

is a *slowly varying* function at infinity (see Bingham et al., 1989). This condition ensures that the second moment grows sufficiently slowly and is compatible with convergence to a normal distribution with infinite support but finite variance.

There are also numerous results concerning the convergence of the partial sums S_n when the sequence $\{X_n\}$ exhibits dependence. In particular, generalizations of the classical central limit theorem and stable limit theory have been developed under various forms of weak dependence. For example *mixing conditions*, which quantify the asymptotic independence between distant observations in the sequence (see Ibragimov and Linnik, 1971) and *martingale difference sequences*, where the conditional expectation of each term given the past is zero (see Hall and Heyde, 1980). These results underscore that, even in the presence of dependence, the asymptotic behavior of sums of random variables can often be characterized using classical or generalized probabilistic limits—provided the dependence structure satisfies suitable regularity conditions.

1.1.3 Random Sup Measures and Convergence in Distribution

Writing $S_n(t) = S_{\lfloor nt \rfloor}$ and $M_n(t) = M_{\lfloor nt \rfloor}$, where $\lfloor x \rfloor$ denotes the greatest integer not exceeding x, the processes $(S_n(t))_{t \geq 0}$ and $(M_n(t))_{t \geq 0}$ can be viewed as stochastic processes indexed by continuous time. It is then natural to investigate their functional convergence in an appropriate function space. Typically, this is the Skorokhod space $D[0, \infty)$ endowed with the J_1 topology; see Billingsley, 1999 for foundational results on this convergence.

An alternative and fruitful perspective on the partial maximum process $(M_n(t))_{t \geq 0}$ is to consider it as a *random sup measure*. For a subset $B \subset [0, \infty)$, define

$$M_n(B) := \max_{k/n \in B} X_k, \tag{1.2}$$

so that $M_n(t)$ corresponds to the special case $B = (0, t]$. The framework of random sup measures, which was systematically developed in O'Brien et al., 1990, provides a robust setting for analyzing random sup measures.

A *sup measure* is a map $m : \mathcal{G} \to \bar{\mathbb{R}} := [-\infty, \infty]$, where $\mathcal{G}$ denotes the collection of open subsets of $\mathbb{R}$, satisfying

$$m\left(\bigcup_{\alpha} G_\alpha\right) = \bigvee_{\alpha} m(G_\alpha)$$

for any collection $\{G_\alpha\} \subset \mathcal{G}$. This parallels the definition of a measure, but with the additive operation replaced by the supremum.

A natural way to construct a sup measure is through the *sup integral*: given a function $f : \mathbb{R} \to \bar{\mathbb{R}}$, define

$$m(G) := \bigvee_{t \in G} f(t), \quad G \in \mathcal{G},$$

and denote this operation by $m = i^\vee f$. Although different functions f may yield the same sup measure, there exists a canonical representative—the *sup derivative*, defined by

$$d^\vee m(t) := \bigwedge_{G \ni t} m(G), \quad t \in \mathbb{R}.$$

This sup derivative is upper semi-continuous (usc), and every sup measure m satisfies the identity $m = i^\vee d^\vee m$, establishing a bijection between the space of sup measures and the space of usc functions. The sup measure can be extended to all subsets of $\mathbb{R}$ via

$$m(B) := \bigvee_{t \in B} d^\vee m(t),$$

ensuring consistency with the original definition on open sets.

To study convergence, we define the *sup vague topology*. A sequence of sup measures (m_n) is said to converge sup vaguely to m if:

$$\limsup_n m_n(K) \le m(K) \quad \text{for all compact } K \subset \mathbb{R},$$

and

$$\liminf_n m_n(G) \ge m(G) \quad \text{for all open } G \subset \mathbb{R}.$$

This topology is metrizable and compact, with convergence characterized by the values of sup measures on *continuity sets*. A set $A \subset \mathbb{R}$ is a continuity set for m if $m(\mathrm{int}\, A) = m(\mathrm{clos}\, A)$, i.e., no mass is concentrated on the boundary.

Under this framework, *sup vague convergence* $m_n \to m$ is equivalent to

$$m_n(A) \to m(A)$$

for all bounded continuity sets A. In many cases, it suffices to verify this convergence for bounded open intervals.

The Borel σ-field on the space of sup measures generated by the sup vague topology is the smallest σ-field that renders the maps $m \mapsto m(A)$ measurable for all open sets A, or equivalently for compact sets, compact intervals, or bounded open intervals. Consequently, a mapping $M : (\Omega, \mathcal{F}, \mathbb{P}) \to$ sup measures is a *random sup measure* if and only if $M(A)$ is a random variable for every set A in any of these collections. The process $M_n(B)$ defined in (1.2) thus forms a random sup measure. For a random sup measure M, its probability law is completely determined by its finite-dimensional distributions $(M(A))_{A \in \mathcal{A}}$, for any of the aforementioned collections $\mathcal{A}$.

We now address *convergence in distribution*. For a random sup measure M, define the set of continuity intervals by

$$\mathcal{I}(M) := \{I \in \mathcal{I} : M(I) = M(\mathrm{clos}\, I) \text{ w.p.1}\},$$

where $\mathcal{I}$ denotes the collection of non-empty bounded open intervals. A sequence $\{M_n\}$ of random sup measures converges in distribution to a limit M, written $M_n \xrightarrow{d} M$, if and only if the finite-dimensional distributions of $(M_n(I))_{I \in \mathcal{I}(M)}$ converge weakly to those of $(M(I))_{I \in \mathcal{I}(M)}$.

1.1.4 Infinitely Divisible Processes

A fundamental building block of an infinitely divisible stochastic process is a one-dimensional infinitely divisible random variable. The most powerful analytical tool for studying infinitely divisible random variables is the *Lévy–Khintchine representation*.

Specifically, a real-valued random variable X is said to be infinitely divisible if and only if there exists a uniquely determined characteristic triplet (σ^2, ν, b), where: $\sigma^2 \geq 0$ is the Gaussian component, ν is a Borel measure on $\mathbb{R} \setminus \{0\}$, called the *Lévy measure*, satisfying

$$\nu(\{0\}) = 0, \quad \int_{\mathbb{R}} (1 \wedge x^2)\, \nu(dx) < \infty,$$

$b \in \mathbb{R}$ is a drift term, such that the characteristic function of X is given by for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[e^{i\theta X}\right] = \exp\left\{-\frac{1}{2}\sigma^2\theta^2 + ib\theta + \int_{\mathbb{R}} \left(e^{i\theta x} - 1 - i\theta[\![x]\!]\right)\nu(dx)\right\},$$

where $[\![x]\!]$ denotes the truncated function defined by

$$[\![x]\!] := \begin{cases} x, & \text{if } |x| \leq 1, \\ \operatorname{sgn}(x), & \text{if } |x| > 1. \end{cases}$$

This decomposition separates the contribution of small jumps (through the Gaussian term and the drift) from that of large jumps (through the Lévy measure or the Poissonian term). For comprehensive coverage of infinitely divisible distributions and their properties, see Sato, 1999.

A stochastic process $(X(t))_{t\in T}$, is said to be *infinitely divisible* if and only if there exists a uniquely determined triple (Σ, ν, b), such that for every $\theta \in \mathbb{R}^T$, the joint characteristic function of $X(t)$ has the Lévy–Khintchine representation:

$$\begin{aligned} &\mathbb{E}\exp\left(i\sum_{t\in T}\theta(t)X(t)\right) \\ &= \exp\left\{-\frac{1}{2}Q(\theta) + \int_{\mathbb{R}^T}\left(e^{i\langle\theta,x\rangle} - 1 - i\langle\theta,[\![x]\!]\rangle\right)\nu(dx) + i\langle\theta,b\rangle\right\}, \end{aligned}$$

where $\langle\theta, x\rangle = \sum_{t\in T}\theta(t)x(t)$ denotes the inner product in $\mathbb{R}^T$ and the truncation function $[\![x(t)]\!]$ is defined coordinate-wise. The elements of the triple are defined as follows:

- $Q(\theta) = \sum_{s,t\in T}\Sigma(s,t)\theta(s)\theta(t)$ is a quadratic form associated with a nonnegative definite function $\Sigma : T \times T \to \mathbb{R}$, representing the covariance structure of the Gaussian component;
- ν is a Lévy measure on $\mathbb{R}^T$, governing the jump behavior of the process;
- $b \in \mathbb{R}^T$ is a shift vector;

This formulation generalizes the classical Lévy–Khintchine representation of one-dimensional infinitely divisible random variables to the case of stochastic processes indexed by an arbitrary set T. For more details, see Samorodnitsky, 2016.

An *infinitely divisible random measure* is one of the most fundamental objects in the study of infinitely divisible stochastic processes. It frequently serves as a building block in constructing more complex infinitely divisible processes. To define an infinitely divisible random measure, we begin with a measurable space $(S, \mathcal{S})$. The construction requires the following three components:

- A σ-finite measure γ on S;
- A measure ν on $S \times (\mathbb{R} \setminus \{0\})$, such that the measure

  $$m_0(B) := \iint_{B\times(\mathbb{R}\setminus\{0\})} [\![x]\!]^2 \, \nu(ds, dx) < \infty, \quad \text{for all } B \in \mathcal{S},$$

  is σ-finite;
- A σ-finite signed measure β on S.

Let $\mathcal{S}_0 \subseteq \mathcal{S}$ be the collection of sets $B \in \mathcal{S}$ such that

$$m(B) := \gamma(B) + \|\beta\|(B) + m_0(B) < \infty,$$

where $\|\beta\|$ denotes the total variation measure of β. For sets $B_1, B_2 \in \mathcal{S}_0$, define the covariance function

$$\Sigma(B_1, B_2) := \gamma(B_1 \cap B_2).$$

Next, define a measurable map $\Phi : S \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}^{\mathcal{S}_0}$ by

$$\Phi(s, x)(B) := x \cdot \mathbf{1}_{\{s\in B\}}, \quad \text{for } B \in \mathcal{S}_0.$$

This map induces a Lévy measure μ on $\mathbb{R}^{\mathcal{S}_0}$ by pushforward:

$$\mu := \nu \circ \Phi^{-1}.$$

Finally, define the shift function $b \in \mathbb{R}^{\mathcal{S}_0}$ by

$$b(B) := \beta(B), \quad \text{for } B \in \mathcal{S}_0.$$

The infinitely divisible stochastic process $M = \{M(B) : B \in \mathcal{S}_0\}$ with the triple (Σ, μ, b) characteristic is called an *infinitely divisible random measure* on $(S, \mathcal{S})$ with Gaussian variance measure γ, Lévy measure ν, and shift measure β.

The fundamental properties of such a random measure are summarized in the following:

1. For every $B \in \mathcal{S}_0$, the random variable $M(B)$ is infinitely divisible with characteristic triplet $(\sigma_B^2, \nu_B, \beta_B)$, where:

- $\sigma_B^2 = \gamma(B)$,
- $\mu_B(\cdot) = \nu(B \times \cdot)$,
- $\beta_B = \beta(B)$.

2. The random measure M is *independently scattered*: the random variables $M(B_1), \ldots, M(B_d)$ are independent for every finite collection of disjoint sets $B_1, \ldots, B_d \in \mathcal{S}_0$.

3. The random measure M is *σ-additive* almost surely: for any countable collection of disjoint sets $\{B_j\} \subset \mathcal{S}_0$ such that $\bigcup_j B_j \in \mathcal{S}_0$, we have

$$M\left(\bigcup_j B_j\right) = \sum_j M(B_j) \quad \text{almost surely.}$$

(Note that the exceptional null set in this identity may depend on the specific choice of the sets $\{B_j\}$.)

An infinitely divisible random measure admits a disintegrated representation, which offers a more intuitive and localized understanding of its structure. Specifically, we consider an infinitely divisible stochastic process $M = \{M(B) : B \in \mathcal{S}_0\}$ to be an infinitely divisible random measure characterized by:

- A control measure m on $(S, \mathcal{S})$,
- A local Gaussian variance function $(\sigma^2(s), s \in S)$,
- Local Lévy measures $\rho(s, \cdot)$ on $\mathbb{R} \setminus \{0\}$, for each $s \in S$,
- A local shift function $(b(s), s \in S)$.

The random measure M is an independently scattered, σ-additive random set function such that for each $B \in \mathcal{S}_0$, the random variable $M(B)$ is infinitely divisible with characteristic triplet given by:

$$\sigma_B^2 = \int_B \sigma^2(s)\, m(ds), \quad \mu_B(\cdot) = \int_B \rho(s, \cdot)\, m(ds), \quad b_B = \int_B b(s)\, m(ds).$$

Infinitely divisible random measures are fundamental because they serve as building blocks for a wide variety of infinitely divisible stochastic processes. In particular, many such processes can be constructed through *integration* of deterministic functions with respect to infinitely divisible random measures.

Specifically, a large class of infinitely divisible processes $\{X(t) : t \in T\}$ can be represented in the form:

$$X(t) = \int_S f(t, s)\, M(ds), \quad t \in T, \tag{1.3}$$

where M is an infinitely divisible random measure on a measurable space $(S, \mathcal{S})$, and $\{f(t, \cdot) : t \in T\}$ is a family of measurable, nonrandom functions.

To define the stochastic integral, we begin with the case of simple functions. Suppose $f : S \to \mathbb{R}$ is a simple function of the form:

$$f(s) = \sum_{j=1}^{k} f_j\, \mathbf{1}_{B_j}(s), \quad s \in S,$$

where $f_1, \ldots, f_k \in \mathbb{R}$ and $B_1, \ldots, B_k \in \mathcal{S}_0$ are disjoint sets such that the random measure M is defined on each B_j. Then the stochastic integral of f with respect to M is defined by:

$$I(f) := \int_S f(s)\, M(ds) := \sum_{j=1}^{k} f_j M(B_j).$$

For more general (non-simple) functions f that are *integrable* with respect to the random measure M, the integral is defined as the limit:

$$I(f) := \lim_{n \to \infty} I(f_n),$$

where $\{f_n\}$ is a sequence of simple functions such that $f_n(s) \to f(s)$ for m-almost every $s \in S$, and the convergence holds in probability.

The stochastic integral $I(f)$ with respect to an infinitely divisible random measure M satisfies the following properties:

1. The integral $I(f)$ is an infinitely divisible random variable. Its characteristic function is given by:

$$\mathbb{E}\left[e^{i\theta I(f)}\right] = \exp\Bigg\{ -\frac{\theta^2}{2} \int_S f(s)^2\, \sigma^2(s)\, m(ds) + i\theta \int_S f(s) b(s)\, m(ds) + \int_S \int_{\mathbb{R}\setminus\{0\}} \left(e^{i\theta f(s) x} - 1 - i\theta f(s) [\![x]\!]\right) \rho(s, dx)\, m(ds) \Bigg\},$$

where $\sigma^2(s)$ is the local Gaussian variance, $\rho(s, dx)$ is the local Lévy measure, $b(s)$ is the local shift function. In particular, the characteristic triplet $(\sigma^2(f), \mu_f, b(f))$ of the random variable $I(f)$ is given as follows:

- The Gaussian variance:

$$\sigma^2(f) = \int_S f(s)^2 \, \sigma^2(s) \, m(ds);$$

- The Lévy measure:

$$\mu_f = \nu_f \circ T_f^{-1},$$

where the measure ν_f on $S \times (\mathbb{R} \setminus \{0\})$ is defined by

$$\nu_f(A) = \nu(A \cap \{(s,x) : f(s) \neq 0\}), \quad A \text{ measurable},$$

ν is the Lévy measure of the random measure M, and $T_f : S \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$ is the measurable transformation $T_f(s,x) = f(s)x$;

- The shift parameter:

$$\begin{aligned} b(f) = &\int_S f(s) b(s) \, m(ds) \\ &+ \int_S \int_{\mathbb{R}\setminus\{0\}} \left([\![f(s)x]\!] - f(s) [\![x]\!] \right) \, \rho(s, dx) \, m(ds). \end{aligned}$$

2. The integral operator $I(f)$ is linear. That is, if f and g are integrable functions and $a, b \in \mathbb{R}$, then:

$$I(af + bg) = aI(f) + bI(g) \quad \text{almost surely}.$$

Not only is the integral of an integrable function with respect to an infinitely divisible random measure M an infinitely divisible random variable, but also the family of such integrals defines an infinitely divisible stochastic process. Let M be an infinitely divisible random measure on a measurable space $(S, \mathcal{S})$, with control measure m, local Gaussian variance $\sigma^2(s)$, local Lévy measures $\rho(s, \cdot)$, and local shift function $b(s)$, for $s \in S$. Suppose $f(t, \cdot)$ is integrable for each $t \in T$. Then the stochastic process defined in (1.3) is infinitely divisible. Moreover, it possesses a characteristic triple (Σ_X, ν_X, b_X), where:

- The Gaussian covariance function is given by

$$\Sigma_X(t_1, t_2) = \int_S f(t_1, s) f(t_2, s) \, \sigma^2(s) \, m(ds), \quad t_1, t_2 \in T;$$

- The Lévy measure ν_X is the image measure:

$$\nu_X = \nu \circ H^{-1},$$

where ν is the Lévy measure of the random measure M, and $H : S \times \mathbb{R} \to \mathbb{R}^T$ is defined by

$$H(s, x) := \Big(xf(t, s)\Big)_{t \in T};$$

- The drift function is

$$\begin{aligned} b_X(t) = &\int_S f(t, s) b(s)\, m(ds) \\ &+ \int_S \int_{\mathbb{R}} \left([\![f(t, s)x]\!] - f(t, s)[\![x]\!] \right) \rho(s, dx)\, m(ds), \quad t \in T. \end{aligned}$$

An infinitely divisible stochastic process without a Gaussian component consists solely of a compound Poisson component. Such processes often admit explicit series representations involving the arrival times of a standard Poisson process, as well as an independent sequence of i.i.d. random variables, see Samorodnitsky, 2016.

Let M be a symmetric infinitely divisible random measure on a measurable space $(S, \mathcal{S})$, without a Gaussian component, and with control measure m. Let γ be a probability measure on S that is equivalent to m. Then

$$r(s) := \frac{dm}{d\gamma}(s), \quad s \in S,$$

is strictly positive m-almost everywhere. For each $s \in S$, define the corresponding Lévy measure:

$$\rho_r(s, \cdot) := r(s)\rho(s, \cdot),$$

where $\rho(s, \cdot)$ is the Lévy measure of M. Then each $\rho_r(s, \cdot)$ is a symmetric one-dimensional Lévy measure. Define the generalized inverse of the tail function of $\rho_r(s, \cdot)$ as

$$G(x, s) := \inf \left\{ y > 0 : \rho_r(s, (y, \infty)) \leq \frac{x}{2} \right\}, \quad x > 0.$$

Now, let $\{\varepsilon_n\}_{n \geq 1}$ be a sequence of i.i.d. Rademacher random variables (i.e., taking values ± 1 with equal probability), $\{Y_n\}_{n \geq 1}$ a sequence of i.i.d. S-valued random variables with common distribution γ. Let $\{\Gamma_n\}_{n \geq 1}$ denote the ordered points of a unit-rate Poisson process on $(0, \infty)$. Note that

$$\Gamma_n = e_1 + \cdots + e_n, \quad n = 1, 2, \ldots,$$

where $\{e_j\}_{j\geq 1}$ are i.i.d. standard exponential random variables. Assume all three sequences are independent. Define the stochastic process $\{Y(t)\}_{t\in T}$ by

$$Y(t) := \sum_{n=1}^{\infty} \varepsilon_n \, G(\Gamma_n, Y_n) \, f(t, Y_n), \quad t \in T. \tag{1.4}$$

Then $\{Y(t)\}_{t\in T}$ is a well-defined stochastic process, and it has the same finite-dimensional distributions as the process $\{X(t)\}_{t\in T}$ defined in (1.3).

1.1.5 Multivariate Extreme Value Theory

Consider a sample of d-dimensional random vectors,

$$\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d}), \quad i = 1, \ldots, n,$$

which are i.i.d. with common distribution function F on $\mathbb{R}^d$. Define the component-wise maximum as

$$\mathbf{M}_n = \bigvee_{i=1}^{n} \mathbf{X}_i,$$

where the jth component of $\mathbf{M}_n$ is given by

$$M_{n,j} = \max_{1\leq i\leq n} X_{i,j}, \quad j = 1, \ldots, d.$$

The distribution function of $\mathbf{M}_n$ is given by

$$\mathbb{P}(\mathbf{M}_n \leq \mathbf{x}) = \mathbb{P}(\mathbf{X}_1 \leq \mathbf{x}, \ldots, \mathbf{X}_n \leq \mathbf{x}) = F^n(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

where the inequality is interpreted component-wise.

The multivariate domain-of-attraction problem is to find sequences of vectors $\mathbf{a}_n > \mathbf{0}$ and $\mathbf{b}_n \in \mathbb{R}^d$ such that the normalized maxima $\mathbf{a}_n^{-1}(\mathbf{M}_n - \mathbf{b}_n)$ converge in distribution to a non-degenerate random vector, that is,

$$F^n(\mathbf{a}_n \mathbf{x} + \mathbf{b}_n) \xrightarrow{d} G(\mathbf{x}), \quad \text{as } n \to \infty, \tag{1.5}$$

for some distribution function G on $\mathbb{R}^d$ with non-degenerate marginals. When such sequences $\mathbf{a}_n, \mathbf{b}_n$ exist, we say that F is in the *(max-)domain of attraction* of G, denoted $F \in \mathcal{D}(G)$. The limiting distribution G is then called a *multivariate extreme value distribution*.

A crucial observation follows from the requirement of marginal convergence. Let F_j and G_j denote the jth marginal distribution functions of F and

G, respectively. Since weak convergence of random vectors implies convergence of each marginal, we must have

$$F_j^n(a_{n,j}x_j + b_{n,j}) \xrightarrow{d} G_j(x_j), \quad \text{as } n \to \infty,$$

for each $j = 1, \ldots, d$, where $a_{n,j}$ and $b_{n,j}$ are the jth components of $\mathbf{a}_n$ and $\mathbf{b}_n$, respectively. Therefore, each marginal distribution F_j must lie in the domain of attraction of a univariate extreme value distribution G_j.

A d-variate distribution function G is called *max-stable* if for every positive integer k, there exist vectors $\boldsymbol{\alpha}_k > \mathbf{0}$ and $\boldsymbol{\beta}_k \in \mathbb{R}^d$ such that the following identity in distribution holds:

$$G^k(\boldsymbol{\alpha}_k \mathbf{x} + \boldsymbol{\beta}_k) = G(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

Intuitively, this means that if $\mathbf{Y}, \mathbf{Y}_1, \ldots, \mathbf{Y}_k$ are i.i.d. random vectors with distribution function G, then there exist normalization vectors $\boldsymbol{\alpha}_k > 0$, $\boldsymbol{\beta}_k$ such that

$$\boldsymbol{\alpha}_k^{-1}\left(\bigvee_{i=1}^{k} \mathbf{Y}_i - \boldsymbol{\beta}_k\right) \stackrel{d}{=} \mathbf{Y}, \quad \text{for all } k \geq 1.$$

That is, the distribution of the component-wise maximum of k i.i.d. copies of $\mathbf{Y}$, properly normalized, is again G. It follows immediately from this definition that a max-stable distribution is in its own domain of attraction. In particular, it is an extreme value distribution function. Conversely, as discussed previously, any extreme value distribution arises as the limit of normalized component-wise maxima and therefore satisfies the max-stability property. Hence, the class of max-stable distributions coincides with the class of multivariate extreme value distributions.

An important property of max-stable distributions is that $G^{1/k}$ is again a distribution function for every integer $k \geq 1$. This implies that G is *max-infinitely divisible*. According to Balkema and Resnick, 1977, any max-infinitely divisible distribution function G can be written in the form

$$G(\mathbf{x}) = \exp\left\{-\Lambda\left([-\infty, \infty)^d \setminus [-\infty, \mathbf{x}]\right)\right\},$$

where $[-\infty, \mathbf{x}] = [-\infty, x_1] \times \cdots \times [-\infty, x_d]$, for some measure Λ on $\mathbb{R}^d$, called the *exponent measure* of G. To facilitate the study of the dependence structure of max-stable distributions, it is customary to standardize the margins. That is, we transform each component so that all margins follow a common distribution. While the specific form of the marginal distribution is not essential, a particularly convenient choice is the α-Fréchet distribution.

For simplicity, we restrict our attention to the nonnegative orthant $[0,\infty)^d$. Let Λ denote the exponent measure of a multivariate α-Fréchet distribution. It satisfies the homogeneity property

$$\Lambda(c\cdot A) = c^{-\alpha}\Lambda(A), \quad \text{for all } c>0 \text{ and Borel sets } A\subseteq[0,\infty)^d.$$

This scaling property implies that Λ admits a polar decomposition into radial and angular components. We follow the formulation of Beirlant et al., 2006, Section 8.2.5, which allows the use of different norms for the radial and angular parts. Let $\|\cdot\|_{(r)}$ and $\|\cdot\|_{(s)}$ denote two arbitrary norms on $\mathbb{R}^d$. We define the following one-to-one transformation from $\mathbb{E}_d := [0,\infty)^d\setminus\{\mathbf{0}\}$ to $(0,\infty)\times\mathbb{S}_+^{d-1}$ by

$$\mathbf{x}\mapsto(r,\mathbf{w}) := \left(\|\mathbf{x}\|_{(r)}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{(s)}}\right),$$

where the positive part of the unit sphere is defined as

$$\mathbb{S}_+^{d-1} = \{\mathbf{x}\in[0,\infty)^d:\ \|\mathbf{x}\|_{(s)}=1\}, \tag{1.6}$$

We continue to denote by Λ the pushforward measure of the original exponent measure under this mapping. The polar decomposition of Λ then takes the form:

$$\Lambda(dr,d\mathbf{w}) = c_{(r)}\alpha r^{-\alpha-1}dr\times H(d\mathbf{w}), \tag{1.7}$$

where H is a probability measure on $\mathbb{S}_+^{d-1}$, referred to as the *spectral measure*, and $c_{(r)}$ is the normalizing constant given by

$$c_{(r)} = \Lambda(\{\mathbf{x}\in[0,\infty)^d:\ \|\mathbf{x}\|_{(r)}\geq 1\}). \tag{1.8}$$

The spectral measure H encodes the angular structure of extremes and governs the dependence among components of $\mathbf{X}$. As a consequence of the marginal standardization (to α-Fréchet margins), the following moment constraint must hold:

$$\int_{\mathbb{S}_+^{d-1}}\left(\frac{w_j}{\|\mathbf{w}\|_{(r)}}\right)^{\alpha}H(d\mathbf{w}) = \frac{1}{c_{(r)}}, \quad \text{for } j=1,\dots,d. \tag{1.9}$$

In practice, common choices for the norms include the p-norm, for $p\in(0,\infty)$, $\|\mathbf{x}\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$, and the supremum norm, $\|\mathbf{x}\|_\infty = \max_{j=1}^d|x_j|$.

Let us write the distribution function of the componentwise maximum as $F^n = [1-n^{-1}\{n(1-F\}]^n$ and use the fact that $(1-n^{-1}x_n)^n\to e^{-x}\in[0,1]$ if and only if $x_n\to x\in[0,\infty]$ as $n\to\infty$, we deduce that the convergence in (1.5) holds if and only if

$$\lim_{n\to\infty} n\left\{1 - F(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n)\right\} = -\log G(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \tag{1.10}$$

with the usual convention that $-\log(0) = \infty$. (1.10) admits a natural interpretation in terms of exponent measures. Recall that a multivariate extreme value distribution G is characterized by an exponent measure Λ such that

$$\Lambda([\mathbf{q}, \infty) \setminus [\mathbf{q}, \mathbf{x}]) = -\log G(\mathbf{x}), \quad \mathbf{x} \geq \mathbf{q},$$

where $\mathbf{q}$ denotes the lower endpoint of G. In particular, for the jth marginal distribution G_j, we let q_j denote its lower endpoint. Observe that $-\log G(\mathbf{x})$ is finite if $\mathbf{x} > \mathbf{q}$, and infinite otherwise. Consequently, the exponent measure Λ assigns finite mass only to Borel sets in $[\mathbf{q}, \infty)$ that are bounded away from $\mathbf{q}$. Now define a sequence of measures (Λ_n) on $[\mathbf{q}, \infty) \setminus \{\mathbf{q}\}$ via:

$$\Lambda_n(B) := n\,\mathbb{P}(\mathbf{X}_{1,n} \in B),$$

where the normalized and thresholded variable $\mathbf{X}_{i,n}$ is defined as

$$\mathbf{X}_{i,n} := \left(\frac{\mathbf{X}_i - \mathbf{b}_n}{\mathbf{a}_n}\right) \vee \mathbf{q}.$$

From this, it follows that

$$\Lambda_n([\mathbf{q}, \infty) \setminus [\mathbf{q}, \mathbf{x}]) = n\left\{1 - F(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n))\right\}, \quad \mathbf{x} \in [\mathbf{q}, \infty).$$

Hence, (1.10) may be rewritten as:

$$\Lambda_n([\mathbf{q}, \infty) \setminus [\mathbf{q}, \mathbf{x}]) \to \Lambda([\mathbf{q}, \infty) \setminus [\mathbf{q}, \mathbf{x}]), \quad \mathbf{x} \in [\mathbf{q}, \infty),$$

which suggests weak convergence of the measures Λ_n to Λ on $[\mathbf{q}, \infty) \setminus \{\mathbf{q}\}$. More precisely, this convergence is *vague convergence*, denoted:

$$\Lambda_n \xrightarrow{v} \Lambda \quad \text{on } [\mathbf{q}, \infty) \setminus \{\mathbf{q}\}.$$

By definition, $\Lambda_n \xrightarrow{v} \Lambda$ if

$$\Lambda_n(B) \to \Lambda(B), \quad \text{for all Borel sets } B \subset [\mathbf{q}, \infty) \setminus \{\mathbf{q}\}$$

with compact closure and such that the boundary ∂B satisfies $\Lambda(\partial B) = 0$. Note that such Borel sets B have compact closure in $[\mathbf{q}, \infty) \setminus \{\mathbf{q}\}$ if and only if there exists $\mathbf{x} > \mathbf{q}$ such that $B \subset [\mathbf{q}, \infty) \setminus [\mathbf{q}, \mathbf{x}]$. For more information on vague convergence of measures, see Resnick, 1987. In summary, (1.5) is equivalent to vague convergence of the rescaled excess measure Λ_n toward the exponent measure Λ.

Let G be a multivariate α-Fréchet distribution. Then the convergence in (1.5) is equivalent to the vague convergence of measures

$$u\mathbb{P}\left(u^{-1/\alpha}\mathbf{X} \in \cdot\right) \xrightarrow{v} \Lambda(\cdot), \quad \text{as } u \to \infty, \tag{1.11}$$

on the punctured nonnegative orthant $\mathbb{E}_d := [0, \infty)^d \setminus \{\mathbf{0}\}$. By the polar decomposition of the exponent measure Λ, the convergence in (1.11) implies a weak limit theorem for the angular component of $\mathbf{X}$. Specifically, we have the following convergence in distribution on $\mathbb{S}_+^{d-1}$ as the threshold $u \to \infty$:

$$\mathbb{P}\left(\frac{\mathbf{X}}{\|\mathbf{X}\|_{(s)}} \in A, \middle|, \|\mathbf{X}\|_{(r)} \geq u\right) \xrightarrow{d} H(A), \tag{1.12}$$

for Borel sets $A \subseteq \mathbb{S}_+^{d-1}$ satisfying $H(\partial A) = 0$ where

$$H(A) = \frac{1}{c_{(r)}} \Lambda\left(\left\{\mathbf{x} \in \mathbb{E}_d : \frac{\mathbf{x}}{\|\mathbf{x}\|_{(s)}} \in A, \ \|\mathbf{x}\|_{(r)} \geq 1\right\}\right).$$

1.1.6 Spherical Clustering

The spherical clustering algorithms considered thus far operate exclusively on the unit sphere $\mathbb{S}_+^{d-1}$ defined with respect to the 2-norm (Euclidean norm), i.e., $\|\cdot\|_{(s)} = \|\cdot\|_2$ in (1.6). We do not adopt this restriction unless we are discussing specific examples, in order to maintain generality. The space $\mathbb{S}_+^{d-1}$ is equipped with the subspace topology inherited from $\mathbb{R}^d$. To facilitate clustering on $\mathbb{S}_+^{d-1}$, we introduce a dissimilarity measure D, which plays a central role in the analysis.

Definition 1. *A dissimilarity measure D on $\mathbb{S}_+^{d-1}$ is a continuous function $D : \mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1} \to [0,1]$ satisfying the following properties for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{S}_+^{d-1}$: (i) $D(\mathbf{w}_1, \mathbf{w}_2) = 0$ if and only if $\mathbf{w}_1 = \mathbf{w}_2$, (ii) $D(\mathbf{w}_1, \mathbf{w}_2) = D(\mathbf{w}_2, \mathbf{w}_1)$.*

Remark 1.1. *Without loss of generality, we assume that D is normalized such that its image covers the entire interval $[0,1]$. A function D satisfying properties (i) and (ii) is commonly known as a semimetric, which differs from a true metric by not satisfying the triangle inequality. Given the compactness of $\mathbb{S}_+^{d-1}$ and the continuity of D, convergence in D is topologically equivalent to convergence in $\mathbb{S}_+^{d-1}$: that is, $\mathbf{w}_n \to \mathbf{w}$ in $\mathbb{S}_+^{d-1}$ if and only if $D(\mathbf{w}_n, \mathbf{w}) \to 0$. Moreover, the* *collection of D-balls*

$$B_D(\mathbf{w}, r) := \{\mathbf{u} \in \mathbb{S}_+^{d-1} : D(\mathbf{w}, \mathbf{u}) < r\}, \quad \mathbf{w} \in \mathbb{S}_+^{d-1},\, r > 0,$$

forms a topological basis for $\mathbb{S}_+^{d-1}$*; see, e.g., Galvin and Shore, 1984. We also define a dual dissimilarity measure* $D^\dagger$ *by*

$$D^\dagger(\mathbf{w}_1, \mathbf{w}_2) := \sup_{\mathbf{w} \in \mathbb{S}_+^{d-1}} |D(\mathbf{w}, \mathbf{w}_1) - D(\mathbf{w}, \mathbf{w}_2)| \,. \tag{1.13}$$

This function is again a continuous semimetric on $\mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1}$*, satisfies* $D^\dagger \geq D$*, and is surjective onto* $[0, 1]$*. Moreover, a relaxed triangle inequality holds:*

$$D(\mathbf{w}_1, \mathbf{w}_3) \leq D(\mathbf{w}_1, \mathbf{w}_2) + D^\dagger(\mathbf{w}_2, \mathbf{w}_3). \tag{1.14}$$

In addition, if $D(\mathbf{w}_n, \mathbf{w}) \to 0$ *as* $n \to \infty$*, then* $D^\dagger(\mathbf{w}_n, \mathbf{w}) \to 0$ *as well.*

Common dissimilarity measures used in practice are often semimetrics. For the 2-norm unit sphere $\mathbb{S}_+^{d-1}$, the cosine dissimilarity used in spherical k-means clustering (Janßen and Wan, 2020) is

$$D_{\cos}(\mathbf{w}_1, \mathbf{w}_2) := 1 - \mathbf{w}_1^\top \mathbf{w}_2, \quad \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{S}_+^{d-1} \subset \mathbb{R}^d. \tag{1.15}$$

The k-principal components dissimilarity from Fomichov and Ivanovs, 2023 is given by

$$D_{\mathrm{pc}}(\mathbf{w}_1, \mathbf{w}_2) := 1 - \left(\mathbf{w}_1^\top \mathbf{w}_2\right)^2. \tag{1.16}$$

Both dissimilarities are semimetrics and offer computational advantages. For $D = D_{\cos}$ or D_{pc}, it follows from elementary inequalities that the dual satisfies

$$D^\dagger(\mathbf{w}_2, \mathbf{w}_3) \leq c\|\mathbf{w}_2 - \mathbf{w}_3\|_2, \quad \text{with } c = 1 \text{ or } 2, \text{ respectively.}$$

To formalize clustering, we adopt the notion of *multisets* (see, e.g., Kettleborough and Rayward-Smith, 2013). A multiset W on $\mathbb{S}_+^{d-1}$ allows repeated elements. Its support, denoted $\mathrm{supp}(W)$, is the usual set of unique elements in W. For example, if $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{S}_+^{d-1}$ are distinct, then $W = \{\mathbf{w}_1, \mathbf{w}_1, \mathbf{w}_2\}$ has support $\mathrm{supp}(W) = \{\mathbf{w}_1, \mathbf{w}_2\}$. A multiset is characterized by its multiplicity function $m_W : \mathbb{S}_+^{d-1} \to \{0, 1, \ldots\}$, where $m_W(\mathbf{w})$ denotes how many times $\mathbf{w}$ appears in W. A set is simply a multiset with multiplicities 0 or 1. If $\mathbf{w} \in W$, this means $\mathbf{w} \in \mathrm{supp}(W)$.

For two multisets W_1, W_2 with multiplicity functions m_1, m_2, their union $W_1 \cup W_2$ and intersection $W_1 \cap W_2$ are defined pointwise via $m_1 \vee m_2$ and $m_1 \wedge m_2$, respectively. We write $W_1 \subset W_2$ if $m_1 \leq m_2$. If $\mathrm{supp}(W)$ is finite, then for any function f on $\mathbb{S}_+^{d-1}$, we define

$$\sum_{\mathbf{w} \in W} f(\mathbf{w}) := \sum_{\mathbf{w} \in \text{supp}(W)} m_W(\mathbf{w}) f(\mathbf{w}),$$

so that the cardinality of W is $|W| := \sum_{\mathbf{w} \in \text{supp}(W)} m_W(\mathbf{w})$. We also define

$$D(\mathbf{w}, W) := \inf_{\mathbf{s} \in \text{supp}(W)} D(\mathbf{w}, \mathbf{s}).$$

Now suppose W is a multiset on $\mathbb{S}_+^{d-1}$ with $|W| < \infty$, and fix $k \in \mathbb{Z}_+$ with $k \leq |W|$. Let $A_k^* = \{\mathbf{a}_1^*, \ldots, \mathbf{a}_k^*\}$ be a multiset of k elements in $\mathbb{S}_+^{d-1}$ minimizing total dissimilarity:

$$\sum_{\mathbf{w} \in W} D(\mathbf{w}, A_k^*) = \inf \left\{ \sum_{\mathbf{w} \in W} D(\mathbf{w}, A) : \text{supp}(A) \subset \mathbb{S}_+^{d-1}, |A| = k \right\}. \tag{1.17}$$

By continuity of D and compactness of $\mathbb{S}_+^{d-1}$, such a minimizer exists, although it may not be unique. When $|\text{supp}(W)| \geq k$, the infimum is attained with k distinct centers.

We now define the notion of a k-clustering.

Definition 2. *A k-clustering of a multiset W on $\mathbb{S}_+^{d-1}$ with respect to a dissimilarity measure D is a pair $(A_k^*, \mathfrak{C}_k)$, where: $A_k^* = \{\mathbf{a}_1^*, \ldots, \mathbf{a}_k^*\}$ is a multiset minimizing* (1.17)*, and $\mathfrak{C}_k = \{C_1, \ldots, C_k\}$ is a partition of W into multisets such that for each $i \in \{1, \ldots, k\}$ and every $\mathbf{w} \in C_i$, we have*

$$D(\mathbf{w}, A_k^*) = D(\mathbf{w}, \mathbf{a}_i^*).$$

We refer to A_k^* as the set of centers, and the C_i's as clusters.

Remark 1.2. *A k-clustering always exists for any finite multiset W, although it need not be unique, even when A_k^* is. Ties in dissimilarity may lead to ambiguity in cluster assignments. Nonetheless, one can always select clusters so that each C_i is nonempty when $k \leq |W|$.*

For $D = D_{\cos}$ or D_{pc}, the above formulation recovers spherical k-means (Janßen and Wan, 2020) and k-PC clustering (Fomichov and Ivanovs, 2023), respectively. Solving (1.17) exactly is computationally intractable in general; thus, heuristic methods such as Lloyd-type algorithms are typically employed in practice. For theoretical purposes, however, we assume that an exact k-clustering is available.

When W is a random multiset, we assume that both the elements of A_k^* and the indicator variables $\mathbf{1}\{\mathbf{w} \in C_j\}$ for $\mathbf{w} \in W$, $j \in \{1, \ldots, k\}$, are measurable.

1.1.7 Spherical Clustering for Multivariate Extremes

Following Janßen and Wan, 2020 and Fomichov and Ivanovs, 2023, we connect spherical clustering with the analysis of multivariate extremes. With $n \in \mathbb{Z}_+$, let $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, be i.i.d. copies of a random vector $\mathbf{X}$ that is marginally standardized and belongs to the domain of attraction of α-Fréchet distribution, with spectral measure H on $\mathbb{S}_+^{d-1}$. We assume that the random vector $\mathbf{X}$ has been marginally standardized so that each component exhibits a standard α-Pareto-type tail behavior, specifically:

$$\lim_{x\to\infty} x^\alpha \mathbb{P}(X_1 > x) = \cdots = \lim_{x\to\infty} x^\alpha \mathbb{P}(X_d > x) = 1, \tag{1.18}$$

where $\alpha > 0$ is a known tail index, commonly chosen as $\alpha = 1$ or $\alpha = 2$ in the literature. While this condition can be generalized to allow for slowly varying functions (see Bingham et al., 1989), we adopt the simplified form in (1.18) for ease of exposition.

Let ℓ_n be an intermediate sequence satisfying $\ell_n \to \infty$ and $\ell_n/n \to 0$ as $n \to \infty$. Define the extremal subsample on $\mathbb{S}_+^{d-1}$ by

$$W_n = \left\{ \frac{\mathbf{X}_i}{\|\mathbf{X}_i\|_{(s)}} : \|\mathbf{X}_i\|_{(r)} \geq \left(\frac{n}{\ell_n}\right)^{1/\alpha},\ i \in \{1, \ldots, n\} \right\}. \tag{1.19}$$

In words, W_n consists of those sample points whose $\|\cdot\|_{(r)}$ norms exceed a high threshold, projected onto the $\|\cdot\|_{(s)}$-norm unit sphere $\mathbb{S}_+^{d-1}$. The choice of ℓ_n and the regular variation of $\mathbf{X}$ imply

$$\mathbb{E}|W_n| = n\mathbb{P}\left(\|\mathbf{X}_1\|_{(r)} \geq \left(\frac{n}{\ell_n}\right)^{1/\alpha}\right) \sim \ell_n c_{(r)} \to \infty, \tag{1.20}$$

as $n \to \infty$, where $c_{(r)}$ is as defined in (1.8). Moreover, the set $\{\mathbf{x} \in \mathbb{E}_d : \|\mathbf{x}\|_{(r)} \geq x\}$ is a Λ-continuity set for all $x > 0$ due to the homogeneity of Λ. Then, by a triangular-array version of the Strong Law of Large Numbers (see, e.g., Hsu and Robbins, 1947), we have

$$\begin{aligned}\frac{|W_n|}{\ell_n} =& \frac{1}{\ell_n}\sum_{i=1}^n \mathbf{1}\left\{\|\mathbf{X}_i\|_{(r)} \geq \left(\frac{n}{\ell_n}\right)^{1/\alpha}\right\} \\ \to& \Lambda\left(\{\mathbf{x} \in \mathbb{E}_d : \|\mathbf{x}\|_{(r)} \geq 1\}\right) = c_{(r)},\end{aligned} \tag{1.21}$$

almost surely as $n \to \infty$.

We now define the empirical spectral measure on $\mathbb{S}_+^{d-1}$ as

$$H_n = \frac{1}{|W_n|} \sum_{\mathbf{w} \in W_n} \delta_{\mathbf{w}}, \tag{1.22}$$

with the convention that H_n is the zero measure when $|W_n| = 0$. We now establish the consistency of this empirical measure.

Proposition 1.1. *Suppose* $\mathbf{X}$ *satisfies conditions* (1.18) *and* (1.11)*, with spectral measure* H *on* $\mathbb{S}_+^{d-1}$ *as in* (1.7)*. Let* W_n *be defined as in* (1.19)*, and* H_n *as in* (1.22)*. Then for any Borel set* $S \subset \mathbb{S}_+^{d-1}$ *that is a continuity set for* H*, we have*

$$H_n(S) \to H(S), \quad \textit{almost surely as } n \to \infty.$$

Proof. The result follows from a triangular-array strong law of large numbers applied to (1.11), (1.12), (1.20), and (1.21). □

We next consider applying the k-clustering procedure from Definition 2 to the subsample W_n. As indicated by Proposition 1.1, W_n is an increasingly accurate approximation of the spectral measure H. When H is discrete with finitely many atoms, clustering W_n allows for accurate recovery of both the locations and the masses of these atoms. The following corollary establishes the consistency of the resulting estimators.

Corollary 1.2. *Suppose* $\mathbf{X}$ *is as in Proposition 1.1, and the spectral measure* H *is of the discrete form*

$$H = \sum_{i=1}^{k} p_i \delta_{\mathbf{a}_i}, \tag{1.23}$$

with $\mathbf{a}_i \in \mathbb{S}_+^{d-1}$ *distinct and* $p_i > 0$, $\sum_{i=1}^k p_i = 1$*. Let* W_n *be the extremal subsample in* (1.19)*, and let* $(A_{k,n}, \mathfrak{C}_{k,n})$ *be a* k*-clustering of* W_n *as in Definition 2, with dissimilarity measure* D*. Define the estimated cluster proportions by*

$$p_{i,n}^k = \frac{|C_{i,n}^k|}{|W_n|}, \quad \textit{if } |W_n| > 0, \quad \textit{and } p_{i,n}^k := 0 \textit{ otherwise}. \tag{1.24}$$

Then there exist bijections $\pi_n : \{1, \dots, k\} \to \{1, \dots, k\}$ *such that*

$$\mathbf{a}_{\pi_n(i),n}^k \to \mathbf{a}_i, \quad p_{\pi_n(i),n}^k \to p_i, \quad \textit{for all } i \in \{1, \dots, k\},$$

almost surely as $n \to \infty$.

Proof. The convergence $\mathbf{a}^k_{\pi_n(i),n} \to \mathbf{a}_i$ follows from Janßen and Wan, 2020, Theorem 3.1, which establishes convergence in Hausdorff distance between $A_{k,n}$ and $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$, along with Proposition 1.1; see also Janßen and Wan, 2020, Section 4.

To prove the convergence of $p^k_{\pi_n(i),n}$, define

$$r_A := \sup\{r > 0 : B(\mathbf{a}_i, r),\ i = 1, \ldots, k \text{ are disjoint}\} > 0. \tag{1.25}$$

Fix $\epsilon \in (0, r_A/3)$. Then, almost surely, for large enough n, the dissimilarities satisfy $D^\dagger(\mathbf{a}^k_{\pi_n(i),n}, \mathbf{a}_i) < \epsilon$. By the triangle inequality (1.14), we obtain the inclusions

$$B_D(\mathbf{a}_i, \epsilon) \subset B_D(\mathbf{a}^k_{\pi_n(i),n}, 2\epsilon) \subset B_D(\mathbf{a}_i, 3\epsilon).$$

Since the sets $B_D(\mathbf{a}_i, 3\epsilon)$ are disjoint, Definition 2 implies:

$$B_D(\mathbf{a}_i, \epsilon) \cap W_n \subset C^k_{\pi_n(i),n} \subset \left(B_D(\mathbf{a}_i, \epsilon) \cup \bigcap_{j \neq i} B_D(\mathbf{a}_j, \epsilon)^c \right) \cap W_n.$$

Consequently,

$$H_n(B_D(\mathbf{a}_i, \epsilon)) \le p^k_{\pi_n(i),n} \le H_n \left(B_D(\mathbf{a}_i, \epsilon) \cup \bigcap_{j \neq i} B_D(\mathbf{a}_j, \epsilon)^c \right).$$

Both bounds converge almost surely to p_i as $n \to \infty$ by Proposition 1.1. $\square$

Remark 1.3. *In contrast to Janßen and Wan, 2020, Proposition 3.3, we work directly under the marginal standardization assumption in* (1.18) *and omit the empirical marginal transformations used in Janßen and Wan, 2020, Eq. (3.5) for simplicity. Nonetheless, our consistency result for order selection (Theorem 3.5) can be extended to the setting of Janßen and Wan, 2020 by leveraging their corresponding results.*

1.2 Literature Review

1.2.1 Joint Sum-and-Max Limit

We now turn to the joint asymptotic behavior of the partial sum S_n and the partial maximum M_n. The study of their joint distribution is of particular interest, as it reveals the interplay between typical and extreme behaviors in a sequence of random variables. Understanding the joint convergence provides a more comprehensive probabilistic description than marginal analyses alone. For instance, in risk management or insurance mathematics, both the accumulated claims (sum) and the largest claim (maximum) are of critical interest, and their dependence structure may significantly impact modeling and inference.

Clearly, joint convergence in distribution implies the marginal convergence of each component. The case where $\{X_n\}_{n\in\mathbb{N}}$ are i.i.d. was thoroughly studied by Chow and Teugels, 1978. They showed that, under suitable normalization, the pair (S_n, M_n) converges jointly in distribution to a non-degenerate limit (S, M) as $n \to \infty$ if and only if each component converges marginally. Moreover, they established a striking result: S_n and M_n are asymptotically independent—meaning the limiting variables S and M are independent—unless the common distribution F of X_n has a heavy tail on the positive side. Specifically, dependence in the limit arises only when the tail $1 - F(x)$ is regularly varying with index $-\alpha$ for some $\alpha \in (0, 2)$, i.e., $F \in \mathcal{D}(\Phi_\alpha) \cap \mathcal{D}(\alpha, p)$ with $p > 0$.

This result has a natural interpretation. When F is light-tailed, the influence of any individual X_n is asymptotically negligible relative to the sum S_n, and the maximum M_n has little effect on the sum, leading to asymptotic independence. In contrast, when F is heavy-tailed, the sum S_n is often dominated by the same extreme values that define M_n, inducing a non-trivial dependence structure in the limit.

Furthermore, Chow and Teugels, 1978 extended these results to the functional setting. They proved that the process $(S_n(t), M_n(t))_{t\geq 0}$, when properly normalized, converges in distribution in a Skorokhod space to a joint limit process $(S(t), M(t))_{t\geq 0}$. In the heavy-tailed case, the processes $S(t)$ and $M(t)$ exhibit dependence, reflecting the dominance of extreme values over entire time intervals. Marginally, $S(t)$ is an α-stable Lévy process and $M(t)$ is an α-Fréchet extremal process.

What happens when the sequence $\{X_n\}$ exhibits dependence? Under suitable weak dependence conditions—typically referred to as short-range dependence—it is often found that the joint distribution of normalized (S_n, M_n) mimics the behavior seen in the i.i.d. case. For example, if $\{X_n\}$ is strongly mixing and has finite variance, then S_n and M_n are asymptotically independent; see Anderson and Turkman, 1991; Hsing, 1995 for precise formulations and results. This mirrors the i.i.d. scenario, where individual terms are asymptotically negligible in the sum and do not influence the maximum significantly.

However, the situation becomes more nuanced when $\{X_n\}$ has a heavy right tail, i.e., when $1 - F(x)$ is regularly varying at infinity. In this regime, dependence in $\{X_n\}$ can significantly alter the joint limiting behavior of (S_n, M_n). Specifically, a key issue arises when large observations from both tails cluster in time: the dependence structure may induce cancellation effects in the partial sums, thereby invalidating the asymptotic dependence typically observed in the i.i.d. case. This type of phenomenon emphasizes that dependence not only affects marginal distributions but also plays a crucial role in shaping joint extremal behavior.

Despite these complications, when such cancellation is excluded—either structurally or via specific assumptions—the normalized sum and maximum often still exhibit joint convergence to the same class of limits as in the i.i.d. setting. This was rigorously analyzed in Anderson and Turkman, 1995, where conditions were given to preclude pathological cancellations while retaining the heavy-tail-driven dependence structure.

More recently, Krizmanić, 2020 extended this line of inquiry to functional convergence. There, the joint process $(S_n(t), M_n(t))_{t\geq 0}$ was shown to converge under conditions that effectively control the dependence among extreme events. Specifically, the assumptions ensured that extreme values form clusters which can be treated as asymptotically independent blocks. Moreover, by requiring that extremes within a cluster have the same sign, the analysis rules out potential cancellations that could otherwise obscure the heavy-tail effects.

There have been relatively few studies on the joint asymptotic distribution of (S_n, M_n) under strong dependence in $\{X_n\}$, often referred to as *long-range dependence*. Notably, much of the existing theory focuses on the case where $\{X_n\}$ is a Gaussian sequence. For such processes, it has been observed that the critical decay rate of the covariance function significantly influences the limiting joint behavior of (S_n, M_n). Specifically, the rate $(1/\log n)$ serves as a boundary between asymptotic independence and dependence.

When the covariance satisfies $\mathrm{Cov}(X_{n+1}, X_1) = o(1/\log n)$, and under some additional regularity conditions, it has been shown that S_n and M_n are asymptotically independent; see Ho and Hsing, 1996; Ho and McCormick, 1999. This mirrors the behavior found in weakly dependent or short-range dependent sequences with light-tailed distributions. In contrast, when

$$\lim_{n\to\infty} \mathrm{Cov}(X_{n+1}, X_1) \log n \in (0, \infty],$$

the partial sum S_n and the maximum M_n become asymptotically dependent. In this regime, the long-range memory embedded in the covariance structure allows extreme values to influence the aggregate behavior of the sum, leading to non-negligible contributions from the maximum to the sum even asymptotically. The limiting joint distribution of (S_n, M_n) under such long-range dependent Gaussian structures was rigorously studied in Ho and Hsing, 1996; Ho and McCormick, 1999; McCormick and Qi, 2000.

On the other hand, the joint sum-and-maximum limit theorem for long-range dependent $\{X_n\}$ with heavy-tailed distributions has not, to the best of our knowledge, been rigorously investigated. To clarify, it is important to define precisely what is meant by long-range dependence in this context. Notably, the threshold between short- and long-range dependence may differ depending on whether we are analyzing the sum or the maximum of the sequence.

We adopt the *phase-transition framework* proposed by Samorodnitsky, 2016, which characterizes long-range dependence through a qualitative shift in both the normalization required for convergence and the nature of the limiting process itself. Under this viewpoint, long-range dependence manifests when the normalizing sequence or the limit diverges significantly from the i.i.d. setting, indicating stronger memory in the process.

A key insight from this perspective is that long-range dependence tends to appear earlier in the sum than in the maximum. For example, in the Gaussian setting, long-range dependence for the sum S_n already emerges when the covariance $\mathrm{Cov}(X_{n+1}, X_1)$ decays as $n^{-\rho}$ with $\rho \in (0, 1)$. In contrast, the maximum M_n exhibits long-range dependence only when the covariance decays as slowly as $(1/\log n)$—the critical threshold discussed previously.

A similar phenomenon occurs in the heavy-tailed setting. The sum process can be substantially influenced by the clustering of extremes even when the dependence is relatively weak, while the maximum typically requires stronger dependence—such as persistent clustering of high-magnitude events—to deviate from classical extreme value behavior. In this work, we emphasize that the term long-range dependence refers to a dependence structure that simultaneously affects the asymptotic behavior of both the sum and the maximum.

In particular, we focus on a class of stationary infinitely divisible processes $\{X_n\}_{n\in\mathbb{N}}$ with regularly varying tails, whose dependence structure arises from a null-recurrent Markov chain characterized by a memory parameter $\beta \in (0, 1)$. This modeling framework, originally introduced in Rosiński and Samorodnitsky, 1996, has since attracted substantial attention due to its capacity to encapsulate long-range dependence both from the perspective of partial sums and partial maxima.

1.2.2 On estimation and order selection for multivariate extremes via clustering

Multivariate extreme value theory (EVT) concerns the statistical behavior of concurrent extreme events across multiple variables; see Beirlant et al., 2006; de Haan and Ferreira, 2006. A common approach in this theory involves standardizing the marginal distributions and examining the angular distribution of extreme observations—those with large norms. Under the multivariate maximum domain of attraction assumption, this angular distribution converges to a limit on the unit sphere, referred to as the spectral measure (or angular measure).

Due to the inherently small sample size of extreme events, the challenge of high dimensionality becomes particularly acute in multivariate EVT. As highlighted in the review article Engelke and Ivanovs, 2021, a central focus of recent work has been on employing parsimonious modeling strategies to mitigate this issue. A notable and interpretable example is the class of discrete spectral measures, where the angular distribution is concentrated on a finite set of directions. Despite its simplicity, Fougères et al., 2013 established that any extremal dependence structure can be approximated arbitrarily well by such discrete spectral measures. Furthermore, several classes of parametric models—including heavy-tailed max-linear and sum-linear models (e.g., Einmahl et al., 2012) and the more recent transformed-linear model Cooley and Thibaud, 2019—are characterized by having discrete spectral measures.

More recently, several authors have proposed applying clustering algorithms on the unit sphere as a parsimonious summary of the angular structure of multivariate extremes. Einmahl et al., 2012 and Janßen and Wan, 2020 employed the spherical k-means algorithm with cosine dissimilarity Dhillon and Modha, 2001 and explored its connection to estimating max-linear factor models. Fomichov and Ivanovs, 2023 proposed the spherical k-principal component (k-PC) clustering method, based on a refined cosine dissimilarity, and demonstrated its ability to detect spectral mass concentrated on lower-dimensional faces of the sphere. Medina et al., 2024 utilized the spectral clustering algorithm (Ng et al., 2001) on a k-nearest neighbor graph constructed from angular components of extreme samples, relating it to sum-linear models.

These works reveal a natural link between discrete spectral measures and spherical clustering: each atom of the spectral measure can be viewed as a cluster center, with extreme data points clustering around them. This intuition has been formalized in Janßen and Wan, 2020; Medina et al., 2024, where consistency results for recovering the spectral measure via clustering were established (with Janßen and Wan, 2020's results also covering the k-PC method of Fomichov and Ivanovs, 2023). Since the underlying parameters of max-linear and sum-linear factor models are directly determined by the spectral measure, consistent estimation of the spectral measure directly enables consistent inference for the model parameters.

However, in all existing theoretical analyses connecting clustering algorithms to discrete spectral measures, the number of atoms—equivalently, the number of clusters (referred to as the order)—is assumed to be known. In practice, ad hoc methods such as the elbow or scree plot are often used to guide order selection (Fomichov and Ivanovs, 2023; Janßen and Wan, 2020; Medina et al., 2024). These methods rely on visual inspection and lack rigorous theoretical justification.

In this paper, we advance the study of clustering-based inference for multivariate extremes with discrete spectral measures. Our contributions are threefold: 1. We develop a novel method for selecting the number of clusters (order), which is both theoretically consistent and practically simple to implement. Our method builds on the classical silhouette method (Hruschka et al., 2004; Rousseeuw, 1987), with a key innovation: we introduce a penalty term to the simplified average silhouette width to discourage both small cluster sizes and small dissimilarities between cluster centers. This adjustment enhances sensitivity to overestimation of the number of clusters and allows for consistent estimation of the true order—even in models where likelihood-based criteria are unavailable, such as max-linear factor models (Einmahl et al., 2012; Yuen and Stoev, 2014). 2. We provide a large deviation-type result on the quality of spectral measure estimation via clustering methods like spherical k-means and k-PC. This offers a quantitative perspective on the convergence properties of clustering-based spectral inference under the multivariate extreme value framework. 3. We describe how the discrete spectral measure estimation can be directly translated into parameter estimates for heavy-tailed max-linear and sum-linear factor models. Through simulations and real-data examples, we demonstrate the performance of our order selection method and its application to factor model inference.

Chapter 2

The Setup

2.1 A Class of Long-Range Dependent Processes Generated by Conservative Flows

We adopt a model essentially following that of Samorodnitsky and Wang, 2019. Let $\{Y_n\}_{n\in\mathbb{N}_0}$ denote an irreducible, aperiodic, null-recurrent Markov chain on $\mathbb{Z}$, with state space $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$. The sample paths $(x_0, x_1, \ldots)$ of the chain belong to the measurable space $(E, \mathcal{E}) := (\mathbb{Z}^{\mathbb{N}_0}, \mathcal{C}(\mathbb{Z}^{\mathbb{N}_0}))$, where $\mathcal{C}(\mathbb{Z}^{\mathbb{N}_0})$ is the cylindrical σ-field. Let $(\pi_i)_{i\in\mathbb{Z}}$ denote the unique invariant measure of the chain, normalized so that $\pi_0 = 1$. For each initial state $i \in \mathbb{Z}$, let P_i be the law of the Markov chain started at $Y_0 = i$. Define a σ-finite infinite measure μ on $(E, \mathcal{E})$ by

$$\mu(A) := \sum_{i\in\mathbb{Z}} \pi_i P_i(A), \quad A \in \mathcal{E}.$$

Let $T : E \to E$ be the left-shift operator given by $T(x_0, x_1, x_2, \ldots) = (x_1, x_2, \ldots)$. Then μ is T-invariant, i.e., $\mu \circ T^{-1} = \mu$.

We consider the stationary process defined by

$$X_n = \int_E f \circ T^n(s)\, M(ds), \quad n \in \mathbb{N}, \tag{2.1}$$

where $f : E \to \mathbb{R}$ is a measurable function and M is a homogeneous symmetric infinitely divisible random measure on $(E, \mathcal{E})$, without a Gaussian component. Specifically, the random measure M is characterized by the following: the control measure is the σ-finite measure μ on $(E, \mathcal{E})$, the local Gaussian variance is identically zero: $\sigma^2(s) = 0$, the local Lévy measure is constant: $\rho(s, \cdot) = \rho(\cdot)$, the local drift is zero: $b(s) = 0$, where ρ is a symmetric Lévy measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying $\rho(-B) = \rho(B)$ for all $B \in \mathcal{B}(\mathbb{R})$. Under these assumptions, the distribution of M is characterized by:

$$\mathbb{E}\left[e^{i\theta M(A)}\right] = \exp\left\{-\mu(A)\int_{\mathbb{R}}(1-\cos(\theta y))\,\rho(dy)\right\}, \quad \theta \in \mathbb{R},$$

for any $A \in \mathcal{E}$ with $\mu(A) < \infty$. The resulting sequence $\{X_n\}_{n\in\mathbb{N}}$ is stationary, infinitely divisible, and has symmetric marginal distributions.

Let $\mathbf{x} := (x_0, x_1, \ldots)$ denote an element of the space E, and define the measurable set

$$A_0 := \{\mathbf{x} \in E : x_0 = 0\},$$

which satisfies $\mu(A_0) = \pi_0 = 1$. For the integrand function f in (2.1), we shall, for simplicity and following the convention in Samorodnitsky and Wang, 2019, assume that

$$f = \mathbf{1}_{A_0}.$$

Since $M(A)$ and $M(B)$ are independent for disjoint sets A and B, the dependence structure of the process $\{X_n\}_{n\in\mathbb{N}}$ is entirely determined by the flow of sets $\{T^{-n}A_0\}_{n\in\mathbb{N}}$, or equivalently, by the ergodic properties of the transformation T. Define the *wandering rate* sequence by

$$w_n := \mu\left(\bigcup_{k=1}^{n} T^{-k}A_0\right), \quad n \in \mathbb{N}. \tag{2.2}$$

This sequence is closely related to the *first entrance time* into A_0, defined as

$$\varphi_{A_0}(\mathbf{x}) := \inf\{n \in \mathbb{N} : x_n = 0\}, \quad \mathbf{x} \in E.$$

It follows that $w_n = \sum_{k=1}^{n} P_0(\varphi_{A_0} \geq k)$. Throughout this paper, we assume that the tail distribution of the first entrance time into A_0 after the Markov chain makes its first departure from A_0 satisfies

$$P_0(\varphi_{A_0} > n) \in \mathrm{RV}_\infty(-\beta) \quad \text{for some } \beta \in (0,1),$$

where $\mathrm{RV}_\infty(-\beta)$ denotes the class of regularly varying functions at infinity with index $-\beta$. Similarly, we use RV_0 to denote regular variation at zero. This assumption on the tail of the entrance time distribution can equivalently be expressed in terms of the wandering rate sequence

$$w_n = \sum_{k=1}^{n} P_0(\varphi_{A_0} \geq k) \sim \frac{n\,P_0(\varphi_{A_0} > n)}{1-\beta} \in \mathrm{RV}_\infty(1-\beta), \tag{2.3}$$

as established in Owada and Samorodnitsky, 2015. This assumption induces long-range dependence in the process.

The marginal distribution of X_n is governed by the Lévy measure ρ, which consequently determines the domain of attraction of the process. We assume that ρ satisfies the regular variation condition at infinity:

$$\rho((x,\infty)) \in \mathrm{RV}_\infty(-\alpha), \quad \text{for some } \alpha > 0. \tag{2.4}$$

Moreover, in the case $0 < \alpha < 2$, following the framework of Bai et al., 2020, we impose an additional condition on the behavior of ρ near the origin:

$$\rho((x,\infty)) = O(x^{-\alpha_0}) \quad \text{as } x \to 0, \quad \text{for some } \alpha_0 \in (0,2). \tag{2.5}$$

This ensures sufficient control over both the tails and the small jump behavior of the Lévy measure.

In our setting, the process $\{X_n\}_{n\in\mathbb{N}}$ defined in (2.1) admits a series representation for each fixed $n \in \mathbb{N}$. For each n, let $\{U_j^{(n)}\}_{j\in\mathbb{N}}$ be a sequence of i.i.d. E-valued random variables with common distribution μ_n defined by

$$\frac{d\mu_n}{d\mu}(\mathbf{x}) = \frac{\mathbf{1}_{\{\bigcup_{k=1}^n T^{-k}A_0\}}(\mathbf{x})}{\mu\left(\bigcup_{k=1}^n T^{-k}A_0\right)}, \quad \mathbf{x} \in E.$$

Due to the shift-invariance of μ, i.e., $\mu(T^{-1}\cdot) = \mu(\cdot)$, and noting that by (2.2), $\mu\left(\bigcup_{k=1}^n T^{-k}A_0\right) = w_n$, it follows that for each $k = 1, \dots, n$ and $j \in \mathbb{N}$,

$$\mathbb{P}\left(U_j^{(n)} \in T^{-k}A_0\right) = \frac{\mu(T^{-k}A_0)}{w_n} = \frac{\mu(A_0)}{w_n} = \frac{1}{w_n}.$$

Let $\{\varepsilon_n\}_{n\geq 1}$ be a sequence of i.i.d. Rademacher random variables and $\{\Gamma_n\}_{n\geq 1}$ denote the ordered points of a unit-rate Poisson process on $(0,\infty)$. Hence, as in (1.4), the finite-dimensional distributions $(X_k)_{k=1,\dots,n}$ admit the following representation in distribution:

$$(X_k)_{k=1,\dots,n} \overset{d}{=} \left(\sum_{j=1}^\infty \varepsilon_j\, \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right) \mathbf{1}_{\{U_j^{(n)} \in T^{-k}A_0\}}\right)_{k=1,\dots,n}, \tag{2.6}$$

where $\rho^{\leftarrow}$ denotes the generalized inverse of the tail of the Lévy measure ρ defined for $y > 0$ by

$$\rho^{\leftarrow}(y) := \inf\left\{x > 0 : \rho((x,\infty)) \leq y\right\}. \tag{2.7}$$

We review known results concerning the limit theorems for partial sums and partial maxima of the process $\{X_n\}_{n\in\mathbb{N}}$ defined in (2.1). The limiting processes are closely connected to stable subordinators, for which we refer to Bertoin, 1999 for a comprehensive treatment. A *subordinator* is a non-decreasing Lévy process characterized by stationary and independent increments and càdlàg sample paths. It starts at zero almost surely. A *β-stable subordinator* $(S_\beta(t))_{t\geq 0}$ has a stability index $\beta \in (0,1)$. For each $t > 0$, the random variable $S_\beta(t)$ follows a one-sided β-stable distribution supported on $[0,\infty)$, with Laplace transform given by

$$\mathbb{E}\left[e^{-\lambda S_\beta(t)}\right] = \exp\left(-t\lambda^\beta\right), \quad \lambda \geq 0.$$

The associated Lévy measure $\nu(dx)$ of S_β has the form

$$\nu(dx) = \frac{\beta}{\Gamma(1-\beta)} x^{-\beta-1}\,\mathbf{1}_{(0,\infty)}(x)\,dx.$$

The sample paths of S_β are strictly increasing and right-continuous almost surely.

The inverse stable subordinator, also known as the *Mittag-Leffler process*, is defined as

$$M_\beta(t) := S_\beta^{\leftarrow}(t) := \inf\{u \geq 0 : S_\beta(u) \geq t\}, \quad t \geq 0.$$

This process is continuous and non-decreasing almost surely. Alternatively, the right-continuous inverse can be used:

$$S_\beta^{\rightarrow}(t) := \inf\{u \geq 0 : S_\beta(u) > t\}.$$

The closure of the range $\{S_\beta(t) : t \geq 0\}$ defines the *β-stable regenerative set*, denoted by $\mathcal{R}_\beta$. This random closed set takes values in $\mathfrak{F}([0,\infty))$, the space of closed subsets of $[0,\infty)$ equipped with the Fell topology; for background on the theory of random sets, we refer to Molchanov, 2005.

Define the partial sum process by

$$S_n(t) := \sum_{k=1}^{\lfloor nt \rfloor} X_k, \quad t \geq 0. \tag{2.8}$$

It is known that the limiting behavior of the normalized process $(S_n(t))_{t\geq 0}$ depends critically on the tail behavior of X_k.

If X_k has finite variance, which is equivalent to the Lévy measure ρ satisfying

$$\int_{\mathbb{R}} x^2 \, \rho(dx) < \infty, \tag{2.9}$$

then the partial sum process satisfies the functional central limit theorem

$$\frac{1}{n w_n^{-1/2}} \left(S_n(t)\right)_{t \geq 0} \Rightarrow c_\beta \left(B_H(t)\right)_{t \geq 0} \quad \text{as } n \to \infty,$$

in the space $D[0, \infty)$ equipped with the Skorokhod J_1-topology. Here, w_n is the wandering rate defined in (2.2), and $B_H(t)$ denotes fractional Brownian motion with Hurst index $H = (1 + \beta)/2$, satisfying $\mathbb{E}[B_H(t)^2] = t^{2H}/2$, $t \geq 0$; see Samorodnitsky, 2016, Theorem 9.4.7. The constant c_β is given by

$$c_\beta^2 = \int_{\mathbb{R}} x^2 \, \rho(dx) \cdot \frac{\Gamma(1 + 2\beta)}{\Gamma(2 - \beta)\Gamma(2 + \beta)} \cdot \mathbb{E}[S_\beta(1)^{-2\beta}], \tag{2.10}$$

where $S_\beta(1)$ denotes the β-stable subordinator evaluated at time 1.

Note that the normalization sequence $(n w_n^{-1/2}) \in \mathrm{RV}_\infty(H)$ due to (2.3), and thus grows faster than $n^{1/2}$, which is the classical rate for the i.i.d. case. This reflects the long-range dependence induced by the conservative flow

When the marginal distributions of X_k exhibit infinite variance, and the Lévy measure ρ satisfies the regular variation condition for some $\alpha \in (0, 2)$ as in (2.4), along with the boundedness condition at the origin as in (2.5), it follows that the generalized inverse $\rho^{\leftarrow}$ of the tail of the Lévy measure (defined in (2.7)) satisfies $\rho^{\leftarrow}(y) \in \mathrm{RV}_0(-1/\alpha)$. Under these assumptions, the normalized partial sum process

$$\frac{1}{\rho^{\leftarrow}(w_n^{-1}) n w_n^{-1}} \left(S_n(t)\right)_{t \geq 0} \Rightarrow \frac{1}{\Gamma(2 - \beta)} C_\alpha^{-1/\alpha} \left(Y_{\alpha,\beta}(t)\right)_{t \geq 0}, \quad \text{as } n \to \infty,$$

in the Skorokhod space $D[0, \infty)$ equipped with the J_1 topology. The normalization sequence belongs to the class $\mathrm{RV}_\infty\left(\beta + \frac{1-\beta}{\alpha}\right)$. The constant C_α is given by

$$C_\alpha = \begin{cases} \{\Gamma(1 - \alpha)\}^{-1} \cos^{-1}\left(\frac{\pi\alpha}{2}\right), & \text{if } \alpha \neq 1, \\ \frac{2}{\pi}, & \text{if } \alpha = 1, \end{cases} \tag{2.11}$$

as shown in Owada and Samorodnitsky, 2015, Theorem 5.1 and Example 5.5 and Bai et al., 2020.

To describe the limit process $(Y_{\alpha,\beta}(t))_{t \geq 0}$, where $0 < \alpha < 2$ and $0 < \beta < 1$, let $(\Omega', \mathcal{F}', \mathbb{P}')$ be an auxiliary probability space independent of the underlying one. Let $(M_\beta(t, \omega'))_{t \geq 0}$ denote a Mittag-Leffler process defined on

$(\Omega', \mathcal{F}', \mathbb{P}')$, and let $\nu(dx) = (1-\beta)x^{-\beta}dx$ be a measure on $(0,\infty)$. Then the limiting process is given by

$$Y_{\alpha,\beta}(t) := \int_{\Omega'\times[0,\infty)} M_\beta((t-x)_+, \omega')\, dZ_{\alpha,\beta}(\omega', x), \quad t \geq 0,$$

where $Z_{\alpha,\beta}$ is a symmetric α-stable (SαS) random measure on $\Omega' \times [0,\infty)$ with control measure $m := \mathbb{P}' \otimes \nu$, characterized by

$$\mathbb{E}\left[\exp\left(i\theta Z_{\alpha,\beta}(A)\right)\right] = \exp\left(-|\theta|^\alpha m(A)\right), \quad \theta \in \mathbb{R},$$

for all $A \in \mathcal{F}' \otimes \mathcal{B}([0,\infty))$ with $m(A) < \infty$.

The process $(Y_{\alpha,\beta}(t))_{t\geq 0}$ is represented as a stochastic integral with respect to an infinitely divisible random measure, and hence it admits a series representation as in (1.4). For any subordinator σ, define its (right-continuous) inverse process, also known as the *local time process*, by

$$L_\sigma(x) := \inf\{t \geq 0 : \sigma(t) > x\}, \quad x \geq 0. \tag{2.12}$$

Let $\{\varepsilon_j\}_{j\in\mathbb{N}}$ be a sequence of i.i.d. Rademacher random variables, and $\{\Gamma_j\}_{j\in\mathbb{N}}$ denote the ordered arrival times of a standard Poisson process on $[0,\infty)$, as introduced earlier. Let $\{\sigma_j\}_{j\in\mathbb{N}}$ be a sequence of i.i.d. standard β-stable subordinators, and let $\{V_j\}_{j\in\mathbb{N}}$ be i.i.d. random variables on $[0,1]$ with common distribution function $\mathbb{P}(V \leq x) = x^{1-\beta}$, $x \in [0,1]$. All sequences above are assumed mutually independent. Let $L_j := L_{\sigma_j}$ denote the local time process of σ_j. Now, define the following series representation of the limiting process:

$$S(t) := (2C_\alpha)^{1/\alpha} \sum_{j=1}^{\infty} \varepsilon_j \, \Gamma_j^{-1/\alpha} \, L_j\left((t - V_j)_+\right), \quad t \in [0,1], \tag{2.13}$$

where C_α is the constant defined in (2.11). The series in (2.13) converges almost surely; see Bai et al., 2020. Moreover, the process $(Y_{\alpha,\beta}(t))_{t\in[0,1]}$, $Y_{\alpha,\beta}$ restricted to time interval $[0,1]$, is equal in law to $(S(t))_{t\in[0,1]}$, see Samorodnitsky, 2016, Example 3.4.4.

Next, we review the limit theorems for the partial maximum process. For each $n \in \mathbb{N}$, define

$$M_n(B) := \max_{k\in(nB)\cap\mathbb{N}} X_k, \quad B \in \mathcal{G}([0,\infty)), \tag{2.14}$$

with the convention $\max_\emptyset := -\infty$, and where $\mathcal{G}([0,\infty))$ denotes the collection of open subsets of $[0,\infty)$ under the subspace topology.

Suppose the marginal law of X_k satisfies the regular variation condition (cf. (2.4)) for some $\alpha > 0$, and assume the additional regularity condition:

$$\sup_{n\in\mathbb{N}} \frac{n\,\mathbb{P}_0(\varphi_A = n)}{\mathbb{P}_0(\varphi_A > n)} < \infty. \tag{2.15}$$

Under these assumptions, Samorodnitsky and Wang, 2019 established the following functional limit theorem:

$$\frac{1}{\rho^{\leftarrow}(w_n^{-1})} \left(M_n(B)\right)_{B\in\mathcal{G}([0,\infty))} \Rightarrow \left(\eta_{\alpha,\beta}(B)\right)_{B\in\mathcal{G}([0,\infty))}, \quad \text{as } n\to\infty,$$

in the space of sup measures $SM([0,\infty))$ equipped with the sup vague topology. In the above theorem, the normalization sequence $\rho^{\leftarrow}(w_n^{-1}) \in \mathrm{RV}_\infty((1-\beta)/\alpha)$.

To define the limiting random sup measure $\eta_{\alpha,\beta}$, consider a Poisson point process on $[0,\infty)\times[0,\infty)\times\mathfrak{F}([0,\infty))$ with mean measure $\alpha u^{-(1+\alpha)}du\,(1-\beta)v^{-\beta}dv\,\mathbb{P}_{\mathcal{R}_\beta}$, where $\mathbb{P}_{\mathcal{R}_\beta}$ is the law of the β-stable regenerative set $\mathcal{R}_\beta$. Let $(U_j, W_j, F_j)_{j\in\mathbb{N}}$ be a measurable enumeration of the points of the Poisson process. Define the shifted regenerative sets:

$$\widetilde{F}_j := W_j + F_j := \{W_j + x : x \in F_j\}, \quad j\in\mathbb{N}.$$

Then, for $B \in \mathcal{G}([0,\infty))$, the limit random sup measure is given by

$$\eta_{\alpha,\beta}(B) := \sup_{t\in B}\left(\sum_{j=1}^{\infty} U_j \mathbf{1}_{\{t\in\widetilde{F}_j\}}\right). \tag{2.16}$$

It is known that if $\beta\in(0,1/2]$, then $\eta_{\alpha,\beta}(B)$ follows an α-Fréchet distribution marginally for each B. However, this property fails when $\beta\in(1/2,1)$, indicating a significant shift in the extremal structure of the process in the strongly dependent regime.

If we restrict the random sup measure $\eta_{\alpha,\beta}$ to a compact interval, say $[0,1]$, we can utilize a particularly convenient measurable enumeration of the points in the associated Poisson point process. Define the closed range of a subordinator σ as

$$\mathcal{R}_\sigma := \overline{\{\sigma(t) : t\geq 0\}}, \tag{2.17}$$

where the closure is taken in $[0,\infty)$. Let $\{\Gamma_j\}_{j\in\mathbb{N}}$ denote the arrival times of a standard Poisson process on $(0,\infty)$, and let $\{\sigma_j\}_{j\in\mathbb{N}}$ be a sequence of i.i.d. standard β-stable subordinators, independent of $\{\Gamma_j\}$. Let $\{V_j\}_{j\in\mathbb{N}}$ be an i.i.d. sequence of random variables on $[0,1]$ with distribution function $\mathbb{P}(V\leq x) =$

$x^{1-\beta}$, independent of the previous sequences. Define $\mathcal{R}_j := \mathcal{R}_{\sigma_j}$, and let

$$\widetilde{\mathcal{R}}_j := V_j + \mathcal{R}_j = \{V_j + x : x \in \mathcal{R}_j\}, \quad j \in \mathbb{N},$$

be the regenerative set $\mathcal{R}_j$ shifted by V_j. Note that since $0 \in \mathcal{R}_j$ almost surely, we also have $V_j \in \widetilde{\mathcal{R}}_j$. We now define the random sup measure as

$$M(B) := \sup_{t\in B}\left(\sum_{j=1}^{\infty} \Gamma_j^{-1/\alpha} \mathbf{1}_{\{t \in \widetilde{\mathcal{R}}_j\}}\right), \quad B \in \mathcal{G}([0,1]).$$

Then the family $(M(B))_{B\in\mathcal{G}([0,1])}$ has the same finite-dimensional distributions as $(\eta_{\alpha,\beta}(B))_{B\in\mathcal{G}([0,1])}$. This equivalence in distribution arises from the fact that the original Poisson point process $(U_j, W_j, F_j)_{j\in\mathbb{N}}$ restricted to $[0,\infty)\times[0,1]\times\mathfrak{F}([0,\infty))$ can be interpreted as a Poisson point process $(U_j)_{j\in\mathbb{N}}$ on $[0,\infty)$, marked by two independent sequences $(V_j)_{j\in\mathbb{N}}$ and $(\mathcal{R}_j)_{j\in\mathbb{N}}$, which are i.i.d. random elements in $[0,1]$ and the space of regenerative sets, respectively. Moreover, we may enumerate the points of the Poisson random measure in decreasing order of the first coordinate $(U_j)_{j\in\mathbb{N}}$, which leads to a natural representation $\{\Gamma_j^{-1/\alpha}\}_{j\in\mathbb{N}}$.

Let us introduce the set

$$I_S := \bigcap_{j\in S} \widetilde{\mathcal{R}}_j \cap [0,1], \quad \emptyset \neq S \subset \mathbb{N},$$

with the convention $I_\emptyset := [0,1]$. It is known that the intersection I_S is nonempty almost surely only if $1 \leq |S| \leq \ell_\beta$, where $\ell_\beta := \max\{k \in \mathbb{N} : k < (1-\beta)^{-1}\}$; see Samorodnitsky and Wang, 2019, Corollary B.3. An alternative but equivalent representation of $M(B)$ that is useful in proofs is given by

$$M(B) = \sup_{S\subset\mathbb{N}}\left(\mathbf{1}_{\{I_S \cap B \neq \emptyset\}} \sum_{j\in S} \Gamma_j^{-1/\alpha}\right), \quad B \in \mathcal{G}([0,1]), \tag{2.18}$$

where the supremum is almost surely attained at some (random) finite set S with $|S| \leq \ell_\beta$; see Samorodnitsky and Wang, 2019.

2.2 Heavy-tailed Factor Models

As observed by Einmahl et al., 2012 and Janßen and Wan, 2020, k-clustering algorithms can be naturally related to the estimation of certain factor-like models frequently encountered in the analysis of multivariate extremes. Let

$$B = (b_{ij})_{i=1,\dots,d;\, j=1,\dots,k} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_k \end{bmatrix},$$

where each $\mathbf{b}_j = (b_{1j}, \dots, b_{dj})^\top \in [0,\infty)^d$, with $j \in \{1, \dots, k\}$, are k distinct, nonzero column vectors. We assume that no row or column of B is identically zero (as such redundancy would allow reduction of d or k).

Let $\mathbf{Z} = (Z_1, \dots, Z_k)^\top$ be a vector of i.i.d. positive random variables with regularly varying tails:

$$\mathbb{P}(Z_1 > z) \sim z^{-\alpha}, \quad \text{as } z \to \infty, \quad \text{for some } \alpha \in (0, \infty).$$

We consider two models linking $\mathbf{X} \in \mathbb{R}^d$ to $\mathbf{Z}$:

1. Sum-linear model:

$$\mathbf{X} = B\mathbf{Z} = \left(\sum_{j=1}^{k} b_{1j} Z_j, \dots, \sum_{j=1}^{k} b_{dj} Z_j \right)^\top. \tag{2.19}$$

2. Max-linear model:

$$\mathbf{X} = B \odot \mathbf{Z} = \left(\bigvee_{j=1}^{k} b_{1j} Z_j, \dots, \bigvee_{j=1}^{k} b_{dj} Z_j \right)^\top, \tag{2.20}$$

where $\odot$ denotes the matrix operation with summation replaced by the maximum. Due to the exchangeability of the components of $\mathbf{Z}$, the distribution of $\mathbf{X}$ under either model is identifiable only up to a permutation of the columns of B; that is, for any permutation π on $\{1, \dots, k\}$, the distribution of $\mathbf{X}$ remains unchanged under $B \mapsto B_\pi = \begin{bmatrix} \mathbf{b}_{\pi(1)} & \cdots & \mathbf{b}_{\pi(k)} \end{bmatrix}$.

Both models (2.19) and (2.20) are known to satisfy multivariate regular variation (MRV), with a discrete spectral measure of the form

$$p_j = \frac{\|\mathbf{b}_j\|_{(r)}^\alpha}{\sum_{\ell=1}^k \|\mathbf{b}_\ell\|_{(r)}^\alpha}, \qquad \mathbf{a}_j = \frac{\mathbf{b}_j}{\|\mathbf{b}_j\|_{(s)}}, \qquad j = 1, \dots, k, \tag{2.21}$$

where $\|\cdot\|_{(r)}$ and $\|\cdot\|_{(s)}$ are norms used in the radial-angular decomposition. This result follows from the "single large jump" heuristic: when $\|\mathbf{X}\|_{(r)}$ is large, it is overwhelmingly due to a single large Z_j; see, e.g., Einmahl et al., 2012; Medina et al., 2024. These works often assume $\|\cdot\|_{(r)} = \|\cdot\|_{(s)}$ and $\alpha = 1$, but generalizations are straightforward.

To ensure marginal standardization as in (1.18) or (1.9), the following constraint must hold:

$$\sum_{j=1}^{k} b_{ij}^{\alpha} = 1, \qquad i = 1, \ldots, d. \tag{2.22}$$

The models can be further extended by including a noise component. For example, one may consider

$$\mathbf{X} = B\mathbf{Z} + \boldsymbol{\varepsilon} \quad \text{or} \quad \mathbf{X} = (B \odot \mathbf{Z}) \vee \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)^\top$ consists of i.i.d. positive noise terms. As long as the tails of ε_i are lighter than those of Z_j, the asymptotic properties remain valid; see Einmahl et al., 2012.

The transformed-linear model of Cooley and Thibaud, 2019 also falls within this framework. Importantly, when fitting such models in practice, one typically focuses on the extremal subset of the data (e.g., those observations exceeding a high threshold; see (1.19)), rather than the entire sample.

Chapter 3

Main Results

3.1 Theory

3.1.1 Joint Convergence of Processes with Finite Variance

For simplicity, we present our results on the unit interval $[0, 1]$; however, they can be extended to the half-line $[0, \infty)$ without essential modifications. Let $\mathcal{G}([0, 1])$ denote the open subsets of $[0, 1]$ with respect to the subspace topology inherited from $\mathbb{R}$.

Theorem 3.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a stationary process defined as in* (2.1) *with $f = \mathbf{1}_{A_0}$. Define the partial sum process $(S_n(t))_{t\in[0,1]}$ as in* (2.8) *and the partial maximum process $(M_n(B))_{B\in\mathcal{G}([0,1])}$ as in* (2.14)*. Suppose the regular variation conditions* (2.3) *and* (2.4) *hold for some $0 < \beta < 1$ and $\alpha \geq 2$, respectively. Additionally, assume the finite variance condition* (2.9) *and the regularity condition* (2.15) *are satisfied. Then, as $n \to \infty$, the following joint convergence holds:*

$$\begin{pmatrix} \frac{1}{nw_n^{-1/2}} \left(S_n(t)\right)_{t\in[0,1]} \\ \frac{1}{\rho^{\leftarrow}(w_n^{-1})} \left(M_n(B)\right)_{B\in\mathcal{G}([0,1])} \end{pmatrix} \Rightarrow \begin{pmatrix} c_\beta \left(B_H(t)\right)_{t\in[0,1]} \\ \left(\eta_{\alpha,\beta}(B)\right)_{B\in\mathcal{G}([0,1])} \end{pmatrix}$$

weakly in the product space $D[0, 1] \times SM[0, 1]$, where $D[0, 1]$ is equipped with the Skorokhod J_1-topology and $SM[0, 1]$ with the sup-vague topology. Here, w_n is defined by (2.2)*, $\rho^{\leftarrow}$ is as in* (2.7)*, c_β is the constant in* (2.10)*, B_H is a fractional Brownian motion with Hurst parameter $H = (1 + \beta)/2$ satisfying $\mathbb{E}[B_H(t)^2] = t^{2H}/2$, and $\eta_{\alpha,\beta}$ is the random sup measure defined in* (2.16)*.*

Furthermore, the processes B_H and $\eta_{\alpha,\beta}$ appearing in the limit are independent.

For simplicity, we do not treat the case $\alpha = 2$ when $\{X_n\}$ has infinite variance. This subtle case requires modifying the normalization $n^{-1}w_n^{1/2}$ by an additional slowly varying diverging factor. Aside from this adjustment, we expect the limit theorem to remain qualitatively similar to Theorem 3.1—specifically, the sum component should still converge to a fractional Brownian motion in the limit.

3.1.2 Joint Convergence of Processes with Infinite Variance

Theorem 3.2. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a stationary process defined by* (2.1) *with $f = \mathbf{1}_{A_0}$. Suppose* (2.3) *and* (2.4) *hold for $0 < \beta < 1$ and $0 < \alpha < 2$, respectively. Let $(S_n(t))_{t\in[0,1]}$ be defined in* (2.8) *and $M_n(B))_{B\in\mathcal{G}([0,1])}$ be defined in* (2.14)*. Assume that* (2.5) *holds for some $\alpha_0 < 2$ and* (2.15) *holds. Then*

$$\begin{pmatrix} \frac{1}{\rho^{\leftarrow}(w_n^{-1})nw_n^{-1}}(S_n(t))_{t\in[0,1]} \\ \frac{1}{\rho^{\leftarrow}(w_n^{-1})}(M_n(B))_{B\in\mathcal{G}([0,1])} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{1}{\Gamma(2-\beta)}C_\alpha^{-1/\alpha}(S(t))_{t\in[0,1]} \\ (M(B))_{B\in\mathcal{G}([0,1])} \end{pmatrix}$$

weakly in $D[0,1] \times SM[0,1]$ as $n \to \infty$, where w_n as in (2.2) *and $\rho^{\leftarrow}$ as in* (2.7)*, C_α is as in* (2.11)*, and the limit process S and random sup measure M are as in* (2.13) *and* (2.18)*.*

The following proposition demonstrates a key contrast with Theorem 3.1: the limiting processes in Theorem 3.2 are no longer independent.

Proposition 3.3. *Let $(S(t))_{t\in[0,1]}$ be defined as in* (2.13)*, and let $(M(B))_{B\in\mathcal{G}([0,1])}$ be defined as in* (2.18)*. Then $S(t)$ and $M(B)$ are dependent.*

3.1.3 Joint Convergence of Subordinators with Their Local Times and Ranges

We present a result stating that the weak convergence of strictly increasing subordinators implies the joint weak convergence of the subordinator, its associated local time process, and its closed range. For our purposes, we also incorporate an independent random shift into the framework.

Proposition 3.4. *Let σ and $\{\sigma_n\}_{n\in\mathbb{N}}$ be subordinators such that σ is strictly increasing on $[0,\infty)$ almost surely. Let V and $\{V_n\}_{n\in\mathbb{N}}$ be non-negative random variables. Suppose that σ_n is independent of V_n for each $n \in \mathbb{N}$, and σ is independent of V. Let L and L_n denote the local time processes* (2.12) *associated with σ and σ_n, respectively, and let $\mathcal{R}$ and $\mathcal{R}_n$ denote their closed ranges* (2.17)*. Assume that*

$$\sigma_n(1) \xrightarrow{d} \sigma(1) \quad \text{and} \quad V_n \xrightarrow{d} V \quad \text{as } n \to \infty.$$

Then, for any $t, y > 0$, the following joint weak convergence holds:

$$\begin{pmatrix} (V_n + \sigma_n(s))_{s\in[0,t]} \\ (L_n((x-V_n)_+))_{x\in[0,y]} \\ V_n + \mathcal{R}_n \end{pmatrix} \overset{d}{\Rightarrow} \begin{pmatrix} (V + \sigma(s))_{s\in[0,t]} \\ (L((x-V)_+))_{x\in[0,y]} \\ V + \mathcal{R} \end{pmatrix}$$

in $D[0,t] \times D[0,y] \times \mathfrak{F}([0,\infty))$, where the function spaces are equipped with the uniform and Fell topologies, respectively. In particular,

$$\begin{pmatrix} (\sigma_n(s))_{s\in[0,t]} \\ (L_n(x))_{x\in[0,y]} \\ \mathcal{R}_n \end{pmatrix} \overset{d}{\Rightarrow} \begin{pmatrix} (\sigma(s))_{s\in[0,t]} \\ (L(x))_{x\in[0,y]} \\ \mathcal{R} \end{pmatrix}$$

in $D[0,t] \times D[0,y] \times \mathfrak{F}([0,\infty))$.

3.1.4 Order Selection via Penalized Silhouette

Suppose W is a multiset on $\mathbb{S}_+^{d-1}$, and let $1 \le k \le |W| < \infty$. Consider a k-clustering of W, denoted by $A_k^* = \{\mathbf{a}_1^*, \dots, \mathbf{a}_k^*\}$ and $\mathfrak{C}_k = \{C_1, \dots, C_k\}$, with respect to a dissimilarity measure D as defined in Definition 2. For $\mathbf{w} \in W$, define

$$a(\mathbf{w}) = D(\mathbf{w}, A_k^*), \quad b(\mathbf{w}) = \max_{1\le i\le k} D(\mathbf{w}, A_k^* \setminus \{\mathbf{a}_i^*\}),$$

which respectively denote the dissimilarity of $\mathbf{w}$ to its closest center (i.e., the center of the cluster to which it belongs) and to its second-closest center. When $k = 1$, we define $b(\mathbf{w}) := 1$. The *(simplified) average silhouette width (ASW)* (Hruschka et al., 2004) of the clustering is then given by

$$\bar{S} = \bar{S}(W; A_k^*) = \frac{1}{|W|}\sum_{\mathbf{w}\in W} \frac{b(\mathbf{w}) - a(\mathbf{w})}{b(\mathbf{w})} = 1 - \frac{1}{|W|}\sum_{\mathbf{w}\in W} \frac{a(\mathbf{w})}{b(\mathbf{w})}. \quad (3.1)$$

A well-clustered dataset typically exhibits small $a(\mathbf{w})$ values relative to $b(\mathbf{w})$ for most $\mathbf{w} \in W$. Hence, $\bar{S}$ is commonly used to select the number of clusters k by maximizing it over a range of values. However, when applying ASW to multivariate extremes with a discrete spectral measure, we find that its performance may be unsatisfactory. Specifically, ASW tends to become insensitive when the number of clusters exceeds the true number k, i.e., the number of atoms in the spectral measure. In particular, we observe two problematic behaviors of ASW: 1. It often treats a small fraction of isolated points as a separate cluster. 2. It may split a single cluster center into multiple nearby centers.

Motivated by these observations, we propose adding a penalty term that discourages both small cluster sizes and small dissimilarities between cluster centers. Recall that for a k-clustering $(A_k^*, \mathfrak{C}_k)$ of a multiset W, the set $A_k^* = \{\mathbf{a}_1^*, \dots, \mathbf{a}_k^*\}$ consists of the cluster centers, and $\mathfrak{C}_k = \{C_1, \dots, C_k\}$ denotes the corresponding partition of W. Define

$$P_t = P_t(W; A_k^*, \mathfrak{C}_k) := 1 - \left(\min_{i=1,\dots,k} \frac{|C_i|}{|W|/k}\right)^t \left(\min_{1\le i<j\le k} D(\mathbf{a}_i^*, \mathbf{a}_j^*)\right)^t, \tag{3.2}$$

where $t \geq 0$ is a tuning parameter. For $k = 1$, we define the second minimum as 1. This penalty term discourages the formation of very small clusters and cluster centers that are too close together. Both quantities lie in $[0, 1]$, and the penalty increases as either quantity decreases. The *penalized ASW* is then defined as

$$\begin{aligned} &S_t(W; A_k^*, \mathfrak{C}_k) \\ :=&\bar{S} - P_t = \left(\min_{i=1,\dots,k} \frac{|C_i|}{|W|/k}\right)^t \left(\min_{1\le i<j\le k} D(\mathbf{a}_i^*, \mathbf{a}_j^*)\right)^t - \frac{1}{|W|}\sum_{\mathbf{w}\in W} \frac{a(\mathbf{w})}{b(\mathbf{w})}. \end{aligned}$$

When $t = 0$, we recover the original ASW: $S_0 = \bar{S}$. As t increases, the penalty term P_t grows and S_t decreases.

We now present a consistency result for using penalized ASW in selecting the number of clusters in multivariate extreme models with discrete spectral measures.

Theorem 3.5. *Suppose* $\mathbf{X}$ *satisfies conditions* (1.18) *and* (1.11)*, and has a discrete spectral measure of the form*

$$H = \sum_{i=1}^{k} p_i \delta_{a_i},$$

where $\mathbf{a}_i \in \mathbb{S}_+^{d-1}$ *are distinct and* $p_i > 0$ *with* $\sum_{i=1}^k p_i = 1$*. Let* W_n *denote the extremal subsample as in* (1.19)*, with* $\ell_n \to \infty$ *and* $\ell_n/n \to 0$*. Let* $(A_{m,n}, \mathfrak{C}_{m,n})$ *denote an* m*-clustering of* W_n *as in Definition 2, with respect to dissimilarity measure* D *from Definition 1. Let* r_A *be as defined in* (1.25)*, and define*

$$p_{\min} := \min_{1\le i\le k} p_i. \tag{3.3}$$

Then for any $t \in (0, t_0)$, *where*

$$t_0 := \frac{\ln(1 - r_A p_{\min})}{\ln(r_A k p_{\min})}, \tag{3.4}$$

we have for all $m \neq k$

$$\liminf_{n\to\infty} \left[S_t(W_n; A_{k,n}, \mathfrak{C}_{k,n}) - S_t(W_n; A_{m,n}, \mathfrak{C}_{m,n})\right] \geq \Delta_t \quad \textit{almost surely},$$

where $\Delta_t := (r_A k p_{\min})^t - (1 - r_A p_{\min}) > 0$ *for* $t \in (0, t_0)$.

This theorem implies that if the tuning parameter t is chosen within an appropriate range, then with high probability as $n \to \infty$, the penalized ASW is uniquely maximized at the true number of clusters k. In practice, we recommend plotting the penalized ASW S_t as a function of $m \in \{1, 2, \dots\}$ over a small range of t values. Start with $t \approx 0$ and gradually increase it. If spurious clusters (e.g., tiny clusters or clusters with very close centers) exist, the curve will typically bend or drop at the correct number of clusters. The "elbow" point in this plot can then be selected as the optimal value k. As a quick illustration, we follow a simulation setup of $(d = 6, k = 6)$ below to simulate a max-linear factor model. Penalized ASW S_t (vertical axis) for spherical k-means clustering is plotted as a function of test order m (horizontal axis); see Figure 3.1 . Increasing t to very large values is generally uninformative and not advised. Developing a data-driven method to select t remains an open and important direction for future research.

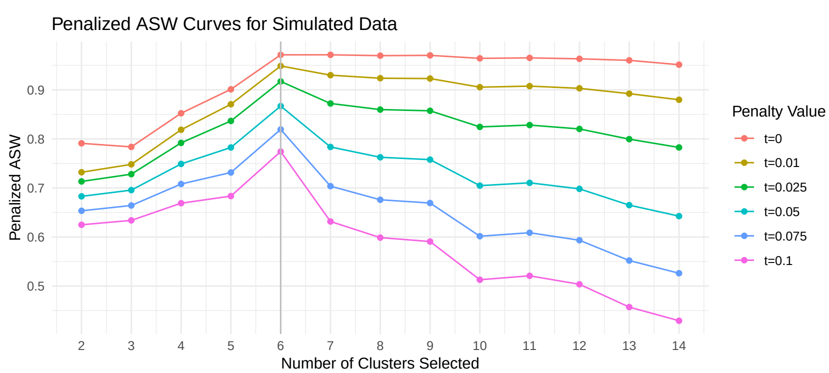

Figure 3.1: A simulation instance taken from $d = 6$, $k = 6$ setup.

3.1.5 Large Deviation Analysis of Clustering-based Spectral Estimation

In this section, we provide a quantitative refinement of the consistency result stated in Corollary 1.2 by deriving large-deviation-type bounds. This analysis builds on several key estimates developed in the proof of Theorem 3.5 and offers additional insight into the convergence behavior of the clustering-based estimator.

As a preliminary step, we establish a Chernoff–Hoeffding-type bound for the sum of a Binomial number of independent Bernoulli random variables. This result may be of independent interest and plays a central role in the subsequent probabilistic analysis.

Lemma 3.6. *Suppose B_i, $i \in \mathbb{Z}_+$, are independent Bernoulli random variables with $\mathbb{P}(B_i = 1) = q_1 \in (0,1)$ and N is a Binomial(n, q_2) random variable which is independent of B_i's, $n \in \mathbb{Z}_+$. Then we have for any $r \in (0, 1-q_1)$,*

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} B_i > q_1 + r\right) \le \exp\left\{nq_2\left[e^{-\mathcal{D}(q_1+r\|q_1)} - 1\right]\right\}$$
$$\le \exp\left\{nq_2\left(e^{-2r^2} - 1\right)\right\}, \tag{3.5}$$

and for any $r \in (0, q_1)$,

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} B_i < q_1 - r\right) \le \exp\left\{nq_2\left[e^{-\mathcal{D}(q_1-r\|q_1)} - 1\right]\right\}$$
$$\le \exp\left\{nq_2\left(e^{-2r^2} - 1\right)\right\}, \tag{3.6}$$

where $\mathcal{D}(x \parallel y) = x\ln(x/y) + (1-x)\ln\{(1-x)/(1-y)\}$ if $x, y \in (0,1)$ (the Kullback–Leibler divergence between two Bernoulli distributions). Here $\sum_{i=1}^{m} B_i/m$ is understood as 0 when $m = 0$.

Proof. We only prove the (3.5) and the proof of (3.6) is similar. It follows from a version of Hoeffding's inequality for Binomial (Hoeffding, 1963, Equation (2.1)) that for any $m \ge 0$,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} B_i > q_1 + r\right) \le e^{-m\mathcal{D}(q_1+r\|q_1)}.$$

Hence

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} B_i > q_1 + r\right) &\le \sum_{m=0}^{n}\binom{n}{m} q_2^m e^{-m\mathcal{D}(q_1+r\|q_1)}(1-q_2)^{n-m} \\
&= \left[q_2\left\{e^{-\mathcal{D}(q_1+r\|q_1)} - 1\right\} + 1\right]^n \\
&\le \exp\left\{nq_2\left[e^{-\mathcal{D}(q_1+r\|q_1)} - 1\right]\right\},
\end{aligned}$$

where in the last inequality we have used the inequality $x + 1 \le \exp(x)$, $x \in \mathbb{R}$. To obtain the second inequality in (3.5), it suffices to note that in view of Hoeffding, 1963, Equation (2.3), one has $\mathcal{D}(q_1 + r \parallel q_1) \ge 2r^2$. $\square$

Remark 3.1. *Note that when r is small, this simplified bound is approximately $\exp(-2nq_2r^2)$, a form identical to the usual Hoeffding's inequality (recall nq_2 is the effective sample size here).*

Let $H = \sum_{i=1}^{k} p_i\delta_{\mathbf{a}_i}$ be as defined in (1.23). Let $(A_{k,n}, \mathfrak{C}_{k,n})$, where $A_{k,n} = (\mathbf{a}_{1,n}^k, \ldots, \mathbf{a}_{k,n}^k)$ and $\mathfrak{C}_{k,n} = \left\{C_{1,n}^k, \ldots, C_{k,n}^k\right\}$, form a k-clustering of the extremal subsample W_n as in (1.19). By Corollary 1.2, there exists permutation π, such that $\mathbf{a}_{\pi_n(i),n}^k$ and $p_{\pi_n(i),n}^k$ in (1.24) are consistent estimators for $\mathbf{a}_i$ and p_i, respectively. Note that an accurate estimation can be interpreted as that for small $x, y > 0$, $D(\mathbf{a}_{\pi(i),n}^k, \mathbf{a}_i) < x$ and $|p_{\pi(i),n}^k - p_i| < y$ for all $i \in \{1, \ldots, k\}$. Now consider the complement "large deviation" event

$$E(x, y) = \bigcap_{\pi}\bigcup_{i=1}^{k}\left\{|\mathbf{a}_{\pi(i),n}^k - \mathbf{a}_i| > x\right\} \cup \left\{|p_{\pi(i),n}^k - p_i| > y\right\}. \qquad (3.7)$$

where the intersection $\cap_\pi$ is over all permutations $\pi : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$. We have the following result.

Proposition 3.7. *Suppose $\mathbf{X}$ satisfying* (1.18) *and* (1.11) *has a spectral measure of the following form $H = \sum_{i=1}^{k} p_i\delta_{\mathbf{a}_i}$, where $\mathbf{a}_i$'s are distinct points on $\mathbb{S}_+^{d-1}$, and $p_i > 0$, $p_1 + \cdots + p_k = 1$. Let W_n denote the extremal subsample as in* (1.19)*, and a k-clustering of W_n formed by $\left(A_{k,n} = (\mathbf{a}_{1,n}^k, \ldots, \mathbf{a}_{k,n}^k), \mathfrak{C}_{k,n} = \left\{C_{1,n}^k, \ldots, C_{k,n}^k\right\}\right)$ as defined in Definition 2 with respect to a dissimilarity measure D defined in Definition 1. Let $E(x, y)$ be the event defined in 3.7. Then for any $x, y > 0$,*

$$\limsup_{n} \frac{1}{c_{(r)}\ell_n} \ln \mathbb{P}(E(x, y)) \le \exp\left(-2\Delta(x, y)^2\right) - 1$$

where

$$\Delta(x, y) = \begin{cases} \max\{y/c_k, p_{min}x/(k + x)\}, & x < \epsilon_0, y < c_k p_{\min}\epsilon_0/(k + \epsilon_0), \\ p_{\min}\epsilon_0/(k + \epsilon_0), & otherwise, \end{cases}$$

where $\epsilon_0 := \sup\{\epsilon > 0 : r_A > \epsilon + r_A^\dagger(\epsilon)\}$ *and* $c_k := (k \vee 2 - 1)$.

Proof. If $H_n(B_D(\mathbf{a}_i,\epsilon)) = |W_n \cap B_D(\mathbf{a}_i,\epsilon)|/|W_n| \geq p_i - \delta$ for all $i \in \{1,\dots,k\}$, by Lemmas A.8 and A.10, as long as (A.28) holds, there exists a permutation $\pi : \{1,\dots,k\} \mapsto \{1,\dots,k\}$, such that $D(\mathbf{a}^k_{\pi(i),n}, \mathbf{a}_i) < \epsilon'$ and $|p^k_{\pi(i),n} - p_i| \leq c_k\delta$ for all $i \in \{1,\dots,k\}$. Hence under (A.28), whenever $\epsilon' \leq x$ or $c_k\delta \leq y$,

$$\begin{aligned}\mathbb{P}(E(x,y)) &\leq \mathbb{P}\left(\bigcap_{\pi}\bigcup_{i=1}^{k}\left\{D(\mathbf{a}^k_{\pi(i),n},\mathbf{a}_i) > \epsilon'\right\} \cup \left\{|p^k_{\pi(i),n} - p_i| > c_k\delta\right\}\right)\\ &\leq \mathbb{P}\left(\bigcup_{i=1}^{k}\{H_n(B_D(\mathbf{a}_i,\epsilon)) < p_i - \delta\}\right),\end{aligned}$$

where H_n is the empirical spectral measure in (1.22). Observe that for any $i \in \{1,\dots,k\}$,

$$\begin{aligned}&\left(|W_n|, \left(\mathbf{1}\{\mathbf{X}_j/\|\mathbf{X}_j\|_{(s)} \in B_D(\mathbf{a}_i,\epsilon),\ \|\mathbf{X}_j\|_{(r)} \geq (n/\ell_n)^{1/\alpha}\}\right)_{j=1,\dots,n}\right)\\ &\stackrel{d}{=} (N, (B_j)_{j=1,\dots,n}),\end{aligned}$$

where N and B_j's are as in Lemma 3.6 with respective parameters q_1 and q_2 given as follows:

$$\begin{aligned}q_1 = q_1(i,\epsilon,n) :=& \frac{\mathbb{P}\left(\mathbf{X}_1/\|\mathbf{X}_1\|_{(s)} \in B_D(\mathbf{a}_i,\epsilon),\ \|\mathbf{X}_1\|_{(r)} \geq (n/\ell_n)^{1/\alpha}\right)}{\mathbb{P}\left(\|\mathbf{X}_1\|_{(r)} \geq (n/\ell_n)^{1/\alpha}\right)}\\ \to& p_i \qquad (3.8)\end{aligned}$$

as $n \to \infty$, where the last convergence holds due to (1.12) and the fact that $B_D(\mathbf{a}_i,\epsilon)$'s are disjoint under $\epsilon < r_A$, and

$$q_2 = q_2(n) = \mathbb{P}\left(\|\mathbf{X}_1\|_{(r)} \geq (n/\ell_n)^{1/\alpha}\right) \sim c_{(r)}(n/\ell_n) \qquad (3.9)$$

as $n \to \infty$. Now applying Lemma 3.6, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{k}\{H_n(B_D(\mathbf{a}_i,\epsilon)) < p_i - \delta\}\right) &\leq \sum_{i=1}^{k}\mathbb{P}\left(H_n(B_D(\mathbf{a}_i,\epsilon)) < p_i - \delta\right)\\ &\leq k\exp\left(nq_2(n)\left[\exp\{-2\delta^2\} - 1\right]\right).\end{aligned} \qquad (3.10)$$

Therefore in view of also (3.8) and (3.9), we have

$$\limsup_n \frac{1}{\ell_n} \ln \mathbb{P}\left(E(x,y)\right) \le c_{(r)} \left\{\exp(-2\delta^2) - 1\right\}.$$

The next step is to determine the largest value of δ as possible. Recall $\epsilon_0 = \sup\{\epsilon' > 0 : r_A > \epsilon' + r_A^\dagger(\epsilon')\}$. Then when $\epsilon' \in (0, \epsilon_0)$, for all ϵ small enough we have $r_A > \epsilon' + 2r_A^\dagger(\epsilon) + r_A^\dagger(\epsilon')$, namely, (A.28) holds. Hence by taking $\epsilon \downarrow 0$ in (A.25), we get from $\epsilon' < \epsilon_0$ the restriction $\delta < p_{\min}\epsilon_0/(k+\epsilon_0)$. Similarly, from $\epsilon' \le x$ we get the restriction $\delta < p_{min}x/(k+x)$. In addition, from $c_k\delta \le y$ we get the restriction $\delta \le y/c_k$. At least one of the last two conditions should be satisfied. Therefore,

$$\begin{cases} \delta < p_{\min}\epsilon_0/(k+\epsilon_0), & \text{if } x \ge \epsilon_0, \\ \delta < p_{\min}\epsilon_0/(k+\epsilon_0), & \text{if } x < \epsilon_0, y \ge c_k p_{\min}\epsilon_0/(k+\epsilon_0), \\ \delta < \max\{y/c_k, p_{min}x/(k+x)\}, & \text{if } x < \epsilon_0, y < c_k p_{\min}\epsilon_0/(k+\epsilon_0). \end{cases}$$

The result then follows. □

Remark 3.2. *The large-deviation-type estimates in Proposition 3.7 show that the probability $\mathbb{P}(E(x,y))$ decays exponentially in the expected extremal subsample size $c_{(r)}\ell_n$. Notably, the structure of the deviation function $\Delta(x,y)$ reveals important qualitative insights: the difficulty of the clustering-based estimation—as reflected in the rate of decay of these error probabilities—depends negatively on $p_{\min}$ and r_A (since ϵ_0 increases with r_A), and positively on the number of clusters k. In other words, the estimation becomes more accurate when the true discrete spectral measure has fewer atoms, the atoms are more well-separated under the dissimilarity measure D, or the mass associated with each atom is relatively large.*

We also have the following result which states that in the context of Theorem 3.5, the probability of false order election tends to 0 exponentially fast.

Proposition 3.8. *Suppose $\mathbf{X}$ satisfying* (1.18) *and* (1.11) *has a discrete spectral measure of the form $H = \sum_{i=1}^k p_i\delta_{\mathbf{a}_i}$, $k \in \mathbb{Z}_+$ where $\mathbf{a}_i$'s are distinct points on $\mathbb{S}_+^{d-1}$, and $p_i > 0$, $p_1 + \cdots + p_k = 1$. Let W_n denote the extremal subsample as in* (1.19)*, and $(A_{m,n}, \mathfrak{C}_{m,n})$, $m \in \mathbb{Z}_+$, form an m-clustering of W_n as defined in Definition 2 with respect to a dissimilarity measure D defined in Definition 1. Let r_A be defined as in* (1.25)*, $p_{\min}$ be defined as in* (3.3) *and t_0 be defined as in*

(3.4). *Then fix* $t \in (0, t_0)$,

$$\limsup_n \frac{\ln \mathbb{P}\left(\{S_t(W_n; A_{k,n}, \mathfrak{C}_{k,n}) \leq S_t(W_n; A_{m,n}, \mathfrak{C}_{m,n}) \textit{ for all } m \neq k\}\right)}{c_{(r)}\ell_n}$$
$$\leq \exp\left(-2\delta_t(k, p_{\min}, r_A)^2\right) - 1$$

where $\delta_t(k, p_{\min}, r_A) > 0$ *is the solution* δ *of the equation* $[k(p_{\min} - \delta)r_A]^t - k\delta = (k^2\delta)^t \vee ([1 - (p_{\min} - \delta)r_A]\mathbf{1}\{k \geq 2\})$.

Proof. Writing $S_t(m) = S_t(W_n; A_{m,n}, \mathfrak{C}_{m,n})$, we have

$$\mathbb{P}\left(\{S_t(k) \leq S_t(m),\ m \neq k\}\right) \leq \mathbb{P}\left(\{S_t(k) \leq S_t(m),\ m \neq k\} \cap E_n(\epsilon, \delta)\right) + \mathbb{P}\left(E_n(\epsilon, \delta)^c\right),$$

where $E_n(\epsilon, \delta)$ is in (A.31). Combining the inequalities regarding $\bar{S}$ in the proof of Proposition A.11, and the inequalities regarding P_t in the proof of Proposition A.12, the event in the first probability on the right-hand side above is empty as long as $\delta > 0$ satisfies

$$[k(p_{\min} - \delta)r_A]^t - k\delta > (k^2\delta)^t \vee ([1 - (p_{\min} - \delta)r_A]\mathbf{1}\{k \geq 2\})$$

and ϵ is sufficiently small (depending on δ). Note that the inequality above holds when δ is sufficiently small due to $0 < t < t_0 = \ln(1 - r_A p_{\min})/\ln(r_A k p_{\min})$, and its left-hand side is decreasing (to negative values) and its right-hand side is increasing with as δ increases to $p_{\min}$. Then for any $\delta \in (0, \delta_t(k, p_{\min}, r_A))$, we have in view of (3.9) and (3.10) that

$$\lim_n \frac{1}{c_{(r)}\ell_n} \ln \mathbb{P}\left(\{S_t(k) \leq S_t(m),\ m \neq k\}\right) \leq \exp(-2\delta^2) - 1.$$

The proof is concluded by letting $\delta \uparrow \delta_t(k, p_{\min}, r_A)$. $\square$

3.2 Methods

3.2.1 Order Selection and Coefficient Estimation

Due to the discrete nature of the spectral measure, the likelihood functions corresponding to the models (2.19) and (2.20) are not accessible; see, e.g., Einmahl et al., 2012; Yuen and Stoev, 2014. In particular, even outside the framework of extremes, the max-linear model (2.20) lacks a smooth density, rendering standard likelihood-based model selection tools, such as information criteria, inapplicable.

Nevertheless, these factor-based models (including (2.19) and (2.20)) possess a discrete spectral measure of the form (1.23). Therefore, the penalized ASW criterion offers a viable alternative for selecting the number of factors k, with consistency established in Theorem 3.5.

From this point on, we assume that the order k is known. A central question is whether the spectral measure—estimated via k-clustering—can be translated into an estimate of the coefficient matrix $B = (\mathbf{b}_1, \ldots, \mathbf{b}_k)$ in the sum-linear or max-linear model, while satisfying the marginal constraint (2.22).

Combining the representation (2.21) for the spectral measure with the constraint (2.22), we obtain a system of $kd + d - 1$ equations for the kd unknowns in B: specifically, $(k-1)$ equations for the weights p_j (due to their sum being one), $(d-1)k$ equations from the unit-norm conditions on the $\mathbf{a}_j$, and d equations from (2.22). Hence, the system is typically overdetermined. As a result, when p_j and $\mathbf{a}_j$ are estimated via clustering, the system may not admit an exact solution, although the relations are asymptotically satisfied (see Corollary 1.2).

To address this, we propose a simple and consistent method to recover an estimate of B that satisfies the marginal constraint (2.22). Observe that for both models (2.19) and (2.20), the exponent measure Λ concentrates on rays $\{t\mathbf{b}_j : t > 0\}$, $j = 1, \ldots, k$. Each spectral mass point $\mathbf{a}_j = \mathbf{b}_j / \|\mathbf{b}_j\|_{(s)}$ on the $\|\cdot\|_{(s)}$-norm sphere corresponds to the point $\mathbf{b}_j / \|\mathbf{b}_j\|_\alpha = \mathbf{a}_j / \|\mathbf{a}_j\|_\alpha$ on the α-norm sphere. An appealing property of using the α-norm is the following identity, derived from (2.22):

$$\sum_{j=1}^{k} \|\mathbf{b}_j\|_\alpha^\alpha = \sum_{i=1}^{d} \sum_{j=1}^{k} b_{ij}^\alpha = d.$$

Thus, under the choice $\|\cdot\|_{(r)} = \|\cdot\|_\alpha$ in (2.21), we obtain the direct relation: $p_j d = \|\mathbf{b}_j\|_\alpha^\alpha$, which leads to the expression:

$$\mathbf{b}_j = (p_j d)^{1/\alpha} \cdot \frac{\mathbf{a}_j}{\|\mathbf{a}_j\|_\alpha}, \qquad j = 1, \ldots, k. \tag{3.11}$$

In particular, if the angular norm $\|\cdot\|_{(s)}$ is chosen as the α-norm (i.e., $\|\cdot\|_{(s)} = \|\cdot\|_\alpha$), then $\|\mathbf{a}_j\|_\alpha = 1$ and the expression above simplifies accordingly.

This suggests a natural procedure: estimate p_j and $\mathbf{a}_j$ via k-clustering on the α-norm sphere, plug them into (3.11) to obtain preliminary estimates $\widehat{\mathbf{b}}_j$, and collect them into the matrix $\widehat{B} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_k)$. Let $\widehat{B} = (\mathbf{r}_1^\top, \ldots, \mathbf{r}_d^\top)^\top$, where $\mathbf{r}_i^\top$ denotes the ith row of $\widehat{B}$. To enforce the marginal constraint (2.22), we rescale each row to unit α-norm:

$$\widetilde{\mathbf{r}}_i = \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_\alpha}, \qquad i = 1, \ldots, d.$$

The resulting matrix $\widetilde{B}$, with rows $\widetilde{\mathbf{r}}_i^\top$, satisfies (2.22) by construction.

Finally, by Corollary 1.2 and a continuous mapping argument, this procedure yields a consistent estimator of the coefficient matrix B, up to permutation of the columns.

3.3 Simulations

In this section, we present simulation studies to illustrate the performance of the penalized ASW criterion introduced in Section 3.1.4. Following the setup of Janßen and Wan, 2020, Section 4, we simulate data from the max-linear factor model (2.20), with randomly generated coefficient matrices B. The latent factors Z_j are taken to be i.i.d. standard Fréchet random variables with shape parameter $\alpha = 1$.

We consider four different combinations of data dimension d and true number of factors k. For each combination, the coefficient vectors $\mathbf{b}_j$ are generated according to the procedures described below. Due to the standardization condition (2.22), it suffices to specify $\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}$. Let $\{U_i\}$ denote i.i.d. uniform random variables on $[0, 1]$:

- $d = 4$, $k = 2$: $\mathbf{b}_1 = (U_1, U_2, U_3, U_4)^\top/2$.
- $d = 4$, $k = 6$:

$$\begin{aligned} &\mathbf{b}_1 = (U_1, U_2, U_3, U_4)^\top/3, \quad \mathbf{b}_2 = (U_5, 0, U_6, 0)^\top/3, \\ &\mathbf{b}_3 = (0, U_7, 0, U_8)^\top/3, \quad \mathbf{b}_4 = (U_9, U_{10}, 0, 0)^\top/3, \\ &\mathbf{b}_5 = (0, 0, U_{11}, U_{12})^\top/3. \end{aligned}$$

- $d = 6$, $k = 6$:

$$\begin{aligned} &\mathbf{b}_1 = (U_1, \ldots, U_6)^\top/3, \quad \mathbf{b}_2 = (U_7, 0, U_8, 0, U_9, 0)^\top/3, \\ &\mathbf{b}_3 = (0, U_{10}, 0, U_{11}, 0, U_{12})^\top/3, \quad \mathbf{b}_4 = (U_{13}, U_{14}, U_{15}, 0, 0, 0)^\top/3, \\ &\mathbf{b}_5 = (0, 0, 0, U_{13}, U_{14}, U_{15})^\top/3. \end{aligned}$$

- $d = 10$, $k = 6$:

$$\begin{aligned}
&\mathbf{b}_1 = (U_1, \ldots, U_{10})^\top/2, \quad \mathbf{b}_2 = (U_{11}, U_{12}, 0, \ldots, 0)^\top/2,\\
&\mathbf{b}_3 = (0, 0, U_{13}, U_{14}, 0, \ldots, 0)^\top/2,\\
&\mathbf{b}_4 = (0, \cdots, 0, U_{15}, U_{16}, 0, \cdots, 0)^\top/2,\\
&\mathbf{b}_5 = (0, \ldots, 0, U_{17}, U_{18}, U_{19}, U_{20})^\top/2.
\end{aligned}$$

For each of the above configurations, we generate 100 independent random coefficient matrices B. From each simulated model, we draw a dataset of size 10,000. We then extract the 1,000 observations with the largest ℓ_2 norms, project them onto the ℓ_2 unit sphere (i.e., we set $\|\cdot\|_{(r)} = \|\cdot\|_{(s)} = \|\cdot\|_2$), and apply spherical clustering and penalized ASW analysis on this extremal subsample.

For clustering, we use the spherical k-means algorithm from the R package `skmeans` (Hornik et al., 2012), and the k-principal directions clustering (k-PC) algorithm using the implementation from the supplementary material of Fomichov and Ivanovs, 2023.

Figures 3.2 through 3.5 visualize the results for each (d, k) setup. Each matrix plot corresponds to one setup and consists of 100 columns (one for each simulation). The upper half of each plot shows results for spherical k-means; the lower half shows those for k-PC. Each row within these halves corresponds to a different value of the penalty parameter t. A cell's color indicates the estimated number of clusters m that maximized the penalized ASW: White indicates correct order identification ($m = k$). Red shades indicate underestimation ($m < k$). Blue shades indicate overestimation ($m > k$). A bar plot to the right of each matrix summarizes the proportion of simulations (out of 100) where the estimated order matched the true order.

Across all setups, we observe that the unpenalized ASW ($t = 0$) tends to overestimate the number of factors—sometimes substantially. As the penalty parameter t increases, this overestimation bias is significantly corrected, and the success rate of correct order identification improves markedly over a range of moderate t values. It is important to note that the success rate shown for each t is aggregated over all 100 datasets using that fixed t. As discussed in Section 3.1.4, further improvements may be achieved by adaptively selecting t per dataset using visual inspection of the ASW curves.

We also find that the k-PC method generally outperforms spherical k-means in terms of order identification accuracy. One particularly challenging scenario is the $(d = 4, k = 6)$ configuration, where the cluster centers often lie in close proximity, making the estimation problem more difficult.

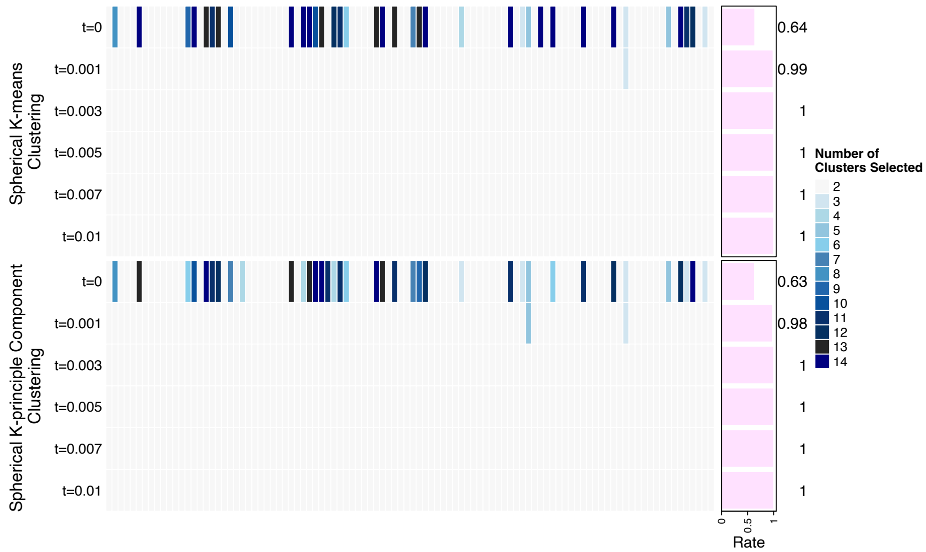

Figure 3.2: Simulation result visualization for the setup $d = 4, k = 2$.

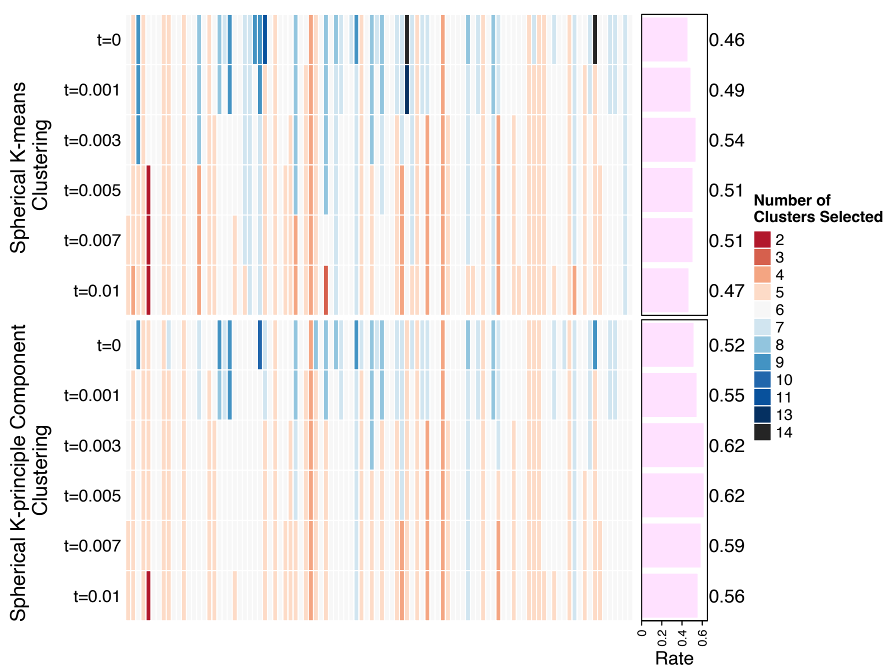

Figure 3.3: Simulation result visualization for the setup $d = 4, k = 6$.

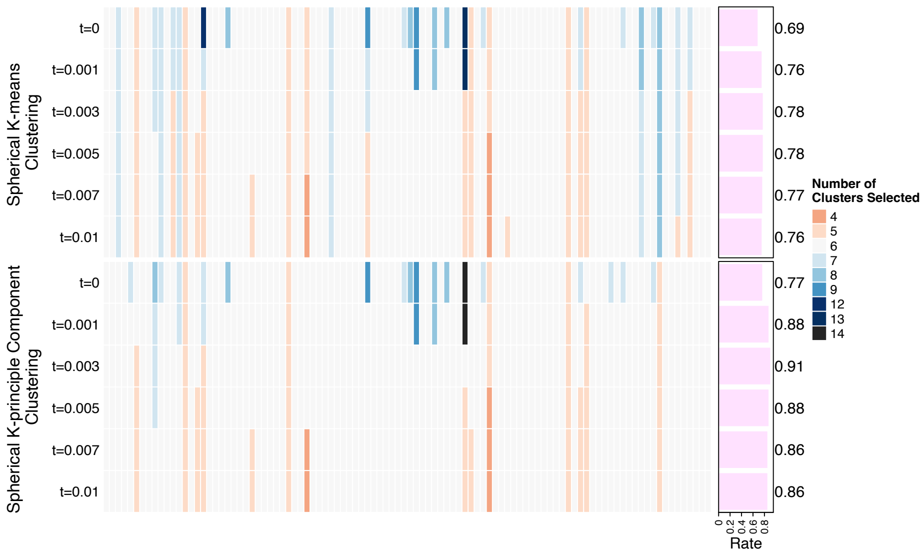

Figure 3.4: Simulation result visualization for the setup $d = 6, k = 6$.

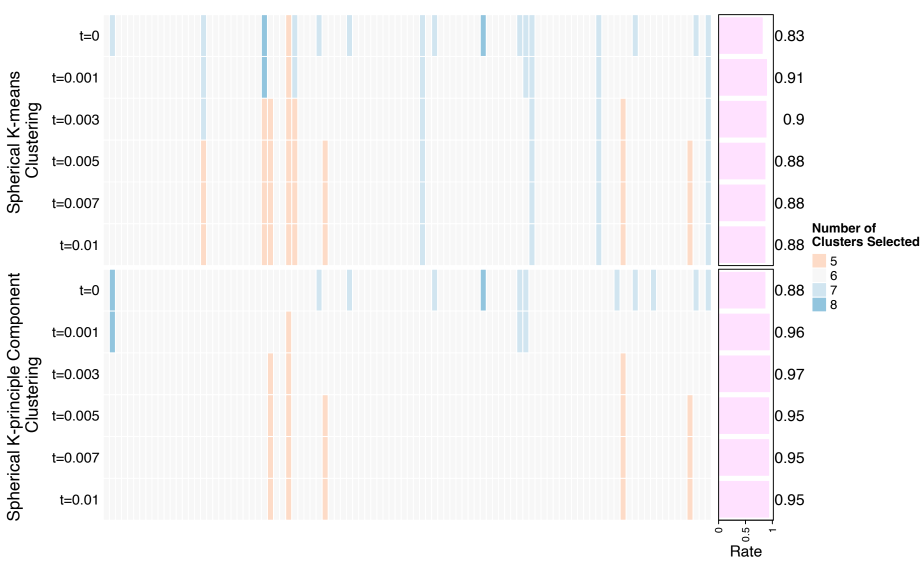

Figure 3.5: Simulation result visualization for the setup $d = 10, k = 6$.

Chapter 4

Applications

We demonstrate the application of the penalized ASW method, introduced in Section 3.1.4, for selecting the number of factors k, as well as the conversion of clustering-based spectral estimates into factor coefficient matrices, as discussed in Section 3.2.1. The analysis is conducted using the spherical k-principal directions clustering (k-PC) algorithm, where the dissimilarity measure D is defined as in (1.16).

We focus exclusively on the k-PC method for two main reasons. First, the simulation results in Section 3.3 suggest that penalized ASW combined with k-PC yields superior empirical performance in identifying the correct order k. Second, as argued in Fomichov and Ivanovs, 2023, the k-PC algorithm is particularly well-suited for detecting groups of concomitant extremes—that is, subsets of variables that tend to become simultaneously large. This property allows for more meaningful comparisons between the selected order k and external or domain-specific knowledge about the underlying data structure.

Let the observed dataset be denoted by $\{\mathbf{x}_i\}_{i=1}^n$, where each observation $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top \in [0, \infty)^d$. As a preprocessing step, we marginally standardize the data to roughly satisfy the standardization condition (1.9) with tail index $\alpha = 2$. Specifically, for each margin $j \in \{1, \ldots, d\}$, we define the empirical distribution function as

$$\hat{F}_j(x) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_{ij} < x\},$$

which ensures that $\hat{F}_j(x_{ij}) < 1$ for all i. The standardized data is then given by the transformed vectors $\{\widetilde{\mathbf{x}}_i\}_{i=1}^n$, where

$$\widetilde{x}_{ij} := \left\{-\log\left[\hat{F}_j(x_{ij})\right]\right\}^{-1/2}.$$

If $\hat{F}_j$ were the true marginal CDF, then $\widetilde{x}_{ij}$ would approximately follow a standard Fréchet distribution with tail index $\alpha = 2$.

To identify clusters of extremal behavior, we adopt a procedure analogous to that used in the simulation study of Section 3.3. We select the top 10% of the transformed observations $\{\widetilde{\mathbf{x}}_i\}$ with the largest ℓ_2-norms. This extremal subsample is then projected onto the ℓ_2-unit sphere. That is, we work with norms $\|\cdot\|_{(r)} = \|\cdot\|_{(s)} = \|\cdot\|_2$, ensuring that subsequent clustering is performed on the angular component of the extreme observations.

This setup enables us to apply the penalized ASW procedure to estimate the number of extremal dependence components and, via the methodology in Section 3.2.1, recover an interpretable factor representation of the multivariate extremes.

4.1 Air Pollution Data

We illustrate our methodology using an air pollution dataset from the R package `texmex` (Southworth et al., 2024), originally published as supplementary material to Heffernan and Tawn, 2004. The dataset contains daily measurements of air pollutant levels recorded in Leeds, U.K., city center between 1994 and 1998. The data are separated into two subsets based on seasonality: the summer dataset includes 578 observations recorded from April to July, the winter dataset includes 532 observations recorded from November to February.

Each observation consists of the daily maximum concentrations of five pollutants: Ozone, NO2, NO, SO2, and PM10. These data were previously analyzed in Janßen and Wan, 2020, where spherical k-means clustering was applied to the analysis of multivariate extremes. In our analysis, we apply the spherical k-PC clustering algorithm from Fomichov and Ivanovs, 2023, in conjunction with the penalized ASW criterion introduced in Section 3.1.4. The extremal subsample is constructed by selecting the top 10% of transformed data (as described in the beginning of this section) with respect to their ℓ_2-norms and projecting them onto the ℓ_2-unit sphere.

Figures 4.1 and 4.3 display the penalized ASW values as functions of the number of clusters m, with curves corresponding to various values of the tuning parameter t. Following the visual selection procedure discussed in Section 3.1.4, we identify the optimal order as $k = 5$ for the summer data and $k = 3$ for the winter data. For the summer data, $k = 4$ may also be a reasonable choice, suggesting mild ambiguity. These results are consistent with, though slightly different from, the order selections reported in Janßen and Wan, 2020, where $k = 5$ was chosen for the summer data and $k = 4$ for the winter data using elbow plot diagnostics (see Janßen and Wan, 2020, Fig. 1). It is worth noting that $k = 3$ for the winter data also appears plausible from their elbow plots. One explanation for the slight discrepancy lies in the clustering algorithm used: we employ the spherical k-PC method, whereas Janßen and Wan, 2020 applied spherical k-means.

To further interpret the clustering results, we estimate the factor coefficient matrix B based on the selected order k, using the procedure from Section 3.2.1, with $\|\cdot\|_{(r)} = \|\cdot\|_{(s)} = \|\cdot\|_2$ and tail index $\alpha = 2$. The resulting coefficient matrices are visualized in Figures 4.2 and 4.4.

For the summer data (Figure 4.2), the estimated factor directions exhibit sharp alignment with coordinate axes, suggesting a near-independence structure in the tail behavior of the five pollutants. This pattern aligns with the concept of asymptotic independence, which is common in environmental data and discussed in Beirlant et al., 2006, Chapter 8.

In contrast, for the winter data (Figure 4.4), one of the estimated factors clearly captures a cluster involving NO, NO2, and PM10, indicating asymptotic dependence among these three pollutants. This observation is consistent with the findings of Heffernan and Tawn, 2004, who also reported dependence between these variables. Thus, the identified factor structure for the winter dataset provides empirical support for the selected order $k = 3$, which groups these pollutants into a common extremal component.

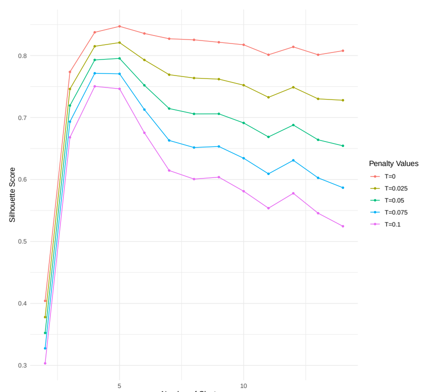

Figure 4.1: Penalized ASW curves for summer air pollution data (top 10% norms).

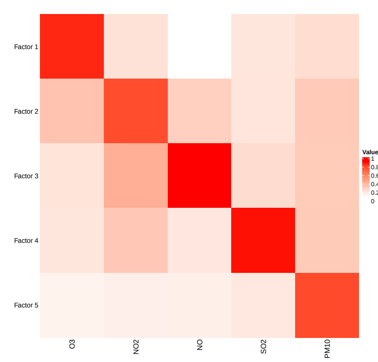

Figure 4.2: Estimated $B^\top$ for summer pollution data (top 10% norms).

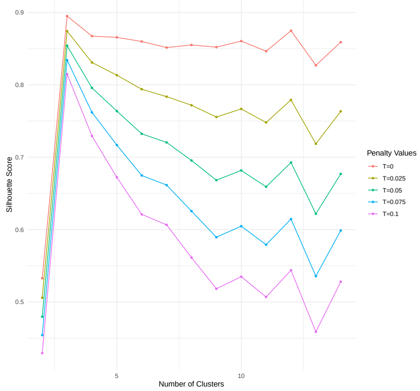

Figure 4.3: Penalized ASW curves for winter air pollution data (top 10% norms).

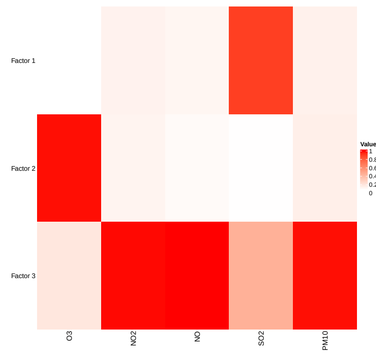

Figure 4.4: Estimated $B^\top$ for winter pollution data (top 10% norms).

4.2 River Discharge Data

The river discharge dataset comprises 16,386 daily discharge records from 13 monitoring stations across North America, obtained from the Global Runoff Data Centre German Federal Institute of Hydrology, n.d. These stations, listed in Table 4.1 and mapped in Fig. 4.5, are situated along five major rivers: the Willamette, Mississippi, Williamson, Hudson, and Broad Rivers.

Table 4.1: Clustering of 13 river discharge stations based on concomitant Extremes.

Station Name	River Name	Factor (Cluster) Index
SALEM, OR	WILLAMETTE RIVER	4
PORTLAND, OR	WILLAMETTE RIVER	4
HARRISBURG, OR	WILLAMETTE RIVER	4
ST.PAUL, MN	MISSISSIPPI RIVER	1
AITKIN, MN	MISSISSIPPI RIVER	1
THEBES, IL	MISSISSIPPI RIVER	6
CHESTER, IL	MISSISSIPPI RIVER	6
CHILOQUIN, OR	WILLIAMSON RIVER	2
GREEN ISLAND, NY	HUDSON RIVER	5
FORT EDWARD, NY	HUDSON RIVER	5
NORTH CREEK, NY	HUDSON RIVER	5
NEAR CARLISLE, SC	BROAD RIVER	3
NEAR BELL, GA	BROAD RIVER	3

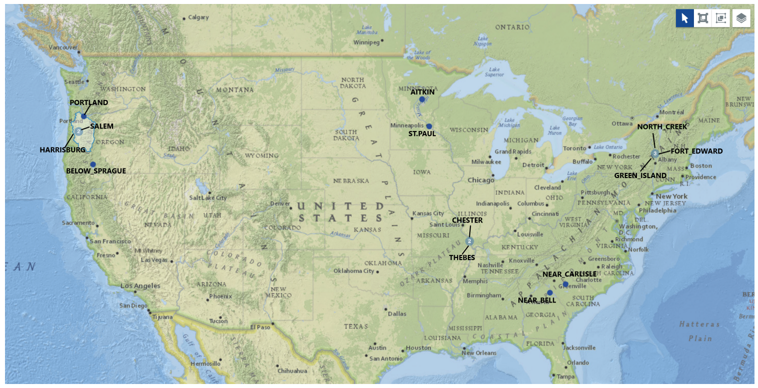

Figure 4.5: Geographical locations of the 13 river discharge stations.

Following the approach described in Section 3.2.1, Fig. 4.6 displays the penalized average silhouette width (ASW) curves, from which an optimal factor order of 6 is suggested. The corresponding factor matrix B, derived from spectral estimation using $\|\cdot\|_{(s)} = \|\cdot\|_{(r)} = \|\cdot\|_2$ and $\alpha = 2$, is visualized in Fig. 4.7. For each row of B, we identify the factor index (equivalently, the cluster index in Fig. 4.7) corresponding to the maximum value. These indices are reported in the final column of Table 4.1, providing a rough categorization of stations based on groups of concomitant extremes.

The clustering results show strong agreement with geographical intuition: stations along the same river are generally grouped together. A notable exception involves the four stations on the Mississippi River, which are split into two distinct clusters. This division is geographically coherent, as the stations fall into two widely separated regions—Minnesota (MN) and Illinois (IL)—justifying the observed partition.

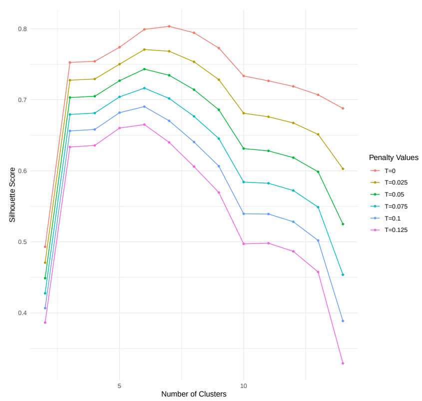

Figure 4.6: Penalized ASW curves for river discharge data (top 10% norms).

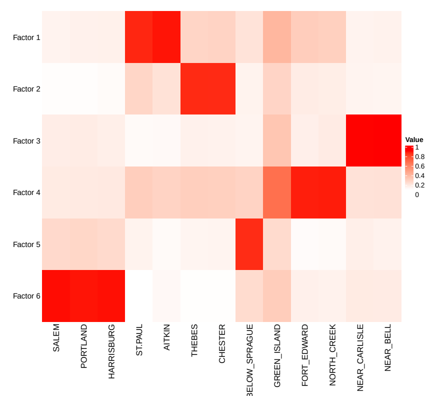

Figure 4.7: Estimated $B^\top$ for river discharge data (top 10% norms).

Chapter 5

Conclusion and Future Work

In this thesis, we established joint sum-and-max limit theorems for a class of stationary infinitely divisible processes exhibiting both long-range dependence and heavy tails. Our results reveal a striking dichotomy depending on whether the marginal variance is finite or infinite. In the finite-variance case, we proved that the normalized partial sum and maximum processes converge jointly to a pair of asymptotically independent limit objects: a fractional Brownian motion and a random sup measure. In contrast, in the infinite-variance case, we demonstrated the emergence of asymptotic dependence in the limit, where the sum converges to a symmetric α-stable process and the maximum to a random sup measure, with the dependence structure intricately captured via the local time and range of a stable subordinator.

Our methodology includes a novel joint convergence result for subordinators, their local times, and ranges, which we expect to be of independent interest. This technical development was essential to describing the dependent limit structure in the infinite-variance regime.

Looking ahead, several extensions of this work are worth exploring. A natural question is whether similar joint convergence results hold in more general dependence frameworks beyond the null-recurrent Markov chain model. Moreover, understanding the impact of different normalization schemes and generalizing our framework to include non-symmetric or non-stationary settings are promising directions for future investigation.

Following recent developments in the literature, we investigate the estimation of multivariate extreme value models with a discrete spectral measure using spherical clustering techniques. Our primary contribution is a novel method for selecting the model order—that is, the number of clusters—that consistently recovers the true number of spectral atoms. This is achieved by augmenting the widely used simplified average silhouette width (ASW) with an additional penalty term that discourages small cluster sizes and small dissimilarities between cluster centers. As a by-product, our approach also facilitates order selection in max-linear factor models. The proposed method is simple to implement and demonstrates strong empirical performance.

In addition, we carry out a large-deviation-type analysis for the estimation of discrete spectral measures via clustering. This analysis sheds light on the convergence behavior of clustering-based estimators in the multivariate extremes setting. We further illustrate how these estimators can be employed for parameter inference in heavy-tailed factor models.

Finally, we outline several directions for future research. First, the tuning parameter t in the penalty term (3.2) is currently selected through visual inspection. Developing a data-driven approach for selecting t would be valuable and may require a deeper understanding of the consistency result in Theorem 3.5. Second, alternative clustering evaluation criteria to ASW, such as the cross-validation method based on algorithmic instability proposed by Wang, 2010, could offer promising alternatives in the context of multivariate extremes. Third, it remains an open problem to determine an appropriate threshold l_n in (1.19) when clustering extreme observations; the methodology developed in Wan and Davis, 2019 may offer useful insights here.

Appendix A

Proofs

A.1 Finite Variance Case

Proof of Theorem 3.1. Observe that tightness in the product space follows if tightness holds in each marginal space. The weak convergence—and hence tightness—of $n^{-1}w_n^{1/2}S_n$ in $D[0,1]$ follows from Samorodnitsky, 2016, Theorem 9.4.7. On the other hand, the normalized empirical sup measure $(\rho^{\leftarrow}(w_n^{-1}))^{-1}M_n$ is automatically tight since the space $SM[0,1]$ is compact (Vervaat, 1988, Theorem 4.2). By Proposition A.6 and Remark A.1, it remains to verify convergence of the finite-dimensional distributions (fdd) in $[0,1] \times \mathcal{I}$, where $\mathcal{I}$ denotes the collection of all non-empty open subintervals of $[0,1]$, in the sense of Definition 4.

Fix $m > 0$. As in Samorodnitsky and Wang, 2019, Section 5, each X_k in (2.1) admits the decomposition

$$X_k = X_{k,m}^{(1)} + X_{k,m}^{(2)},$$

where

$$X_{k,m}^{(j)} := \int_E f \circ T^k(s)\, M_m^{(j)}(ds), \quad j = 1, 2,$$

and $M_m^{(1)}$ and $M_m^{(2)}$ are two independent homogeneous symmetric infinitely divisible random measures. The Lévy measure of $M_m^{(1)}$ is ρ restricted to $\{|x| \leq m\}$, while the Lévy measure of $M_m^{(2)}$ is ρ restricted to $\{|x| > m\}$. Define for $j = 1, 2$:

$$S_{n,m}^{(j)}(t) := \sum_{k=1}^{\lfloor nt \rfloor} X_{k,m}^{(j)}, \quad t \in [0,1],$$

and

$$M_{n,m}^{(j)}(B) := \max_{k \in nB \cap \mathbb{N}} X_{k,m}^{(j)}, \quad B \in \mathcal{G}([0,1]),$$

with the convention that $\max \emptyset := -\infty$. By Samorodnitsky, 2016, Theorem 9.4.7, we have

$$n^{-1} w_n^{1/2} S_{n,m}^{(1)} \xrightarrow{fdd} c_{\beta,m} B_H, \quad \text{as } n \to \infty,$$

in $[0,1]$, where

$$c_{\beta,m}^2 = \frac{\Gamma(1+2\beta)}{\Gamma(2-\beta)\Gamma(2+\beta)} \mathbb{E}\left(\{S_\beta(1)\}^{-2\beta}\right) \int_{-m}^{m} x^2 \rho(dx).$$

By a slight extension of Samorodnitsky and Wang, 2019, Theorem 5.1 to accommodate the more general regular variation assumption on ρ (see also the proof of Theorem 3.2), we obtain

$$(\rho^{\leftarrow}(w_n^{-1}))^{-1} M_{n,m}^{(2)} \xrightarrow{fdd} \eta_{\alpha,\beta}, \quad \text{in } \mathcal{I}, \quad \text{as } n \to \infty.$$

Note that the limit law above is independent of the truncation parameter m, which reflects the fact that the extremal behavior is determined solely by the tail behavior of the joint distribution of $\{X_{k,m}^{(2)}\}_{k=1,\ldots,n}$, which in turn depends only on the tail of the Lévy measure ρ. Since $S_{n,m}^{(1)}$ and $M_{n,m}^{(2)}$ are based on independent components, we conclude:

$$\begin{pmatrix} n^{-1} w_n^{1/2} S_{n,m}^{(1)} \\ (\rho^{\leftarrow}(w_n^{-1}))^{-1} M_{n,m}^{(2)} \end{pmatrix} \xrightarrow{fdd} \begin{pmatrix} c_{\beta,m} B_H \\ \eta_{\alpha,\beta} \end{pmatrix}, \quad \text{as } n \to \infty,$$

in $[0,1] \times \mathcal{I}$, where B_H and $\eta_{\alpha,\beta}$ are independent. Since $c_{\beta,m} \to c_\beta$ as $m \to \infty$, we obtain

$$\begin{pmatrix} c_{\beta,m} B_H \\ \eta_{\alpha,\beta} \end{pmatrix} \xrightarrow{fdd} \begin{pmatrix} c_\beta B_H \\ \eta_{\alpha,\beta} \end{pmatrix}, \quad \text{in } [0,1] \times \mathcal{I}, \quad \text{as } m \to \infty.$$

The desired convergence in fdd now follows by a standard triangular approximation argument (see, e.g., Billingsley, 1999, Theorem 3.2), once we verify the following negligibility conditions for any $t \in [0,1]$, $B \in \mathcal{G}([0,1])$), and $\epsilon > 0$:

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(n^{-1} w_n^{1/2} \left|S_n(t) - S_{n,m}^{(1)}(t)\right| > \epsilon\right) = 0 \tag{A.1}$$

and

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}\left((\rho^{\leftarrow}(w_n^{-1}))^{-1} \left|M_n(B) - M_{n,m}^{(2)}(B)\right| > \epsilon\right) = 0. \tag{A.2}$$

To verify (A.1), observe that by Samorodnitsky, 2016, Theorem 9.4.7 again,

$$n^{-1}w_n^{1/2}\left(S_n(t)-S_{n,m}^{(1)}(t)\right)=n^{-1}w_n^{1/2}S_{n,m}^{(2)}(t)\xrightarrow{fdd}(c_\beta-c_{\beta,m})B_H(t),$$

as $n\to\infty$, from which (A.1) follows because $c_{\beta,m}\to c_\beta$ as $m\to\infty$.

For (A.2), we assume n is large enough that $nB\cap\mathbb{N}\neq\emptyset$. For any real sequences $(\alpha_k)_{k=1}^n$ and $(\beta_k)_{k=1}^n$, we have:

$$\left|\max_{1\le k\le n}(\alpha_k+\beta_k)-\max_{1\le k\le n}\alpha_k\right|\le\max_{1\le k\le n}|\beta_k|.$$

Hence,

$$\begin{aligned}&\limsup_{n\to\infty}\mathbb{P}\left((\rho^{\leftarrow}(w_n^{-1}))^{-1}\left|M_n(B)-M_{n,m}^{(2)}(B)\right|>\epsilon\right)\\ \le&\limsup_{n\to\infty}\mathbb{P}\left(\max_{k\in nB\cap\mathbb{N}}\left|X_{k,m}^{(1)}\right|>\rho^{\leftarrow}(w_n^{-1})\epsilon\right).\end{aligned}$$

By stationarity and a union bound:

$$\le\limsup_{n\to\infty}n\cdot\mathbb{P}\left(\left|X_{1,m}^{(1)}\right|>\rho^{\leftarrow}(w_n^{-1})\epsilon\right)=0,$$

since $\rho^{\leftarrow}(w_n^{-1})\in\mathrm{RV}_\infty((1-\beta)/\alpha)$ and $X_{1,m}^{(1)}$ has tails satisfying

$$\mathbb{P}\left(\left|X_{1,m}^{(1)}\right|>x\right)=o\left(e^{-\delta x\log x}\right),\quad x\to\infty,$$

for some $\delta>0$, as is known for infinitely divisible distributions with bounded Lévy measures (Sato, 1999, Theorem 26.1). □

A.2 Infinite Variance Case

We now proceed to introduce a Poissonization construction that plays a central role in the proof. For each $j,n\in\mathbb{N}$, define the set of scaled entrance times, starting from a random point $U_j^{(n)}$, as

$$\widetilde{\mathcal{R}}_{j,n}:=\left\{\frac{k}{n}:k=1,\dots,n,\,U_j^{(n)}\in T^{-k}A_0\right\}.$$

Due to the construction of $U_j^{(n)}$, the set $\widetilde{\mathcal{R}}_{j,n}$ is almost surely non-empty, with (random) cardinality $|\widetilde{\mathcal{R}}_{j,n}|$. Thus, we can write

$$\widetilde{\mathcal{R}}_{j,n}=V_{j,n}+\{\tau_{j,n}(0),\tau_{j,n}(1),\dots,\tau_{j,n}(|\widetilde{\mathcal{R}}_{j,n}|-1)\}\subset n^{-1}\{1,\dots,n\},$$

where

$$V_{j,n} := \frac{1}{n} \min \left\{ k = 1, \dots, n : U_j^{(n)} \in T^{-k} A_0 \right\} \tag{A.3}$$

is the scaled first entrance time, and $\tau_{j,n}(0) := 0 < \tau_{j,n}(1) < \cdots < \tau_{j,n}(|\widetilde{\mathcal{R}}_{j,n}| - 1)$ are the successive scaled entrance times relative to $V_{j,n}$.

In view of the Markov chain construction, the sequence $\{n\tau_{j,n}(i)\}_i$ forms a $\mathbb{N}$-valued renewal process (i.e., a random walk with i.i.d. steps in $\mathbb{N}$) starting at $i = 0$ and stopped at $i = |\widetilde{\mathcal{R}}_{j,n}| - 1$. Moreover, $\lim_{n\to\infty} |\widetilde{\mathcal{R}}_{j,n}| = \infty$. The inter-arrival times of $\{n\tau_{j,n}(i)\}_i$, before stopping, are i.i.d. with common probability mass function $P_0(\varphi_{A_0} = k)$, for $k = 1, 2, \dots$. For technical convenience, we extend $\tau_{j,n}(i)$ to all $i \in \mathbb{N}_0$ by appending i.i.d. inter-arrivals drawn from the same distribution, maintaining independence across different j's.

Next, for each $j, n \in \mathbb{N}$, let $\{N_{j,n}(t)\}_{t\geq 0}$ be a Poisson process with intensity

$$\gamma_n := \frac{n w_n^{-1}}{\Gamma(2-\beta)} \in \mathrm{RV}_\infty(\beta),$$

and assume it is independent of all other random elements. Define the non-decreasing processes $\sigma_{j,n}$ as

$$\{\sigma_{j,n}(t) : t \geq 0\} := \{\tau_{j,n}(N_{j,n}(t)) : t \geq 0\}.$$

Then each $\sigma_{j,n}$ is a non-decreasing compound Poisson process and hence a subordinator. Let $L_{j,n}$ and $\mathcal{R}_{j,n}$ denote the local time and range of $\sigma_{j,n}$. Note that the scaled range satisfies

$$\widetilde{\mathcal{R}}_{j,n} = (V_{j,n} + \mathcal{R}_{j,n}) \cap [0,1]. \tag{A.4}$$

For each subset $S \subset \mathbb{N}$, define the intersection of the shifted random sets as

$$I_{S,n} := \begin{cases} \bigcap_{j\in S} \widetilde{\mathcal{R}}_{j,n}, & \text{if } S \neq \emptyset, \\ n^{-1}\{1, \dots, n\}, & \text{if } S = \emptyset. \end{cases} \tag{A.5}$$

Now fix a level $\ell \in \mathbb{N}$, and define the following quantities central to the analysis. First, introduce the process

$$S^*_{n,\ell}(t) := \sum_{j=1}^{\ell} \varepsilon_j \, \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right) L_{j,n}((t - V_{j,n})_+), \quad t \in [0,1], \tag{A.6}$$

and the maximum functional

$$M^*_{n,\ell}(B) := \max_{S\subset\{1,\dots,\ell\}} \mathbf{1}_{\{I_{S,n}\cap B=\emptyset\}} \sum_{j\in S} \mathbf{1}_{\{\varepsilon_j=1\}}\, \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right), \quad B\in\mathcal{G}([0,1]). \tag{A.7}$$

As per convention, any summation or maximization over an empty index set is interpreted as zero. We note that equations (A.6) and (A.7) serve as approximations of the partial sum and partial maximum processes, respectively, as will be made precise in Lemmas A.3 and A.4 below. To facilitate this approximation, we also define the truncated versions of the processes $S(t)$ and $M(B)$ from (2.13) and (2.18), respectively, as follows:

$$S_\ell(t) := (2C_\alpha)^{1/\alpha} \sum_{j=1}^{\ell} \varepsilon_j \Gamma_j^{-1/\alpha} L_j\Big((t-V_j)^+\Big), \quad t\in[0,1], \tag{A.8}$$

and

$$M_\ell(B) := \max_{S\subset\{1,\dots,\ell\}} \mathbf{1}_{\{I_S\cap B=\emptyset\}} \sum_{j\in S} \Gamma_j^{-1/\alpha}, \quad B\in\mathcal{G}([0,1]). \tag{A.9}$$

Proposition A.1. *Fix an integer $\ell > 0$. Let $S^*_{n,\ell}$ and $M^*_{n,\ell}$ be defined as in equations* (A.6) *and* (A.7)*, respectively, and let S_ℓ and M_ℓ be their corresponding limits defined in equations* (A.8) *and* (A.8)*. Then, as $n\to\infty$,*

$$\frac{1}{\rho^{\leftarrow}(w_n^{-1})}\begin{pmatrix} S^*_{n,\ell} \\ M^*_{n,\ell}\end{pmatrix} \xrightarrow{f.d.d.} \begin{pmatrix} C^{-1/\alpha} S_\ell \\ M_\ell\end{pmatrix} \quad \text{in } \mathcal{T}_0 := [0,1]\times\mathcal{I},$$

where $\mathcal{I}$ denotes the collection of all non-empty open subintervals of $[0,1]$, and the convergence is in the sense of finite-dimensional distributions on $\mathcal{T}_0$.

Proof. Since $\{P_0(\varphi_{A_0} > n)\}_n \in \mathrm{RV}_\infty(-\beta)$, we may write

$$P_0(\varphi_{A_0} > n) := n^{-\beta} f(n),$$

where $f(n)$ is slowly varying at infinity. Recall that each $\tau_{j,n}$ is an increasing random walk with i.i.d. steps that are integer-valued, non-negative, and regularly varying with index $\beta\in(0,1)$. Let $\{\sigma_j\}, \{L_j\}, \{V_j\}$ be as defined earlier. A classical result on convergence to stable laws (see, e.g., Section 1.a in Chow and Teugels, 1978 yields

$$\frac{n\tau_{j,n}(\lfloor\gamma_n\rfloor)}{\gamma_n^{1/\beta} f^{-1/\beta}(\gamma_n^{1/\beta})} \Rightarrow (\Gamma(1-\beta))^{1/\beta}\sigma_j(1),$$

as $n \to \infty$, where $\overline{f^{-1/\beta}}$ denotes the de Bruijn conjugate of $f^{-1/\beta}$, satisfying the property

$$\lim_{n\to\infty} f^{-1/\beta}(n)\overline{f^{-1/\beta}}(nf^{-1/\beta}(n)) = 1$$

(see, e.g., Theorem 1.5.13 in Bingham et al., 1989). Moreover, using the asymptotic behavior in (2.3), we have

$$\gamma_n^{1/\beta} = (nw_n^{-1})^{1/\beta}(\Gamma(2-\beta))^{-1/\beta} \sim nf^{-1/\beta}(n)(\Gamma(1-\beta))^{-1/\beta}, \quad \text{as } n \to \infty.$$

Hence, it follows that

$$\left\{\gamma_n^{1/\beta}\overline{f^{-1/\beta}}\left(\gamma_n^{1/\beta}\right)\right\}^{-1} n \to (\Gamma(1-\beta))^{1/\beta}, \quad \text{and thus} \quad \tau_{j,n}(\lfloor\gamma_n\rfloor) \Rightarrow \sigma_j(1).$$

Since $N_{j,n}(1)/\gamma_n \xrightarrow{P} 1$, and $N_{j,n}(1)$ is independent of $\tau_{j,n}$, a standard argument for replacing deterministic time arguments by independent random times yields

$$\sigma_{j,n}(1) = \tau_{j,n}(N_{j,n}(1)) \Rightarrow \sigma_j(1), \quad \text{as } n \to \infty.$$

In addition, by Theorem 5.4 of Samorodnitsky and Wang, 2019, we have

$$V_{j,n} \xrightarrow{d} V_j, \quad \text{as } n \to \infty,$$

with $V_{j,n}$ as defined in (A.3). Applying Proposition 3.4 and using independence across j, we obtain

$$\begin{pmatrix} (L_{j,n}((x - V_{j,n})_+))_{x\in[0,1]} \\ (V_{j,n} + \mathcal{R}_{j,n}) \cap [0,1] \end{pmatrix}_{j=1,\dots,\ell} \Rightarrow \begin{pmatrix} (L_j((x - V_j)_+))_{x\in[0,1]} \\ (V_j + \mathcal{R}_j) \cap [0,1] \end{pmatrix}_{j=1,\dots,\ell},$$

weakly in $D([0,1])^\ell \times \mathfrak{F}([0,1])^\ell$ as $n \to \infty$. Furthermore, by Theorem 5.4 of Samorodnitsky and Wang, 2019, for each $S \subset \{1,\dots,\ell\}$,

$$I_{S,n} := \bigcap_{j\in S}(V_{j,n}+\mathcal{R}_{j,n})\cap[0,1] \Rightarrow I_S := \bigcap_{j\in S}(V_j+\mathcal{R}_j)\cap[0,1], \quad \text{in } \mathfrak{F}([0,1]).$$

By Lemma A.2 below (which extends Samorodnitsky and Wang, 2019, 26, Theorem 2.1), we have

$$\begin{pmatrix} (L_{j,n}((x - V_{j,n})_+))_{x\in[0,1]} \\ I_{S,n} \end{pmatrix}_{\substack{j=1,\dots,\ell \\ S\subset\{1,\dots,\ell\}}} \Rightarrow \begin{pmatrix} (L_j((x - V_j)_+))_{x\in[0,1]} \\ I_S \end{pmatrix}_{\substack{j=1,\dots,\ell \\ S\subset\{1,\dots,\ell\}}},$$

weakly in $D([0,1])^\ell \times \mathfrak{F}([0,1])^{2^\ell}$. The result then follows from the above joint convergence and the facts:

$$\frac{\rho^{\leftarrow}(\Gamma_j/(2w_n))}{\rho^{\leftarrow}(w_n^{-1})} \to (\Gamma_j/2)^{-1/\alpha}, \quad j = 1, \dots, \ell,$$

due to regular variation of $\rho^{\leftarrow}$, and

$$(M_\ell(B))_{B\in\mathcal{G}([0,1])} \overset{d}{=} \left(\max_{S\subset\{1,\dots,\ell\}} \mathbf{1}_{\{\mathcal{I}_S\cap B=\emptyset\}} \sum_{j\in S} \mathbf{1}_{\{\varepsilon_j=1\}}(\Gamma_j/2)^{-1/\alpha} \right)_{B\in\mathcal{G}([0,1])},$$

where the equality in law follows from the fact that the thinned Poisson process $\{\Gamma_j/2\}_{j\in\mathbb{N},\varepsilon_j=1}$ has the same distribution as $\{\Gamma_j\}_{j\in\mathbb{N}}$.

□

Lemma A.2. *Let $\{A_k\}_{k=1}^m$ and $\{A_k^{(n)}\}_{n\in\mathbb{N},k=1,\dots,m}$ be random closed sets in $\mathfrak{F} = \mathfrak{F}(\mathbb{R}^d)$, for some fixed $m \in \mathbb{N}$. Let $x^{(n)}$ and x be random elements in a separable metric space E. Suppose that the following joint weak convergence holds:*

$$\begin{pmatrix} x^{(n)} \\ (A_k^{(n)})_{k=1,\cdots,m} \end{pmatrix} \Rightarrow \begin{pmatrix} x \\ (A_k)_{k=1,\cdots,m} \end{pmatrix} \quad \text{in } E\times\mathfrak{F}^m \text{ as } n\to\infty. \tag{A.10}$$

For any non-empty index set $I \subset \{1,\dots,m\}$, define the intersections:

$$A_I^{(n)} := \bigcap_{k\in I} A_k^{(n)}, \qquad A_I := \bigcap_{k\in I} A_k.$$

In addition, let $A_\emptyset^{(n)}$ and $A_\emptyset$ be non-random elements of $\mathfrak{F}$ such that $A_k^{(n)} \subset A_\emptyset^{(n)}$ and $A_k \subset A_\emptyset$ for all $k = 1,\dots,m$ and all $n \in \mathbb{N}$. Assume further that for every $I \subset \{1,\dots,m\}$, the intersections satisfy the marginal convergence:

$$A_I^{(n)} \Rightarrow A_I \quad \text{in } \mathfrak{F} \text{ as } n\to\infty.$$

Then we have the joint convergence:

$$\begin{pmatrix} x^{(n)} \\ (A_I^{(n)})_{I\subset\{1,\dots,m\}} \end{pmatrix} \Rightarrow \begin{pmatrix} x \\ (A_I)_{I\subset\{1,\dots,m\}} \end{pmatrix} \quad \text{in } E\times\mathfrak{F}^{2^m} \text{ as } n\to\infty. \tag{A.11}$$

Proof. The only substantive difference between Samorodnitsky and Wang, 2019, Theorem 2.1(b) and the present lemma is the inclusion of the auxiliary components $x^{(n)}$ and x. Note that the product space $E\times\mathfrak{F}^m$ is itself a separable metric space, since both E and $\mathfrak{F}$ are separable. Therefore, by the Skorokhod representation theorem, the weak convergence in (A.10) can be upgraded to almost sure convergence on a suitable probability space. The remainder of the proof then proceeds by adapting the arguments from Samorodnitsky and Wang, 2019, Theorem 2.1(b), thereby establishing a convergence-in-probability version of (A.11). □

Recall the series representation of the sequence $(X_k)_{k=1,\dots,n}$ as given in equation (2.6). For each fixed $\ell \in \mathbb{N}$, we define the truncated partial sum process by

$$S_{n,\ell}(t) := \sum_{j=1}^{\ell} \varepsilon_j \, \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right) \sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}}, \quad t \in [0,1], \tag{A.12}$$

and the corresponding truncated partial maximum process by

$$M_{n,\ell}(B) := \max_{k \in nB \cap \mathbb{N}} \sum_{j=1}^{\ell} \varepsilon_j \, \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right) \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}}, \quad B \in \mathcal{G}([0,1]). \tag{A.13}$$

Lemma A.3. *For any $t \in [0,1]$, $\ell \in \mathbb{N}$, and $\epsilon > 0$, Then,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\Gamma(2-\beta)}{\rho^{\leftarrow}(w_n^{-1})\, n w_n^{-1}} S_{n,\ell}(t) - \frac{1}{\rho^{\leftarrow}(w_n^{-1})} S^*_{n,\ell}(t)\right| > \epsilon\right) = 0.$$

where $S_{n,\ell}$ and $S^*_{n,\ell}$ are as in (A.12) *and* (A.6)*, respectively.*

Proof. Using the relation $\gamma_n = n w_n^{-1}/\Gamma(2-\beta)$, it suffices to show that for each $j = 1, \dots, \ell$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right)}{\rho^{\leftarrow}(w_n^{-1})} \left|\frac{1}{\gamma_n} \sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}} - L_{j,n}\left((t - V_{j,n})_+\right)\right| > \epsilon\right) = 0,$$

for any fixed $\epsilon > 0$. Note that, by the regular variation of $\rho^{\leftarrow}$, we have

$$\frac{\rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right)}{\rho^{\leftarrow}(w_n^{-1})} \to \left(\frac{\Gamma_j}{2}\right)^{-1/\alpha} \quad \text{as } n \to \infty.$$

Thus, to establish the desired limit, it suffices to prove that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{\gamma_n} \sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}} - L_{j,n}\left((t - V_{j,n})_+\right)\right| > \epsilon\right) = 0.$$

Define

$$Q_{j,n}(t) := \sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}}.$$

Observe that $Q_{j,n}(t) \geq 1$ if and only if $V_{j,n} \leq t$, in which case

$$\tau_{j,n}(Q_{j,n}(t) - 1) \leq t - V_{j,n} < \tau_{j,n}(Q_{j,n}(t)).$$

This implies

$$\begin{aligned} L_{j,n}((t - V_{j,n})_+) &= \inf\{s \geq 0 : \tau_{j,n}(N_{j,n}(s)) > (t - V_{j,n})_+\} \\ &= \inf\{s \geq 0 : N_{j,n}(s) = Q_{j,n}(t) \vee 1\}. \end{aligned}$$

Let $\{T_{j,n}(i)\}_{i\in\mathbb{N}}$ be the inter-arrival times of the Poisson process $N_{j,n}$; then these are i.i.d. exponential random variables with mean γ_n^{-1}. Consequently,

$$L_{j,n}((t - V_{j,n})_+) = \sum_{i=1}^{Q_{j,n}(t)\vee 1} T_{j,n}(i).$$

We now estimate the deviation:

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{\gamma_n} \left| \sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}} - L_{j,n}((t - V_{j,n})_+) \right| > \epsilon \right) \\ \leq &\mathbb{P}\left(\left| \frac{Q_{j,n}(t)}{\gamma_n} - \sum_{i=1}^{Q_{j,n}(t)} T_{j,n}(i) \right| > \epsilon \right) + \mathbb{P}\left(T_{j,n}(1) > \epsilon\right). \end{aligned}$$

The second term vanishes as $n \to \infty$ since $\mathrm{Var}(T_{j,n}(1)) = \gamma_n^{-2} \to 0$. For the first term, by Chebyshev's inequality and the independence of the $T_{j,n}(i)$, we obtain

$$\begin{aligned} \mathbb{P}\left(\left| Q_{j,n}(t) - \sum_{i=1}^{Q_{j,n}(t)} T_{j,n}(i) \right| > \epsilon \right) \leq &\frac{\mathrm{Var}(T_{j,n}(1))\, \mathbb{E}[Q_{j,n}(t)]}{\epsilon^2} \\ = &\frac{\gamma_n^{-2}}{\epsilon^2} \cdot \frac{n}{w_n}, \end{aligned}$$

which converges to zero as $n \to \infty$. □

Lemma A.4. *For any non-empty open interval $B \subset [0, 1]$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\left| M_{n,\ell}(B) - M^*_{n,\ell}(B) \right| > 0 \right) = 0,$$

where $M_{n,\ell}$ is as defined in (A.13), *and $M^*_{n,\ell}$ is as defined in* (A.7).

Proof. Recall $\mathcal{R}_{j,n}$ from (A.4) and $I_{S,n}$ from (A.5). Letting $S^c := \{1, \dots, \ell\} \setminus S$, define the set

$$I^*_{S,n} := I_{S,n} \cap \bigcap_{j \in S^c} \mathcal{R}^c_{j,n}.$$

Thus, each (rescaled) time point in $I^*_{S,n}$ is contained precisely in those intervals $\mathcal{R}_{j,n}$ for which $j \in S$, and in none of those with $j \in S^c$. Now define the event

$$A_n(B) := \bigcup_{S \subset \{1,\dots,\ell\}} \left(\{I_{S,n} \cap B \neq \emptyset\} \cap \left\{ I^*_{S,n} \cap B = \emptyset \right\} \right).$$

By Lemma 5.5 of Samorodnitsky and Wang, 2019, we have the identity

$$M_{n,\ell}(B) = M^*_{n,\ell}(B) \quad \text{on the complement } A_n(B)^c,$$

and moreover, $\lim_{n\to\infty} \mathbb{P}(A_n(B)) = 0$. The result follows immediately.

□

Proof of Theorem 3.2. Tightness on the product space follows if it holds for each marginal component. The tightness of the normalized partial sum process S_n in the J_1-topology of $D[0,1]$ has been established in Owada and Samorodnitsky, 2015; see also Bai et al., 2020. The normalized partial maxima M_n is automatically tight, since the space $SM[0,1]$ is compact. In view of Proposition A.6 and Remark A.1, it remains to verify the convergence of finite-dimensional distributions in the index set $T_0 := [0,1] \times \mathcal{I}$, where $\mathcal{I}$ denotes the collection of all non-empty open subintervals of $[0,1]$. This will be achieved using a triangular approximation argument (cf. Theorem 3.2 in Billingsley, 1999).

Note that, as $\ell \to \infty$, the truncated processes $S_\ell(t) \to S(t)$ and $M_\ell(B) \to M(B)$ almost surely, for every $t \in [0,1]$ and $B \in \mathcal{G}([0,1])$, where S_ℓ and M_ℓ are defined in (A.8) and (A.9) as truncated versions of S and M, respectively. The triangular approximation argument is completed by Proposition A.1 and Lemmas A.3 and A.4, provided we can establish the negligibility of the tail contributions beyond the truncation level. Specifically, we require that for any $\epsilon > 0$,

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\rho^{\leftarrow}(w_n^{-1}) n w_n^{-1}} \left| \sum_{k=1}^{\lfloor nt \rfloor} \sum_{j=\ell+1}^{\infty} \varepsilon_j \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right) \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}} \right| > \epsilon \right) = 0 \tag{A.14}$$

and

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\rho^{\leftarrow}(w_n^{-1})} \max_{k \in nB \cap \mathbb{N}} \left| \sum_{j=\ell+1}^{\infty} \varepsilon_j \, \rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right) \mathbf{1}_{\{U_j^{(n)} \in T^{-k} A_0\}} \right| > \epsilon \right) = 0 \tag{A.15}$$

Relation (A.14) is a special case of relation (66) in Bai et al., 2020. To handle (A.15), we apply a union bound and a version of Markov's inequality. For any $r > 0$, the probability on the left-hand side of (A.15) is bounded above by

$$\begin{aligned} & n\,\mathbb{P}\left(\frac{1}{\rho^{\leftarrow}(w_n^{-1})}\left|\sum_{j=\ell+1}^{\infty}\varepsilon_j\,\rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right)\mathbf{1}_{\{U_j^{(n)}\in T^{-k}A_0\}}\right|>\epsilon\right) \\ \le & n\,\epsilon^{-r}\,\mathbb{E}\left(\frac{1}{\rho^{\leftarrow}(w_n^{-1})}\left|\sum_{j=\ell+1}^{\infty}\varepsilon_j\,\rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right)\mathbf{1}_{\{U_j^{(n)}\in T^{-k}A_0\}}\right|\right)^r . \end{aligned} \tag{A.16}$$

By independence and using the Khintchine inequality for Rademacher variables, the expectation in (A.16) is further bounded by

$$n\,\epsilon^{-r}C_r\left(\sum_{j=\ell+1}^{\infty}\frac{\mathbb{E}\,[\rho^{\leftarrow}(\Gamma_j/(2w_n))^2]}{\rho^{\leftarrow}(w_n^{-1})^2}\,\mathbb{P}(U_j^{(n)}\in T^{-k}A_0)\right)^{r/2} .$$

From inequality (82) in the proof of (66) in Bai et al., 2020, for large j, we have

$$\frac{\mathbb{E}[\rho^{\leftarrow}(\Gamma_j/(2w_n))^2]}{\rho^{\leftarrow}(w_n^{-1})^2}\le C\mathbb{E}\left((\Gamma_j/2)^{-1/\alpha_0}+(\Gamma_j/2)^{-(1/\alpha)-\delta}\right)^2\le Cj^{-2\gamma},$$

for some generic positive constant C and a small $\delta > 0$, where $\alpha_0 \in (0, 2)$ is as in (2.5) and $\gamma := \min\{1/\alpha_0, 1/\alpha + \delta\}$. Since $\mathbb{P}(U_j^{(n)} \in T^{-k}A) = w_n^{-1}$, it follows that

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{\rho^{\leftarrow}(w_n^{-1})}\max_{k\in nB\cap\mathbb{N}}\left|\sum_{j=\ell+1}^{\infty}\varepsilon_j\,\rho^{\leftarrow}\left(\frac{\Gamma_j}{2w_n}\right)\mathbf{1}_{\{U_j^{(n)}\in T^{-k}A_0\}}\right|>\epsilon\right) \\ & \le n\,C\,w_n^{-r/2}\sum_{j=\ell+1}^{\infty}j^{-2\gamma}. \end{aligned}$$

Finally, since $(w_n) \in \mathrm{RV}_\infty(1-\beta)$, choosing $r > 2/(1-\beta)$ guarantees that

$$\lim_{n\to\infty} nw_n^{-r/2} = 0.$$

Using the fact that $\sum_{j=\ell+1}^{\infty} j^{-2\gamma} \to 0$ as $\ell \to \infty$ (since $2\gamma > 1$), we conclude that (A.15) holds. This completes the triangular approximation argument.

□

A.3 Dependence of Limits in the Infinite Variance Case

Proof of Proposition 3.3. To establish the dependence, it suffices to show that $S(1)$ and $M([0,1])$ are dependent. We demonstrate this by verifying that the *tail dependence coefficient*:

$$\lim_{x\to\infty} \mathbb{P}(|S(1)| > x \mid M([0,1]) > x) \neq 0.$$

Let

$$Z_j := (2C_\alpha)^{1/\alpha}\varepsilon_j L_j(1 - V_j), \quad j \in \mathbb{N},$$

where the marginal distribution of $L_j(1)$ follows the Mittag-Leffler law and thus admits finite moments of all orders. Then, we can write

$$S(1) = \sum_{j=1}^{\infty} Z_j \Gamma_j^{-1/\alpha}.$$

Moreover, due to the strict ordering $\Gamma_1 < \Gamma_2 < \cdots$ and the fact that $I_S \neq \emptyset$ almost surely for any $|S| \leq \ell_\beta$, we have

$$M([0,1]) = \sup_{\substack{S\subset\mathbb{N} \\ 1\leq|S|\leq\ell_\beta}} \left(\sum_{j\in S} \Gamma_j^{-1/\alpha}\right) = \sum_{j=1}^{\ell_\beta} \Gamma_j^{-1/\alpha}.$$

We claim that, as $x \to \infty$, the marginal tail behavior of $|S(1)|$ satisfies

$$\mathbb{P}(|S(1)| > x) \sim \mathbb{P}(|Z_1|\Gamma_j^{-1/\alpha} > x) \sim x^{-\alpha}\mathbb{E}[|Z_1|^\alpha]. \tag{A.17}$$

To justify (A.17), note that by orthogonality $\mathbb{E}[Z_i Z_j] = 0$ for $i \neq j$, and independence, we compute:

$$\mathbb{E}\left[\left(\sum_{j=\ell}^{\infty} Z_j \Gamma_j^{-1/\alpha}\right)^2\right] = \sum_{j=\ell}^{\infty} \mathbb{E}[Z_j^2 \Gamma_j^{-2/\alpha}] \leq C\mathbb{E}[Z_1^2] \sum_{j=\ell}^{\infty} j^{-2/\alpha} < \infty, \tag{A.18}$$

for all large enough ℓ, where C is a generic positive constant, and we have applied a bound for negative moments of Γ_j. A Markov inequality then yields:

$$\mathbb{P}\left(\left|\sum_{j=\ell}^{\infty} Z_j \Gamma_j^{-1/\alpha}\right| > x\right) \leq Cx^{-2}.$$

For fixed $j \in \mathbb{N}$, using Breiman's lemma and asymptotics of the gamma distribution of Γ_j, we get:

$$\mathbb{P}(|Z_j|\Gamma_j^{-1/\alpha} > x) \sim \mathbb{E}[|Z_j|^{j\alpha}]\,\mathbb{P}(\Gamma_j^{-1/\alpha} > x) \sim \frac{\mathbb{E}[|Z_j|^{j\alpha}]}{j!} x^{-j\alpha}. \quad \text{(A.19)}$$

Combining (A.18) and (A.19), and using Lemma 4.2.4 from Samorodnitsky, 2016, the claim in (A.17) follows.

On the other hand, by Proposition 3.3 in Samorodnitsky and Wang, 2019, we have:

$$\mathbb{P}(M([0,1]) > x) \sim \mathbb{P}(\Gamma_1^{-1/\alpha} > x) \sim x^{-\alpha}.$$

Now consider the joint tail. For any $\epsilon \in (0,1)$, using union bounds, triangular inequalities, and (A.19), we have:

$$\begin{aligned}
&\mathbb{P}(|S(1)| > x, M([0,1]) > x) \\
\leq &\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > (1-\epsilon)x, \Gamma_1^{-1/\alpha} > (1-\epsilon)x) \\
&+ \mathbb{P}\left(\left|\sum_{j=2}^{\infty} Z_j \Gamma_j^{-1/\alpha}\right| > \epsilon x\right) + \mathbb{P}\left(\sum_{j=2}^{\ell_\beta} \Gamma_j^{-1/\alpha} > \epsilon x\right) \\
= &\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > (1-\epsilon)x, \Gamma_1^{-1/\alpha} > (1-\epsilon)x) + o(x^{-\alpha}),
\end{aligned}$$

as $x \to \infty$, due to (A.18) and (A.19). Note that

$$\begin{aligned}
&\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > x, \Gamma_1^{-1/\alpha} > x) \\
= &\mathbb{P}(\mathbf{1}_{\{|Z_1|\leq 1\}}|Z_1|\Gamma_1^{-1/\alpha} > x) + \mathbb{P}(|Z_1| > 1)\mathbb{P}(\Gamma_1^{-1/\alpha} > x) \\
\sim &\left(\mathbb{E}[|Z_1|^\alpha \mathbf{1}_{\{|Z_1|\leq 1\}}] + \mathbb{P}(|Z_1| > 1)\right) x^{-\alpha},
\end{aligned}$$

where Breiman's lemma justifies the last relation. Putting this together:

$$\limsup_{x\to\infty} \frac{\mathbb{P}(|S(1)| > x, M([0,1]) > x)}{\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > x, \Gamma_1^{-1/\alpha} > x)} \leq (1-\epsilon)^{-\alpha}.$$

Similarly, for the lower bound, for any $\epsilon \in (0,1)$,

$$\begin{aligned}
&\mathbb{P}(|S(1)| > x, M([0,1]) > x) \\
\geq &\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > (1+\epsilon)x, \Gamma_1^{-1/\alpha} > (1+\epsilon)x) - o(x^{-\alpha}),
\end{aligned}$$

so

$$\liminf_{x\to\infty} \frac{\mathbb{P}(|S(1)| > x, M([0,1]) > x)}{\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > x, \Gamma_1^{-1/\alpha} > x)} \geq (1+\epsilon)^{-\alpha}.$$

Letting $\epsilon \to 0$, we conclude that:

$$\lim_{x\to\infty} \frac{\mathbb{P}(|S(1)| > x, M([0,1]) > x)}{\mathbb{P}(|Z_1|\Gamma_1^{-1/\alpha} > x, \Gamma_1^{-1/\alpha} > x)} = 1,$$

which implies that

$$\begin{aligned}&\lim_{x\to\infty} \mathbb{P}(|S(1)| > x \mid M([0,1]) > x)\\ =&\mathbb{E}[|Z_1|^\alpha \mathbf{1}_{\{|Z_1|\le 1\}}] + \mathbb{P}(|Z_1| > 1) \neq 0.\end{aligned}$$

This completes the proof.

□

A.4 Joint Convergence on Subordinators

Lemma A.5. *Suppose $f, f_n : [0,\infty) \to [0,\infty)$, $n \in \mathbb{N}$, are non-decreasing, unbounded, and right-continuous functions with right-continuous inverses $f^{\to}$ and $f_n^{\to}$, respectively. Let F and F_n denote the closed ranges of f and f_n, respectively. Assume additionally that f is strictly increasing and that the local uniform convergence*

$$\sup_{0\le s\le t} |f_n(s) - f(s)| \to 0 \quad \textit{as } n \to \infty \tag{A.20}$$

holds for every $t \ge 0$. Then the following conclusions hold:

1. For every $y \in [0,\infty)$,

$$\sup_{0\le x\le y} |f_n^{\to}(x) - f^{\to}(x)| \to 0 \quad \textit{as } n \to \infty, \tag{A.21}$$

that is, the inverses converge uniformly on compact intervals.

2. For any sequence $x_n \to x \in \mathbb{R}$,

$$|\rho(x_n, F_n) - \rho(x, F)| \to 0 \quad \textit{as } n \to \infty, \tag{A.22}$$

where for a non-empty set $A \subseteq \mathbb{R}$, the distance function ρ is defined by $\rho(x, A) := \inf_{u\in A} |x - u|$.

Proof. Since f is strictly increasing, its right-continuous inverse $f^{\to}$ is continuous (see, e.g., Whitt, 2002, Lemma 13.6.5). The local uniform convergence of the inverses, i.e., (A.21), then follows from Whitt, 2002, Corollary 13.6.4.

We now prove (A.22). Since $|\rho(x_n, F_n) - \rho(x, F_n)| \le |x_n - x|$, it suffices to consider the case $x_n \equiv x$. Fix $x \in \mathbb{R}$. Because f is non-decreasing and unbounded, there exists $t > 0$ such that $f(t) > x$. By monotonicity, we may write

$$\rho(x, F) = \inf_{0 \le s \le t} |x - f(s)|.$$

On the other hand, the convergence $f_n(t) \to f(t)$ as $n \to \infty$, implied by (A.20), ensures that for all sufficiently large n, we have $f_n(t) > x$. Thus,

$$\rho(x, F_n) = \inf_{0 \le s \le t} |x - f_n(s)|, \quad \text{for all large } n.$$

By the triangle inequality, we then obtain

$$\rho(x, F) - \sup_{0 \le s \le t} |f(s) - f_n(s)| \le \rho(x, F_n) \le \rho(x, F) + \sup_{0 \le s \le t} |f(s) - f_n(s)|.$$

Letting $n \to \infty$ and using the convergence in (A.20), we conclude that $\rho(x, F_n) \to \rho(x, F)$, which completes the proof of (A.22).

□

Proof of Proposition 3.4. The key step in the proof is the following coupling result, which is a consequence of Theorem 15.17 in Kallenberg, 2002: under the assumption that $\sigma_n(1) \xrightarrow{d} \sigma(1)$ as $n \to \infty$, there exist versions $\widetilde{\sigma}_n \overset{d}{=} \sigma_n$ such that for all $t \ge 0$,

$$\Delta_n(t) := \sup_{0 \le s \le t} |\widetilde{\sigma}_n(s) - \sigma(s)| \xrightarrow{\mathbb{P}} 0.$$

By the Skorokhod representation theorem, we can further assume that on a possibly extended probability space, there exist random variables $\widetilde{V}_n \overset{d}{=} V_n$ and $\widetilde{V} \overset{d}{=} V$ such that $\widetilde{V}_n \to \widetilde{V}$ almost surely, and both $\widetilde{V}_n$ and $\widetilde{V}$ are independent of $\widetilde{\sigma}_n$ and σ, respectively.

Now fix an arbitrary subsequence $S \subset \mathbb{N}$. We claim that there exists a further subsequence $S' \subset S$ such that

$$\mathbb{P}\left(\lim_{n \in S'} \Delta_n(t) = 0 \text{ for all } t > 0\right) = 1.$$

This follows from the standard sub-subsequence argument: for each $k \in \mathbb{N}$, there exists a subsequence $S_k \subset S_{k-1} \subset \cdots \subset S_1 \subset S$ such that $\Delta_n(k) \to 0$ almost surely along $n \in S_k$. Using the monotonicity of $\Delta_n(t)$ in t, we can extract a diagonal subsequence $S' = \{n_k\}_{k \in \mathbb{N}}$ such that the convergence holds for all $t > 0$.

Let $\widetilde{L}_{n_k}$ denote the right-continuous inverse (i.e., the local time) of $\widetilde{\sigma}_{n_k}$. Then for each $y \geq 0$,

$$
\begin{aligned}
&\sup_{0\leq x\leq y}\left|\widetilde{L}_{n_k}\Big((x-\widetilde{V}_{n_k})_+\Big)-L\Big((x-\widetilde{V})_+\Big)\right| \\
&\leq \sup_{0\leq x\leq y}\left|\widetilde{L}_{n_k}\Big((x-\widetilde{V}_{n_k})_+\Big)-L\Big((x-\widetilde{V}_{n_k})_+\Big)\right| \\
&\quad+\sup_{0\leq x\leq y}\left|L\Big((x-\widetilde{V}_{n_k})_+\Big)-L\Big((x-\widetilde{V})_+\Big)\right| \\
&\leq \sup_{0\leq x\leq y}\left|\widetilde{L}_{n_k}\Big(x\Big)-L\Big(x\Big)\right|+\sup_{0\leq x\leq y}\left|L\Big((x-\widetilde{V}_{n_k})_+\Big)-L\Big((x-\widetilde{V})_+\Big)\right|.
\end{aligned}
$$

Applying Lemma A.5 (specifically, its uniform convergence (A.21)) and using the uniform continuity of L on $[0, y]$, we conclude that

$$
\mathbb{P}\left(\lim_{k\to\infty}\sup_{0\leq x\leq y}\left|\widetilde{L}_{n_k}\Big((x-\widetilde{V}_{n_k})_+\Big)-L\Big((x-\widetilde{V})_+\Big)\right|=0\right)=1.
$$

Next, applying conclusion (A.22) from Lemma A.5 yields

$$
\mathbb{P}\left(\lim_{k\to\infty}\rho\left(x,\widetilde{V}_{n_k}+\widetilde{\mathcal{R}}_{n_k}\right)=\rho(x,\widetilde{V}+\mathcal{R})\text{ for all } x\in[0,\infty)\right)=1,
$$

where $\widetilde{\mathcal{R}}_{n_k}$ and $\mathcal{R}$ denote the closed ranges of $\widetilde{\sigma}_{n_k}$ and σ, respectively, and $\rho(x, A)$ denotes the distance from point x to set A. Since convergence in the Fell topology is characterized by this type of pointwise convergence of distance functions (see, e.g., Theorem 2.2(iii) in Salinetti and Wets, 1981), we obtain

$$
\widetilde{V}_{n_k}+\widetilde{\mathcal{R}}_{n_k}\xrightarrow{a.s.}\widetilde{V}+\mathcal{R}\quad\text{in } \mathfrak{F}([0,\infty)).
$$

Thus, we have established the almost sure convergence along any subsubsequence and hence convergence in probability of the triplet consisting of the shifted subordinator, its local time, and its range. Since the coupled sequences have the same distributions as the original ones, the desired joint weak convergence follows.

□

A.5 Criterion for Weak Convergence

It is a classical strategy to establish weak convergence of stochastic processes by verifying the convergence of finite-dimensional distributions (fdd) and proving tightness; see, for example, the treatment of weak convergence in the Skorokhod space in Billingsley, 1999, Chapter 3. In the present work, we also consider random elements whose codomain extends beyond the Skorokhod space—specifically, into the space of sup measures $SM[0,1]$. Unsurprisingly, the same "fdd + tightness" principle continues to apply in this broader setting. To elucidate this methodology, we begin with a general result and then specify it to our framework. Throughout, we denote by $\mathcal{B}(E)$ the Borel σ-algebra on a topological space E.

Suppose T is the index set of the processes (typically time), and let U be a separable metric space serving as the state space. Let $\mathcal{M} \subseteq U^T$ be a function space consisting of mappings $x : T \to U$, which itself is equipped with a metric under which it becomes a Polish space (i.e., complete and separable). For any finite collection $t_1, \ldots, t_d \in T$, define the (multi-)projection map

$$\pi_{t_1,\ldots,t_d} : \mathcal{M} \to U^d, \quad \pi_{t_1,\ldots,t_d}(x) = (x(t_1), \ldots, x(t_d)),$$

and assume that each single-time projection $\pi_t : \mathcal{M} \to U$ is measurable for every $t \in T$.

In our application, the function spaces of interest are $D[0,1]$, endowed with the J_1-topology, and $SM[0,1]$, equipped with the sup-vague topology. Both of these are metrizable and form Polish spaces; see Billingsley, 1999, Section 12 for $D[0,1]$, and Vervaat, 1988, Remark 5.6 for $SM[0,1]$. We set the index set $T = [0,1] \times \mathcal{G}([0,1])$, and define the product space $\mathcal{M} = D[0,1] \times SM[0,1]$, endowed with the corresponding product metric. The state space is $U = \mathbb{R} \times \overline{\mathbb{R}}$. Measurability of the projection mappings π_t then follows from standard results—see Billingsley, 1999, Section 12 for $D[0,1]$ and the definition of the sup-vague topology, as well as the proof of Vervaat, 1988, Theorem 11.1 for $SM[0,1]$.

Let $\xi = (\xi(t))_{t \in T}$ be a stochastic process taking values in a Polish space $\mathcal{M}$, and let $\mathbb{P}_\xi$ denote its law on $\mathcal{M}$. We define the following subset of continuity indices:

$$T_\xi := \{t \in T : \mathbb{P}_\xi \left(\pi_t : \mathcal{M} \to U \text{ is continuous}\right) = 1\}.$$

Following Vervaat, 1988, we introduce key notions related to weak convergence in $\mathcal{M}$. Throughout, we write $\stackrel{d}{=}$ to denote equality in law.

Definition 3. *(Law-Determining and Convergence-Determining Sets). A subset $T_0 \subset T$ is said to be law determining if the following holds: for any two processes ξ_1, ξ_2 with values in $\mathcal{M}$, if for every $d \in \mathbb{N}$ and $t_1, \ldots, t_d \in T_0$,*

$$\pi_{t_1,\ldots,t_d}\xi_1 \stackrel{d}{=} \pi_{t_1,\ldots,t_d}\xi_2 \quad \textit{on } U^d,$$

then $\xi_1 \overset{d}{=} \xi_2$ as random elements in $\mathcal{M}$. That is, the finite-dimensional distributions indexed by T_0 uniquely determine the law on $\mathcal{M}$.

A subset $T_0 \subset T$ is said to be convergence determining if, for any pair of processes ξ_1, ξ_2 with values in $\mathcal{M}$, the set $T_0 \cap T_{\xi_1} \cap T_{\xi_2}$ is law determining.

Definition 4. *(Finite-Dimensional Convergence). We say $\xi_n \xrightarrow{fdd} \xi$ in a subset $T_0 \subset T$, if for every $d \in \mathbb{N}$ and $t_1, \ldots, t_d \in T_0$,*

$$\pi_{t_1,\ldots,t_d}\xi_n \Rightarrow \pi_{t_1,\ldots,t_d}\xi \quad \textit{in distribution on } U^d.$$

When T_0 is omitted, it is understood to be the full index set T.

The family $\{\xi_n\}_{n\in\mathbb{N}}$ is said to be *tight* in $\mathcal{M}$ if for every $\epsilon > 0$, there exists a compact subset $K \subset \mathcal{M}$ such that

$$\inf_n \mathbb{P}(\xi_n \in K) \geq 1 - \epsilon.$$

Proposition A.6. *Suppose $T_0 \subset T$ is a convergence determining set in the sense of Definition 3. Then the weak convergence $\xi_n \Rightarrow \xi$ in $\mathcal{M}$ holds if and only if*

$$\xi_n \xrightarrow{fdd} \xi \quad \textit{in } T_\xi \quad \textit{and} \quad \{\xi_n\} \textit{ is tight in } \mathcal{M}.$$

Moreover, if $T_0 \subset T_\xi$, one may replace T_ξ with T_0 in the statement above.

Proof. "If" part. Suppose that $\{\xi_n\}$ is tight in $\mathcal{M}$, and that $\xi_n \xrightarrow{\text{fdd}} \xi$ in T_ξ. By Prokhorov's theorem (see, e.g., Kallenberg, 2002, Theorem 16.3), any subsequence of $\{\xi_n\}$ admits a further subsequence which converges weakly in $\mathcal{M}$ to a random element $\xi^* \in \mathcal{M}$. To establish the desired convergence $\xi_n \Rightarrow \xi$, it suffices to show that $\mathbb{P}_{\xi^*} = \mathbb{P}_\xi$.

Let $t_1, \ldots, t_d \in T_{\xi^*}$ for arbitrary $d \in \mathbb{N}$. By definition of T_{ξ^*}, the projection mapping $\pi_{t_1,\ldots,t_d} : \mathcal{M} \to U^d$ is continuous $\mathbb{P}_{\xi^*}$-almost surely. Then, by the continuous mapping theorem (see, e.g., Kallenberg, 2002, Theorem 4.27), we obtain

$$\pi_{t_1,\ldots,t_d}\xi_n \Rightarrow \pi_{t_1,\ldots,t_d}\xi^*$$

along the chosen sub-subsequence. On the other hand, by assumption,

$$\pi_{t_1,\ldots,t_d}\xi_n \Rightarrow \pi_{t_1,\ldots,t_d}\xi$$

for all $t_1, \ldots, t_d \in T_0 \cap T_\xi \cap T_{\xi^*}$. By uniqueness of limits in distribution, this implies

$$\pi_{t_1,\ldots,t_d}\xi^* \overset{d}{=} \pi_{t_1,\ldots,t_d}\xi, \quad \text{for all } t_1, \ldots, t_d \in T_0 \cap T_\xi \cap T_{\xi^*}.$$

Since T_0 is convergence determining, this implies $\xi^* \stackrel{d}{=} \xi$, i.e., $\mathbb{P}_{\xi^*} = \mathbb{P}_\xi$. Hence, every subsequence of $\{\xi_n\}$ contains a further subsequence converging weakly to $\mathbb{P}_\xi$, and thus the full sequence $\xi_n \Rightarrow \xi$ in $\mathcal{M}$.

"Only if" part. Weak convergence $\xi_n \Rightarrow \xi$ in $\mathcal{M}$ implies tightness of $\{\xi_n\}$ (again by Kallenberg, 2002, Theorem 16.3). Moreover, for any $d \in \mathbb{N}$ and $t_1, \ldots, t_d \in T_\xi$, the mapping $\pi_{t_1,\ldots,t_d}$ is continuous almost surely with respect to $\mathbb{P}_\xi$, and hence the continuous mapping theorem yields

$$\pi_{t_1,\ldots,t_d}\xi_n \Rightarrow \pi_{t_1,\ldots,t_d}\xi.$$

Therefore, $\xi_n \xrightarrow{\text{fdd}} \xi$ in T_ξ.

$\square$

Remark A.1. *We claim that the subset $T_0 = [0,1] \times \mathcal{I}$ is convergence determining in the sense of Definition 3, where $\mathcal{I}$ denotes the collection of all non-empty open subintervals of $[0,1]$. To justify this, consider two arbitrary random elements $\xi_1 = (Z_1, M_1)$ and $\xi_2 = (Z_2, M_2)$, each taking values in $\mathcal{M} = D[0,1] \times SM[0,1]$. By results in Billingsley, 1999, Section 12 and the proof of Vervaat, 1988, Theorem 12.2, there exist subsets $\mathcal{J}_i \subset [0,1]$ and $\mathcal{I}_i \subset \mathcal{I}$, for $i = 1,2$, such that the complements $[0,1] \setminus \mathcal{J}_i$ and $\mathcal{I} \setminus \mathcal{I}_i$ are countable. Moreover, these subsets can be selected so that projection maps evaluated at any point in $\mathcal{J}_i$ (for $D[0,1]$) or $\mathcal{I}_i$ (for $SM[0,1]$) are continuous with respect to the marginal law of Z_i or M_i, respectively, for $i = 1,2$. This implies that $\mathcal{J}_i \times \mathcal{I}_i \subset T_{\xi_i}$, $i = 1,2$.*

Now, consider the intersection

$$T^* := (\mathcal{J}_1 \cap \mathcal{J}_2) \times (\mathcal{I}_1 \cap \mathcal{I}_2) \subset T_0 \cap T_{\xi_1} \cap T_{\xi_2}.$$

To conclude that T_0 is convergence determining, it suffices to show that T^* is law determining in the sense of Definition 3. By Dynkin's π-λ theorem, it is enough to show that the π-system

$$\left\{\pi^{-1}_{t_1,\ldots,t_d}(U) : d \in \mathbb{N},\ t_1, \ldots, t_d \in T^*,\ U \in \mathcal{B}(U^d)\right\}$$

generates the Borel σ-field $\mathcal{B}(\mathcal{M})$. This follows from the known results in each component space: for $\mathcal{M} = D[0,1]$, with $T^* = \mathcal{J}_1 \cap \mathcal{J}_2$, the result holds by Billingsley, 1999, Theorem 12.5; and for $\mathcal{M} = SM[0,1]$, with $T^* = \mathcal{I}_1 \cap \mathcal{I}_2$, the result follows from Vervaat, 1988, Theorem 11.1. Since both $D[0,1]$ and $SM[0,1]$ are separable spaces, the product Borel σ-field satisfies

$$\mathcal{B}(D[0,1] \times SM[0,1]) = \mathcal{B}(D[0,1]) \otimes \mathcal{B}(SM[0,1]),$$

as noted in Kallenberg, 2002, Lemma 1.2.

We also observe that projection mappings in $D[0,1]$ *are almost surely continuous with respect to the limit laws* B_H *(Theorem 3.1) and* S *(Theorem 3.2), by virtue of Billingsley, 1999, Theorem 12.5, since both the fractional Brownian motion* B_H *and the stable process* S *admit versions with continuous paths almost surely (see Owada and Samorodnitsky, 2015, Theorem 3.3). Similarly, projection mappings in* $SM[0,1]$ *are almost surely continuous with respect to the limit random sup-measure laws, as shown in the proof of Samorodnitsky and Wang, 2019, Proposition 5.2. Therefore, the projection mapping on the product space* $D[0,1] \times SM[0,1]$ *is almost surely continuous with respect to the joint limit law of* ξ *in either Theorem 3.1 or 3.2. It follows that* $T_0 = [0,1] \times \mathcal{I} \subset T_\xi$*, completing the argument.*

A.6 Consistency of Order Selection via Penalized Silhouette

We begin by establishing deterministic bounds related to the k-clustering framework introduced in Definition 2 and the average silhouette width (ASW) criterion in (3.1). Proposition 1.1 suggests that if the true spectral measure H consists of finitely many atoms, then by selecting the subset S as a union of neighborhoods surrounding each atom, almost all points in the extremal subsample W_n (defined in (1.19)) will, for sufficiently large n, lie near one of these atoms with respect to the dissimilarity measure D. This observation motivates the analysis of scenarios where the majority of points in a finite multiset are concentrated around a finite set of centers.

Throughout this section, fix distinct points $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\} =: A \subset \mathbb{S}_+^{d-1}$, with $k \in \mathbb{Z}_+$, and associated weights $p_i > 0$ satisfying $\sum_{i=1}^k p_i = 1$. Let D be the dissimilarity measure defined in Definition 1. We say a finite multiset $W \subset \mathbb{S}_+^{d-1}$ satisfies the concentration condition $\mathcal{A}(\epsilon, \delta)$ for given $\epsilon, \delta > 0$ if the following hold: $k \leq |W| \wedge \delta^{-1}$; $\epsilon \in (0, r_A)$, with r_A defined in (1.25); $\delta \in (0, p_{\min})$, where $p_{\min}$ is defined in (3.3); and

$$\frac{|W \cap B_D(\mathbf{a}_i, \epsilon)|}{|W|} \geq p_i - \delta, \quad \text{for all } i \in \{1, \ldots, k\}. \tag{A.23}$$

In proving the subsequent lemmas, it is necessary to control the dual dissimilarity $D^\dagger$ between points lying within the same D-neighborhood. To this end, we define a uniform upper bound for the dual dissimilarity. Given the set

$A = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ and $s > 0$, define

$$r_A^\dagger(s) = \sup\left\{D^\dagger(\mathbf{a}_i, \mathbf{w}) : i \in \{1, \ldots, k\},\ \mathbf{w} \in B_D(\mathbf{a}_i, s)\right\}. \tag{A.24}$$

Remark A.2. *Note that $r_A^\dagger(s) > 0$ for all $s > 0$, and $r_A^\dagger(s) \to 0$ as $s \to 0$; see Remark 1.1. In particular, for sufficiently small $\epsilon > 0$, we always have $r_A^\dagger(\epsilon) < r_A$.*

When a multiset W satisfies the concentration condition $\mathcal{A}(\epsilon, \delta)$ but is partitioned into fewer than k clusters, then necessarily, at least two of the underlying true clusters are merged. This merging leads to reduced separability between the resulting clusters, which can be effectively detected via the ASW criterion. The following lemma formalizes this intuition.

Lemma A.7. *Suppose a multiset W satisfies the condition $\mathcal{A}(\epsilon, \delta)$ and $1 \leq m < k$. Let $(A_m^*, \mathfrak{C}_m)$ be an m-clustering of W as defined in Definition 2. Then the (unpenalized) ASW $\bar{S}$ satisfies*

$$\bar{S} = \bar{S}(W; A_m^*, \mathfrak{C}_m) \leq 1 - (p_{\min} - \delta)\left(r_A - r_A^\dagger(\epsilon)\right),$$

where $r_A^\dagger$ is as in (A.24).

Proof. Since $m < k$ and $B_D(\mathbf{a}_i, r_A)$'s are disjoint, $i \in \{1, \ldots, k\}$, there exists $\ell \in \{1, \ldots, k\}$ such that $B_D(\mathbf{a}_\ell, r_A) \cap A_m^* = \emptyset$. Hence for any $\mathbf{w} \in W \cap B_D(\mathbf{a}_\ell, \epsilon)$, we have by the triangular inequality (1.14) that

$$a(\mathbf{w}) = D(\mathbf{w}, A_m^*) \geq D(\mathbf{a}_\ell, A_m^*) - D^\dagger(\mathbf{w}, \mathbf{a}_\ell) \geq r_A - r_A^\dagger(\epsilon).$$

Then since $b(\mathbf{w}) \leq 1$, we have

$$\begin{aligned}\frac{1}{|W|}\sum_{\mathbf{w}\in W}\frac{a(\mathbf{w})}{b(\mathbf{w})} \geq \frac{1}{|W|}\sum_{\mathbf{w}\in W} a(\mathbf{w}) &\geq \frac{|W \cap B_D(\mathbf{a}_\ell, \epsilon)|}{|W|}\left(r_A - r_A^\dagger(\epsilon)\right)\\ &\geq (p_{\min} - \delta)\left(r_A - r_A^\dagger(\epsilon)\right),\end{aligned}$$

which implies the desired result. □

The next lemma establishes that if the multiset W satisfies the concentration condition $\mathcal{A}(\epsilon, \delta)$ and is partitioned into at least k clusters, then for each true center $\mathbf{a}_i \in A$, there exists at least one cluster center that lies within a small D-neighborhood of $\mathbf{a}_i$. In other words, no true cluster is left unrepresented among the fitted clusters, provided the number of clusters is no less than the true number k.

Lemma A.8. *Suppose a multiset W satisfies the condition $\mathcal{A}(\epsilon, \delta)$. Let $(A_m^*, \mathfrak{C}_m)$, where $A_m^* = \{\mathbf{a}_1^*, \ldots, \mathbf{a}_m^*\}$, be an m-clustering of W as defined in Definition 2, $m \geq k$. Then for any $i \in \{1, \ldots, k\}$, there exists $j \in \{1, \ldots, m\}$, such that $D\left(\mathbf{a}_j^*, \mathbf{a}_i\right) < \epsilon'$, where*

$$\epsilon' = \epsilon'(\epsilon, \delta) = \frac{(1 - k\delta)\epsilon + k\delta}{p_{\min} - \delta} + r_A^\dagger(\epsilon). \tag{A.25}$$

In particular, when $m = k$ and $\epsilon' < r_A$, there exists a bijection $\pi : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$, such that $D\left(\mathbf{a}_{\pi(i)}^*, \mathbf{a}_i\right) < \epsilon'$ for all $i \in \{1, \ldots, k\}$.

Proof. We prove the first claim by contradiction. Suppose there exists $i \in \{1, \ldots, k\}$ such that $D(\mathbf{a}_j^*, \mathbf{a}_i) \geq \epsilon'$ for all $j \in \{1, \ldots, m\}$. Then for any $\mathbf{w} \in W \cap B_D(\mathbf{a}_i, \epsilon)$, we have by the triangular inequality (1.14) that

$$D(\mathbf{w}, A_m^*) \geq D(\mathbf{a}_i, A_m^*) - D^\dagger(\mathbf{w}, \mathbf{a}_i) \geq \epsilon' - r_A^\dagger(\epsilon).$$

Hence combining this and (A.23),

$$\begin{aligned}\frac{1}{|W|} \sum_{\mathbf{w} \in W} D(\mathbf{w}, A_m^*) \geq & \frac{1}{|W|} \sum_{\mathbf{w} \in W \cap B_D(\mathbf{a}_i, \epsilon)} D(\mathbf{w}, A_m^*) \\ \geq & (p_i - \delta)\left(\epsilon' - r_A^\dagger(\epsilon)\right) \geq (p_{\min} - \delta)\left(\epsilon' - r_A^\dagger(\epsilon)\right).\end{aligned} \tag{A.26}$$

Next, suppose that a multiset S on $\mathbb{S}_+^{d-1}$ contains A and $|S| = m$, which is only possible when $m \geq k$ as assumed. Then we have $D(\mathbf{w}, S) \leq D(\mathbf{w}, A)$. Set $U_\epsilon := W \cap \left(\cup_{i=1}^k B_D(\mathbf{a}_i, \epsilon)\right)$, we have that

$$\frac{1}{|W|} \sum_{\mathbf{w} \in W} D(\mathbf{w}, S) \leq \frac{1}{|W|} \left\{ \sum_{\mathbf{w} \in U_\epsilon} D(\mathbf{w}, A) + \sum_{\mathbf{w} \in W \setminus U_\epsilon} 1 \right\} < (1 - k\delta)\epsilon + k\delta, \tag{A.27}$$

where the last inequality is obtained by maximizing $|W \setminus U_\epsilon|$ with the constraint (A.23). Now in view of (1.17), the first expression in (A.26) is less than or equal to the first expression in (A.27), and hence these two inequalities imply:

$$\epsilon' < \{(1 - k\delta)\epsilon + k\delta\}/(p_{\min} - \delta) + r_A^\dagger(\epsilon),$$

which contradicts the choice of ϵ'.

For the second claim, note that $B_D(\mathbf{a}_i, r_A)$'s are disjoint, $i \in \{1, \ldots, k\}$. So if $\epsilon' < r_A$, it is impossible that $D\left(\mathbf{a}_j^*, \mathbf{a}_i\right) < \epsilon'$ and $D\left(\mathbf{a}_j^*, \mathbf{a}_{i'}\right) < \epsilon'$ hold simultaneously when $i \neq i'$. The conclusion then follows. $\square$

As a consequence of the previous lemma, if the multiset W is concentrated around k centers but is partitioned into more than k clusters, then one of two outcomes must occur: either some of the resulting clusters will have small cardinality, or at least two cluster centers will be close to each other with respect to the dissimilarity measure D. This is formalized in the following lemma.

Lemma A.9. *Suppose a multiset W satisfies the condition $\mathcal{A}(\epsilon, \delta)$. Assume additionally that ϵ' in* (A.25) *satisfies $\epsilon' < r_A$. Let $(A_m^*, \mathfrak{C}_m)$, where $A_m^* = \{\mathbf{a}_1^*, \ldots, \mathbf{a}_m^*\}$ and $\mathfrak{C}_m = \{C_1, \ldots, C_m\}$, be an m-clustering of W as defined in Definition 2, $m > k$. Then either of the following happens:*

$$\min_{i=1,\ldots,m} \frac{|C_i|}{|W|} \leq k\delta \quad \textit{or} \quad \min_{1 \leq i < j \leq m} D(\mathbf{a}_i^*, \mathbf{a}_j^*) \leq \epsilon' + 2r_A^\dagger(\epsilon) + r_A^\dagger(\epsilon').$$

Proof. Since $B_D(\mathbf{a}_i, \epsilon')$, $i \in \{1, \ldots, k\}$, are disjoint (because $\epsilon' < r_A$), by Lemma A.8, we can, without loss of generality, assume that $\mathbf{a}_i^* \in B_D(\mathbf{a}_i, \epsilon')$, $i \in \{1, \ldots, k\}$. We now divide into two cases as follows.
Case 1: there exists one $j \in \{k+1, \ldots, m\}$ (fixed below in the discussion of this case) which satisfies $D(\mathbf{a}_j^*, A) > \epsilon' + 2r_A^\dagger(\epsilon)$. Then for any $i \in \{1, \ldots, k\}$ and any $\mathbf{w} \in W \cap B_D(\mathbf{a}_i, \epsilon)$, we have by the triangular inequality (1.14) that $D(\mathbf{w}, \mathbf{a}_i^*) \leq D(\mathbf{a}_i, \mathbf{a}_i^*) + D^\dagger(\mathbf{a}_i, \mathbf{w}) \leq \epsilon' + r_A^\dagger(\epsilon)$, and hence

$$D(\mathbf{w}, \mathbf{a}_j^*) \geq D(\mathbf{a}_j^*, \mathbf{a}_i) - D^\dagger(\mathbf{w}, \mathbf{a}_i) > \epsilon' + 2r_A^\dagger(\epsilon) - r_A^\dagger(\epsilon) \geq D(\mathbf{w}, \mathbf{a}_i^*).$$

This in view of Definition 2 implies that $W \cap B_D(\mathbf{a}_i, \epsilon) \subset W \cap C_j^c$ for all $i \in \{1, \ldots, k\}$. Therefore, we have by (A.23) that

$$\min_{i=1,\ldots,m} \frac{|C_i|}{|W|} \leq \frac{|C_j|}{|W|} \leq \frac{|W \cap \bigcap_{i=1,\ldots,k} B_D(\mathbf{a}_i, \epsilon)^c|}{|W|} \leq k\delta.$$

Case 2: for any $j \in \{k+1, \ldots, m\}$, we have $D(\mathbf{a}_j^*, \mathbf{a}_i) \leq \epsilon' + 2r_A^\dagger(\epsilon)$ for some $i \in \{1, \ldots, k\}$. Then for any such pair of j and i, we have

$$D(\mathbf{a}_i^*, \mathbf{a}_j^*) \leq D(\mathbf{a}_i, \mathbf{a}_j^*) + D^\dagger(\mathbf{a}_i, \mathbf{a}_i^*) \leq \epsilon' + 2r_A^\dagger(\epsilon) + r_A^\dagger(\epsilon').$$

□

The next lemma states that if the multiset W is concentrated around k centers and is partitioned into exactly k clusters, then the unpenalized average silhouette width $\bar{S}$ will favor this clustering by yielding a high score—close to the ideal value of 1. Moreover, under such a configuration, all clusters will have sufficiently large sizes, and the corresponding cluster centers will be well-separated with respect to the dissimilarity measure D.

Lemma A.10. *Let* $(A_k^* = \{\mathbf{a}_1^*, \ldots, \mathbf{a}_k^*\}, \mathfrak{C}_k = \{C_1, \ldots, C_k\})$ *be a* k*-clustering of* W *as defined in Definition 2. Suppose a multiset* W *satisfies the condition* $\mathcal{A}(\epsilon, \delta)$*. Suppose in addition*

$$r_A > \epsilon' + 2r_A^\dagger(\epsilon) + r_A^\dagger(\epsilon') \tag{A.28}$$

with ϵ' *in* (A.25)*. Then the (unpenalized) ASW* $\bar{S}$ *satisfies*

$$\bar{S} = \bar{S}(W; A_k^*, \mathfrak{C}_k) \geq 1 - (1 - k\delta)\frac{\epsilon' + r_A^\dagger(\epsilon)}{r_A - r_A^\dagger(\epsilon) - r_A^\dagger(\epsilon')} - k\delta.$$

In addition, with the same permutation $\pi : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$ *found in Lemma A.8, we have*

$$\frac{|C_{\pi(i)}|}{|W|} \geq p_i - \delta \text{ for each } i \quad \text{and} \quad \min_{1 \leq i < j \leq k} D(\mathbf{a}_i^*, \mathbf{a}_j^*) \geq r_A - 2r_A^\dagger(\epsilon'),$$

where when $k = 1$, $\min_{1 \leq i < j \leq k} D(\mathbf{a}_i^*, \mathbf{a}_j^*)$ *is understood as 1, and the inequalities still hold.*

Proof. Since $r_A > \epsilon'$, by Lemma A.8, there exists a permutation $\pi : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$, such that $D(\mathbf{a}_i, \mathbf{a}_{\pi(i)}^*) < \epsilon'$, $i \in \{1, \ldots, k\}$. Then for each i and any $\mathbf{w} \in B_D(\mathbf{a}_i, \epsilon)$, we have by the triangular inequality (1.14) that

$$D(\mathbf{w}, \mathbf{a}_{\pi(i)}^*) \leq D(\mathbf{a}_i, \mathbf{a}_{\pi(i)}^*) + D^\dagger(\mathbf{w}, \mathbf{a}_i) < \epsilon' + r_A^\dagger(\epsilon), \tag{A.29}$$

and for $j \neq i$ that

$$D\left(\mathbf{w}, \mathbf{a}_{\pi(j)}^*\right) \geq D\left(\mathbf{a}_i, \mathbf{a}_j\right) - D^\dagger(\mathbf{a}_j, \mathbf{a}_{\pi(j)}^*) - D^\dagger(\mathbf{w}, \mathbf{a}_i) \geq r_A - r_A^\dagger(\epsilon') - r_A^\dagger(\epsilon), \tag{A.30}$$

where if $k = 1$, the left-hand side $D\left(\mathbf{w}, \mathbf{a}_{\pi(j)}^*\right)$ in (A.30) is understood as 1, and the inequality still holds. Writing as before $U_\epsilon = \bigcup_{1 \leq i \leq k} B_D(\mathbf{a}_i, \epsilon) \cap W$. In view of (A.23) and the inequalities above, we have

$$\begin{aligned}\bar{S} =& \frac{1}{|W|}\left\{\sum_{\mathbf{w} \in U_\epsilon} + \sum_{\mathbf{w} \in W \setminus U_\epsilon}\right\}\left(1 - \frac{a(\mathbf{w})}{b(\mathbf{w})}\right) \\ \geq& \left(1 - \frac{\epsilon' + r_A^\dagger(\epsilon)}{r_A - r_A^\dagger(\epsilon) - r_A^\dagger(\epsilon')}\right)(1 - k\delta) + 0,\end{aligned}$$

which implies the first claim.

For the second claim, in view of Definition 2, (A.28), (A.29), (A.30), we have $W \cap B_D(\mathbf{a}_i, \epsilon) \subset C_{\pi(i)}$, $i \in \{1, \ldots, k\}$. Hence by (A.23),

$$\frac{|C_{\pi(i)}|}{|W|} \geq \frac{|W \cap B_D(\mathbf{a}_i, \epsilon)|}{|W|} \geq p_i - \delta.$$

Furthermore, for any $1 \leq i < j \leq k$ and $k > 1$,

$$D(\mathbf{a}^*_{\pi(i)}, \mathbf{a}^*_{\pi(j)}) \geq D(\mathbf{a}_i, \mathbf{a}_j) - D^\dagger\left(\mathbf{a}_i, \mathbf{a}^*_{\pi(i)}\right) - D^\dagger\left(\mathbf{a}_j, \mathbf{a}^*_{\pi(j)}\right) \geq r_A - 2r^\dagger_A(\epsilon').$$

□

We begin by stating a result concerning the average silhouette width $\bar{S}$ defined in (3.1), in the case where the number of clusters is less than or equal to k, the true number of atoms in the underlying discrete spectral measure.

Proposition A.11. *Suppose* $\mathbf{X}$ *satisfying* (1.18) *and* (1.11) *has a discrete spectral measure of the form* $H = \sum_{i=1}^k p_i \delta_{\mathbf{a}_i}$, *where* $\mathbf{a}_i$*'s are distinct points on* $\mathbb{S}^{d-1}_+$, *and* $p_i > 0$, $p_1 + \cdots + p_k = 1$. *Let* W_n *denote the extremal subsample as in* (1.19), *and* $(A_{m,n}, \mathfrak{C}_{m,n})$ *form an* m*-clustering of* W_n *as defined in Definition 2 with respect to a dissimilarity measure* D *defined in Definition 1. If* $m < k$, *then almost surely,*

$$\limsup_n \bar{S}(W_n; A_{m,n}, \mathfrak{C}_{m,n}) \leq 1 - r_A p_{\min},$$

where r_A *is as in* (1.25) *and* $p_{\min}$ *is as in* (3.3). *If* $m = k$, *then almost surely,*

$$\lim_n \bar{S}(W_n; A_{k,n}, \mathfrak{C}_{k,n}) = 1.$$

Proof. Note that $\epsilon' \to 0$ as $\epsilon, \delta \to 0$, we may choose them small enough such that (A.28) is satisfied. Define the event

$$E_n(\epsilon, \delta) = \{|W_n \cap B_D(\mathbf{a}_i, \epsilon)| \geq |W_n|(p_i - \delta),\ i \in \{1, \ldots, k\}\}. \quad \text{(A.31)}$$

By Proposition 1.1 with $S = B_D(\mathbf{a}_i, \epsilon)$ and the choice $\epsilon < r_A$, we have each $H_n(S) = |W_n \cap B_D(\mathbf{a}_i, \epsilon)|/|W_n|$ converges almost surely to $H(S) = p_i$, $i \in \{1, \ldots, k\}$. Hence, with probability 1, the event $E_n(\epsilon, \delta)$ happens eventually as $n \to \infty$, namely, $\mathbb{P}(\liminf_n \mathbf{1}\{E_n(\epsilon, \delta)\} = 1) = 1$. Since W_n satisfies the condition $\mathcal{A}(\epsilon, \delta)$ on $E_n(\epsilon, \delta)$, by Lemmas A.7 and A.10, for almost every outcome ω in the sample space Ω, when n is sufficiently large, we have when

$m < k$ that

$$\bar{S}(W_n; A_{m,n}, \mathfrak{C}_{m,n})\mathbf{1}\{E_n(\epsilon,\delta)\}$$
$$\leq \left\{1 - (p_{\min} - \delta)\left(r_A - r_A^{\dagger}(\epsilon)\right)\right\}\mathbf{1}\{E_n(\epsilon,\delta)\}$$

and

$$\bar{S}(W_n; A_{k,n}, \mathfrak{C}_{k,n})\mathbf{1}\{E_n(\epsilon,\delta)\}$$
$$\geq \left\{1 - (1-k\delta)\frac{\epsilon' + r_A^{\dagger}(\epsilon)}{r_A - r_A^{\dagger}(\epsilon) - r_A^{\dagger}(\epsilon')} - k\delta\right\}\mathbf{1}\{E_n(\epsilon,\delta)\}.$$

The desired results follow if one takes $\limsup_n$ and $\liminf_n$ respectively in the two inequalities above, and then lets $\delta, \epsilon \to 0$ (see also Remark A.2). □

Next, we present a result concerning the penalty term P_t defined in (3.2), in the setting where the number of clusters is greater than or equal to k.

Proposition A.12. *Suppose* $\mathbf{X}$ *satisfying* (1.18) *and* (1.11) *has a discrete spectral measure of the form* $H = \sum_{i=1}^k p_i \delta_{\mathbf{a}_i}$, *where* $\mathbf{a}_i$*'s are distinct points on* $\mathbb{S}_+^{d-1}$, *and* $p_i > 0$, $p_1 + \cdots + p_k = 1$. *Let* W_n *denote the extremal subsample as in* (1.19), *and* $(A_{m,n}, \mathfrak{C}_{m,n})$ *form an* m*-clustering of* W_n *as defined in Definition 2 with respect to a dissimilarity measure* D *defined in Definition 1. Suppose* $t > 0$. *If* $m > k$, *we have almost surely*

$$\lim_n P_t(W_n; A_{m,n}, \mathfrak{C}_{m,n}) = 1.$$

If $m = k$, *we have almost surely*

$$\limsup_n P_t(W_n; A_{k,n}, \mathfrak{C}_{k,n}) \leq 1 - (r_A k p_{\min})^t,$$

where r_A *is as in* (1.25) *and* $p_{\min}$ *is as in* (3.3).

Proof. The argument is similar to that of Proposition A.11. In particular, under the restriction to the event $E_n(\epsilon,\delta)$ in (A.31), we have by Lemma A.9 that for $m > k$

$$P_t(W_n; A_{m,n}, \mathfrak{C}_{m,n}) \geq 1 - (k^2\delta)^t \vee \left(\epsilon' + 2r_A^{\dagger}(\epsilon) + r_A^{\dagger}(\epsilon')\right)^t,$$

and by Lemma A.10 that

$$P_t(W_n; A_{k,n}, \mathfrak{C}_{k,n}) \leq 1 - [k(p_{\min} - \delta)(r_A - 2r_A^{\dagger}(\epsilon'))]^t.$$

We omit the rest of the details. □

Now we are ready to prove Theorem 3.5.

Proof of Theorem 3.5: Putting together Propositions A.11 and A.12, and using the facts that $\bar{S} \in [0,1]$ and $P_t \in [0,1]$, we have almost surely that

$$\begin{cases} \limsup_n S_t(W_n; A_{m,n}, \mathfrak{C}_{m,n}) \leq 1 - r_A p_{\min}, & \text{if } m < k; \\ \liminf_n S_t(W_n; A_{k,n}, \mathfrak{C}_{k,n}) \geq (r_A k p_{\min})^t, & \text{if } m = k; \\ \limsup_n S_t(W_n; A_{m,n}, \mathfrak{C}_{m,n}) \leq 0, & \text{if } m > k. \end{cases}$$

Therefore, the desired claim follows. □

Bibliography

Anderson, C. W., & Turkman, K. F. (1991). The joint limiting distribution of sums and maxima of stationary sequences. *Journal of Applied Probability*, *28*(1), 33–44. https://doi.org/10.2307/3214738

Anderson, C. W., & Turkman, K. F. (1995). Sums and maxima of stationary sequences with heavy tailed distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, *57*(1), 1–10.

Bai, S., Owada, T., & Wang, Y. (2020). A functional non-central limit theorem for multiple-stable processes with long-range dependence. *Stochastic Processes and their Applications*, *130*(9), 5768–5801. https://doi.org/10.1016/j.spa.2020.04.007

Balkema, A. A., & Resnick, S. I. (1977). Max-infinite divisibility. *Journal of Applied Probability*, *14*(2), 309–319.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2006). *Statistics of extremes: Theory and applications*. John Wiley & Sons.

Bertoin, J. (1999). Subordinators: Examples and applications. In P. Bernard (Ed.), *Lectures on probability theory and statistics* (pp. 1–91, Vol. 1717). Springer. https://doi.org/10.1007/BFb0096057

Billingsley, P. (1999). *Convergence of probability measures* (2nd). Wiley.

Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1989). *Regular variation*. Cambridge University Press.

Chow, T. L., & Teugels, J. L. (1978). The sum and the maximum of i.i.d. random variables. *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 81–92.

Cooley, D., & Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, *106*(3), 587–604.

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction*. Springer. https://doi.org/10.1007/0-387-34471-3

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, *42*, 143–175.

Einmahl, J. H., Krajina, A., & Segers, J. (2012). An m-estimator for tail dependence in arbitrary dimensions.

Engelke, S., & Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, *8*, 241–270.

Feller, W. (1971). *An introduction to probability theory and its applications, volume ii* (2nd). John Wiley & Sons.

Fomichov, V., & Ivanovs, J. (2023). Spherical clustering in detection of groups of concomitant extremes. *Biometrika*, *110*(1), 135–153.

Fougères, A.-L., Mercadier, C., & Nolan, J. P. (2013). Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, *116*, 109–129.

Galvin, F., & Shore, S. (1984). Completeness in semimetric spaces. *Pacific Journal of Mathematics*, *113*(1), 67–75.

German Federal Institute of Hydrology. (n.d.). Global runoff data centre (grdc) portal.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic Press.

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *66*(3), 497–546.

Ho, H.-C., & Hsing, T. (1996). On the asymptotic joint distribution of the sum and maximum of stationary normal random variables. *Journal of Applied Probability*, *33*(1), 138–145. https://doi.org/10.2307/3215271

Ho, H.-C., & McCormick, W. P. (1999). Asymptotic distribution of sum and maximum for gaussian processes. *Journal of Applied Probability*, *36*(4), 1031–1044. https://doi.org/10.1239/jap/1031

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13–30. https://doi.org/10.1080/01621459.1963.10500830

Hornik, K., Feinerer, I., Kober, M., & Buchta, C. (2012). Spherical k-means clustering. *Journal of Statistical Software*, *50*(10), 1–22. https://doi.org/10.18637/jss.v050.i10

Hruschka, E. R., de Castro, L. N., & Campello, R. J. (2004). Evolutionary algorithms for clustering gene-expression data. *Fourth IEEE International Conference on Data Mining (ICDM'04)*, 403–406.

Hsing, T. (1995). A note on the asymptotic independence of the sum and maximum of strongly mixing stationary random variables. *The Annals of Probability*, *23*(2), 938–947. https://doi.org/10.1214/aop/1176988296

Hsu, P.-L., & Robbins, H. (1947). Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences*, *33*(2), 25–31.

Ibragimov, I. A., & Linnik, Y. V. (1971). *Independent and stationary sequences of random variables*. Wolters-Noordhoff.

Janßen, A., & Wan, P. (2020). K-means clustering of extremes. *Electronic Journal of Statistics*, *14*(1), 1211–1233.

Kallenberg, O. (2002). *Foundations of modern probability* (2nd). Springer.

Kettleborough, G., & Rayward-Smith, V. J. (2013). Optimising sum-of-squares measures for clustering multisets defined over a metric space. *Discrete Applied Mathematics*, *161*(16–17), 2499–2513.

Krizmanić, D. (2020). On joint weak convergence of partial sum and maxima processes. *Stochastics*, *92*(6), 876–899. https://doi.org/10.1080/17442508.2019.1652543

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer.

McCormick, W. P., & Qi, Y. (2000). Asymptotic distribution for the sum and maximum of gaussian processes. *Journal of Applied Probability*, *37*(4), 958–971. https://doi.org/10.1239/jap/1014843076

Medina, M. A., Davis, R. A., & Samorodnitsky, G. (2024). Spectral learning of multivariate extremes. *Journal of Machine Learning Research*, *25*, 1–36.

Molchanov, I. (2005). *Theory of random sets*. Springer. https://doi.org/10.1007/3-540-25638-7

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NeurIPS)*, *14*.

O'Brien, G. L., Torfs, P. J. J. F., & Vervaat, W. (1990). Stationary self-similar extremal processes. *Probability Theory and Related Fields*, *87*(1), 97–119. https://doi.org/10.1007/BF01217748

Owada, T., & Samorodnitsky, G. (2015). Functional central limit theorem for heavy tailed stationary infinitely divisible processes generated by conservative flows. *The Annals of Probability*, *43*(1), 240–285. https://doi.org/10.1214/13-AOP901

Resnick, S. I. (1987). *Extreme values, regular variation, and point processes*. Springer-Verlag.

Rosiński, J., & Samorodnitsky, G. (1996). Classes of mixing stable processes. *Bernoulli*, *2*(4), 365–377. https://doi.org/10.2307/3318412

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.

Salinetti, G., & Wets, R. J.-B. (1981). On the convergence of closed-valued measurable multifunctions. *Transactions of the American Mathematical Society*, *266*(1), 275–289. https://doi.org/10.2307/1998407

Samorodnitsky, G. (2016). *Stochastic processes and long range dependence* (Vol. 26). Springer.

Samorodnitsky, G., & Wang, Y. (2019). Extremal theory for long range dependent infinitely divisible processes. *The Annals of Probability*, *47*(4), 2529–2562. https://doi.org/10.1214/18-AOP1301

Sato, K. (1999). *Lévy processes and infinitely divisible distributions* (Vol. 68). Cambridge University Press.

Southworth, H., Heffernan, J. E., & Metcalfe, P. D. (2024). *Texmex: Statistical modelling of extreme values* [R package version 2.4.8]. https://CRAN.R-project.org/package=texmex

Vervaat, W. (1988). *Random upper semicontinuous functions and extremal processes* (Research Report No. R 8801). Department of Mathematical Statistics, University of Amsterdam.

Wan, P., & Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, *22*(1), 131–166.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, *97*(4), 893–904.

Whitt, W. (2002). *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer.

Yuen, R., & Stoev, S. (2014). Crps m-estimation for max-stable models. *Extremes*, *17*, 387–410.