Point process convergence of random walks and the estimation of multivariate heavy–tailed distributions

Jorge Yslas

PhD Thesis

This thesis has been submitted to the PhD School of the Faculty of Science, University of Copenhagen

July 31, 2020

Department of Mathematical Sciences
Faculty of Science
University of Copenhagen

Jorge Yslas
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København Ø
Denmark
jorge@math.ku.dk
jorge.yslas1@gmail.com

Supervisor:	Prof. Thomas Mikosch University of Copenhagen
Co-supervisor:	Prof. Mogens Bladt University of Copenhagen
Assessment committee:	Prof. Olivier Wintenberger Université Pierre & Marie Curie Paris (Sorbonne)
	Prof. Søren Asmussen Aarhus University
	Prof. Jostein Paulsen (chairman) University of Copenhagen

ISBN: 978-87-7125-032-9

Abstract

This thesis is concerned with point process convergence for sequences of random walks and the estimation of inhomogeneous phase–type distributions.

First, we study point process convergence for sequences of i.i.d. random walks. The objective is to derive asymptotic theory for the extremes of these random walks. We show convergence of the maximum random walk to the Gumbel or the Fréchet distributions. We make heavily use of precise large deviation results for sums of i.i.d. random variables. In particular, we show convergence of the maximum random walk to the Gumbel distribution under the existence of a $(2 + \delta)$th moment, and as a consequence, we derive the joint convergence of the off–diagonal entries in sample covariance and correlation matrices of a high–dimensional sample whose dimension increases with the sample size.

Then, we provide a fitting procedure for the class of inhomogeneous phase–type distributions introduced in [3]. We propose a multivariate extension of inhomogeneous phase–type distributions as functionals of elements of Kulkarni's multivariate phase–type class [73], and study parameter estimation for the resulting new and flexible class of multivariate distributions.

Resumé

Denne afhandling omhandler punktproceskonvergens for følger af random walks og estimation af inhomogene fasetypefordelinger.

Først studerer vi punktproceskonvergens for følger af i.i.d. random walks. Målsætningen er at udlede asymptotisk teori for ekstremaer af disse random walks. Vi viser konvergens af den maksimale random walk til Gumbel- eller Fréchetfordelinger. Vi benytter i høj grad præcise large deviations resultater for summer af i.i.d. stokastiske variable. Navnlig viser vi konvergens af den maksimale random walk til Gumbelfordelingen under antagelse af eksistens af et $(2 + \delta)$'te moment, og afledt heraf udleder vi simultan konvergens af ikke-diagonal elementerne i empirisk kovarians- og korrelationsmatricer for en højdimensionel stikprøve hvis dimension vokser med stikprøvestørrelsen.

Derpå giver vi en estimationsprocedure for klassen af inhomogene fasetypefordelinger introduceret i [3]. Vi foreslår en flerdimensionel udvidelse af inhomogene fasetypefordelinger som funktionaler af elementer af Kulkarnis flerdimensionelle fasetypeklasse [73] og studerer parameterestimation for denne nye og fleksible klasse af flerdimensionelle fordelinger.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Thomas Mikosch for his guidance during this process. Your continuous support has been invaluable to me and I could not have wished for a better supervisor. I also owe my gratitude to my co-supervisor Prof. Mogens Bladt who has been supporting me since my Master studies. I thank them for all the effort they put into my development as a researcher. Their advice and expertise have always been invaluable to me.

I would also like to express my deep gratitude to my co-authors Prof. Hansjörg Albrecher and Dr. Johannes Heiny. Collaboration with them has been not only fruitful but also enjoyable. Special thanks to Prof. Hansjörg Albrecher who also hosted me at the University of Lausanne for three months. Thank you for welcoming me in your group and for the fascinating discussions.

Thanks to all my colleagues and friends at the University of Copenhagen. You made my time at the university an extraordinary experience. I am extremely grateful to Laura who has been an exceptional source of support during the most challenging parts of this process.

I acknowledge the support by Danmarks Frie Forskningsfond Grant No 9040-00086B.

Last but not least, I would like to thank the PhD Committee for spending their precious time to review and assess this thesis.

Jorge Yslas *Copenhagen, July* 2020.

Summary

This PhD thesis provides results on point process convergence for sequences of random walks and estimation methods for inhomogeneous phase–type distributions. Its main part consists of the following three research papers. They were written from August 2017 until July 2020.

[P1] Mikosch, T., and Yslas, J. Gumbel and Fréchet convergence of the maxima of independent random walks. *Advances in Applied Probability* 52(1) (2020), 213–236. [pdf]

[P2] Heiny, J., Mikosch, T., and Yslas, J. Point process convergence for the off–diagonal entries of sample covariance matrices. *To appear in Annals of Applied Probability.*

[P3] Albrecher, H., Bladt, M., and Yslas J. Fitting inhomogeneous phase–type distributions to data: The univariate and the multivariate case. *Submitted for publication.*

Jorge Yslas *Copenhagen, July* 2020.

Contents

Chapter 1

Introduction

In this chapter we introduce some of the mathematical tools needed for the formulation and proofs of the results in the subsequent chapters.

1.1 Some elements from extreme value theory

The main objective of extreme value theory is the study of rare/extreme events. From a mathematical point of view, an event is extreme if its probability is very small. An important aim of this field of probability theory is to find general rules and connections that apply not only to one specific model/distribution but rather to whole classes of models. This task is similar to the central limit problem in classical probability theory where one was searching for the possible limit distributions of (centered and normalized) sample means of i.i.d. data, conditions on the summands to converge to a given limit distribution as well as possible centering and normalizing constants.

The limit behavior of the maxima of i.i.d. random variables belongs to the same kind of problems. A seminal solution to it was given by Gnedenko [54] in 1943; the main result is commonly referred to as Fisher–Tippett Theorem; see Fisher and Tippet [49]. It specifies the possible types of the limit distribution of centered and normalized maxima of an i.i.d. sequence. We will formulate this result next.

Consider an i.i.d. sequence (X_i) of random variables with generic element X, distribution/distribution function F and right tail $\overline{F} = 1 - F$, and set

$$M_n = \max(X_1, \ldots, X_n)\,, \qquad n \geq 1\,.$$

We call a non–degenerate distribution H *extreme value distribution* if there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$, $n \geq 1$, such that the sequence of the maxima (M_n) satisfies the limit relation

$$c_n^{-1}(M_n - d_n) \xrightarrow{\mathrm{d}} Y \sim H\,, \qquad n \to \infty\,. \tag{1.1}$$

In this case, there exists $\gamma \in \mathbb{R}$ such that H is of the type of the one–parametric family of distributions given by

$$H_\gamma(x) = \exp\big(-(1+\gamma x)_+^{-1/\gamma}\big)\,, \quad x \in \mathbb{R}\,, \tag{1.2}$$

where for any $y \in \mathbb{R}$, $y_+ = \max(y, 0)$. This is the content of the Fisher–Tippett Theorem. The parametric family (1.2) is also called *Jenkinson–von–Mises representation* of an extreme value distribution or *generalized extreme value distribution* (GEV). According to the sign of γ, one distinguishes between three cases for H_γ:

$\gamma > 0$; In this case,

$$H_\gamma(x) = \exp\big(-(1+\gamma x)^{-1/\gamma}\big)\,, \quad x > -1/\gamma\,.$$

This distribution is of the type of the Fréchet distribution $\Phi_\alpha(x) = \exp(-x^{-\alpha})$, $x > 0$, with $\alpha = 1/\gamma$. We will refer to the *Fréchet*–case.

$\gamma = 0$; In this case, we understand H_0 as the limit

$$H_0(x) = \lim_{\gamma \to 0} H_\gamma(x) = \exp(-\exp(-x)) = \Lambda(x)\,, \quad x \in \mathbb{R}$$

which corresponds to the Gumbel distribution; we refer to the *Gumbel*–case.

$\gamma < 0$; In this case,

$$H_\gamma(x) = \exp\big(-(1+\gamma x)^{1/|\gamma|}\big)\,, \quad x < 1/|\gamma|\,.$$

The distribution is of the type of the Weibull distribution $\Psi_\alpha(x) = \exp(-(-x)^\alpha)$, $x < 0$, with $\alpha = 1/|\gamma|$; we refer to the *Weibull*–case.

1.1.1 Maximum domains of attraction

A related problem is the characterization of the distributions F satisfying (1.1) for a given extreme value distribution $H = H_\gamma$. This leads to the notion of *maximum domain of attraction* of H ($F \in \mathrm{MDA}(H)$).

The maximum domain of attraction of an extreme value distribution is determined by the conditional behavior of a single random variable, given that it exceeds a certain (high) threshold. This is the content of a famous result proved independently by Pickands [100] and Balkema and de Haan [14]: a distribution $F \in \mathrm{MDA}(H_\gamma)$ if and only if there exists a measurable function $a : \mathbb{R} \to \mathbb{R}_+$ such that

$$\lim_{u \uparrow x_F} \frac{\overline{F}(u + a(u)\,x)}{\overline{F}(u)} = \overline{G}_\gamma(x) := -\log(H_\gamma(x))\,, \qquad x \in \mathbb{R}\,, \tag{1.3}$$

where $x_F := \sup\{x \in \mathbb{R} : F(x) < 1\}$ is the *upper endpoint* of the distribution F. When restricted to positive x, $\overline{G}_\gamma(x)$ is indeed the tail of a distribution G_γ, called *generalized Pareto distribution* (GPD). This result is often reformulated in terms of the *excess distribution function* $F_u(x) := \mathbb{P}(X \le x + u \,|\, X > u)$ as follows:

$$\overline{F}_u(x) \approx \overline{G}_{\gamma,\beta(u)}(x)\,, \quad x > 0\,,$$

for some function β, where $G_{\gamma,\sigma}(x) := G_\gamma(x/\sigma)$ and u is large enough. This approximation is the theoretical basis for a major statistical tool: the *Peaks–Over–Threshold method* (POT).

In this thesis we will deal only with the Fréchet and Gumbel extreme value distributions. Next, we characterize the domains of attraction of these two distributions.

The domain of attraction of the Fréchet distribution is closely connected with the concept of *regular variation*. Recall that a measurable function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is called *regularly varying with index* $\rho \in \mathbb{R}$, if $\lim_{t\to\infty} f(xt)/f(t) = x^\rho$ for all $x > 0$. We write $f \in \mathrm{RV}_\rho$ and call f *slowly varying* if $\rho = 0$. Regularly varying functions can be represented as the product of a power function with a slowly varying function, i.e., $f \in \mathrm{RV}_\rho$ if and only if $f(x) = L(x)\,x^\rho$ for some slowly varying function L. Regularly varying functions have been extensively studied in real analysis and probability theory; we refer to Bingham et al. [21] for a comprehensive account.

A distribution $F \in \mathrm{MDA}(\Phi_\alpha)$ for some $\alpha > 0$ if and only if $\overline{F} \in \mathrm{RV}_{-\alpha}$ (see [46, Section 3.3.1]), and then

$$c_n^{-1} M_n \xrightarrow{\mathrm{d}} Y \sim \Phi_\alpha\,, \qquad n \to \infty\,, \tag{1.4}$$

where (c_n) can be chosen such that $n\,\mathbb{P}(X > c_n) \to 1$. Thus, the maximum domain of attraction of Φ_α consists of the distributions F whose right tail is regularly varying with index $-\alpha$. This class of distributions contains very heavy–tailed distributions in the sense that $\mathbb{E}(X_+^k) = \infty$, for any $k > \alpha$.

Example 1.1 (Pareto distribution)**.** The Pareto distribution $F(x) = 1 - x^{-\alpha}$, $x > 1$, $\alpha > 0$, satisfies $F \in \mathrm{MDA}(\Phi_\alpha)$ and (1.4) holds with $c_n = n^{1/\alpha}$.

The domain of attraction of the Gumbel distribution can be described in terms of the so–called von Mises functions: a distribution function F is called a *von Mises function* if there exists $z < x_F$ such that F has the representation

$$\overline{F}(x) = c \exp\Big(- \int_z^x \frac{1}{a(t)} dt \Big), \quad z < x < x_F\,,$$

where c is a positive constant and $a(\cdot)$ is a positive and absolutely continuous function with density $a'(x) \to 0$ as $x \to x_F$. Recall that two distribution functions F and G are called *tail–equivalent* if they have the same right endpoint and $\lim_{x \to x_F} \overline{F}(x)/\overline{G}(x) = c$ for some positive constant c. Now, $F \in \mathrm{MDA}(\Lambda)$ if and only if F is tail–equivalent to a von Mises function (see [46, Section 3.3.3]). Moreover,

$$c_n^{-1}(M_n - d_n) \xrightarrow{\mathrm{d}} Y \sim \Lambda\,, \qquad n \to \infty\,, \tag{1.5}$$

where (d_n) can be chosen such that $n\,\mathbb{P}(X > d_n) \to 1$ and $c_n = a(d_n)$.

This domain of attraction contains distribution functions with very different tail behaviors, from heavy–tailed distributions (such as the lognormal) to light–tailed distributions (such as the exponential distribution). A distinctive difference with respect to the Fréchet case is that $\mathbb{E}(X_+^k) < \infty$, for any $k > 0$. It is also possible to have distributions with finite and infinite right endpoints x_F, but for the purposes of this thesis we will only deal with distributions with infinite right endpoint.

Example 1.2 (Normal distribution)**.** The standard normal distribution $\Phi \in \mathrm{MDA}(\Lambda)$ and satisfies (1.5) with $c_n = 1/d_n$ and

$$d_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2(2\log n)^{1/2}}\,. \tag{1.6}$$

Since $d_n \sim \sqrt{2\log n}$ we can replace c_n in (1.5) by $1/\sqrt{2\log n}$ while d_n cannot be replaced by $\sqrt{2\log n}$.

We conclude by recalling the characterization (1.3) of a maximum domain of attraction. Switching from the continuous threshold u to a sequence (d_n) satisfying $n\,\mathbb{P}(X > d_n) \to 1$ we have $F \in \mathrm{MDA}(H_\gamma)$ if and only if for any $x \in \mathbb{R}$, a positive measurable function $a(\cdot)$ and $c_n = a(d_n)$,

$$n\,\mathbb{P}(X > d_n + x\,c_n) \to -\log(H_\gamma(x))\,, \qquad n \to \infty\,. \tag{1.7}$$

This characterization creates a link with the notion of *vague convergence* and, in turn, combines extreme value theory, large deviation theory and the convergence of point processes. In this thesis, we will frequently make use of this connection.

1.1.2 Vague convergence

Let E be a locally compact topological space with countable base. This is a technical assumption, and we will work on sets that satisfy this condition. We equip E with the Borel σ–field $\mathcal{E}$, i.e., the σ–field generated by the open sets of E. Recall that a measure μ on $\mathcal{E}$ is *Radon* if $\mu(A) < \infty$ for all compact sets $A \subset E$. Now define

$$M_+(E) = \{\mu : \mu \text{ is a non-negative Radon measure on } E\}$$

and

$$C_K^+(E) = \{f : E \to \mathbb{R}_+ : f \text{ is continuous with compact support}\}.$$

Let (μ_n) be a sequence of measures on $M_+(E)$ and $\mu \in M_+(E)$. We say that μ_n *converges vaguely* to μ, and we write $\mu_n \xrightarrow{v} \mu$, if

$$\int_E f(x)d\mu_n(x) \to \int_E f(x)d\mu(x), \qquad n \to \infty, \qquad f \in C_K^+(E).$$

A practical equivalent way of proving vague convergence is to show $\mu_n(A) \to \mu(A)$ for all relatively compact sets $A \in \mathcal{E}$ for which $\mu(\partial A) = 0$, where ∂A is the boundary of A; see [103, Proposition 3.12]. In particular, if $E \subset \mathbb{R}$ it suffices to show that $\mu_n(a,b] \to \mu(a,b]$ for all intervals $(a,b] \subset E$, which are also μ–continuity sets.

Example 1.3. We know that $F \in \mathrm{MDA}(H_\gamma)$ if and only if there exist sequences (c_n) of positive numbers and (d_n) of real numbers such that (1.7) holds. Keeping in mind the comment about the restriction of vague convergence to intervals $(a,b]$, we conclude that (1.7) is equivalent to

$$\mu_n(\cdot) := n\,\mathbb{P}(c_n^{-1}(X - d_n) \in \cdot) \xrightarrow{v} \nu_\gamma(\cdot), \qquad n \to \infty \tag{1.8}$$

in $M_+(\mathrm{supp}(H_\gamma))$, where ν_γ is given by $\nu_\gamma(x,\infty) = -\log(H_\gamma(x))$, $x \in \mathrm{supp}(H_\gamma)$.

The characterization of MDA in (1.8) will be particularly useful when dealing with the weak convergence of point processes; see Section 1.1.3. Moreover, the abstract MDA–condition (1.8) allows one to extend MDA–conditions from the univariate to the multivariate case, for example the regular variation property for $\mathrm{MDA}(\Phi_\alpha)$ to a multivariate setting.

1.1.3 Point processes and their weak convergence

Point processes techniques give a thorough understanding of many structural results in extreme value theory. In contrast to the classical techniques described above, point processes allow for the treatment of extremes of sequences more general than i.i.d. in a straightforward and unified way. One can think of a point process N simply as a random distribution of points X_i in space. More specifically, the state space where the points live is denoted by E, and E is equipped with the Borel σ–field $\mathcal{E}$. A point process is conveniently written by using the *Dirac measure* ε_x for $x \in \mathcal{E}$:

$$\varepsilon_x(A) = \begin{cases} 1 & \text{if} \quad x \in A, \\ 0 & \text{if} \quad x \notin A, \end{cases} \qquad A \in \mathcal{E}.$$

Consider a sequence of points $(x_i)_{i\geq 1}$ in E. Then

$$m = \sum_{i=1}^{\infty} \varepsilon_{x_i} ,$$

defines a *counting measure* on $\mathcal{E}$. If $m(K) < \infty$ for all compact sets $K \in \mathcal{E}$, then m is called a *point measure*. Let $M_p(E)$ be the space of all point measures on E equipped with the σ–algebra $\mathcal{M}_p(E)$, the smallest σ–algebra making all the evaluation maps $m \mapsto m(A)$ measurable for all $A \in \mathcal{E}$. Then, a *point process* on E is a measurable map

$$N : [\Omega, \mathcal{F}, P] \to [M_p(E), \mathcal{M}_p(E)] .$$

The distribution of N is fully determined by its Laplace functional Ψ_N given by

$$\Psi_N(f) := \mathbb{E}\Big[\exp\Big\{ -\int_E f dN \Big\}\Big] = \int_{M_p(E)} \exp\Big\{ -\int_E f(x) dm(x) \Big\}\, dP_N(m) ,$$

for non–negative measurable functions f on E.

Perhaps the most important point process in the context of extreme value theory is the *Poisson random measure*.

Example 1.4 (Poisson random measure (PRM))**.** A point process N is called a *Poisson process* or a *Poisson random measure* with mean measure μ (we write PRM(μ)) if N satisfies:

(a) For any $A \in \mathcal{E}$,

$$\mathbb{P}(N(A) = k) = \begin{cases} \mathrm{e}^{-\mu(A)}\, \dfrac{(\mu(A))^k}{k!} & \text{if} \quad \mu(A) < \infty , \\ 0 & \text{if} \quad \mu(A) = \infty , \end{cases} \qquad k \geq 0 .$$

(b) For any $k \geq 1$, if $A_1, \ldots, A_k$ are mutually disjoint sets in $\mathcal{E}$ then $N(A_1), \ldots, N(A_k)$ are independent random variables.

The name *mean measure* is justified by the fact that $\mathbb{E}(N(A)) = \mu(A)$. Note also that a PRM($\mu$) is fully determined by the knowledge of its mean value function μ. This is an analogous property to the conventional Poisson distribution, whose distribution is determined by its mean value. Moreover, the Laplace functional of a PRM(μ) is given by

$$\Psi_N(f) = \exp\left\{ -\int_E \left(1 - \mathrm{e}^{-f(x)}\right) d\mu(x) \right\} .$$

The weak convergence of point processes is a basic tool in the study of the asymptotic theory of extremes. We say that the point processes (N_n) *converge weakly* to a point process N in $M_p(E)$ (we write $N_n \xrightarrow{\mathrm{d}} N$) if and only if the corresponding Laplace functionals converge for every $f \in C_K^+(E)$, i.e.,

$$\Psi_{N_n}(f) \to \Psi_N(f) .$$

The following point process limit result (Resnick [102, Theorem 5.3]) is of particular interest in this thesis.

Proposition 1.5. *Let $(X_{ni})_{n=1,2,\ldots;i=1,2,\ldots}$ be a triangular array of row–wise i.i.d. random elements on some state space $E \subset \mathbb{R}^d$ equipped with the Borel σ–field $\mathcal{E}$. Let μ be a Radon measure on $\mathcal{E}$. Then*

$$N_n = \sum_{i=1}^{n} \varepsilon_{X_{ni}} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

holds for some PRM(μ) N *if and only if*

$$n\,\mathbb{P}(X_{n1} \in \cdot\,) \xrightarrow{v} \mu(\cdot\,)\,, \qquad n \to \infty\,.$$

Example 1.6. If $F \in \mathrm{MDA}(H_\gamma)$, Proposition 1.5 and (1.7) imply that

$$N_n = \sum_{i=1}^{n} \varepsilon_{c_n^{-1}(X_i - d_n)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

for a PRM(μ) N with mean measure given by $\mu(x,\infty) = -\log(H(x))$.

The weak convergence of point processes and random measures has been treated in detail in the literature. For a compressive reading we refer to, e.g., Kallenberg [70]. For an account of point process techniques in extreme value theory we refer to Resnick [103].

In this thesis we are interested in the point process convergence of sequences with a complex random structure. In this context, precise large deviations results will prove useful. This is the content of the next section.

1.2 Precise large deviations

We consider an i.i.d. sequence (X_i) of random variables with generic element X and distribution F and define the corresponding partial sum process

$$S_0 = 0\,, \qquad S_n = X_1 + \cdots + X_n\,, \qquad n \geq 1\,.$$

We refer to vanishing probabilities of the type $\mathbb{P}\big(S_n > a_n\big) \to 0$, $n \to \infty$, as *large deviation probabilities*. Large deviation probabilities have been studied in detail in the literature. For instance, S.V. Nagaev [90] proved a result about the large deviations of a finite variance random walk (S_n) in the case of $\overline{F} \in \mathrm{RV}_{-\alpha}$ for some $\alpha > 2$. We formulate this result next to give an idea of *precise large deviations*.

Theorem 1.7 (Nagaev's theorem [86, 90])**.** *Consider an i.i.d. sequence (X_i) of random variables with $\mathbb{E}[X] = 0$, $\mathrm{var}(X) = 1$ and $\mathbb{E}[|X|^{2+\delta}] < \infty$ for some $\delta > 0$. Assume that $\overline{F} \in RV_{-\alpha}$ for some $\alpha > 2$. Then uniformly for $x \geq \sqrt{n}$ as $n \to \infty$,*

$$\mathbb{P}(S_n > x) = \overline{\Phi}(x/\sqrt{n})\,(1 + o(1)) + n\,\overline{F}(x)\,(1 + o(1))\,.$$

In particular, for any positive constant $c_1 < \alpha - 2$

$$\sup_{1 < x/\sqrt{n} < \sqrt{c_1 \log n}} \Big| \frac{\mathbb{P}(S_n > x)}{\overline{\Phi}\big(x/\sqrt{n}\big)} - 1 \Big| \to 0\,, \qquad n \to \infty\,,$$

and for any constant $c_2 > \alpha - 2$,

$$\sup_{x/\sqrt{n} > \sqrt{c_2 \log n}} \Big| \frac{\mathbb{P}(S_n > x)}{n\,\mathbb{P}(X > x)} - 1 \Big| \to 0\,, \qquad n \to \infty\,.$$

Recall that a distribution function F on $(0,\infty)$ is called *subexponential* if $\mathbb{P}(S_n > x) \sim n\overline{F}(x)$ as $x \to \infty$ for any $n \geq 2$. We write $\mathcal{S}$ for the class of subexponential distributions.

This result says that there is an exact threshold where large deviation probabilities for centered partial sums of an i.i.d. sequence are approximated by the tail of a normal distribution to the left and behave in a subexponential way to the right. Similar results exist in the literature for subexponential distributions F; we refer to Section 2.3 below for the relevant literature. However, the regions where the normal and subexponential approximations hold are in general separated from each other by a region where neither of these approximations hold.

Example 1.8. Consider i.i.d. copies $(S_n^{(i)})_{i=1,2,\ldots}$ of S_n satisfying the condition of Theorem 1.7. We also introduce an integer sequence (p_n) such that $p = p_n \to \infty$ as $n \to \infty$. Take (a_n) such that $n\,\mathbb{P}(X > a_n) \to 1$ as $n \to \infty$, and let (a_{np}) be (a_n) evaluated on np. If (p_n) is such that $a_{np} \geq \sqrt{(\alpha - 2 + \delta) n \log n}$ for any small positive δ, then Theorem 1.7 and the definition of (a_{np}) yield for positive x,

$$p\,\mathbb{P}(a_{np}^{-1} S_n > x) \sim np\,\mathbb{P}(a_{np}^{-1} X > x) \sim x^{-\alpha}\,, \qquad n \to \infty\,.$$

Thus, Proposition 1.5 implies that

$$N_p = \sum_{i=1}^{p} \varepsilon_{a_{np}^{-1} S_n^{(i)}} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

for a PRM(μ) N with mean measure $\mu(x,\infty) = x^{-\alpha}$, $x > 0$. Moreover, a continuous mapping argument implies that

$$\max_{i=1,\ldots,p} a_{np}^{-1} S_n^{(i)} \xrightarrow{\mathrm{d}} Y \sim \Phi_\alpha\,, \quad n \to \infty\,.$$

The convergence of point processes whose points are i.i.d. partial sums processes is treated in detail in Chapter 2.

This example shows the usefulness of precise large deviation results combined with point process convergence. In the subsequent section we consider a related example where sum processes figure but these sum processes are in general dependent.

1.3 Sample covariance matrices

Consider p–dimensional random vectors $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{pt})'$, $t = 1, \ldots, n$, whose components $(X_{it})_{i,t \geq 1}$ are i.i.d. random variables with generic element X. We also assume that $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$ if these expectations are finite. We are interested in the (non–normalized) $p \times p$ sample covariance matrix $\mathbf{S}$

$$\mathbf{S} = \mathbf{S}_n = \sum_{t=1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t'$$

with entries

$$S_{ij} = \sum_{t=1}^{n} X_{it} X_{jt}\,, \quad i, j = 1, \ldots, p\,.$$

The sample covariance matrix is of crucial importance in multivariate statistics, for instance in principal component analysis, multivariate regression and hypothesis testing. In many contemporary applications, one is faced with large data sets where both the dimension of the observations and the sample size are large. Thus, asymptotic methods based on limit theorems have proven to be useful tools. This is the case, for instance, in the study of the largest eigenvalues and their corresponding eigenvectors. In limit theory for the eigenvalues $\lambda_1, \ldots, \lambda_p$, one frequently assumes that the dimension $p = p_n$ is some integer sequence tending to infinity as $n \to \infty$. The most common condition in the literature is

$$p_n/n \to \gamma \in (0, \infty)\,, \quad n \to \infty\,. \tag{1.9}$$

Under this assumption, and the moment condition $E(X^2) < \infty$, the *empirical spectral distribution* function

$$F_{n^{-1}\mathbf{S}} = \frac{1}{p}\#\{1 \le j \le p : n^{-1}\lambda_j \le x\}$$

converges weakly with probability 1 to the so–called *Marčenko–Pastur law* F_γ (see Debashis et al. [97]). If $\gamma \in (0, 1]$, F_γ has density

$$f_\gamma(x) = \begin{cases} \frac{1}{2\pi x\gamma}\sqrt{(b-x)(x-a)}, & \text{if } a \le x \le b\,, \\ 0, & \text{otherwise}\,, \end{cases}$$

where $a = (1 - \sqrt{\gamma})^2$ and $b = (1 + \sqrt{\gamma})^2$. If $\gamma > 1$, the Marčenko–Pastur law is a mixture of a point mass at 0 and the density function $f_{1/\gamma}$ with weights $1 - 1/\gamma$ and $1/\gamma$, respectively. The Marčenko–Pastur law describes the global behavior of the eigenvalues, and the moment condition $\mathbb{E}(X^2) < \infty$ is crucial for this result to hold. The moment condition $\mathbb{E}(X^4) < \infty$ plays a similar role when studying the largest eigenvalues of $\mathbf{S}$. We write $\lambda_{(1)} \ge \cdots \ge \lambda_{(p)}$ for the ordered eigenvalues of $\mathbf{S}$. If (1.9) holds, Geman [53] showed that

$$\frac{\lambda_{(1)}}{n} \xrightarrow{\text{a.s.}} (1 + \sqrt{\gamma})^2\,, \quad n \to \infty\,.$$

This means that $\lambda_{(1)}/n$ converges to the right endpoint of the Marčenko–Pastur law. These results are sometimes regarded as "light–tailed" in the sense that they require finite second/fourth moments of X. In the "heavy–tailed" case (i.e., infinite fourth moment), Heiny and Mikosch [60], under the additional condition of regular variation on the tails

$$\mathbb{P}(X > x) \sim p_+ \frac{L(x)}{x^\alpha} \quad \text{and} \quad \mathbb{P}(X < -x) \sim p_- \frac{L(x)}{x^\alpha}\,, \quad x \to \infty\,, \tag{1.10}$$

for some $\alpha \in (0, 4)$ and $p_\pm$ non–negative constants such that $p_- + p_+ = 1$, proved that

$$a_{np}^{-2}\|\mathbf{S} - \operatorname{diag}(\mathbf{S})\|_2 \xrightarrow{\mathbb{P}} 0\,, \qquad n \to \infty\,,$$

where $\|\boldsymbol{A}\|_2$ denotes the spectral norm of a $p \times p$ symmetric matrix $\boldsymbol{A}$, $\operatorname{diag}(\boldsymbol{A})$ consists of the diagonal of $\boldsymbol{A}$, (a_k) is any sequence satisfying $k\,\mathbb{P}(|X| > a_k) \to 1$ as $k \to \infty$, and $p_n = n^\beta \ell(n)$ for some $\beta \in (0, 1]$ and a slowly varying function ℓ. According to Weyl's inequality (see Bhatia [17]), the eigenvalues of $\mathbf{S}$ satisfy the relation

$$a_{np}^{-2} \sup_{i=1,\ldots,p} \left|\lambda_{(i)}(\mathbf{S}) - \lambda_{(i)}(\operatorname{diag}(\mathbf{S}))\right| \le a_{np}^{-2}\|\mathbf{S} - \operatorname{diag}(\mathbf{S})\|_2 \xrightarrow{\mathbb{P}} 0\,, \quad n \to \infty\,.$$

Thus, limit theory for the order statistics of $\mathbf{S}$ is reduced to limit theory for the order statistics of the i.i.d. partial sums $S_{ii} = \sum_{t=1}^{n} X_{it}^2$, $i = 1, \ldots, p$, which are also the ordered eigenvalues $\lambda_{(i)}(\mathrm{diag}(\mathbf{S}))$. This theory is completely described by the Fréchet convergence of the point processes with points $(a_{np}^{-2} S_{ii})_{i=1,\ldots,p}$, for which Nagaev–type large deviation results are essential.

On the other hand, limit theory for the extremes of the off–diagonal elements of $\mathbf{S}$ is less developed. The situation is more complicated because the points S_{ij} are typically dependent. The largest off–diagonal entry of a sample covariance has been studied in the literature and theoretical developments are mainly due to Jiang. For instance, in [66], he analyzed the asymptotic distributions of

$$W_n := \frac{1}{n} \max_{1 \le i < j \le p} |S_{ij}|$$

under the assumption $p/n \to \gamma \in (0, \infty)$. If $\mathbb{E}[|X|^{30+\delta}] < \infty$ for some $\delta > 0$, he proved that

$$\lim_{n\to\infty} \mathbb{P}(nW_n^2 - 4\log p + \log\log p \le x) = \exp\Big(-\tfrac{1}{\sqrt{8\pi}} \, \mathrm{e}^{-x/2} \Big), \quad x \in \mathbb{R},$$

The limiting law is a non–standard Gumbel distribution. Under the same assumptions Jiang [66] also derived the limit

$$\lim_{n\to\infty} \sqrt{\frac{n}{\log p}} W_n = 2 \qquad \text{a.s.} \tag{1.11}$$

However, results concerning the joint limit behavior of the entries were lacking. In Chapter 3, we prove limit theory for the point processes of scaled and centered points (S_{ij}) in a more general framework than used for the results above, and as byproduct we deduce results on their extreme behavior.

The type of limit distributions for the point processes considered in this thesis strongly depends on the centering and normalizing constants employed, whose choice is determined by the behavior of the tail $\overline{F}$ of the underlying random variables above certain thresholds. A similar phenomenon occurs in extreme value statistics, where the estimators obtained from classical tools like POT method and Hill estimator, significantly depend on the threshold selected. Moreover, to completely describe the behavior of given data, one typically needs to fit tail and body separately. On the other hand, the alternative approach of selecting a specific parametric distribution is often too rigid for modeling the tail behavior of general distributions. In this thesis we propose inhomogeneous phase–type distributions as alternative models where one can fit tail and body at the same time, due to their denseness in the class of all distributions with support on the positive real numbers and genuinely heavy tails. Thus, we need to introduce first conventional phase–type distributions.

1.4 Phase–type distributions

The origins of phase–type distributions can be traced back to the pioneering work of Erlang [47] and Jensen [65]. Erlang [47] employed a distribution based on i.i.d. sequential hidden phases with exponential holding times to model the duration of telephone calls. Such a distribution was one of the first attempts to generalize the exponential distribution and went on to be known as the Erlang distribution. The idea of hidden stages was then generalized by Jensen [65] in terms of Markov jump processes. This class of distributions is, in fact, what we now know as phase–type distributions. However, it was not until

the late seventies that Neuts and co–workers established much of the modern theory available today; a large portion of their work is summarized in Neuts [92]. We now present a precise definition of this class of distributions.

Let $(J_t)_{t\geq 0}$ denote a Markov jump process on a state–space $\{1,\ldots,p,p+1\}$, where the states $1,\ldots,p$ are transient and state $p+1$ is absorbing. Then $(J_t)_{t\geq 0}$ has an intensity matrix of the form

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{t} \\ \boldsymbol{0} & 0 \end{pmatrix},$$

where $\boldsymbol{T}$ is a sub–intensity matrix of dimensions $p \times p$ and $\boldsymbol{t}$ is a p–dimensional column vector. Since the rows of $\boldsymbol{\Lambda}$ sum to zero, we have that $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$, where $\boldsymbol{e}$ is the p–dimensional column vector of ones. Let $\pi_k = \mathbb{P}(J_0 = k)$, $k = 1,\ldots,p$, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_p)$ and assume that $\mathbb{P}(J_0 = p+1) = 0$. Then we say that the time until absorption

$$Y = \inf\{t \geq 0 \mid J_t = p+1\}$$

has a phase–type distribution with representation $(\boldsymbol{\pi}, \boldsymbol{T})$ and we write $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. The density f_Y and distribution function F_Y of $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ are given by

$$\begin{aligned} f_Y(y) &= \boldsymbol{\pi} e^{\boldsymbol{T}y}\boldsymbol{t}\,, \quad y > 0\,, \\ F_Y(y) &= 1 - \boldsymbol{\pi} e^{\boldsymbol{T}y}\boldsymbol{e}\,, \quad y > 0\,, \end{aligned}$$

where the exponential of a matrix $\boldsymbol{A}$ is defined by

$$\exp(\boldsymbol{A}) = \sum_{n=0}^{\infty} \frac{\boldsymbol{A}^n}{n!}\,.$$

The class of phase–type distributions has been widely used in the area of Applied Probability, since they often provide exact (or even explicit) solutions to some complex stochastic models. This is the case, for instance, in ruin theory where the use of phase–type distributions allows one to derive explicit expressions for ruin probabilities. This class of distributions is also dense in the set of distributions with support on the positive real numbers (in the sense of weak convergence). This means that any distribution with support on $\mathbb{R}_+$ can be approximated arbitrarily well by a phase–type distribution (of sufficiently large dimension). Moreover, statistical inference for phase–type distributions has been a topic of research for a long time. Maximum likelihood estimation was first proposed by Asmussen et al. [10] using an expectation–maximization (EM) algorithm whereas a Markov chain Monte Carlo (MCMC) based approach was suggested in Bladt et al. [22]. For a comprehensive reading on phase–type distributions we refer to Bladt and Nielsen [24].

One of the downsides of phase–type distributions is that despite of their denseness, they are light–tailed (exponential type) and consequently they fail to capture heavy–tail behavior. In particular, we have for $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$

$$\overline{F}_Y(y) \sim c y^{k-1} \mathrm{e}^{-\lambda y}\,, \quad y \to \infty\,,$$

where $-\lambda$ is the largest real eigenvalue of $\boldsymbol{T}$, k is the dimension of the Jordan block of λ and c is a positive constant; see Asmussen et al. [10, Section 5.7].

Some approaches to remedy this problem have been introduced in the literature. For example, in Bladt et al. [26] the class of PH distributions was extended to allow for

infinite–dimensional matrices (denoted by NPH) which led to a class of distributions with phase–type like properties and a genuinely heavy tail. The estimation of such distributions was considered in Bladt and Rojas-Nandayapa [27]. More recently, Albrecher et al. [3] considered the transformation of the time scales of each state of the underlying Markov process, leading to inhomogeneous phase–type (IPH) distributions. Their approach has the advantage of preserving the finite dimensionality, and for suitable transformations offer classes of genuinely heavy–tailed distributions. Another advantage of this construction is that one can apply standard techniques from univariate phase–type fitting in very specific cases, however a more general fitting procedure was lacking. In Chapter 4 we consider the class of inhomogeneous phase–type distributions as models where heavy tails (and lighter than exponential tails) are present, and develop fitting procedures for this class in the uni– and multivariate cases. To extend our analysis to the multivariate case, we need to introduce first the notion of multivariate phase–type models.

1.4.1 Multivariate phase–type distributions

Various multivariate phase–type distributions have been introduced in the literature. The first multivariate distribution with phase–type marginals (denoted by MPH) was proposed by Assaf et al. [11]. Their approach consisted of detecting the hitting times of different subsets of the state–space of a Markov jump process. Later, this class was extended by Kulkarni [73] to the MPH* class. Their definition is based on linear rewards of the sojourn times in the states of the Markov jump process underlying a phase–type distribution. The latest proposal (denoted by MVPH) is given in Bladt et al. [23], in which the authors considered vectors whose linear combination of entries are phase–type distributed. These classes satisfy MPH $\subset$ MPH* $\subset$ MVPH, and each one of them has their own distinctive advantages and disadvantages. However, for the purposes of this thesis we deal only with the MPH* class due to its probabilistic appeal and flexibility. More specifically, the MPH* class is constructed as follows. Let $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ be a phase–type distributed random variable with underlying time–homogeneous Markov jump process $(J_t)_{t\geq 0}$. Let $\boldsymbol{r}_j = (r_j(1), \ldots, r_j(p))'$ be non–negative p–dimensional column vectors, $j = 1, \ldots, d$, and let $\boldsymbol{R}$ be a $p \times d$ *reward matrix* defined by

$$\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_d)\ .$$

The component $r_j(k)$ can be interpreted as the rate at which a reward is obtained while J_t is in state k. Now define a vector $\boldsymbol{Y} = (Y^{(1)}, \ldots, Y^{(d)})'$ such that

$$Y^{(j)} = \int_0^Y r_j\left(J_t\right) dt\,, \qquad j = 1, \ldots, d\,.$$

Then $Y^{(j)}$ is the total reward obtained according to $\boldsymbol{r}_j$ prior to absorption. We say that the random vector $\boldsymbol{Y}$ has a multivariate phase–type distribution of the MPH* (or Kulkarni) type, and we write $\boldsymbol{Y} \sim \mathrm{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$. We refer to [24] for a comprehensive text on multivariate phase–type distributions.

The MPH* class of distributions is dense within the set of distributions with support in $\mathbb{R}_+^d$. Their marginals and linear combinations are phase–type distributed and explicit expressions for their parameters are available; see [24, Theorem 8.1.6]. These properties (among others) make this class very flexible and appealing from a stochastic modeling perspective. A disadvantage is that there is no explicit expression for joint densities: in Kulkarni [73] the density is characterized by a system of partial differential equations, and

in Breuer [28] a semi–explicit form is deduced in the bivariate case. The following type of MPH* distributions does lead to an explicit joint density: Let $\boldsymbol{Y} = (Y^{(1)}, \ldots, Y^{(d)})' \sim \mathrm{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ with

$$\boldsymbol{\pi} = (\boldsymbol{\alpha}, \mathbf{0}, \ldots, \mathbf{0}) \,, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}_{22} & \boldsymbol{T}_{23} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{T}_{dd} \end{pmatrix} , \tag{1.12}$$

and

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{e} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{e} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{e} \end{pmatrix} , \tag{1.13}$$

where $\boldsymbol{T}_{jj}$ are sub–intensity matrices of dimensions p_j, $j = 1, \ldots, d$, and $\boldsymbol{T}_{j(j+1)}$ are non–negative matrices such that $-\boldsymbol{T}_{jj}\,\boldsymbol{e} = \boldsymbol{T}_{j(j+1)}\,\boldsymbol{e}$, $j = 1, \ldots, d-1$. Then the joint density of $\boldsymbol{Y}$ is given by

$$f_{\boldsymbol{Y}}\left(y^{(1)}, \ldots, y^{(d)}\right) = \boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11} y^{(1)}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22} y^{(2)}} \cdots \boldsymbol{T}_{(d-1)d} \mathrm{e}^{\boldsymbol{T}_{dd} y^{(d)}} (-\boldsymbol{T}_{dd})\,\boldsymbol{e}\,,$$

Another advantage of this subclass of MPH* distributions is that they are also dense in the set of distributions on $\mathbb{R}_+^d$; see [6, Remark 7]. However, they are always tail–independent, making them less suitable for certain applications, for example in quantitative risk management.

Despite their appealing properties for applications, the literature on estimation of multivariate phase–type distributions is sparse. Ahlström et al. [1] introduced an algorithm in the bivariate case for the subclass of MPH* described above. Breuer [28] proposed an EM algorithm for parameter estimation in the general case by exploiting some properties of the MPH* class and the existing results for the univariate case. However, the method of Breuer was not actually implemented and contains an inconsistency in the M–step for the estimation of the reward matrix $\boldsymbol{R}$, which is amended in Chapeter 4 of this thesis.

1.5 Inhomogeneous phase–type distributions

Recently, Albrecher and Bladt [3] considered an extension of the construction principle of phase–type distributions by allowing the Markov jump process to be time–inhomogeneous. Thus, one gains a lot of flexibility in terms of the tail behavior of the absorption times, while carrying over some of the computational tools and advantages of the PH class. The construction goes as follows. Let $(J_t)_{t\geq 0}$ denote a time–inhomogeneous Markov jump process on the state–space $\{1, \ldots, p, p+1\}$, where the states $1, \ldots, p$ are transient and state $p+1$ is absorbing. Then $(J_t)_{t\geq 0}$ has an intensity matrix of the form

$$\boldsymbol{\Lambda}(t) = \begin{pmatrix} \boldsymbol{T}(t) & \boldsymbol{t}(t) \\ \mathbf{0} & 0 \end{pmatrix} , \quad t \geq 0\,,$$

where $\boldsymbol{T}(t)$ is a $p \times p$ matrix and $\boldsymbol{t}(t)$ is a p–dimensional column vector. Here, $\boldsymbol{t}(t) = -\boldsymbol{T}(t)\,\boldsymbol{e}$, $t \geq 0$. Let $\pi_k = \mathbb{P}(J_0 = k)$, $k = 1, \ldots, p$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ and assume that

$\mathbb{P}(J_0 = p+1) = 0$. Then we say that the time until absorption

$$\tau = \inf\{t \geq 0 \mid J_t = p+1\}$$

has an inhomogeneous phase–type distribution with representation $(\boldsymbol{\pi}, \boldsymbol{T}(t))$ and we write $\tau \sim \text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}(t))$. The density f and distribution function F of τ are given by

$$f(x) = \boldsymbol{\pi} \prod_0^x (\boldsymbol{I} + \boldsymbol{T}(u)du)\boldsymbol{t}(x)\,, \quad x > 0\,,$$

$$F(x) = 1 - \boldsymbol{\pi} \prod_0^x (\boldsymbol{I} + \boldsymbol{T}(u)du)\boldsymbol{e}\,, \quad x > 0\,,$$

where

$$\prod_s^t (\boldsymbol{I} + \boldsymbol{T}(u)du) = \boldsymbol{I} + \sum_{k=1}^{\infty} \int_s^t \int_s^{u_k} \cdots \int_s^{u_2} \boldsymbol{T}(u_1) \cdots \boldsymbol{T}(u_k) du_1 \cdots du_k\,;$$

see [106].

We are particularly interested in the case where $\boldsymbol{T}(t) = \lambda(t)\,\boldsymbol{T}$ with a known non–negative real function $\lambda(t)$ and a sub–intensity matrix $\boldsymbol{T}$. Then we write $\tau \sim \text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$ and the density and distribution functions are given by

$$\begin{aligned} f(x) &= \lambda(x)\,\boldsymbol{\pi} \exp\left(\int_0^x \lambda(t)dt\ \boldsymbol{T}\right)\boldsymbol{t}, \\ F(x) &= 1 - \boldsymbol{\pi} \exp\left(\int_0^x \lambda(t)dt\ \boldsymbol{T}\right)\boldsymbol{e}. \end{aligned}$$

Note that if $\lambda(t) = 1$, then τ has a conventional phase–type distribution.

This subclass of inhomogeneous phase–type distributions is of particular interest since its member distributions can be expressed as the distributions of transformed (conventional) phase–type random variables. Specifically, if $X \sim \text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$, then there exists a function g such that

$$X \sim g(Y)\,, \tag{1.14}$$

where $Y \sim \text{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ and g is defined in terms of its inverse function through

$$g^{-1}(x) = \int_0^x \lambda(t)dt\,.$$

For further reading on inhomogeneous phase–type distributions and motivations for their use in modeling, we refer to Albrecher and Bladt [3].

As illustrated in [3], various IPH distributions can be expressed as classical distributions with matrix–valued parameters. This is the case, for instance, for the Pareto, Weibull and GEV distributions. Moreover, these classes of distributions are all dense in the class of distributions with support on $\mathbb{R}_+$ and inherit the computational advantages of the PH class, while providing genuinely heavy tails. For the representation of such distributions we make use of functional calculus. If h is an analytic function and $\boldsymbol{A}$ is a matrix, define

$$h(\boldsymbol{A}) = \frac{1}{2\pi i} \oint_\gamma h(z)(z\boldsymbol{I} - \boldsymbol{A})^{-1} dz,$$

where γ is a simple path enclosing the eigenvalues of $\boldsymbol{A}$; see [24, Section 3.4] for further details.

Example 1.9 (Matrix–Pareto distributions)**.** Let $X = \beta(e^Y - 1)$, where $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ and $\beta > 0$. Then

$$\overline{F}_X(x) = \boldsymbol{\pi}\left(\frac{x}{\beta} + 1\right)^{\boldsymbol{T}} \boldsymbol{e}, \quad f_X(x) = \boldsymbol{\pi}\left(\frac{x}{\beta} + 1\right)^{\boldsymbol{T}-\boldsymbol{I}} \boldsymbol{t}\frac{1}{\beta}.$$

We refer to the distribution of X as a *matrix–Pareto* distribution. Asymptotically, it is easy to see that $\overline{F}_X(x) \sim L(x)x^{-\lambda}$ as $x \to \infty$, where $L(\cdot)$ is a slowly varying function and $-\lambda$ is the largest real eigenvalue of $\boldsymbol{T}$. In other words $F_X \in \mathrm{MDA}(\Phi_\lambda)$. Moreover, if $\beta = 1$, Albrecher et al. [3] noticed that we can fit a matrix–Pareto distribution to i.i.d. data $x_1, ..., x_N$ by simply fitting a phase–type distribution $\mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ to the transformed data $\log(1 + x_1), ..., \log(1 + x_N)$ using an EM–algorithm; see [10].

Example 1.10 (Matrix–Weibull distributions)**.** Let $X = Y^{1/\beta}$, where $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ and $\beta > 0$. Then

$$\overline{F}_X(x) = \boldsymbol{\pi} e^{\boldsymbol{T}x^\beta} \boldsymbol{e}\,, \quad f_X(x) = \boldsymbol{\pi} e^{\boldsymbol{T}x^\beta} \boldsymbol{t}\beta x^{\beta-1}\,.$$

It is easy to see that $\overline{F}_X(x) \sim x^\gamma e^{-\lambda x^\beta}$ as $x \to \infty$. This distribution is called *matrix–Weibull* since the scale parameter of the usual Weibull distribution is now replaced by a matrix. Moreover, $F_X \in \mathrm{MDA}(\Lambda)$.

The parameter estimation in this last example and more generally in the case when the transformation depends on parameters is more subtle, and it will be treated in detail in Chapter 4 along with a multivariate extension.

1.6 Contribution of this thesis

In this section we summarize the results from the upcoming chapters.

1.6.1 Gumbel and Fréchet convergence for the point process of i.i.d. random walks

In Chapter 2 we will consider i.i.d. copies $(S_n^{(i)})_{i=1,2,...}$ of a random walk S_n with i.i.d. step sizes X_i with distribution F, and an integer sequence (p_n) such that $p = p_n \to \infty$ as $n \to \infty$. We also assume that $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) = 1$ if these expectations are finite. We show Gumbel and Fréchet convergence of the centered and normalized maxima $\max_{i=1,...,p} S_n^{(i)}$. More generally, we prove the weak convergence of point processes of the form

$$N_p = \sum_{i=1}^{p} \varepsilon_{\tilde{c}_n^{-1}(S_n^{(i)} - \tilde{d}_n)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

for suitable constants $\tilde{c}_n > 0$ and $\tilde{d}_n \in \mathbb{R}$ toward a $\mathrm{PRM}(\mu)$. As discussed in Section 1.1.3 this means proving relationships of the type

$$p\,\mathbb{P}(S_n > \tilde{c}_n x + \tilde{d}_n) \to \mu(x, \infty)\,,$$

for which precise large deviations results prove to be handy. We assume that F belongs to $\mathrm{MDA}(\Lambda)$ or $\mathrm{MDA}(\Phi_\alpha)$, and we borrow the known normalizing and centering constants from these MDAs.

Gumbel convergence via normal approximations to large deviation probabilities for small x.

We assume that a large deviation approximation to the standard normal distribution Φ holds, that is

$$\sup_{0\le x<\gamma_n} \Big| \frac{\mathbb{P}(S_n/\sqrt{n} > x)}{\overline{\Phi}(x)} - 1 \Big| \to 0\,, \qquad n\to\infty\,. \tag{1.15}$$

for some $\gamma_n \to \infty$. We recall that $\Phi \in \mathrm{MDA}(\Lambda)$. Then

$$p\,\mathbb{P}\Big(\frac{S_n}{\sqrt{n}} > d_p + x/d_p\Big) \to \mathrm{e}^{-x} = -\log\Lambda(x)\,, \qquad n\to\infty\,, \qquad x\in\mathbb{R}\,, \tag{1.16}$$

holds for any integer sequence $p_n \to \infty$ such that $p_n < \exp(\gamma_n^2/2)$ and (d_p) is defined in (1.6). This limit relationship is equivalent to the following point process convergence on the state space $\mathbb{R}$

$$N_p = \sum_{i=1}^{p} \varepsilon_{d_p\left(\frac{S_n^{(i)}}{\sqrt{n}} - d_p\right)} \stackrel{\mathrm{d}}{\to} N\,,$$

where N is $\mathrm{PRM}(-\log\Lambda)$. Moreover, (1.16) is also equivalent to Gumbel convergence of the maximum random walk

$$d_p \max_{i=1,\dots,p} \big(S_n^{(i)}/\sqrt{n} - d_p\big) \stackrel{\mathrm{d}}{\to} Y \sim \Lambda\,, \qquad n\to\infty\,.$$

Gumbel convergence via the subexponential approximation to large deviation probabilities for very large x

We assume that $F \in \mathrm{MDA}(\Lambda) \cap \mathcal{S}$, the subexponential approximation

$$\sup_{x>\gamma_n} \Big| \frac{\mathbb{P}(S_n - \mathbb{E}[S_n] > x)}{n\,\mathbb{P}(X>x)} - 1 \Big| \to 0\,,$$

holds for some $\gamma_n \to \infty$, and for sufficiently large n and an integer sequence $p_n \to \infty$,

$$d_{np} + x\,c_{np} > \gamma_n\,, \qquad \text{for any } x<0\,, \tag{1.17}$$

where (d_{np}) and (c_{np}) are the centering and normalizing constants (d_n) and (c_n), respectively, for $F\in\mathrm{MDA}(\Lambda)$ (see (1.5)), evaluated at np. Then

$$p\,\mathbb{P}\big(S_n - \mathbb{E}[S_n] > d_{np} + x\,c_{np}\big) \to \mathrm{e}^{-x}\,, \qquad x\in\mathbb{R}\,, \qquad n\to\infty\,, \tag{1.18}$$

holds. This limit relationship is equivalent to the following point process convergence on the state space $\mathbb{R}$

$$N_p = \sum_{i=1}^{p} \varepsilon_{c_{np}^{-1}(S_n^{(i)} - \mathbb{E}[S_n] - d_{np})} \stackrel{\mathrm{d}}{\to} N\,, \qquad n\to\infty\,,$$

where N is $\mathrm{PRM}(-\log\Lambda)$. Moreover, (1.18) is also equivalent to Gumbel convergence of the maximum random walk

$$\max_{i=1,\dots,p} c_{np}^{-1}\big(S_n^{(i)} - \mathbb{E}[S_n] - d_{np}\big) \stackrel{\mathrm{d}}{\to} Y\sim\Lambda\,, \qquad n\to\infty\,.$$

Note that (1.17) implicitly determines the relationship of p_n with n, and a sufficient condition for such inequality to hold is $d_{np} \ge (1+\delta)\gamma_n$ for any small $\delta>0$ and large n.

Fréchet convergence via the subexponential approximations to large deviation probabilities for large x

We assume that X is regularly varying with index $\alpha > 0$ in the sense of (1.10). We choose a normalizing sequence (a_n) such that $n\,\mathbb{P}(|X| > a_n) \to 1$ as $n \to \infty$, and a centering sequence (d_n) such that

$$d_n = \begin{cases} 0\,, & \alpha \in (0,1) \cup (1,\infty)\,, \\ n\,\mathbb{E}[X\,\mathbf{1}(|X| \le a_n)]\,, & \alpha = 1\,. \end{cases}$$

We also assume that $p_n \to \infty$ is an integer sequence which satisfies the additional conditions

$$\begin{cases} a_{np} \ge \sqrt{(\alpha - 2 + \delta) n \log n} \text{ for some small } \delta > 0 & \text{if } \alpha > 2\,, \\ \lim_{n\to\infty} \sup_{x > a_{np}} p^{\delta} \dfrac{n}{x^2}\, \mathbb{E}[X^2\,\mathbf{1}(|X| \le x)] = 0 \text{ for some small } \delta > 0 & \text{if } \alpha = 2\,. \end{cases} \tag{1.19}$$

Then, by a Nagaev–type large deviation result, the following limit relation

$$p\,\mathbb{P}(\pm a_{np}^{-1}(S_n - d_n) > x) \to p_{\pm} x^{-\alpha}\,, \qquad x > 0\,, \qquad n \to \infty\,,$$

holds. This limit relationship is equivalent to

$$N_p = \sum_{i=1}^{p} \varepsilon_{a_{np}^{-1}(S_n^{(i)} - d_n)} \xrightarrow{\mathrm{d}} N\,,$$

where N is PRM(μ_α) with intensity

$$\mu_\alpha(dx) = |x|^{-\alpha-1}\big(p_+ \mathbf{1}(x > 0) + p_- \mathbf{1}(x < 0)\big)\,dx\,.$$

Moreover, a continuous mapping argument implies

$$\lim_{n\to\infty} \mathbb{P}\Big(0 < a_{np}^{-1} \max_{i=1,\dots,p} S_n^{(i)} \le x\,, -y < a_{np}^{-1} \min_{i=1,\dots,p} S_n^{(i)}\Big) = \Phi_\alpha^{p_+}(x)\Phi_\alpha^{p_-}(y)\,, \quad x, y > 0\,,$$

where $\Phi_\alpha^c(x) = \exp(-c\,x^{-\alpha})$, $c \ge 0$. Note again that (1.19) implicitly determines the relationship of p_n with n. If $\alpha > 2$, then (1.19) holds if $p \ge n^{(\alpha/2)-1+\gamma'}$ for any choice of $\gamma' > 0$. If $\alpha = 2$ and $\mathrm{var}(X) < \infty$, then (1.19) is satisfied for any sequence $p_n \to \infty$ and $\delta < 1$. If $\alpha = 2$ and $\mathrm{var}(X) = \infty$, then (1.19) holds if $p/n^{\gamma} \to \infty$ for any small $\gamma > 0$.

These results clarify which rates of growth are possible for $p_n \to \infty$ and their importance for the form of the limit law of the maxima of $(S_n^{(i)})$. More specifically, we can summarize our findings as follows: in the case of the MDA of the Gumbel distribution the maxima may converge to the Gumbel distribution due to two distinct mechanisms: the normal approximation at medium–high thresholds and the subexponential approximation at high–level thresholds. In the case of the MDA of the Fréchet distribution two distinct limit distributions are possible: the Gumbel distribution due to normal approximation at medium–high thresholds and the Fréchet distribution due to subexponential approximation at high–level thresholds provided the distribution has finite second moment. If this last moment condition is not satisfied only the Fréchet convergence is possible.

In Chapter 2 we derive various extensions of these results. Specifically, we consider the case when p_n is replaced by $k_n = [n/r_n] \to \infty$ for some integer sequence $r_n \to \infty$ and n is replaced by r_n. This means that we show weak convergence for point processes with points $S_{r_n i} - S_{r_n(i-1)}$, the sum of the i-th block $X_{r_n(i-1)+1}, \dots, X_{r_n i}$, $i = 1, \dots, k_n$.

Moreover, we clarify which rates of growth are possible for the block sizes $r_n \to \infty$. We also consider the case when the index of the point process is random, and we show that analogs of all point process convergences above hold with corresponding growth rates for p_n; see Section 2.4.4. For Fréchet convergence, we also discuss extensions to stationary regularly varying sequences (see Section 2.4.3.3), i.i.d. multivariate regularly varying sequences (see Section 2.4.3.4), and i.i.d. random sums (see Section 2.4.3.5).

1.6.2 Gumbel convergence for the point process of the off-diagonal elements (S_{ij})

In Chapter 3 we will prove limit theory for the point processes of the scaled and centered points (S_{ij}) introduced in Section 1.3. More specifically, we derive asymptotic theory for the point processes

$$\widetilde{N}_n = \sum_{1\le i<j\le p} \varepsilon_{\widetilde{d}_p(S_{ij}/\sqrt{n}-\widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,,$$

for suitable constants $(\widetilde{d}_p)$ and a PRM N with mean measure μ on $\mathbb{R}$ such that $\mu(x,\infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$. As a byproduct, we study the joint asymptotic distribution of functionals of a fixed number of the order statistics of the points S_{ij}. In particular, we show that the maximum entry converges to the Gumbel distribution provided X has suitably many finite moments. The main challenge is that the S_{ij} are not independent. It turns out that, despite their non-trivial dependence, the maximum behaves like the maximum of i.i.d. copies. Therefore we will first present further results on i.i.d. random walk points $(S_n^{(i)})$. We assume one of the following three moment conditions:

(C1) There exists $s > 2$ such that $\mathbb{E}[|X|^s] < \infty$.

(C2) There exists an increasing differentiable function g on $(0,\infty)$ such that $\mathbb{E}[\exp(g(|X|))] < \infty$, $g'(x) \le \tau g(x)/x$ for sufficiently large x and some $\tau < 1$, and $\lim_{x\to\infty} g(x)/\log x = \infty$.

(C3) There exists a constant $h > 0$ such that $\mathbb{E}[\exp(h\,|X|)] < \infty$.

Note that conditions (C1)–(C3) are increasing in strength, i.e., (C3) $\Rightarrow$ (C2) $\Rightarrow$ (C1). Under these conditions large deviation approximations to the standard normal distribution of the type (1.15) hold, where $\gamma_n \to \infty$ are sequences depending on the conditions (C1)–(C3). This implicitly determines the rate of growth of $p = p_n \to \infty$. Specifically, assume that $p = p_n \to \infty$ satisfies

- $p = O(n^{(s-2)/2})$ if (C1) holds.
- $p = \exp(o(g_n^2 \wedge n^{1/3}))$ where g_n is the solution of the equation $g_n^2 = g(g_n\sqrt{n})$, if (C2) holds.
- $p = \exp(o(n^{1/3}))$ if (C3) holds.

Then, by (1.16), we have

$$p\,\mathbb{P}(S_n/\sqrt{n} > d_p + x/d_p) \to \mathrm{e}^{-x}, \quad n\to\infty, \quad x\in\mathbb{R}\,, \tag{1.20}$$

where d_p if given by (1.6). Moreover, relation (1.20) is equivalent to either of the following two limit relations:

$$\mathbb{P}\Big(d_p \max_{i=1,\dots,p}(S_n^{(i)}/\sqrt{n} - d_p) \le x\Big) \to \Lambda(x)\,, \qquad x\in\mathbb{R}\,, \qquad n\to\infty\,.$$

and

$$N_n = \sum_{i=1}^{p} \varepsilon_{d_p\,(S_n^{(i)}/\sqrt{n} - d_p)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,, \tag{1.21}$$

where N is a PRM with mean measure $\mu(x,\infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$.

We now consider the sample covariance matrix $\mathbf{S} = (S_{ij})_{i,j=1,\ldots,p}$. Since the summands of S_{ij} are i.i.d. products $X_{it}X_{jt}$ we need to adjust the conditions (C2) and (C3) to this situation while (C1) remains unchanged.

(C2') There exists an increasing differentiable function g on $(0,\infty)$ such that $\mathbb{E}[\exp(g(|X_{11}X_{12}|))] < \infty$, $g'(x) \le \tau g(x)/x$ for sufficiently large x and some $\tau < 1$, and $\lim_{x\to\infty} g(x)/\log x = \infty$.

(C3') There exists a constant $h > 0$ such that $\mathbb{E}[\exp(h\,|X_{11}X_{12}|)] < \infty$.

Now assume that $p = p_n \to \infty$ satisfies:

- $p = O(n^{(s-2)/4})$ if (C1) holds.
- $p = \exp(o(g_n^2 \wedge n^{1/3}))$, where g_n is the solution of the equation $g_n^2 = g(g_n\sqrt{n})$, if (C2') holds.
- $p = \exp(o(n^{1/3}))$ if (C3') holds.

Define $\widetilde{d}_p = d_{p(p-1)/2}$. Then the following point process convergence holds:

$$N_n^S := \sum_{1\le i<j\le p} \varepsilon_{\widetilde{d}_p(S_{ij}/\sqrt{n} - \widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,,$$

where N is the PRM defined in (1.21) and has representation

$$N = \sum_{i=1}^{\infty} \varepsilon_{-\log\Gamma_i}\,,$$

$\Gamma_i = E_1 + \cdots + E_i$, $i \ge 1$, for i.i.d. standard exponential sequence (E_i). Next we consider the order statistics of S_{ij}, $1 \le i < j \le p$:

$$\min_{1\le i<j\le p} S_{ij} =: S_{(p(p-1)/2)} \le \cdots \le S_{(1)} := \max_{1\le i<j\le p} S_{ij}\,.$$

Then a continuous mapping argument implies joint convergence of the upper and lower order statistics: for any $k \ge 1$,

$$\begin{aligned} \widetilde{d}_p\big(S_{(i)}/\sqrt{n} - \widetilde{d}_p\big)_{i=1,\ldots,k} &\xrightarrow{\mathrm{d}} (-\log\Gamma_i)_{i=1,\ldots,k}\,,\\ \widetilde{d}_p\big(S_{(i)}/\sqrt{n} + \widetilde{d}_p\big)_{i=p(p-1)/2,\ldots,p(p-1)/2-k+1} &\xrightarrow{\mathrm{d}} (\log\Gamma_i)_{i=1,\ldots,k}\,. \end{aligned}$$

Moreover, the properly normalized maxima and minima are asymptotically independent, that is for any $x, y \in \mathbb{R}$ we have as $n \to \infty$,

$$\mathbb{P}\Big(\widetilde{d}_p(S_{(1)}/\sqrt{n} - \widetilde{d}_p) \le x\,, \widetilde{d}_p(S_{(p(p-1)/2)}/\sqrt{n} + \widetilde{d}_p) \le y\Big) \to \Lambda(x)(1 - \Lambda(-y))\,.$$

An immediate consequence is

$$\frac{S_{(1)}}{\sqrt{n \log p}} \xrightarrow{\mathbb{P}} 2 \quad \text{and} \quad \frac{S_{(p(p-1)/2)}}{\sqrt{n \log p}} \xrightarrow{\mathbb{P}} -2\,,$$

which is in agreement with (1.11).

We also also derive point process convergence for the sample correlation matrix $\boldsymbol{R} = (R_{ij})_{i,j=1,\ldots,p}$ with $R_{ij} = S_{ij}/\sqrt{S_{ii}S_{jj}}$. Under the same assumptions on $p = p_n \to \infty$ as for $\mathbf{S}$, the following point process convergence holds,

$$N_n^R := \sum_{1 \le i < j \le p} \varepsilon_{\widetilde{d}_p(\sqrt{n} R_{ij} - \widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,,$$

where N is the PRM defined in (1.21). Moreover, the results for the order statistics of (R_{ij}) carry over from those for the order statistics of (S_{ij}). In Section 3.4, we extend our results to hypercubic random matrices of the form $\sum_{t=1}^{n} \mathbf{X}_t \otimes \cdots \otimes \mathbf{X}_t$ and we briefly discuss some statistical applications such as threshold-based estimation and independence tests.

1.6.3 Parameter estimation for inhomogeneous phase–type distributions

The univariate case

In Chapter 4 we will provide an EM algorithm for parameter estimation of inhomogeneous phase–type distributions that are obtained as transformations of (conventional) phase–type distribution. The main assumption is that the transformations are parametric functions. More specifically, we consider an i.i.d. sample $x_1, \ldots, x_N$ from an inhomogeneous phase–type distribution with representation $X \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda(\,\cdot\,;\boldsymbol{\beta}))$, where $\lambda(\,\cdot\,;\boldsymbol{\beta})$ is a parametric non–negative function depending on the vector $\boldsymbol{\beta}$. We then know that $X \overset{d}{=} g(Y;\boldsymbol{\beta})$ with $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$; see (1.14). In particular $g^{-1}(X;\boldsymbol{\beta}) \overset{d}{=} Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. The EM algorithm for fitting $\mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda(\,\cdot\,;\boldsymbol{\beta}))$ then works as follows:

0. Initialize with some "arbitrary" $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta})$.

1. Transform the data into $y_i = g^{-1}(x_i;\boldsymbol{\beta})$, $i = 1, ..., N$ and apply the E– and M–steps of the conventional EM algorithm of Asmussen [10] by which we obtain the estimators $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}})$.

2. Compute

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log(f_X(x_i;\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \boldsymbol{\beta}))\,.$$

3. Assign $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\beta}})$ and GOTO 1.

We show that the likelihood function increases for each iteration, and hence converges to a (possibly local) maximum.

The multivariate case

There are various possibilities for how to extend inhomogeneous phase–type distributions to more than one dimension. In Chapter 4 we suggest one particular approach, which consists of taking transformations of MPH* distributions. Specifically, we let $\boldsymbol{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ and define $\boldsymbol{X} := (g_1(Y^{(1)}), \ldots, g_d(Y^{(d)}))'$, where $g_j : \mathbb{R}_+ \to \mathbb{R}_+$ are increasing and differentiable functions for $j = 1, \ldots, d$, then we say that $\boldsymbol{X}$ has an *inhomogeneous MPH** distribution. Several of its properties follow directly from the definition:

a) The marginals $X^{(j)} = g_j(Y^{(j)})$ are IPH distributed, since each $Y^{(j)}$ is phase–type distributed, $j = 1, \ldots, d$.

b) Since g_j is increasing for all $j = 1, \ldots, d$, the copula of $\boldsymbol{X}$ is the same as the copula of $\boldsymbol{Y}$ (see e.g. [82, Prop.7.7]).

c) For fixed $g_j(\cdot)$, $j = 1, \ldots, d$, this new class is dense in $\mathbb{R}_+^d$ (by the denseness of the MPH* class).

Example 1.11. Let $\boldsymbol{X} = (g_1(Y^{(1)}), \ldots, g_d(Y^{(d)}))'$, where $\boldsymbol{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ and

$$g_j(y) = \beta_j(e^y - 1), \quad \beta_j > 0,\ j = 1, \ldots, d. \tag{1.22}$$

Then we say that $\boldsymbol{X}$ follows a *multivariate matrix–Pareto distribution*. Now, consider a d–dimensional multivariate sample of N i.i.d. observations

$$\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(d)})', \quad i = 1, \ldots, N\,.$$

As in the univariate case, if we assume the simpler transformation $g_j(y) = e^y - 1$, $j = 1, \ldots, d$, then we can use fitting methods of the MPH* class by simply taking the logarithm of the marginal observations. That is, we can apply the algorithms of Breuer [28] and Ahlström et al. [1] to the transformed data $y_i^{(j)} := \log(x_i^{(j)} + 1)$, $i = 1, \ldots, N$, $j = 1, \ldots, d$, to estimate the parameters $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$.

If the transformation is parameter–dependent, the situation is more subtle, and we propose an estimation method that makes use of the joint density. Thus, we consider the particular case when $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)}) \sim \text{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ has parameters of the form (1.12) and (1.13). Then, it is easy to see that the density of $\boldsymbol{X} = (g_1(Y^{(1)}), g_2(Y^{(2)}))'$ is given by

$$f_{\boldsymbol{X}}\left(x^{(1)}, x^{(2)}\right) = \boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11} g_1^{-1}(x^{(1)})} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22} g_2^{-1}(x^{(2)})} (-\boldsymbol{T}_{22}) \boldsymbol{e}\, \frac{1}{g_1'(g_1^{-1}(x^{(1)})) g_2'(g_2^{-1}(x^{(2)}))}\,.$$

We further assume that $g_j(\,\cdot\,; \boldsymbol{\beta}_j)$ is a parametric non–negative function depending on the vector $\boldsymbol{\beta}_j$, $j = 1, 2$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Then, the EM algorithm works as follows:

0. Initialize with some "arbitrary" $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\beta})$.

1. Transform the data into $y_i^{(j)} := g_j^{-1}(x_i^{(j)}; \boldsymbol{\beta}_j)$, $i = 1, \ldots, N$, $j = 1, 2$, and apply the E– and M–steps of Ahlström et al. [1] by which we obtain the estimators $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}})$.

2. Compute

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log(f_{\boldsymbol{X}}(x_i^{(1)}, x_i^{(2)}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}}, \boldsymbol{\beta}))\,.$$

3. Assign $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\beta}})$ and GOTO 1.

Chapter 2

Gumbel and Fréchet convergence of the maxima of independent random walks

Thomas Mikosch & Jorge Yslas
Advances in Applied Probability 52(1) (2020), 213–236

Abstract

We consider point process convergence for sequences of i.i.d. random walks. The objective is to derive asymptotic theory for the largest extremes of these random walks. We show convergence of the maximum random walk to the Gumbel or the Fréchet distributions. The proofs heavily depend on precise large deviation results for sums of independent random variables with a finite moment generating function or with a subexponential distribution.

Keywords: Large deviation; subexponential distribution; regular variation; extreme value theory; Gumbel distribution; Fréchet distribution; maximum random walk.

2.1 Introduction

Let (X_i) be an i.i.d. sequence of random variables with generic element X, distribution F and right tail $\overline{F} = 1 - F$. Define the corresponding partial sum process

$$S_0 = 0\,, \qquad S_n = X_1 + \cdots + X_n\,, \qquad n \geq 1\,.$$

Consider i.i.d. copies $(S_{ni})_{i=1,2,\ldots}$ of S_n. We also introduce an integer sequence (p_n) such that $p = p_n \to \infty$ as $n \to \infty$. We are interested in the limiting behavior of the k largest values among $(S_{ni})_{i=1,\ldots,p}$, in particular in the possible limit laws of the maximum $\max_{i=1,\ldots,p} S_{ni}$. More generally, writing ε_x for Dirac measure at x, we are interested in the limiting behavior of the point processes

$$N_p = \sum_{i=1}^{p} \varepsilon_{c_n^{-1}(S_{ni} - d_n)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,, \tag{2.1}$$

for suitable constants $c_n > 0$ and $d_n \in \mathbb{R}$ toward a Poisson random measure N with Radon mean measure μ (we write PRM(μ)).

Our main motivation for this work comes from random matrix theory, in particular when dealing with sample covariance matrices. Their entries are dependent random walks. However, in various situations the theory can be modified in such a way that it suffices to study independent random walks. We refer to Section 2.4.6 for a discussion.

Relation (2.1) is equivalent to the following limit relations for the tails

$$p_n\, \mathbb{P}\big(c_n^{-1}(S_n - d_n) \in (a, b]\big) \to \mu(a, b]\,,$$

for any $a < b$ provided that $\mu(a, b] < \infty$; see Resnick [102, Theorem 5.3]. These conditions involve precise large deviation probabilities for the random walk (S_n); in Section 2.3 we provide some results which are relevant in this context.

We distinguish between two types of precise large deviation results:

- normal approximation
- subexponential approximation

The normal approximation can be understood as an extension of the central limit theorem for $(S_n/\sqrt{n})$ toward increasing intervals. This approximation causes the maxima of $(S_{ni}/\sqrt{n})$ to behave like the maxima of an i.i.d. normal sequence, i.e., these maxima converge in distribution to the Gumbel distribution. This is in contrast to the subexponential approximation which requires that F is a so–called *subexponential* distribution; see Section 2.2.1. In particular, $\overline{F}$ is heavy–tailed in the sense that the moment generating function does not exist. This fact implies that $\mathbb{P}(S_n > x_n) \sim n\,\overline{F}(x_n)$ for sufficiently fast increasing sequences $x_n \to \infty$. Hence $n\,\overline{F}(x_n)$ dominates $\mathbb{P}(S_n > x_n)$ at sufficiently high levels x_n and, as in limit theory for the maxima of an i.i.d. sequence, $\overline{F}$ determines the type of the limit distribution of the maxima of (S_{ni}) as well as the normalizing and centering constants. In this case we also assume that F belongs to the maximum domain of attraction (MDA) of the Gumbel or Fréchet distributions, and we borrow the known normalizing and centering constants from these MDAs. Thus, in the case of the MDA of the Gumbel distribution the maxima of (S_{ni}) may converge to the Gumbel distribution due to two distinct mechanisms: the normal approximation at medium–high thresholds and the subexponential approximation at high–level thresholds. In the case of the MDA

of the Fréchet distribution, two distinct approximations are possible: Gumbel approximation at medium–high thresholds and Fréchet approximation at high–level thresholds provided the distribution has finite second moment. If this condition is not satisfied only the Fréchet approximation is possible.

The paper is organized as follows. In Section 2.2 we introduce the necessary notions for this paper: subexponential and regularly varying distributions (Section 2.2.1), maximum domain of attraction and relevant distributions in it (Section 2.2.2), point process convergence of triangular arrays toward Poisson random measures (Section 2.2.3), and precise large deviations (Section 2.2.4). Due to the importance of the latter topic we devote Section 2.3 to it and collect some of the known precise large deviation results in the case when the moment generating function is finite in some neighborhood of the origin and for subexponential distributions. The main results of this paper are formulated in Section 2.4. Based on the large deviation results of Section 2.3 we give sufficient conditions for the point process convergence relation (2.1) to hold and we clarify which rates of growth are possible for $p_n \to \infty$. In particular, we consider the case when p_n in (2.1) is replaced by $k_n = [n/r_n]$ for some integer sequence $r_n \to \infty$ and n is replaced by r_n. This means that we are interested in (2.1) when $S_{ni} = S_{r_n i} - S_{r_n(i-1)}$, $i = 1, \ldots, k_n$, are i.i.d. block sums. We also discuss extensions of these results to stationary regularly varying sequences (Section 2.4.3.3) and i.i.d. multivariate regularly varying sequences (Section 2.4.3.4).

2.2 Preliminaries I

2.2.1 Subexponential and regularly varying distributions

We are interested in the class $\mathcal{S}$ of *subexponential distributions* F, i.e., it is a distribution supported on $[0, \infty)$ such that, for any $n \geq 2$,

$$\mathbb{P}(S_n > x) \sim n\,\overline{F}(x)\,, \qquad x \to \infty\,.$$

For an encyclopedic treatment of subexponential distributions, see Foss et al. [50]. In insurance mathematics, $\mathcal{S}$ is considered a natural class of heavy–tailed distributions. In particular, F does not have a finite moment generating function; see Embrechts et al. [46, Lemma 1.3.5].

The regularly varying distributions are another class of heavy–tailed distributions supported on $\mathbb{R}$. We say that X and its distribution F are *regularly varying with index* $\alpha > 0$ if there are a slowly varying function L and constants $p_\pm$ such that $p_+ + p_- = 1$ and

$$F(-x) \sim p_-\, x^{-\alpha}\, L(x) \qquad \text{and} \qquad \overline{F}(x) \sim p_+\, x^{-\alpha}\, L(x)\,, \qquad x \to \infty\,. \tag{2.2}$$

A non–negative regularly varying X is subexponential; see [46, Corollary 1.3.2].

2.2.2 Maximum domains of attraction

We call a non–degenerate distribution H an *extreme value distribution* if there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$, $n \geq 1$, such that the maxima $M_n = \max(X_1, \ldots, X_n)$ satisfy the limit relation

$$c_n^{-1}(M_n - d_n) \xrightarrow{\mathrm{d}} Y \sim H\,, \qquad n \to \infty\,. \tag{2.3}$$

In the context of this paper we will deal with two standard extreme value distributions: the Fréchet distribution $\Phi_\alpha(x) = \exp(-x^{-\alpha})$, $x > 0$, and the Gumbel distribution $\Lambda(x) = \exp(-\exp(-x))$, $x \in \mathbb{R}$. As a matter of fact, the third type of extreme value distribution – the Weibull distribution – cannot appear since (2.3) is only possible for X with finite right endpoint but a random walk is not bounded from above by a constant. We say that the distribution F of X is in the *maximum domain of attraction* of the extreme value distribution H ($F \in \mathrm{MDA}(H)$).

Example 2.1. A distribution $F \in \mathrm{MDA}(\Phi_\alpha)$ for some $\alpha > 0$ if and only if

$$\overline{F}(x) = \frac{L(x)}{x^\alpha}\,, \qquad x > 0$$

(see [46, Section 3.3.1]). Then

$$c_n^{-1} M_n \stackrel{\mathrm{d}}{\to} Y \sim \Phi_\alpha\,, \qquad n \to \infty\,,$$

where (c_n) can be chosen such that $n\,\mathbb{P}(X > c_n) \to 1$.

Example 2.2. A distribution F with infinite right endpoint obeys $F \in \mathrm{MDA}(\Lambda)$ if and only if there exists a positive function $a(x)$ with derivative $a'(x) \to 0$ as $x \to \infty$ such that

$$\lim_{u\to\infty} \frac{\overline{F}(u + a(u)\,x)}{\overline{F}(u)} = \mathrm{e}^{-x}\,, \qquad x \in \mathbb{R}$$

(see [46, Section 3.3.3]). Then

$$c_n^{-1}(M_n - d_n) \stackrel{\mathrm{d}}{\to} Y \sim \Lambda\,, \qquad n \to \infty\,,$$

where (d_n) can be chosen such that $n\,\mathbb{P}(X > d_n) \to 1$ and $c_n = a(d_n)$.

The standard normal distribution $\Phi \in \mathrm{MDA}(\Lambda)$ and satisfies

$$c_n^{-1}(M_n - d_n) \stackrel{\mathrm{d}}{\to} Y \sim \Lambda\,, \qquad n \to \infty\,, \tag{2.4}$$

where $c_n = 1/d_n$ and

$$d_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2(2\log n)^{1/2}}\,. \tag{2.5}$$

Since $d_n \sim \sqrt{2\log n}$ we can replace c_n in (2.4) by $1/\sqrt{2\log n}$ while d_n cannot be replaced by $\sqrt{2\log n}$.

The standard lognormal distribution (i.e., $X = \exp(Y)$ for a standard normal random variable Y) is also in $\mathrm{MDA}(\Lambda)$. In particular, one can choose

$$c_n = d_n/\sqrt{2\log n} \quad \text{and} \quad d_n = \exp\Big(\sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2(2\log n)^{1/2}}\Big) \tag{2.6}$$

(see [46, page 156]).

The standard Weibull distribution has tail $\overline{F}(x) = \exp(-x^{-\tau})$, $x > 0$, $\tau > 0$. We consider a distribution F on $(0,\infty)$ with a Weibull–type tail $\overline{F}(x) \sim c\,x^\beta \exp(-\lambda x^\tau)$ for constants $c, \beta, \lambda, \tau > 0$. Then $F \in \mathrm{MDA}(\Lambda)$ and one can choose

$$c_n = (\lambda\tau)^{-1} s_n^{1/\tau - 1} \quad \text{and} \quad d_n = s_n^{1/\tau} + \frac{1}{\tau} s_n^{1/\tau - 1}\Big(\frac{\beta}{\lambda\tau}\log s_n + \frac{\log c}{\lambda}\Big)\,, \tag{2.7}$$

where $s_n = \lambda^{-1}\log n$ (see [46, page 155]).

2.2.3 Point process convergence of independent triangular arrays

For further use we will need the following point process limit result (Resnick [102, Theorem 5.3]).

Proposition 2.3. *Let $(X_{ni})_{n=1,2,\ldots;i=1,2,\ldots}$ be a triangular array of row–wise i.i.d. random elements on some state space $E \subset \mathbb{R}^d$ equipped with the Borel σ–field $\mathcal{E}$. Let μ be a Radon measure on $\mathcal{E}$. Then*

$$\widetilde{N}_p = \sum_{i=1}^{p} \varepsilon_{X_{ni}} \stackrel{\mathrm{d}}{\to} N\,, \qquad n \to \infty\,,$$

holds for some PRM(μ) N *if and only if*

$$p\,\mathbb{P}(X_{n1} \in \cdot) \stackrel{v}{\to} \mu(\cdot)\,, \qquad n \to \infty\,, \tag{2.8}$$

where $\stackrel{v}{\to}$ denotes vague convergence on E.

2.2.4 Large deviations

Our main goal is to prove the point process convergence (2.1) for i.i.d. sequences (S_{ni}) of partial sum processes ($\mathbb{R}$– or $\mathbb{R}^d$–valued), properly normalized and centered. It follows from Proposition 2.3 that this means to prove relations of the type

$$p\,\mathbb{P}\big(c_n^{-1}(S_n - d_n) \in (a,b]\big) \to \mu(a,b] \quad \text{or} \quad p\,\mathbb{P}\big(c_n^{-1}(S_n - d_n) > a\big) \to \mu(a,\infty)\,,$$

provided $\mu(a,b] + \mu(a,\infty) < \infty$. Since $p = p_n \to \infty$, this means that $\mathbb{P}\big(c_n^{-1}(S_n - d_n) > a\big) \to 0$ as $n \to \infty$. We will refer to these vanishing probabilities as *large deviation probabilities*. In Section 2.3 we consider some of the well–known precise large deviation results in heavy– and light–tail situations.

2.3 Preliminaries II: precise large deviations

In this section we collect some precise large deviation results in the light– and heavy–tailed cases.

2.3.1 Large deviations with normal approximation

We assume $\mathbb{E}[X] = 0$, $\mathrm{var}(X) = 1$ and write Φ for the standard normal distribution. We start with a classical result when X has finite exponential moments.

Theorem 2.4 (Petrov's theorem [98], Theorem 1, Chapter VIII)**.** *Assume that the moment generating function $\mathbb{E}[\exp(h\,X)]$ is finite in some neighborhood of the origin. Then the following tail bound holds for $0 \le x = o(\sqrt{n})$:*

$$\frac{\mathbb{P}(S_n/\sqrt{n} > x)}{\overline{\Phi}(x)} = \exp\Big(\frac{x^3}{\sqrt{n}}\lambda\big(\frac{x}{\sqrt{n}}\big)\Big)\Big[1 + O\big(\frac{x+1}{\sqrt{n}}\big)\Big]\,, \quad n \to \infty\,,$$

where $\lambda(t) = \sum_{k=0}^{\infty} a_k t^k$ is the Cramér series whose coefficients a_k depend on the cumulants of X, and $\lambda(t)$ converges for sufficiently small values $|t|$.

Under the conditions of Theorem 2.4, uniformly for $x = o(n^{1/6})$,

$$\frac{\mathbb{P}(S_n/\sqrt{n} > x)}{\overline{\Phi}(x)} \to 1\,, \qquad n \to \infty\,. \tag{2.9}$$

Theorem 7 in Chapter VIII of Petrov [98] considers the situation of Theorem 2.4 under the additional assumption that the cumulants of order $k = 3, \dots, r+2$ of X vanish for some positive integer r. Then the coefficients $a_0, \dots, a_{r-1}$ in the series $\lambda(t)$ vanish, and it is not difficult to see that (2.9) holds uniformly for $0 \le x = o\big(n^{(r+1)/(2(r+3))}\big)$.

In Section VIII.3 of [98] one also finds necessary and sufficient conditions for (2.9) to hold in certain intervals.

The following result was proved by S.V. Nagaev [89] for $x \in (0, \sqrt{(s/2-1)\log n})$ and improved by R. Michel [83] for $x \in (0, \sqrt{(s-2)\log n})$. The statement of the proposition is sharp under the given moment condition; see Theorem 2.8 below.

Proposition 2.5. *Assume that* $\mathbb{E}[|X|^s] < \infty$ *for some* $s > 2$. *Then* (2.9) *holds uniformly for* $0 \le x \le \sqrt{(s-2)\log n}$.

2.3.2 Large deviations with normal/subexponential approximations

Cline and Hsing [36] (in an unpublished article) discovered that the subexponential class $\mathcal{S}$ of distributions exhibits a completely different kind of large deviation behavior:

Proposition 2.6 (Cline and Hsing [36])**.** *We consider a distribution* F *on* $(0, \infty)$ *with infinite right endpoint. Then the following statements hold.*

1. $F \in \mathcal{S}$ *if and only if*

$$\lim_{x\to\infty} \frac{\overline{F}(x+y)}{\overline{F}(x)} = 1\,, \quad \textit{for any real } y, \tag{2.10}$$

and there exists a sequence $\gamma_n \to \infty$ *such that*

$$\lim_{n\to\infty} \sup_{x>\gamma_n} \frac{\mathbb{P}(S_n > x)}{n\,\overline{F}(x)} \le 1\,.$$

2. *If* $F \in \mathcal{S}$ *then there exists a sequence* $\gamma_n \to \infty$ *such that*

$$\lim_{n\to\infty} \sup_{x>\gamma_n} \Big| \frac{\mathbb{P}(S_n > x)}{n\,\overline{F}(x)} - 1 \Big| = 0\,. \tag{2.11}$$

Remark 2.7. If F satisfies (2.10) we say that F is *long–tailed*, we write $F \in \mathcal{L}$. It is well known that $F \in \mathcal{S}$ implies $F \in \mathcal{L}$; see Embrechts et al. [46, Lemma 1.3.5, page 41]. The converse is not true.

Proposition 2.6 shows that the subexponential class is the one for which heavy–tail large deviations are reasonable to study. Given that we know that F is long–tailed, F is subexponential if and only if a uniform large deviation relation of the type (2.11) holds.

Subexponential and normal approximations to large deviation probabilities were studied in detail in various papers. Among them, large deviations for i.i.d. regularly varying random variables are perhaps studied best. S.V. Nagaev [90] formulated a seminal result about the large deviations of a random walk (S_n) in the case of regularly varying X with finite variance. He dedicated this theorem to his brother A.V. Nagaev who had started this line of research in the 1960s; see for example [86, 87].

Theorem 2.8 (Nagaev's theorem [86, 90])**.** *Consider an i.i.d. sequence (X_i) of random variables with $\mathbb{E}[X] = 0$, $\mathrm{var}(X) = 1$ and $\mathbb{E}[|X|^{2+\delta}] < \infty$ for some $\delta > 0$. Assume that $\overline{F}(x) = x^{-\alpha} L(x)$, $x > 0$, for some $\alpha > 2$ and a slowly varying function L. Then for $x \geq \sqrt{n}$ as $n \to \infty$,*

$$\mathbb{P}(S_n > x) = \overline{\Phi}(x/\sqrt{n})\,(1 + o(1)) + n\,\overline{F}(x)\,(1 + o(1))\,.$$

In particular, if X satisfies (2.2) with constants $p_\pm$, then for any positive constant $c_1 < \alpha - 2$

$$\sup_{1 < x/\sqrt{n} < \sqrt{c_1 \log n}} \Big| \frac{\mathbb{P}(\pm S_n > x)}{\overline{\Phi}\big(x/\sqrt{n}\big)} - 1 \Big| \to 0\,, \qquad n \to \infty\,, \tag{2.12}$$

and for any constant $c_2 > \alpha - 2$,

$$\sup_{x/\sqrt{n} > \sqrt{c_2 \log n}} \Big| \frac{\mathbb{P}(\pm S_n > x)}{n\,\mathbb{P}(|X| > x)} - p_\pm \Big| \to 0\,, \qquad n \to \infty\,.$$

Remark 2.9. If X is regularly varying with index α, $\mathbb{E}[|X|^s]$ is finite (infinite) for $s < \alpha$ ($s > \alpha$). Therefore the normal approximation (2.12) is in agreement with Proposition 2.5.

In the infinite variance regularly varying case this result is complemented by an analogous statement. It can be found in Cline and Hsing [36], Denisov et al. [41].

Theorem 2.10. *Consider an i.i.d. sequence (X_i) of regularly varying random variables with index $\alpha \in (0, 2]$ satisfying (2.2). Assume $\mathbb{E}[X] = 0$ if this expectation is finite. Choose (a_n) such that*

$$n\,\mathbb{P}(|X| > a_n) + \frac{n}{a_n^2} \mathbb{E}[X^2\,\mathbf{1}(|X| \leq a_n)] = 1\,, \qquad n = 1, 2, \ldots,$$

and (γ_n) such that $\gamma_n / a_n \to \infty$ as $n \to \infty$. For $\alpha = 2$, also assume for sufficiently small $\delta > 0$,

$$\lim_{n \to \infty} \sup_{x > \gamma_n} \frac{n}{x^2} \frac{\mathbb{E}[X^2\,\mathbf{1}(|X| \leq x)]}{[n\,\mathbb{P}(|X| > x)]^\delta} = 0\,.$$

Choose (d_n) such that

$$d_n = \begin{cases} 0\,, & \alpha \in (0,1) \cup (1,2]\,, \\ n\,\mathbb{E}[X\,\mathbf{1}(|X| \leq a_n)]\,, & \alpha = 1\,. \end{cases} \tag{2.13}$$

Then the following large deviation result holds:

$$\sup_{x > \gamma_n} \Big| \frac{\mathbb{P}\big(\pm\,(S_n - d_n) > x\big)}{n\,\mathbb{P}(|X| > x)} - p_\pm \Big| \to 0\,, \qquad n \to \infty\,.$$

Remark 2.11. The normalization (a_n) is chosen such that $a_n^{-1}(S_n - d_n) \xrightarrow{\mathrm{d}} Y_\alpha$ for an α–stable random variable Y_α, $\alpha \in (0, 2]$. Therefore $\gamma_n^{-1}(S_n - d_n) \xrightarrow{\mathbb{P}} 0$. In the case $\alpha < 2$, in view of Karamata's theorem (see Bingham et al. [20]), it is possible to choose (a_n) according as $n\,\mathbb{P}(|X| > a_n) \to 1$. The case $\alpha = 2$ is delicate: in this case $\mathrm{var}(X)$ can be finite or infinite. In the former case, (a_n) is proportional to $\sqrt{n}$, in the latter case $(a_n/\sqrt{n})$ is a slowly varying sequence; see Feller [48] or Ibragimov and Linnik [64, Section II.6].

Normal and subexponential approximations to large deviation probabilities also exist for subexponential distributions that have all moments finite. Early on, this was observed by A.V. Nagaev [86, 87, 88]. Rozovskii [104] did not use the name of subexponential distribution, but the conditions on the tails of the distributions he introduced are "close" to subexponentiality; he also allowed for distributions F supported on the whole real line. In particular, A.V. Nagaev and Rozovskii discovered that, in general, the x–regions where the normal and subexponential approximations hold are separated from each other. To make this precise, we call two sequences (ξ_n) and (ψ_n) *separating sequences for the normal and subexponential approximations to large deviation probabilities* if for an i.i.d. sequence (X_i) with variance 1,

$$\sup_{x<\xi_n} \Big| \frac{\mathbb{P}(S_n - \mathbb{E}[S_n] > x)}{\overline{\Phi}(x/\sqrt{n})} - 1 \Big| \to 0\,,$$
$$\sup_{x>\psi_n} \Big| \frac{\mathbb{P}(S_n - \mathbb{E}[S_n] > x)}{n\,\mathbb{P}(X > x)} - 1 \Big| \to 0\,, \qquad n \to \infty\,.$$

A.V. Nagaev and Rozovskii gave conditions under which (ψ_n) and (ξ_n) cannot have the same asymptotic order; i.e., one necessarily has $\psi_n/\xi_n \to \infty$. In particular, in the x–region (ξ_n, ψ_n) neither the normal nor the subexponential approximation holds; Rozovskii [104] also provided large deviation approximations for $\mathbb{P}(S_n > x)$ for these regions involving $\overline{\Phi}(x/\sqrt{n})$ and a truncated Cramér series. Explicit expressions for (ψ_n) and (ξ_n) are in general hard to get. We focus on two classes of subexponential distributions where the separating sequences are known.

- *Lognormal–type tails*, we write $F \in \mathrm{LN}(\gamma)$: for some constants $\beta, \xi \in \mathbb{R}$, $\gamma > 1$ and $\lambda, c > 0$,

$$\overline{F}(x) \sim c\, x^{\beta}\, (\log x)^{\xi}\, \exp\big(-\lambda\, (\log x)^{\gamma}\big)\,, \qquad x \to \infty\,.$$

  In the notation $\mathrm{LN}(\gamma)$ we suppress the dependence on β, ξ, λ, c.

- *Weibull–type tails*, we write $F \in \mathrm{WE}(\tau)$: for some $\beta \in \mathbb{R}$, $\tau \in (0,1)$, $\lambda, c > 0$.

$$\overline{F}(x) \sim c\, x^{\beta}\, \exp\big(-\lambda\, x^{\tau}\big)\,, \qquad x \to \infty\,.$$

  In the notation $\mathrm{WE}(\tau)$ we suppress the dependence on β, λ, c.

The name "Weibull–type tail" is motivated by the fact that the Weibull distribution F with shape parameter $\tau \in (0,1)$ belongs to $\mathrm{WE}(\tau)$. Indeed, in this case $\overline{F}(x) = \exp(-\lambda x^{\tau})$, $x > 0$, for positive parameters λ. Similarly, the lognormal distribution F belongs to $\mathrm{LN}(2)$. This is easily seen by an application of Mill's ratio: for a standard normal random variable Y,

$$\overline{F}(x) = \mathbb{P}(Y > \log x) \sim \frac{\exp\big(-(\log x)^2/2\big)}{\sqrt{2\pi}\, \log x}\,, \qquad x \to \infty\,.$$

$F \in$	ξ_n	ψ_n
$\mathrm{RV}(\alpha)\,,\ \alpha > 2$	$\big((\alpha-2)n\,\log n\big)^{1/2}$	$\big((\alpha-2)n\,\log n\big)^{1/2}$
$\mathrm{LN}(\gamma)\,,\ 1 < \gamma < 2$	$(n\,(\log n)^{\gamma})^{1/2}$	$(n\,(\log n)^{\gamma})^{1/2}$
$\mathrm{LN}(\gamma)\,,\ \gamma \geq 2$	$(n\,(\log n)^{\gamma})^{1/2}/\widetilde{h}_n$	$n^{1/2}(\log n)^{\gamma-1}\,h_n$
$\mathrm{WE}(\tau)\,,\ 0 < \tau \leq 0.5$	$n^{1/(2-\tau)}/\widetilde{h}_n$	$n^{1/(2-2\tau)}\,h_n$
$\mathrm{WE}(\tau)\,,\ 0.5 < \tau < 1$	$n^{2/3}/\widetilde{h}_n$	$n^{1/(2-2\tau)}\,h_n$

Table 2.1: Separating sequences (ξ_n) and (ψ_n) for the normal and subexponential approximations of $\mathbb{P}(S_n - \mathbb{E}[S_n] > x)$. We also assume $\mathrm{var}(X) = 1$. Here $(h_n), (\widetilde{h}_n)$ are any sequences converging to infinity. For completeness, we also include the regularly varying class $\mathrm{RV}(\alpha)$. The table is taken from Mikosch and Nagaev [84].

These classes of distributions have rather distinct tail behavior. It follows from the theory in Embrechts et al. [46, Sections 1.3 and 1.4] that membership of F in $\mathrm{RV}(\alpha)$, $\mathrm{LN}(\gamma)$ or $\mathrm{WE}(\tau)$ implies $F \in \mathcal{S}$. The case $\mathrm{WE}(\tau)$, $0 < \tau < 1$, was already considered by A.V. Nagaev [87, 88].

For the heaviest tails when $F \in \mathrm{LN}(\gamma)$, $1 < \gamma < 2$ one can still choose $\xi_n = \psi_n$. This means that one threshold sequence separates the normal and subexponential approximations to the right tail $\mathbb{P}(S_n - \mathbb{E}[S_n] > x)$. Rozovskii [104] discovered that the classes $\mathrm{LN}(\gamma)$, $\gamma \geq 2$, and $\mathrm{LN}(\gamma)$, $1 < \gamma < 2$ have rather distinct large deviation properties. In the case $\gamma \geq 2$ one cannot choose (ξ_n) and (ψ_n) the same. The class $\mathrm{LN}(\gamma)$ with $1 < \gamma < 2$ satisfies the conditions of Theorem 3b in Rozovskii [104] which implies that

$$\mathbb{P}(S_n - \mathbb{E}[S_n] > x) = \big[\overline{\Phi}(x/\sqrt{n})\mathbf{1}\big(x < \gamma_n\big) + n\overline{F}(x)\mathbf{1}\big(x > \gamma_n\big)\big](1 + o(1))$$

uniformly for x, where $\gamma_n = \big(\lambda 2^{-\gamma+1}\big)^{1/2} n^{1/2} (\log n)^{\gamma/2}$. For $\gamma = 2$ the conditions of Theorem 3a in [104] are satisfied: with $g(x) = \lambda(\log x)^2 - (\beta+2)\log x - \xi \log(\log x) - \log c$ and as $n \to \infty$,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] > x) = \big[\overline{\Phi}(x/\sqrt{n})\,\mathbf{1}(x < \gamma_n) + n\overline{F}(x)\mathrm{e}^{n(g'(x))^2/2}\mathbf{1}(x > \gamma_n)\big](1 + o(1))\,.$$

Direct calculation shows that $\mathbb{P}(S_n - \mathbb{E}[S_n] > \gamma_n) \sim \exp(\lambda)\, n\,\overline{F}(\gamma_n)$ while, uniformly for $x > \gamma_n h_n$, $h_n \to \infty$, we have that $\mathbb{P}(S_n - \mathbb{E}[S_n] > x) \sim n\overline{F}(x)$.

It is interesting to observe that all but one class of subexponential distributions considered in Table 2.1 have the property that $c\,n \in (\psi_n, \infty)$ for any $c > 0$. The exception is $\mathrm{WE}(\tau)$ for $\tau \in (0.5, 1)$. This fact turns the investigation of the tail probabilities $\mathbb{P}(S_n - \mathbb{E}[S_n] > c\,n)$ into a complicated technical problem. The exponential ($\mathrm{WE}(1)$) and superexponential ($\mathrm{WE}(\tau)$), $\tau > 1$, classes do not contain subexponential distributions. The corresponding partial sums exhibit the light-tailed large deviation behavior of Petrov's Theorem 2.4. As a historical remark, Linnik [78] and S.V. Nagaev [89] determined lower separating sequences (ξ_n) for the normal approximation to the tails $\mathbb{P}(S_n - \mathbb{E}[S_n] > x)$ under the assumption that $\overline{F}$ is dominated by the tail of a regular subexponential distribution from the table.

Denisov et al. [41] and Cline and Hsing [36] considered a unified approach to subexponential large deviation approximations for general subexponential and related distributions. In particular, they identified separating sequences (ψ_n) for the subexponential approximation of the tails $\mathbb{P}(S_n - \mathbb{E}[S_n] > x)$ for general subexponential distribu-

tions. Denisov et al. [41] also considered local versions, i.e., approximations to the tails $\mathbb{P}(S_n \in [x, x+T])$ for $T > 0$ as $x \to \infty$.

2.4 Main results

2.4.1 Gumbel convergence via normal approximations to large deviation probabilities for small x.

We assume that $\mathbb{E}[X] = 0$ and $\mathrm{var}(X) = 1$ and the large deviation approximation to the standard normal distribution Φ holds: for some $\gamma_n \to \infty$,

$$\sup_{0 \le x < \gamma_n} \Big| \frac{\mathbb{P}(S_n/\sqrt{n} > x)}{\overline{\Phi}(x)} - 1 \Big| \to 0\,, \qquad n \to \infty\,. \tag{2.14}$$

We recall that $\Phi \in \mathrm{MDA}(\Lambda)$ and (2.4) holds. An analogous relation holds for the maxima of i.i.d. random walks $S_{n1}/\sqrt{n}, \ldots, S_{np}/\sqrt{n}$ as follows from the next result.

Theorem 2.12. *Assume that* (2.14) *is satisfied for some* $\gamma_n \to \infty$. *Then*

$$p\,\mathbb{P}\Big(\frac{S_n}{\sqrt{n}} > d_p + x/d_p\Big) \to \mathrm{e}^{-x}\,, \qquad n \to \infty\,, \qquad x \in \mathbb{R}\,, \tag{2.15}$$

holds for any integer sequence $p_n \to \infty$ *such that* $p_n < \exp(\gamma_n^2/2)$ *and* (d_p) *is defined in* (2.5). *Moreover, for the considered* (p_n), (2.15) *is equivalent to either of the following limit relations:*

1. *For* $\Gamma_i = E_1 + \cdots + E_i$ *and an i.i.d. standard exponential sequence* (E_i) *the following point process convergence holds on the state space* $\mathbb{R}$

$$N_p = \sum_{i=1}^{p} \varepsilon_{d_p\left(\frac{S_{ni}}{\sqrt{n}} - d_p\right)} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{-\log \Gamma_i}\,, \tag{2.16}$$

   where N *is* $\mathrm{PRM}(-\log \Lambda)$ *on* $\mathbb{R}$.

2. *Gumbel convergence of the maximum random walk*

$$d_p \max_{i=1,\ldots,p} \big(S_{ni}/\sqrt{n} - d_p\big) \xrightarrow{\mathrm{d}} Y \sim \Lambda\,, \qquad n \to \infty\,.$$

Proof. In view of Proposition 2.3 it suffices for $N_p \xrightarrow{\mathrm{d}} N$ to show that

$$p\,\mathbb{P}\Big(d_p\big(\frac{S_n}{\sqrt{n}} - d_p\big) > x\Big) = p\,\mathbb{P}\Big(\frac{S_n}{\sqrt{n}} > d_p + x/d_p\Big) \sim p\,\overline{\Phi}(d_p + x/d_p) \to \mathrm{e}^{-x}\,, \qquad x \in \mathbb{R}\,.$$

But this follows from (2.14) and the definition of (d_p) if we assume that $d_p + x/d_p < \gamma_n$, i.e., $p_n < \exp(\gamma_n^2/2)$ such that $p_n \to \infty$.

If $N_p \xrightarrow{\mathrm{d}} N$ a continuous mapping argument implies that

$$\begin{aligned} \mathbb{P}\big(N_p(x,\infty) = 0\big) &= \mathbb{P}\Big(\max_{i=1,\ldots,p} d_p\big(S_{ni}/\sqrt{n} - d_p\big) \le x\Big) \\ &\to \mathbb{P}\big(N(x,\infty) = 0\big) = \Lambda(x)\,, \qquad x \in \mathbb{R}\,, \qquad n \to \infty\,. \end{aligned}$$

On the other hand, for $x \in \mathbb{R}$ as $n \to \infty$,

$$\begin{aligned}\mathbb{P}\Big(\max_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} - d_p\big) \le x\Big) &= \Big(1 - \frac{p\,\mathbb{P}\big(d_p(S_{n1}/\sqrt{n} - d_p) > x\big)}{p}\Big)^p \\ &\to \exp\big(-\mathrm{e}^{-x}\big)\,,\end{aligned}$$

if and only if (2.15) holds. □

Remark 2.13. If one replaces in (2.16) the quantities $(S_{ni}/\sqrt{n})_{i=1,\dots,p}$ by i.i.d. standard normal random variables then this limit relation remains valid. This means that, under (2.14), e.g. under the assumption of a finite moment generating function in some neighborhood of the origin (see Section 2.3), the central limit theorem makes the tails of $(S_{ni}/\sqrt{n})_{i=1,\dots,p}$ almost indistinguishable from those of the standard normal distribution. This is in stark contrast to subexponential distributions, where the characteristics of $\overline{F}(x)$ show up in the tail $\mathbb{P}(S_{ni}/\sqrt{n} > x)$ for large values of x.

2.4.1.1 The extreme values of i.i.d. random walks.

Write

$$S_{n,(p)} \le \cdots \le S_{n,(1)}$$

for the ordered values of $S_{n1}, \dots, S_{np}$ The following result is immediate from Theorem 2.12.

Corollary 2.14. *Assume that the conditions of Theorem 2.12 hold. Then*

$$\sqrt{2\log p}\,\Big(\frac{S_{n,(1)}}{\sqrt{n}} - d_p, \dots, \frac{S_{n,(k)}}{\sqrt{n}} - d_p\Big) \stackrel{\mathrm{d}}{\to} \big(-\log\Gamma_1, \dots, -\log\Gamma_k\big)\,, \qquad n \to \infty\,. \tag{2.17}$$

Moreover, if there is $\gamma_n \to \infty$ such that $\sup_{0\le x<\gamma_n} \big|\mathbb{P}(\pm S_n/\sqrt{n} > x)/\overline{\Phi}(x) - 1\big| \to 0$ as $n \to \infty$, then we have

$$\begin{aligned}&\mathbb{P}\Big(\max_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} - d_p\big) \le x\,, \min_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} + d_p\big) \le y\Big) \\ &\quad\to \Lambda(x)\big(1 - \Lambda(-y)\big)\,, \qquad x, y \in \mathbb{R}\,, \qquad n \to \infty\,.\end{aligned} \tag{2.18}$$

Proof. We observe that $d_p/\sqrt{2\log p} \to 1$. Then (2.16) and the continuous mapping theorem imply that (2.17) holds for any fixed $k \ge 1$.

We observe that

$$\begin{aligned}&\mathbb{P}\Big(\max_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} - d_p\big) \le x\,, \min_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} + d_p\big) \le y\Big) \\ &= \mathbb{P}\Big(\max_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} - d_p\big) \le x\Big) \\ &\quad -\mathbb{P}\Big(\max_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} - d_p\big) \le x\,, \min_{i=1,\dots,p} d_p\big(S_{ni}/\sqrt{n} + d_p\big) > y\Big) \\ &= P_1(x,y) - P_2(x,y)\,.\end{aligned}$$

Of course, $P_1(x,y) \to \Lambda(x)$. On the other hand,

$$P_2(x,y) = \mathbb{P}\Big(\bigcap_{i=1}^p \big\{S_{ni}/\sqrt{n} \le d_p + x/d_p\,, S_{ni}/\sqrt{n} > -d_p + y/d_p\big\}\Big)$$

$$
\begin{aligned}
&= \Big(\mathbb{P}\Big(- d_p + y/d_p < S_{n1}/\sqrt{n} \le d_p + x/d_p\Big)\Big)^p \\
&= \exp\Big(p \log\Big(1 - \mathbb{P}\Big(S_{n1}/\sqrt{n} > d_p + x/d_p\Big) - \mathbb{P}\Big(S_{n1}/\sqrt{n} \le -d_p + y/d_p\Big)\Big)\Big) \\
&\to \exp\big(- (\mathrm{e}^{-x} + \mathrm{e}^{y})\big) = \Lambda(x)\Lambda(-y)\,.
\end{aligned}
$$

The last step follows from a Taylor expansion of the logarithm and Theorem 2.12. This proves (2.18). □

2.4.1.2 Examples.

In this section we verify the assumptions of Theorem 2.12 for various classes of distributions F. We always assume $\mathbb{E}[X] = 0$ and $\mathrm{var}(X) = 1$.

Example 2.15. Assume the existence of the moment generating function of X in some neighborhood of the origin. Petrov's Theorem 2.4 ensures (2.16) for $p \le \exp(o(n^{1/3}))$.

Example 2.16. Assume $\mathbb{E}[|X|^s] < \infty$ for some $s > 2$. Proposition 2.5 ensures that (2.16) for $p \le n^{(s-2)/2}$.

Example 2.17. Assume that X is regularly varying with index $\alpha > 2$. Then we can apply Nagaev's Theorem 2.8 with $\gamma_n = \sqrt{c \log n}$ for any $c < \alpha - 2$ and (2.16) holds for $p \le n^{c/2}$. This is in agreement with Example 2.16.

Example 2.18. Assume that X has a distribution in $\mathrm{LN}(\gamma)$ for some $\gamma > 1$. From Table 2.1, $\gamma_n = o((\log n)^{\gamma/2}$, and (2.16) holds for $p \le \exp(o((\log n)^{\gamma}))$

Example 2.19. Assume that $F \in \mathrm{WE}(\tau)$, $0 < \tau < 1$. Table 2.1 yields $\gamma_n = o(n^{\tau/(2(2-\tau))})$ for $\tau \le 0.5$, hence $p \le \exp(o(n^{\tau/(2-\tau)}))$, and for $\tau \in (0.5, 1)$, $\gamma_n = o(n^{1/6})$ and $p \le \exp(o(n^{1/3}))$.

We summarize these examples in Table 2.2.

Example No	Upper bound for p
2.15 Petrov case	$\exp(o(n^{1/3}))$
2.16 $\mathbb{E}[\|X\|^s] < \infty$, $s > 2$	$n^{(s-2)/2}$
2.17 $\mathrm{RV}(\alpha)$, $\alpha > 2$, $c < \alpha - 2$	$n^{c/2}$
2.18 $\mathrm{LN}(\gamma)$, $\gamma > 1$	$\exp(o((\log n)^{\gamma}))$
2.19 $\mathrm{WE}(\tau)$, $\tau \le 0.5$	$\exp(o(n^{\tau/(2-\tau)}))$
2.19 $\mathrm{WE}(\tau)$, $\tau \in (0.5, 1)$	$\exp(o(n^{1/3}))$

Table 2.2: Upper bounds for p.

2.4.1.3 The extremes of the blocks of a random walk.

We consider a random walk S_n with i.i.d. step sizes X_i with $\mathbb{E}[X] = 0$ and $\mathrm{var}(X) = 1$, and with distribution F, and any integer sequence $r_n \to \infty$ such that $k_n = [n/r_n] \to \infty$ as $n \to \infty$. Set $S_{ni} = S_{r_n i} - S_{r_n(i-1)}$, i.e., this is the sum of the ith block $X_{r_n(i-1)+1}, \ldots, X_{r_n i}$. Then we are in the setting of Theorem 2.12 if we replace p_n by k_n

and n by r_n. We are interested in the following result for the point process of the block sums of S_n with length r_n (see (2.16))

$$N_{k_n} = \sum_{i=1}^{k_n} \varepsilon_{d_{k_n}\left(\frac{S_{r_n i} - S_{r_n(i-1)}}{\sqrt{r_n}} - d_{k_n}\right)} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{-\log \Gamma_i}\,. \tag{2.19}$$

This means we are looking for (r_n) such that $n/r_n < \exp(\gamma^2_{r_n}/2)$. This amounts to the following conditions on (r_n) in Table 2.3:

Example No	Lower bounds for (r_n)
2.15 Petrov case	$r_n/(\log n)^3 \to \infty$
2.17 RV(α), $\alpha > 2$	$r_n > n^{2/(\alpha-\varepsilon)}$ any $\varepsilon \in (0, \alpha-2)$
2.16 $\mathbb{E}[\lvert X\rvert^s] < \infty$, $s > 2$	$r_n > n^{2/s}$
2.18 LN(γ), $\gamma > 1$	$r_n/\exp((2\log n)^{1/\gamma}) \to \infty$
2.19 WE(τ), $\tau \le 0.5$	$r_n/(\log n)^{(2-\tau)/\tau} \to \infty$
2.19 WE(τ), $\tau \in (0.5, 1)$	$r_n/(\log n)^3 \to \infty$

Table 2.3: Lower bounds on the block size r_n

This table shows convincingly that, the heavier the tails, the larger we have to choose the block length r_n. Otherwise, the normal approximation does not function sufficiently well simultaneously for the block sums $S_{r_n i} - S_{r_n(i-1)}$, $i = 1, \ldots, k_n$. In particular, in the regularly varying case we always need that r_n grows polynomially.

Notice that we have from (2.19) in particular

$$\frac{d_{k_n}}{\sqrt{r_n}} \max_{i=1,\ldots,k_n} \left(S_{r_n i} - S_{r_n(i-1)} - \sqrt{r_n}\, d_{k_n}\right) \xrightarrow{\mathrm{d}} -\log \Gamma_1 \sim \Lambda\,, \qquad n \to \infty\,.$$

The normalization $d_{k_n}/\sqrt{r_n}$ is asymptotic to $\sqrt{(2\log k_n)/r_n}$.

2.4.2 Gumbel convergence via the subexponential approximation to large deviation probabilities for very large x

In this section we will exploit the subexponential approximation to large deviation probabilities for subexponential distributions F, i.e.,

$$\sup_{x > \gamma_n} \left| \frac{\mathbb{P}(S_n - \mathbb{E}[S_n] > x)}{n\,\mathbb{P}(X > x)} - 1 \right| \to 0\,, \tag{2.20}$$

and we will also assume that $F \in \mathrm{MDA}(\Lambda)$; see Example 2.2 for the corresponding MDA conditions and the definition of the centering constants (d_n) and the normalizing constants (c_n). Then, in particular, X has all moments finite. In this case, the Gumbel approximation of the point process of the (S_{ni}) is also possible.

Theorem 2.20. *Assume that $F \in \mathrm{MDA}(\Lambda) \cap \mathcal{S}$, the subexponential approximation* (2.20) *holds and for sufficiently large n and an integer sequence $p_n \to \infty$,*

$$d_{np} + x\, c_{np} > \gamma_n\,, \qquad \textit{for any } x < 0\,, \tag{2.21}$$

where (d_{np}) and (c_{np}) are the subsequences of (d_n) and (c_n), respectively, evaluated at np. Then

$$p\,\mathbb{P}\big(S_n - \mathbb{E}[S_n] > d_{np} + x\,c_{np}\big) \to \mathrm{e}^{-x}\,, \qquad x \in \mathbb{R}\,, \qquad n \to \infty\,, \tag{2.22}$$

holds. Moreover, (2.22) *is equivalent to either of the following limit relations:*

1. *Point process convergence to a Poisson process on the state space $\mathbb{R}$*

$$N_p = \sum_{i=1}^{p} \varepsilon_{c_{np}^{-1}(S_{ni} - \mathbb{E}[S_n] - d_{np})} \stackrel{\mathrm{d}}{\to} N\,, \qquad n \to \infty\,, \tag{2.23}$$

   where $N \sim \mathrm{PRM}(-\log \Lambda)$; see Theorem 2.12.

2. *Gumbel convergence of the maximum random walk*

$$\max_{i=1,\ldots,p} c_{np}^{-1}\big((S_{ni} - \mathbb{E}[S_n]) - d_{np}\big) \stackrel{\mathrm{d}}{\to} Y \sim \Lambda\,, \qquad n \to \infty\,. \tag{2.24}$$

Proof. If $d_{np} + x\,c_{np} > \gamma_n$ for every $x < 0$, then it holds for $x \in \mathbb{R}$. Therefore (2.20) applies. Since $F \in \mathrm{MDA}(\Lambda) \cap \mathcal{S}$ and by definition of (c_n) and (d_n), we have

$$p\,\mathbb{P}\big(S_n - \mathbb{E}[S_n] > d_{np} + x\,c_{np}\big) \sim p\,n\,\mathbb{P}(X > d_{np} + x\,c_{np}) \to \mathrm{e}^{-x}\,, \quad x \in \mathbb{R}\,, \qquad n \to \infty\,,$$

proving (2.22). Proposition 2.3 yields the equivalence of (2.23) and (2.22). The equivalence of (2.23) and (2.24) follows from a standard argument. □

Remark 2.21. Since $a(x)$ defined in Example 2.2 has density $a'(x) \to 0$ as $x \to \infty$ we have $a(x)/x \to 0$. On the other hand, $c_n = a(d_n)$ and $d_n \to \infty$ since $F \in \mathcal{S}$. Therefore, for any $x > 0$,

$$d_{np} + x\,c_{np} = d_{np}\Big(1 + x\,\frac{a(d_{np})}{d_{np}}\Big) \sim d_{np}\,.$$

Hence (2.21) holds if $d_{np} \geq (1+\delta)\gamma_n$ for any small $\delta > 0$ and large n.

2.4.2.1 The extreme values of i.i.d. random walks.

Relation (2.23) and a continuous mapping argument imply the following analog of Corollary 2.14. We use the same notation as in Section 2.4.1.1. One can follow the lines of the proof of Corollary 2.14.

Corollary 2.22. *Assume the conditions of Theorem 2.20. Then the following relation holds for $k \geq 1$,*

$$c_{np}^{-1}\Big(S_{n,(1)} - \mathbb{E}[S_n] - d_{np}, \ldots, S_{n,(k)} - \mathbb{E}[S_n] - d_{np}\Big) \stackrel{\mathrm{d}}{\to} \big(-\log \Gamma_1, \ldots, -\log \Gamma_k\big)$$

as $n \to \infty$.

2.4.2.2 Examples.

Theorem 2.20 applies to $F \in \mathrm{LN}(\gamma)$, $\gamma > 1$, and $F \in \mathrm{WE}(\tau)$, $0 < \tau < 1$; see the discussion in Section 2.3.2. However, the calculation of the constants (c_n) and (d_n) is rather complicated for these classes of subexponential distributions. For illustration of the theory we restrict ourselves to two parametric classes of distributions where these constants are known.

Example 2.23. We assume that X has a standard lognormal distribution. From (2.6), Table 2.1 and Remark 2.21 we conclude that we need to verify the condition $\exp\left(\sqrt{2\log(np)}\right) \geq h_n \sqrt{n} \log n$ for a sequence (h_n) increasing to infinity arbitrarily slowly. Calculation shows that it suffices to choose $p_n \to \infty$ such that $p > \exp\left((\log n)^2\right)$.

Example 2.24. We assume that X has a Weibull distribution with tail $\overline{F}(x) = \exp(-x^\tau)$ for some $\tau \in (0,1)$. From (2.7) we conclude that $d_{np} \sim (\log np)^{1/\tau}$. In view of Remark 2.21 and Table 2.1 it suffices to verify that $(\log np)^{1/\tau} \geq h_n\, n^{1/(2-2\tau)}$ for a sequence $h_n \to \infty$ arbitrarily slowly. It holds if $p > n^{-1} \exp\left(\left(h_n n^{1/(2-2\tau)}\right)^\tau\right)$.

2.4.2.3 The extremes of the blocks of a random walk.

We appeal to the notation in Section 2.4.1.3. We are in the setting of Theorem 2.20 if we replace p_n with k_n and n with r_n. We are interested in the following result for the point process of the block sums of S_n with length r_n (see (2.23))

$$N_{k_n} = \sum_{i=1}^{k_n} \varepsilon_{c_n^{-1}\left(S_{r_n i} - S_{r_n (i-1)} - \mathbb{E}[S_{r_n}] - d_n\right)} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{-\log \Gamma_i}\,.$$

We need to verify condition (2.21) which turns into $d_n + c_n\, x > \gamma_{r_n}$. In view of Remark 2.21 it suffices to prove that $d_n > h_n \gamma_{r_n}$ for a sequence $h_n \to \infty$ arbitrarily slowly; see Table 2.1 for some γ_n–values.

We start with a standard lognormal distribution; see (2.6) for the corresponding (c_n) and (d_n). In particular, we need to verify

$$d_n = \exp\left(\sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2(2\log n)^{1/2}}\right) \geq h_n \sqrt{r_n} \log r_n\,.$$

A sufficient condition is $\exp(2\sqrt{2\log n}) > \widetilde{h}_n r_n$ for a sequence $\widetilde{h}_n \to \infty$ arbitrarily slowly. We observe that the left–hand expression is a slowly varying function.

Next we consider a standard Weibull distribution for $\tau \in (0,1)$. The constants (c_n) and (d_n) are given in (2.7). In particular, we need to verify

$$d_n \sim (\log n)^{1/\tau} > h_n r_n^{1/(2-2\tau)}\,.$$

This holds if $(\log n)^{2(1-\tau)/\tau} h_n^{-2(1-\tau)} > r_n$. Again, this is a strong restriction on the growth of (r_n) and is in contrast to the regularly varying case where polynomial growth of (r_n) is possible; see Section 2.4.3.2.

2.4.3 Fréchet convergence via the subexponential approximations to large deviation probabilities for large x

In this section we assume that X is regularly varying with index $\alpha > 0$ in the sense of (2.2). Throughout we choose a normalizing sequence (a_n) such that $n\,\mathbb{P}(|X| > a_n) \to 1$ as $n \to \infty$. The following result is an analog of Theorems 2.12 and 2.20.

Theorem 2.25. *Assume that X is regularly varying with index $\alpha > 0$ and $\mathbb{E}[X] = 0$ if the expectation is finite. Choose a sequence (d_n) such that*

$$d_n = \begin{cases} 0\,, & \alpha \in (0,1) \cup (1,\infty)\,, \\ n\,\mathbb{E}[X\,\mathbf{1}(|X| \le a_n)]\,, & \alpha = 1\,. \end{cases}$$

We assume that $p_n \to \infty$ is an integer sequence which satisfies the additional conditions

$$\begin{cases} a_{np} \ge \sqrt{(\alpha - 2 + \delta) n \log n} \text{ for some small } \delta > 0 & \text{if } \alpha > 2\,, \\ \lim_{n\to\infty} \sup_{x > a_{np}} p^{\delta} \dfrac{n}{x^2}\, \mathbb{E}[X^2\,\mathbf{1}(|X| \le x)] = 0 \text{ for some small } \delta > 0 & \text{if } \alpha = 2\,. \end{cases} \tag{2.25}$$

Then the following limit relation

$$p\,\mathbb{P}(\pm a_{np}^{-1}(S_n - d_n) > x) \to p_{\pm} x^{-\alpha}\,, \qquad x > 0\,, \qquad n \to \infty\,, \tag{2.26}$$

holds. Moreover, (2.26) is equivalent to

$$N_p = \sum_{i=1}^{p} \varepsilon_{a_{np}^{-1}(S_{ni} - d_n)} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{q_i\,\Gamma_i^{-1/\alpha}}\,, \tag{2.27}$$

where (Γ_i) is defined in Theorem 2.12 and (q_i) is an i.i.d. sequence of Bernoulli variables with distribution $\mathbb{P}(q_i = \pm 1) = p_{\pm}$ independent of (Γ_i).

Proof. We start by verifying (2.26). Assume $\alpha < 2$. Then for any sequence $p_n \to \infty$, $a_{np}/a_n \to \infty$. Therefore Theorem 2.10 and the definition of (a_{np}) yield

$$p\,\mathbb{P}(\pm a_{np}^{-1}(S_n - d_n) > x) \sim p\,n\,\mathbb{P}(\pm X > a_{np}\,x) \sim p_{\pm} x^{-\alpha}\,, \qquad n \to \infty\,.$$

If $\alpha > 2$ the same result holds in view of Theorem 2.8 since we assume condition (2.25). If $\alpha = 2$ we can again apply Theorem 2.10 with $\gamma_n = a_{np}$ and use (2.25).

We notice that the limit point process N is PRM(μ_α) with intensity

$$\mu_\alpha(dx) = |x|^{-\alpha-1}\big(p_+ \mathbf{1}(x > 0) + p_- \mathbf{1}(x < 0)\big)\,dx\,. \tag{2.28}$$

An appeal to Proposition 2.3 shows that (2.26) and (2.27) are equivalent. □

Remark 2.26. Assume $\alpha > 2$. Since $a_{np} = (np)^{1/\alpha}\ell(np)$ for a slowly varying function ℓ and $\ell(x) \ge x^{-\gamma/\alpha}$ for any small $\gamma > 0$ and sufficiently large x, (2.25) holds if $p \ge n^{(\alpha/2)-1+\gamma'}$ for any choice of $\gamma' > 0$. Assume $\alpha = 2$ and $\mathrm{var}(X) < \infty$. Then $a_{np} \sim c\sqrt{np}$ and (2.25) is satisfied for any sequence $p_n \to \infty$ and $\delta < 1$. If $\mathrm{var}(X) = \infty$, $a_{np} = (np)^{1/2}\ell(np)$ for a slowly varying function ℓ and $\mathbb{E}[X^2(|X| \le x)]$ is an increasing slowly varying function. Using Karamata bounds for slowly varying functions, we conclude that (2.25) holds if $p/n^{\gamma} \to \infty$ for any small $\gamma > 0$.

2.4.3.1 The extreme values of i.i.d. random walks.

For simplicity, we assume $d_n = 0$. Let N_p^+ be the restriction of N_p to the state space $(0,\infty)$, and $S_{n,(1)}^+$ the maximum of $(S_{n1})_+, \dots, (S_{np})_+$. We also write $\xi = \min\{i \ge 1 : q_i = 1\}$ and assume that ξ is independent of (Γ_i). Then (2.27) and the continuous mapping theorem imply that

$$\mathbb{P}\big(N_p^+(x,\infty) = 0\big) = \mathbb{P}\big(a_{np}^{-1} S_{n,(1)}^+ \le x\big) \xrightarrow{\mathrm{d}} \mathbb{P}\big(\Gamma_\xi^{-1/\alpha} \le x\big) = \Phi_\alpha^{p_+}(x)\,. \tag{2.29}$$

Moreover, we have joint convergence of minima and maxima.

Corollary 2.27. *Assume the conditions of Theorem 2.25 and $d_n = 0$. Then*

$$\lim_{n\to\infty} \mathbb{P}\Big(0 < a_{np}^{-1} \max_{i=1,\dots,p} S_{ni} \le x\,, -y < a_{np}^{-1} \min_{i=1,\dots,p} S_{ni}\Big) = \Phi_\alpha^{p_+}(x)\Phi_\alpha^{p_-}(y)\,, \quad x, y > 0\,.$$

Proof. We have

$$\begin{aligned}
\mathbb{P}\Big(a_{np}^{-1} \max_{i=1,\dots,p} S_{ni} \le x\,, -y < a_{np}^{-1} \min_{i=1,\dots,p} S_{ni}\Big) &= \mathbb{P}\big(N_p\big((x,\infty)\cup(-\infty,-y]\big) = 0\big) \\
&\to \mathbb{P}\big(N\big((x,\infty)\cup(-\infty,-y]\big) = 0\big) \\
&= \exp\big(-(p_+ x^{-\alpha} + p_- y^{-\alpha})\big) \\
&= \Phi_\alpha^{p_+}(x)\Phi_\alpha^{p_-}(y)\,, \qquad n\to\infty\,.
\end{aligned}$$

□

2.4.3.2 The extremes of the blocks of a random walk.

We appeal to the notation of Section 2.4.1.3 and apply Theorem 2.25 in the case when n is replaced by some integer–sequence $r_n \to \infty$ such that $k_n = [n/r_n] \to \infty$ and p_n is replaced by k_n. We also assume for simplicity that $d_n = 0$. Observing that a_{np} turns into $a_{r_n k_n} \sim a_n$, (2.27) turns into

$$N_{k_n} = \sum_{i=1}^{k_n} \varepsilon_{a_n^{-1}(S_{r_n\, i} - S_{r_n\,(i-1)})} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{q_i\, \Gamma_i^{-1/\alpha}}\,, \qquad n\to\infty\,.$$

For simplicity, we assume $\alpha \neq 2$. If $\alpha < 2$ no further restrictions on (r_n) are required. If $\alpha > 2$ we have the additional growth condition $a_n > \sqrt{(\alpha - 2 + \delta) r_n \log r_n}$ for sufficiently large n. Since $a_n = n^{1/\alpha}\ell(n)$ for some slowly varying function ℓ, this amounts to showing that $n^{2/\alpha}\ell^2(n)/(\alpha - 2 + \delta) > r_n \log r_n$. Since any slowly varying function satisfies $\ell(n) \ge n^{-\varepsilon}$ for any $\varepsilon > 0$ and $n \ge n_0(\varepsilon)$ we get the following sufficient condition on the growth of (r_n): for any sufficiently small $\varepsilon > 0$, $n^{2/\alpha - \varepsilon} > r_n$. This condition ensures that (r_n) is significantly smaller than n, and the larger α the more stringent this condition becomes.

An appeal to (2.29) yields in particular

$$\begin{aligned}
\mathbb{P}\Big(a_n^{-1} \max_{i=1,\dots,k_n} (S_{r_n\, i} - S_{r_n\,(i-1)})_+ \le x\Big) &\xrightarrow{\mathrm{d}} \mathbb{P}\big(\Gamma_\xi^{-1/\alpha} \le x\big) = \Phi_\alpha^{p_+}(x)\,, \\
\mathbb{P}\Big(a_n^{-1} \max_{i=1,\dots,k_n} |S_{r_n\, i} - S_{r_n\,(i-1)}| \le x\Big) &\xrightarrow{\mathrm{d}} \mathbb{P}\big(\Gamma_1^{-1/\alpha} \le x\big) = \Phi_\alpha(x)\,, \qquad n\to\infty\,.
\end{aligned}$$

2.4.3.3 Extension to a stationary regularly varying sequence.

In view of classical theory (e.g. Feller [48]) X is regularly varying with index $\alpha \in (0,2)$ if and only if $a_n^{-1}(S_n - d_n) \xrightarrow{\mathrm{d}} \xi_\alpha$ for an α–stable random variable ξ_α where one can choose (a_n) such that $n\,\mathbb{P}(|X| > a_n) \to 1$ and (d_n) as in (2.13). For the sake of argument we also assume $d_n = 0$; this is a restriction only in the case $\alpha = 1$.

If (r_n) is any integer sequence such that $r_n \to \infty$ and $k_n = [n/r_n] \to 0$ then

$$a_n^{-1} S_n = a_n^{-1} \sum_{i=1}^{k_n} (S_{r_n\, i} - S_{r_n\,(i-1)}) + o_{\mathbb{P}}(1) \xrightarrow{\mathrm{d}} \xi_\alpha\,. \tag{2.30}$$

Moreover, since $a_n/a_{r_n} \to \infty$, Theorem 2.10 yields

$$\frac{\mathbb{P}(\pm a_n^{-1} S_{r_n} > x)}{r_n \, \mathbb{P}(|X| > a_n)} \sim \frac{\mathbb{P}(\pm X > x \, a_n)}{\mathbb{P}(|X| > a_n)} \to p_\pm \, x^{-\alpha} \,, \qquad x > 0 \,. \tag{2.31}$$

Classical limit theory for triangular arrays of the row–wise i.i.d. random variables $(S_{r_n i} - S_{r_n (i-1)})_{i=1,\ldots,k_n}$ (e.g. Petrov [98, Theorem 8, Chapter IV]) yields that (2.30) holds if and only if

$$k_n \, \mathbb{P}(a_n^{-1} S_{r_n} \in \cdot) \stackrel{v}{\to} \mu_\alpha(\cdot) \,, \tag{2.32}$$

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} k_n \mathrm{var}\big(a_n^{-1} S_{r_n} \mathbf{1}\big(|S_{r_n}| \le \delta a_n\big)\big) = 0 \,, \tag{2.33}$$

where μ_α is defined in (2.28). We notice that (2.32) is equivalent to (2.31).

An alternative way of proving limit theory for the sum process (S_n) with an α–stable limit ξ_α would be to *assume* the relations (2.32) and (2.33). This would be rather indirect and complicated in the case of i.i.d. (X_i). However, this approach has some merits in the case when (X_i) is a strictly stationary sequence with a regularly varying dependence structure, i.e., its finite–dimensional distributions satisfy a multivariate regular variation condition (see Davis and Hsing [38] or Basrak and Segers [15]), and a weak dependence assumption of the type

$$\mathbb{E}\big[\exp\big(a_n^{-1} i t S_n\big)\big] - \Big(\mathbb{E}\big[\exp\big(a_n^{-1} i t S_{r_n}\big)\big]\Big)^{k_n} \to 0 \,, \quad t \in \mathbb{R} \,, \qquad n \to \infty \,, \tag{2.34}$$

holds. Then $a_n^{-1} S_n \stackrel{\mathrm{d}}{\to} \xi_\alpha$ if and only if $a_n^{-1} \sum_{i=1}^{k_n} S_{ni} \stackrel{\mathrm{d}}{\to} \xi_\alpha$ where $(S_{ni})_{i=1,\ldots,k_n}$ is an i.i.d. sequence with the same distribution as S_{r_n}. Condition (2.34) is satisfied under mild conditions on (X_i), in particular under standard mixing conditions such as α–mixing. Thus one has to prove the conditions (2.32) and (2.33). In the dependent case the limit measure μ_α has to be modified. The following analog of (2.31) holds: there exists a positive number θ_X such that

$$\frac{\mathbb{P}(\pm a_n^{-1} S_{r_n} > x)}{r_n \, \mathbb{P}(|X| > a_n)} \sim \theta_X \, \frac{\mathbb{P}(\pm X > x \, a_n)}{\mathbb{P}(|X| > a_n)} \to \theta_X \, p_\pm \, x^{-\alpha} \,, \qquad x > 0 \,.$$

The quantity θ_X has an explicit structure in terms of the so–called tail chain of the regularly varying sequence (X_i). It has interpretation as a *cluster index* in the context of the partial sum operation acting on (X_i). For details we refer to Mikosch and Wintenberger [85] and the references therein.

2.4.3.4 Extension to the multivariate regularly varying case.

Consider a sequence $(\boldsymbol{X}_i)$ of i.i.d. $\mathbb{R}^d$–valued random vectors with generic element $\boldsymbol{X}$, and define

$$\mathbf{S}_0 = \mathbf{0} \,, \qquad \mathbf{S}_n = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n \,, \qquad n \ge 1.$$

We say that $\boldsymbol{X}$ is regularly varying with index $\alpha > 0$ and a Radon measure μ on $\mathbb{R}^d_{\mathbf{0}} = \mathbb{R}^d \backslash \{\mathbf{0}\}$, and we write $\boldsymbol{X} \in \mathrm{RV}(\alpha, \mu)$, if the following vague convergence relation is satisfied on $\mathbb{R}^d_{\mathbf{0}}$:

$$\frac{\mathbb{P}(x^{-1} \boldsymbol{X} \in \cdot)}{\mathbb{P}(|\boldsymbol{X}| > x)} \stackrel{v}{\to} \mu(\cdot) \,, \qquad x \to \infty \,, \tag{2.35}$$

and μ has the homogeneity property $\mu(t\,\cdot) = t^{-\alpha}\mu(\cdot)$, $t > 0$. We will also use the sequential version of regular variation: for a sequence (a_n) such that $n\mathbb{P}(|\boldsymbol{X}| > a_n) \to 1$, (2.35) is equivalent to

$$n\,\mathbb{P}(a_n^{-1}\boldsymbol{X} \in \cdot) \xrightarrow{v} \mu(\cdot)\,, \qquad n \to \infty\,.$$

For more reading on multivariate regular variation, we refer to Resnick [103, 102].

Hult et al. [63] extended Nagaev's Theorem 2.8 to the multivariate case:

Theorem 2.28 (A multivariate Nagaev-type large deviation result)**.** *Consider an i.i.d. $\mathbb{R}^d$–valued sequence $(\boldsymbol{X}_i)$ with generic element $\boldsymbol{X}$. Assume the following conditions.*

(1) $\mathbf{X} \in \mathrm{RV}(\alpha, \mu)$.

(2) *The sequence of positive numbers (x_n) satisfies*

$$x_n^{-1}\mathbf{S}_n \xrightarrow{\mathbb{P}} \mathbf{0} \ \text{ as } \ n \to \infty\,, \tag{2.36}$$

and, in addition,

$$\begin{cases} \dfrac{x_n^2}{n\mathbb{E}[|\boldsymbol{X}|^2\mathbf{1}(|\boldsymbol{X}| \le x_n)]\log x_n} \to \infty & \alpha = 2 \ \text{ and } \ \mathbb{E}[|\boldsymbol{X}|^2] = \infty\,, \\ \dfrac{x_n^2}{n\log n} \to \infty, & \alpha > 2 \ \text{ or } \ [\alpha = 2 \ \text{ and } \ \mathbb{E}[|\boldsymbol{X}|^2] < \infty]\,. \end{cases} \tag{2.37}$$

Then

$$\frac{\mathbb{P}(x_n^{-1}\mathbf{S}_n \in \cdot)}{n\mathbb{P}(|\boldsymbol{X}| > x_n)} \xrightarrow{v} \mu(\cdot)\,, \qquad n \to \infty\,.$$

Remark 2.29. Condition (2.36) requires that $n\,\mathbb{E}[\boldsymbol{X}]/a_{np} \to \mathbf{0}$ for $\alpha > 1$. It is always satisfied if $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$. Now assume that the latter condition is satisfied if the expectation of $\boldsymbol{X}$ is finite. If $\alpha \in (0, 2)$ we can choose any (p_n) such that $p_n \to \infty$. If $\alpha \ge 2$ and $(np)^{1/\alpha}/n^{0.5+\gamma/\alpha} \to \infty$, equivalently, $p/n^{\alpha/2-1+\gamma} \to \infty$ holds for any small $\gamma > 0$ then (2.37) is satisfied.

The following result extends Theorem 2.25 to the multivariate case.

Theorem 2.30. *Assume that $\boldsymbol{X}$ satisfies the conditions of Theorem 2.28. Consider an integer sequence $p = p_n \to \infty$ and, in addition for $\alpha \ge 2$, that $x_n = a_{np}$ satisfies* (2.37)*. Then the following limit relation holds*

$$N_p = \sum_{i=1}^{p} \varepsilon_{a_{np}^{-1}\mathbf{S}_{ni}} \xrightarrow{\mathrm{d}} N\,,$$

where $(\mathbf{S}_{ni})$ are i.i.d. copies of $\mathbf{S}_n$ and N is $\mathrm{PRM}(\mu)$ on $\mathbb{R}^d_{\mathbf{0}}$.

Proof. In view of Proposition 2.3 it suffices to show that

$$p\,\mathbb{P}(a_{np}^{-1}\mathbf{S}_n \in \cdot) \xrightarrow{v} \mu(\cdot).$$

Assume $\alpha < 2$. Then for any sequence $p_n \to \infty$, $a_{np}/a_n \to \infty$. Therefore Theorem 2.28 and the definition of (a_{np}) imply that for any μ–continuity set $A \subset \mathbb{R}^d_{\mathbf{0}}$,

$$p\,\mathbb{P}(a_{np}^{-1}\mathbf{S}_n \in A) \sim p\,n\,\mathbb{P}(|\boldsymbol{X}| > a_{np})\,\mu(A) \to \mu(A)\,, \qquad n \to \infty\,.$$

If $\alpha \ge 2$ the same result holds by virtue of Theorem 2.28 and the additional condition (2.37). □

Example 2.31. Write

$$\begin{aligned}\mathbf{S}_{ni} &= \big(S_{ni}^{(1)},\ldots,S_{ni}^{(d)}\big)^\top,\\ \mathbf{M}_n &= \big(\max_{i=1,\ldots,p} S_{ni}^{(1)},\ldots,\max_{i=1,\ldots,p} S_{ni}^{(d)}\big)^\top = \big(M_n^{(1)},\ldots,M_n^{(d)}\big)^\top.\end{aligned}$$

For vectors $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$ with non-negative components, we write $\boldsymbol{x}\le\boldsymbol{y}$ for the componentwise ordering, $[\mathbf{0},\boldsymbol{x}]=\{\boldsymbol{y}:\mathbf{0}\le\boldsymbol{y}\le\boldsymbol{x}\}$ and $[\mathbf{0},\boldsymbol{x}]^c=\mathbb{R}_+^d\backslash[\mathbf{0},\boldsymbol{x}]$. We have by Theorem 2.30,

$$\begin{aligned}\mathbb{P}\big(\mathbf{0}\le a_{np}^{-1}\mathbf{M}_n\le\boldsymbol{x}\big) &= \mathbb{P}\big(N_p([\mathbf{0},\boldsymbol{x}]^c)=0\big)\\ &\to \mathbb{P}\big(N([\mathbf{0},\boldsymbol{x}]^c)=0\big)\\ &= \exp\big(-\mu([\mathbf{0},\boldsymbol{x}]^c)\big)=:H(\boldsymbol{x}),\qquad n\to\infty,\end{aligned}$$

for the continuity points of the function $-\log H(\boldsymbol{x})=\mu([\mathbf{0},\boldsymbol{x}]^c)$. If $\mu(\mathbb{R}_+^d\backslash\{\mathbf{0}\})$ is not zero, H defines a distribution on $\mathbb{R}_+^d$ with the property $-\log H(t\boldsymbol{x})=t^{-\alpha}(-\log H(\boldsymbol{x}))$, $t>0$. The non–degenerate components of H are in the type of the Fréchet distribution; H is referred to as a multivariate Fréchet distribution with exponent measure μ.

2.4.3.5 An extension to i.i.d. random sums.

In this section we consider an alternative random sum process:

$$S(t)=\sum_{i=1}^{\nu(t)}X_i,\qquad t\ge 0,$$

where $(\nu(t))_{t\ge0}$ is a process of integer–valued non–negative random variables independent of the i.i.d. sequence (X_i) with generic element X and finite expectation. Throughout we assume that $\lambda(t)=\mathbb{E}[\nu(t)]$, $t\ge0$, is finite but $\lim_{t\to\infty}\lambda(t)=\infty$. We also define

$$m(t)=\mathbb{E}[S(t)]=\mathbb{E}[X]\,\lambda(t).$$

In addition, we assume some technical conditions on the process ν:

N1 $\nu(t)/\lambda(t)\xrightarrow{\mathbb{P}}1$, $t\to\infty$.

N2 There exist $\epsilon,\delta>0$ such that

$$\lim_{t\to\infty}\sum_{k>(1+\delta)\lambda(t)}\mathbb{P}(\nu(t)>k)\,(1+\epsilon)^k=0.$$

These conditions are satisfied for a wide variety of processes ν, including the homogeneous Poisson process on $(0,\infty)$. Klüppelberg and Mikosch [72] proved the following large deviation result for the random sums $S(t)$ ([72] allow for the more general condition of extended regular variation).

Theorem 2.32. *Assume that ν satisfies* **N1, N2** *and is independent of the i.i.d. non-negative sequence (X_i) which is regularly varying with index $\alpha>1$. Then for any $\gamma>0$,*

$$\sup_{x\ge\gamma\lambda(t)}\Big|\frac{\mathbb{P}(S(t)-m(t)>x)}{\lambda(t)\mathbb{P}(X>x)}-1\Big|,\qquad t\to\infty.$$

The same method of proof as in the previous sections in combination with the large deviation result of Theorem 2.32 yields the following statement. As usual, we assume that $(a(t))$ is a function such that $t\,\mathbb{P}(X > a(t)) \to 1$ as $t \to \infty$.

Corollary 2.33. *Assume the condition of Theorem 2.32. Let $(p(t))$ be an integer–valued function such that that $p(t) \to \infty$ as $t \to \infty$ and a growth condition is satisfied for every fixed $\gamma > 0$ and sufficiently large $t \geq t_0$:*

$$a(\lambda(t)p(t)) \geq \gamma\,\lambda(t)\,. \tag{2.38}$$

Then the following limit relation holds for i.i.d. copies S_i of the random sum process S:

$$N_{p(t)} = \sum_{i=1}^{p(t)} \varepsilon_{\frac{S_i(t)-m(t)}{a(\lambda(t)p(t))}} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{\Gamma_i^{-1/\alpha}}\,, \qquad t \to \infty\,,$$

where (Γ_i) is defined in Theorem 2.12.

Proof. In view of Proposition 2.3 the result is proved if we can show that as $t \to \infty$,

$$\begin{aligned} p(t)\,\mathbb{P}((a(\lambda(t)p(t)))^{-1}\big(S(t)-m(t)\big) > x\big) &\sim \lambda(t)\,p(t)\,\mathbb{P}(X > a(\lambda(t)p(t))\,x) \to x^{-\alpha}\,,\\ p(t)\,\mathbb{P}((a(\lambda(t)p(t)))^{-1}\big(S(t)-m(t)\big) < -x\big) &\to 0\,, \qquad x > 0\,. \end{aligned}$$

But this follows by an application of Theorem 2.32 in combination with (2.38) and the regular variation of X. □

Remark 2.34. Since $a(\lambda(t)p(t)) = (\lambda(t)p(t))^{1/\alpha}\ell(\lambda(t)p(t))$ for a slowly varying function ℓ and $\ell(x) \geq x^{-\epsilon/\alpha}$ for any small $\epsilon > 0$ and sufficiently large x, (2.38) holds if $p(t) \geq (\lambda(t))^{\alpha-1+\epsilon'}$ for any choice of $\epsilon' > 0$.

2.4.4 An extension: the index of the point process is random

Let $(P_n)_{n\geq 0}$ be a sequence of positive integer–valued random variables. We assume that there exists a sequence of positive numbers (p_n) such that $p_n \to \infty$ and

$$\frac{P_n}{p_n} \xrightarrow{\mathbb{P}} 1\,, \qquad n \to \infty\,. \tag{2.39}$$

This condition is satisfied for wide classes of integer–valued sequences (P_n), including the renewal counting processes and (inhomogeneous) Poisson processes when calculated at the positive integers. In particular, for renewal processes $p_n \sim c\,n$ provided the inter–arrival times have finite expectation.

We have the following analog of Proposition 2.3.

Proposition 2.35. *Let $(X_{ni})_{n=1,2,\ldots;i=1,2,\ldots}$ be a triangular array of i.i.d. random variables assuming values in some state space $E \subset \mathbb{R}^d$ equipped with the Borel σ–field $\mathcal{E}$. Let μ be a Radon measure on $\mathcal{E}$. If the relation*

$$p_n\,\mathbb{P}\big(X_{n1} \in \cdot\big) \xrightarrow{v} \mu(\cdot)\,, \qquad n \to \infty\,, \tag{2.40}$$

holds on E then

$$\widetilde{N}_p = \sum_{i=1}^{P_n} \varepsilon_{X_{ni}} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,,$$

where N is PRM(μ) *on E.*

Proof. We prove the result by showing convergence of the Laplace functionals. The arguments of a Laplace functional are elements of

$$C_K^+(E) = \{g : E \to \mathbb{R}_+ : g \text{ continuous with compact support}\}\,.$$

For $f \in C_K^+$ we have by independence of the (X_{ni}),

$$\mathbb{E}\Big[\exp\Big(-\int_E f\,d\widetilde{N}_p\Big)\Big] \;=\; \mathbb{E}\Big[\exp\Big(-\sum_{j=1}^{P_n} f(X_{nj})\Big)\Big] = \mathbb{E}\Big[\Big(\mathbb{E}\big[\exp(-f(X_{n1}))\big]\Big)^{P_n}\Big]\,.$$

In view of (2.39) there is a real sequence $\epsilon_n \downarrow 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\big(|P_n/p_n - 1| > \epsilon_n\big) = \mathbb{P}(A_n^c) = 0\,. \tag{2.41}$$

Then

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\mathbb{E}\big[\exp(-f(X_{n1}))\big]\Big)^{P_n}\Big] \\
&\quad= \mathbb{E}\Big[\Big(\mathbb{E}\big[\exp(-f(X_{n1}))\big]\Big)^{P_n}\big(\mathbf{1}(A_n^c) + \mathbf{1}(A_n)\big)\Big] \\
&\quad= I_1 + I_2\,.
\end{aligned}$$

By (2.41) we have $I_1 \le \mathbb{P}(A_n^c) \to 0$ as $n \to \infty$ while

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\mathbb{E}\big[\exp(-f(X_{n1}))\big]\Big)^{(1+\epsilon_n)p_n}\mathbf{1}(A_n)\Big] \\
&\quad\le\; I_2 \le \mathbb{E}\Big[\Big(\mathbb{E}\big[\exp(-f(X_{n1}))\big]\Big)^{(1-\epsilon_n)p_n}\mathbf{1}(A_n)\Big]\,.
\end{aligned} \tag{2.42}$$

In view of Proposition 2.3 and (2.40)

$$\Big(\mathbb{E}\big[\exp(-f(X_{n1}))\big]\Big)^{(1\pm\epsilon_n)p_n} \to \exp\Big(-\int_E (1-\mathrm{e}^{-f(\boldsymbol{x})})\,\mu(d\boldsymbol{x})\Big)\,.$$

The right–hand side is the Laplace functional of a PRM(μ). Now an application of dominated convergence to I_2 in (2.42) yields the desired convergence result. □

An immediate consequence of this result is that all point process convergences in Section 2.4 remain valid if the point processes N_p are replaced by their corresponding analogs $\widetilde{N}_p$ with a random index sequence (P_n) independent of (S_{ni}) and satisfying (2.39). Moreover, the growth rates for $p_n \to \infty$ remain the same.

2.4.5 Extension to the tail empirical process

We assume that (S_{ni}) are i.i.d. copies of a real–valued random walk (S_n). Instead of the point processes considered in the previous sections one can also study the tail empirical process

$$N_p = \frac{1}{k}\sum_{i=1}^{p} \varepsilon_{c_{[p/k]}^{-1}(S_{ni}/\sqrt{n} - d_{[p/k]})}$$

where $k = k_n \to \infty$, $p = p_n \to \infty$ and $p_n/k_n \to \infty$, and (c_n) and (d_n) are suitable normalizing and centering constants. To illustrate the theory we consider two examples.

Example 2.36. Assume the conditions and notation of Theorem 2.12. In this case, choose $c_n = 1/d_n$. Then

$$\begin{aligned} \mathbb{E}[N_p(x,\infty)] &= \frac{p}{k}\,\mathbb{P}\big(S_n/\sqrt{n} > d_{[p/k]} + x/d_{[p/k]}\big) \to \mathrm{e}^{-x}\,, \\ \mathrm{var}\big(N_p(x,\infty)\big) &\le \frac{p}{k^2}\,\mathbb{P}\big(S_n/\sqrt{n} > d_{[p/k]} + x/d_{[p/k]}\big) \to 0\,, \qquad x \in \mathbb{R}, \qquad n \to \infty\,, \end{aligned}$$

provided $p/k < \exp(\gamma_n^2/2)$. It is not difficult to see that

$$N_p \stackrel{\mathbb{P}}{\to} -\log \Lambda\,.$$

Similarly, assume the conditions and the notation of Theorem 2.25 and consider

$$N_p = \frac{1}{k}\sum_{i=1}^{p} \varepsilon_{a^{-1}_{[np/k]}(S_{ni}-d_n)}\,.$$

Then for $x > 0$ as $n \to \infty$,

$$\begin{aligned} \mathbb{E}[N_p(x,\infty)] &= \frac{p}{k}\,\mathbb{P}\big(a^{-1}_{[np/k]}(S_n - d_n) > x\big) \sim \frac{np}{k}\,\mathbb{P}(X > a_{[np/k]}\,x) \\ &\to p_+\,x^{-\alpha} = \mu_\alpha(x,\infty)\,, \\ \mathrm{var}(N_p(x,\infty)) &\to 0\,, \\ \mathbb{E}[N_p(-\infty,-x]] &= \frac{p}{k}\,\mathbb{P}\big(a^{-1}_{[np/k]}(S_n - d_n) \le -x\big) \to p_-\,x^{-\alpha} = \mu_\alpha(-\infty,-x]\,, \\ \mathrm{var}(N_p(-\infty,-x]) &\to 0\,, \end{aligned}$$

provided the modified sequence $p_n/k_n \to \infty$ satisfies the conditions imposed on (p_n) in Theorem 2.25. We notice that the values of μ_α on $(-\infty,-x]$ and (x,∞) determine a Radon measure on $\mathbb{R}\backslash\{0\}$. From these relations we conclude that $N_p \stackrel{\mathbb{P}}{\to} \mu_\alpha$. Then, following the lines of Resnick and Stărică [101, Proposition 2.3], one can for example prove consistency of the Hill estimator based on the sample $(S_{ni})_{i=1,\ldots,p}$. Assuming for simplicity $d_n = 0$, $p_+ > 0$, we write $S_{n,(1)} \ge \cdots \ge S_{n,(k)}$ for the k largest values. Then

$$\frac{1}{k}\sum_{i=1}^{k} \log \frac{S_{n,(i)}}{S_{n,(k)}} \stackrel{\mathbb{P}}{\to} \frac{1}{\alpha}\,.$$

2.4.6 Some related results

The largest values of sequences of i.i.d. normalized and centered partial sum processes play a role in the context of random matrix theory which is also the main motivation for the present work. Consider a double array (X_{it}) of i.i.d. regularly varying random variables with index $\alpha \in (0,4)$ (see (2.2)) and generic element X, and also assume that $\mathbb{E}[X] = 0$ if this expectation is finite. Consider the data matrix

$$\boldsymbol{X} := \boldsymbol{X}_n = (X_{it})_{i=1,\ldots,p;t=1,\ldots,n}$$

and the corresponding sample covariance matrix $\boldsymbol{X}\boldsymbol{X}^\top = (S_{ij})$. Heiny and Mikosch [60] proved that

$$a_{np}^{-2}\|\boldsymbol{X}\boldsymbol{X}^\top - \mathrm{diag}(\boldsymbol{X}\boldsymbol{X}^\top)\|_2 \stackrel{\mathbb{P}}{\to} 0\,, \qquad n \to \infty\,, \tag{2.43}$$

where $\|\boldsymbol{A}\|_2$ denotes the spectral norm of a $p \times p$ symmetric matrix $\boldsymbol{A}$, $\mathrm{diag}(\boldsymbol{A})$ consists of the diagonal of $\boldsymbol{A}$, (a_k) is any sequence satisfying $k\,\mathbb{P}(|X| > a_k) \to 1$ as $k \to \infty$, and $p_n = n^\beta \ell(n)$ for some $\beta \in (0, 1]$ and a slowly varying function ℓ. Write $\lambda_{(1)}(\boldsymbol{A}) \geq \cdots \geq \lambda_{(p)}(\boldsymbol{A})$ for the ordered eigenvalues of $\boldsymbol{A}$. According to Weyl's inequality (see Bhatia [17]), the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$ satisfy the relation

$$a_{np}^{-2} \sup_{i=1,\ldots,p} \left|\lambda_{(i)}(\boldsymbol{X}\boldsymbol{X}^\top) - \lambda_{(i)}(\mathrm{diag}(\boldsymbol{X}\boldsymbol{X}^\top))\right| \leq a_{np}^{-2} \|\boldsymbol{X}\boldsymbol{X}^\top - \mathrm{diag}(\boldsymbol{X}\boldsymbol{X}^\top)\|_2 \xrightarrow{\mathbb{P}} 0\,. \quad (2.44)$$

But of course, $\lambda_{(i)}(\mathrm{diag}(\boldsymbol{X}\boldsymbol{X}^\top))$ are the ordered values of the i.i.d. partial sums $S_{ii} = \sum_{t=1}^n X_{it}^2$, $i = 1, \ldots, p$. In view of (2.44) the asymptotic theory for the largest eigenvalues of the normalized sample covariance matrix $a_{np}^{-2}\boldsymbol{X}\boldsymbol{X}^\top$ (which also needs centering for $\alpha \in (2, 4)$) are determined through the Fréchet convergence of the processes with points $(a_{np}^{-2} S_{ii})_{i=1,\ldots,p}$. Moreover, (2.44) implies the Fréchet convergence of the point processes of the normalized and centered eigenvalues of the sample covariance matrix.

The large deviation approach also works for proving limit theory for the point process of the off–diagonal elements of $\boldsymbol{X}\boldsymbol{X}^\top$ provided X has sufficiently high moments. Heiny et al. [61] prove Gumbel convergence for the point process of the off–diagonal elements $(S_{ij})_{1 \leq i < j \leq p}$. The situation is more complicated because the points S_{ij} are typically dependent. Multivariate extensions of the normal large deviation approximation $0.5p^2\mathbb{P}(d_{p^2/2}(S_{12} - d_{p^2/2}) > x) \to \exp(-x)$ show that the point process of the standardized (S_{ij}) has the same limit Poisson process as if the S_{ij} were independent. Moreover, [61] show that the point process of the diagonal elements (S_{ii}) (under suitable conditions on the rate of $p_n \to \infty$ and under $\mathbb{E}[|X|^s] < \infty$ for $s > 4$) converges to $\mathrm{PRM}(-\log \Lambda)$. This result indicates that the off-diagonal and diagonal entries of $\boldsymbol{X}\boldsymbol{X}^\top$ exhibit very similar extremal behavior. This is in stark contrast to the aforementioned results in [60] where the diagonal entries have Fréchet extremal behavior.

Related results can also be found in Gantert and Höfelsauer [52] who consider real-valued branching random walks and prove a large deviation principle for the position of the right-most particle; see Theorem 3.2 in [52]. The position of the right-most particle is the maximum of a collection of a random number of dependent random walks. In this context, the authors also prove a related large deviation result under the assumption that the considered random walks are i.i.d. . They show that the maximum of these i.i.d. random walks stochastically dominates the maximum of the branching random walks; see Theorem 3.1 and Lemma 5.2 in [52]. An early comparison between maxima of branching and i.i.d. random walks was provided by Durrett [43].

Chapter 3

Point process convergence for the off–diagonal entries of sample covariance matrices

Johannes Heiny, Thomas Mikosch & Jorge Yslas
To appear in Annals of Applied Probability

Abstract

We study point process convergence for sequences of i.i.d. random walks. The objective is to derive asymptotic theory for the extremes of these random walks. We show convergence of the maximum random walk to the Gumbel distribution under the existence of a $(2+\delta)$th moment. We make heavily use of precise large deviation results for sums of i.i.d. random variables. As a consequence, we derive the joint convergence of the off–diagonal entries in sample covariance and correlation matrices of a high–dimensional sample whose dimension increases with the sample size. This generalizes known results on the asymptotic Gumbel property of the largest entry.

Keywords: Gumbel distribution, extreme value theory, maximum entry, sample covariance matrix

3.1 Introduction

3.1.1 Motivation

An accurate probabilistic understanding of covariances and correlations is often the backbone of a thorough statistical data analysis. In many contemporary applications, one is faced with large data sets where both the dimension of the observations and the sample size are large. A major reason lies in the rapid improvement of computing power and data collection devices which has triggered the necessity to study and interpret the sometimes overwhelming amounts of data in an efficient and tractable way. Huge data sets arise naturally in genome sequence data in biology, online networks, wireless communication, large financial portfolios, and natural sciences. More applications where the dimension p might be of the same or even higher magnitude than the sample size n are discussed in [42, 68]. In such a high–dimensional setting, one faces new probabilistic and statistical challenges; see [69] for a review. The sample (auto)covariance matrices will typically be misleading [13, 45]. Even in the null case, i.e., when the components of the time series are i.i.d., it is well–known that the sample covariance matrix poorly estimates the population covariance matrix. The fluctuations of the off–diagonal entries of the sample covariance matrix aggregate, creating an estimation bias which is quantified by the famous Marčenko–Pastur theorem [80]. This paper provides insight into the joint behavior of the off–diagonal entries with a particular focus on their extremes.

Aside from the high dimension, the marginal distributions of the components present another major challenge for an accurate assessment of the dependence. In the literature, one typically assumes a finite fourth moment since otherwise the largest eigenvalue of the sample covariance matrix would tend to infinity when n and p increase. This moment assumption, however, excludes heavy–tailed time series from the analysis. The theory for the eigenvalues and eigenvectors of the sample autocovariance matrices stemming from such time series is quite different from the classical Marčenko–Pastur theory which applies in the light–tailed case. For detailed discussions about classical random matrix theory, we refer to the monographs [13, 110], while the developments in the heavy–tailed case can be found in [12, 39, 37, 60, 107, 108] and the references therein. For applications of extreme value statistics in finance and physics we refer to [29, 51].

In this paper, we study point process convergence for sequences of i.i.d. random walks. We then apply our results to derive the joint asymptotic behavior of the off–diagonal entries of sample covariance and correlation matrices. Based on this joint convergence we propose new independence tests in high dimensions.

3.1.2 The model

We are given p–dimensional random vectors $\mathbf{x}_t = (X_{1t}, \ldots, X_{pt})^\top$, $t = 1, \ldots, n$, whose components $(X_{it})_{i,t\geq 1}$ satisfy the following *standard conditions*:

- (X_{it}) are independent and identically distributed (i.i.d.) random variables with generic element X.
- $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$.

The dimension $p = p_n$ is some integer sequence tending to infinity as $n \to \infty$.

We are interested in the (non–normalized) $p \times p$ sample covariance matrix $\mathbf{S}$ and the sample correlation matrix $\boldsymbol{R}$,

$$\mathbf{S} = \mathbf{S}_n = \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^\top \quad \text{and} \quad \boldsymbol{R} = \boldsymbol{R}_n = (\mathrm{diag}(\mathbf{S}))^{-1/2} \mathbf{S} (\mathrm{diag}(\mathbf{S}))^{-1/2} \tag{3.1}$$

with entries

$$S_{ij} = \sum_{t=1}^{n} X_{it} X_{jt} \quad \text{and} \quad R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}\,, \quad i, j = 1, \ldots, p\,, \tag{3.2}$$

respectively. The dependence on n is often suppressed in our notation.

Our goal is to prove limit theory for the point processes of scaled and centered points (S_{ij}), (R_{ij}). Asymptotic theory for the extremes of these points can be deduced from the limit point process.

3.1.3 State–of–the–art

In the literature, the largest off–diagonal entry of a sample covariance or correlation matrix has been studied, but results concerning the joint behavior of the entries are currently lacking. The theoretical developments are mainly due to Jiang. In [66], he analyzed the asymptotic distributions of

$$W_n := \frac{1}{n} \max_{1 \le i < j \le p} |S_{ij}| \quad \text{and} \quad L_n := \max_{1 \le i < j \le p} |R_{ij}|$$

under the assumption $p/n \to \gamma \in (0, \infty)$. If $\mathbb{E}[|X|^{30+\delta}] < \infty$ for some $\delta > 0$, he proved that

$$\lim_{n \to \infty} \mathbb{P}(n W_n^2 - 4 \log p + \log \log p \le x) = \exp\Big(- \tfrac{1}{\sqrt{8\pi}}\, \mathrm{e}^{-x/2} \Big)\,, \quad x \in \mathbb{R}\,, \tag{3.3}$$

$$\lim_{n \to \infty} \mathbb{P}(n L_n^2 - 4 \log p + \log \log p \le x) = \exp\Big(- \tfrac{1}{\sqrt{8\pi}}\, \mathrm{e}^{-x/2} \Big)\,, \quad x \in \mathbb{R}\,. \tag{3.4}$$

The limiting law is a non–standard Gumbel distribution. Under the same assumptions Jiang [66] also derived the limits

$$\lim_{n \to \infty} \sqrt{\frac{n}{\log p}}\, L_n = 2 = \lim_{n \to \infty} \sqrt{\frac{n}{\log p}}\, W_n \qquad \text{a.s.} \tag{3.5}$$

Several authors managed to relax Jiang's moment condition while keeping the proportionality of p and n. Zhou [111] showed that (3.4) holds if $n^6 \mathbb{P}(|X_{11} X_{12}| > n) \to 0$ as $n \to \infty$. A sufficient condition is $\mathbb{E}[X^6] < \infty$. The papers [75, 76, 77] provide refinements of Zhou's condition. We summarize the distributional assumptions on X for the validity of (3.4) and (3.3) under proportionality of dimension p and sample size n as follows: $\mathbb{E}[X^6] < \infty$ is sufficient, and $\mathbb{E}[|X|^{6-\delta}] < \infty$ for any $\delta > 0$ is necessary. In that sense, finiteness of the sixth moment is the optimal moment assumption.

Interestingly, the optimal moment requirement also depends on the growth of p if p increases at a different rate than n. For the largest off–diagonal entry of the sample correlation matrix, also known as coherence of the random matrix $\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the interplay between dimension and moments was addressed in [79]. If $\mathbb{E}[|X|^s] < \infty$ for $s > 2$ and

$$c_1 n^{(s-2)/4} \le p \le c_2 n^{(s-2)/4}$$

with positive constants c_1, c_2, Theorem 1.1 in [79] shows that (3.4) still applies. Note that, for proportional p and n, this result requires the finiteness of the sixth moment. The larger p relative to n, the more moments of X are needed. If the moment generating function of $|X|$ exists in some neighborhood of zero, (3.4) holds for $p = O\big(\exp(n^\beta)\big)$ for certain $\beta \in (0, 1/3)$; see [33]. Finally, if $(\log p)/n \not\to 0$, various phase transitions appear in the limit distribution of L_n. These were explored in [34] under convenient assumptions on X which yield an explicit formula for the density of R_{12}.

3.1.4 Objective and structure of this paper

Our main objective is to prove limit theory for the point processes of scaled and centered points (S_{ij}), (R_{ij}) in a more general framework than used for the results above. By a continuous mapping argument, the joint asymptotic distribution of functionals of a fixed number of points can easily be deduced from the limit process. In particular, we obtain the asymptotic distribution of the largest and smallest entries.

First, we establish our result for $\mathbf{S}$. Since each S_{ij} is a sum of i.i.d. random variables, we prove a useful large deviation theorem which exploits the asymptotic normal distribution of S_{ij}. Aside from finding suitable assumptions on X, the main challenge is that the S_{ij} are not independent. It turns out that despite their non–trivial dependence, the maximum behaves like the maximum of i.i.d. copies. Therefore we will first solve the problem for i.i.d. random walk points $(S_n^{(i)})$ instead of (S_{ij}). This is done in Section 3.2.

We continue in Section 3.3 with the main results of the paper. Here we derive asymptotic theory for the point processes

$$\widetilde{N}_n = \sum_{1 \le i < j \le p} \varepsilon_{\widetilde{d}_p(S_{ij}/\sqrt{n} - \widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,,$$

for suitable constants $(\widetilde{d}_p)$ and limit Poisson random measure N with mean measure μ on $\mathbb{R}$ such that $\mu(x, \infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$. Throughout this paper ε_x denotes the Dirac measure at x. A continuous mapping theorem implies distributional convergence of finitely many S_{ij}. In particular, the maximum entry S_{ij} converges to the Gumbel distribution provided X has suitably many finite moments. A related result holds with S_{ij} replaced by the corresponding sample correlations $R_{ij} = S_{ij}/\sqrt{S_{ii}S_{jj}}$. In Section 3.4, we extend our results to hypercubic random matrices of the form $\sum_{t=1}^n \mathbf{x}_t \otimes \cdots \otimes \mathbf{x}_t$ and we briefly discuss some statistical applications. The proofs of the main results are presented in Sections 3.5 and 3.6.

3.2 Normal approximation to large deviation probabilities

In this section we collect some precise large deviation results for sums of independent random variables. Throughout this section, (X_i) is an i.i.d. sequence of mean zero, unit variance random variables with generic element X, distribution F and right tail $\overline{F} = 1 - F$. We define the corresponding partial sum process

$$S_0 = 0\,, \qquad S_n = X_1 + \cdots + X_n\,, \qquad n \ge 1\,.$$

Consider i.i.d. copies $(S_n^{(i)})_{i \ge 1}$ of S_n. We also introduce an integer sequence (p_n) such that $p = p_n \to \infty$ as $n \to \infty$. We are interested in the limit behavior of the k largest values among $(S_n^{(i)})_{i=1,\ldots,p}$, in particular in the possible limit laws of the maximum

$\max_{i=1,\dots,p} S_n^{(i)}$. More generally, we are interested in the limit behavior of the point processes N_n,

$$N_n = \sum_{i=1}^{p} \varepsilon_{d_p(S_n^{(i)}/\sqrt{n} - d_p)} \xrightarrow{\mathrm{d}} N\,, \qquad n \to \infty\,, \tag{3.6}$$

toward a Poisson random measure N on $\mathbb{R}$ with mean measure μ given by $\mu(x,\infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$. The sequence (d_p) is chosen such that $p\,\overline{\Phi}(d_p) = p(1 - \Phi(d_p)) \to 1$ as $p \to \infty$ where Φ is the standard normal distribution function. In this paper, we work with

$$d_p = \sqrt{2\log p} - \frac{\log\log p + \log 4\pi}{2(2\log p)^{1/2}}\,. \tag{3.7}$$

A motivation for this choice is that for an i.i.d. sequence (X_i) with distribution function Φ we have

$$\lim_{n\to\infty} \mathbb{P}\Big(d_p\big(\max_{i=1,\dots,p} X_i - d_p\big) \le x\Big) = \exp(-\mathrm{e}^{-x}) = \Lambda(x)\,, \qquad x \in \mathbb{R}\,.$$

The limit distribution function is the standard Gumbel Λ; see Embrechts et al. [46, Example 3.3.29].

By [102, Theorem 5.3], relation (3.6) is equivalent to the following limit relation for the tails

$$p\,\mathbb{P}\big(d_p\,(S_n/\sqrt{n} - d_p) > x\big) \to \mathrm{e}^{-x}\,, \qquad n \to \infty\,, \qquad x \in \mathbb{R}\,, \tag{3.8}$$

and also to convergence of the maximum of the random walks $(S_n^{(i)})_{i=1,\dots,p}$ to the Gumbel distribution:

$$\lim_{n\to\infty} \mathbb{P}\Big(\max_{i=1,\dots,p} d_p(S_n^{(i)}/\sqrt{n} - d_p) \le x\Big) = \Lambda(x)\,, \qquad x \in \mathbb{R}\,. \tag{3.9}$$

Equations (3.8) and (3.9) involve precise large deviation probabilities for the random walk (S_n). To state some results which are relevant in this context, we assume one of the following three moment conditions:

(C1) There exists $s > 2$ such that $\mathbb{E}[|X|^s] < \infty$.

(C2) There exists an increasing differentiable function g on $(0,\infty)$ such that $\mathbb{E}[\exp(g(|X|))] < \infty$, $g'(x) \le \tau g(x)/x$ for sufficiently large x and some $\tau < 1$, and $\lim_{x\to\infty} g(x)/\log x = \infty$.

(C3) There exists a constant $h > 0$ such that $\mathbb{E}[\exp(h\,|X|)] < \infty$.

Note that the conditions (C1)–(C3) are increasing in strength. One has the implications (C3) $\Rightarrow$ (C2) $\Rightarrow$ (C1). The following result explains the connection between the rate of $p_n \to \infty$ in (3.8) and the conditions (C1)–(C3) on the distribution of X.

Theorem 3.1. *Assume the standard conditions on (X_i) and that $p = p_n \to \infty$ satisfies*

- $p = O(n^{(s-2)/2})$ *if* (C1) *holds.*
- $p = \exp(o(g_n^2 \wedge n^{1/3}))$ *where g_n is the solution of the equation $g_n^2 = g(g_n\sqrt{n})$, if* (C2) *holds.*

- $p = \exp(o(n^{1/3}))$ *if* (C3) *holds.*

Then we have

$$p\,\mathbb{P}(S_n/\sqrt{n} > d_p + x/d_p) \sim p\,\overline{\Phi}(d_p + x/d_p) \to \mathrm{e}^{-x}, \quad n \to \infty, \quad x \in \mathbb{R}\,. \tag{3.10}$$

Remark 3.2. The proofs of these results follow from the definition of d_p and precise large deviation bounds of the type

$$\sup_{0 \le y \le \gamma_n} \Big| \frac{\mathbb{P}(S_n/\sqrt{n} > y)}{\overline{\Phi}(y)} - 1 \Big| \to 0\,, \quad n \to \infty\,, \tag{3.11}$$

where $\gamma_n \to \infty$ are sequences depending on the conditions (C1)–(C3). If (C3) holds, one can choose $\gamma_n = o(n^{1/6})$ implying the growth rate $p = \exp(o(n^{1/3}))$. This follows from Petrov's large deviation result [98, Theorem VIII.2]. Under (C2) one can choose $\gamma_n = o(n^{1/6} \wedge g_n)$ implying the growth rate $p = \exp(o(g_n^2 \wedge n^{1/3}))$. This follows from S.V. Nagaev's [87, Theorem 3]. Under (C1) he also derived $\gamma_n = \sqrt{(s/2-1)\log n}$ in [87, Theorem 4]. The best possible range under (C1) is $\gamma_n = \sqrt{(s-2)\log n}$; see Michel [83, Theorem 4].

The aforementioned large deviation results cannot be improved in general unless additional conditions are assumed. For example, under (C3) if the cumulants of X of order $k = 3, \ldots, r+2$ vanish then (3.10) holds for $p = \exp(o(n^{(r+1)/(r+3)}))$. This follows from the fact that one can choose $\gamma_n = o(n^{(r+1)/(2(r+3))})$; see [98, Theorem VIII.7]. In Section VIII.3 of [98] one also finds necessary and sufficient conditions for (3.11) to hold in certain intervals. As a matter of fact, such conditions are not of moment–type. Therefore one cannot expect that necessary and sufficient conditions for (3.10) for general sequences (p_n) can be expressed in terms of moments. There is however a clear relationship between possible rates of (p_n) and the existence of moments: the higher moments exist the larger we can choose (p_n), but the growth cannot be arbitrarily fast.

In passing we mention a sharp large deviation result for a sequence of i.i.d. regularly varying random variables (X_i) with tail index $\alpha > 2$, i.e., a generic element X has tails

$$\mathbb{P}(\pm X > x) \sim p_\pm\, \frac{L(x)}{x^\alpha}\,, \qquad x \to \infty\,, \tag{3.12}$$

where $p_+ + p_- = 1$ and L is slowly varying. Then, due to S.V. Nagaev's results in [90], one has (3.11) with $\gamma_n = \sqrt{c_1 \log n}$ for $c_1 < \alpha - 2$, while for $\xi_n = \sqrt{c_2 \log n}$ and any $c_2 > \alpha - 2$,

$$\sup_{y > \xi_n} \Big| \frac{\mathbb{P}(\pm S_n/\sqrt{n} > y)}{n\,\overline{F}(y\sqrt{n})} - p_\pm \Big| \to 0\,, \qquad n \to \infty\,. \tag{3.13}$$

There exists a small but increasing literature on precise large deviation results; we refer to [41, 104] and the references therein.

Now consider i.i.d. copies $(S_n^{(i)})_{i \ge 1}$ of (S_n). The following result is an immediate consequence of Theorem 3.1.

Theorem 3.3. *Assume the conditions of Theorem 3.1. Relation* (3.10) *is equivalent to either of the following two limit relations:*

$$\mathbb{P}\Big(d_p \max_{i=1,\ldots,p} (S_n^{(i)}/\sqrt{n} - d_p) \le x \Big) \to \Lambda(x)\,, \qquad x \in \mathbb{R}\,, \qquad n \to \infty\,. \tag{3.14}$$

and

$$N_n = \sum_{i=1}^{p} \varepsilon_{d_p\,(S_n^{(i)}/\sqrt{n} - d_p)} \xrightarrow{\mathrm{d}} N = \sum_{i=1}^{\infty} \varepsilon_{-\log \Gamma_i}\,, \qquad n \to \infty\,. \tag{3.15}$$

where $\Gamma_i = E_1 + \cdots + E_i$, $i \geq 1$, *and* (E_i) *is i.i.d. standard exponential, i.e.,* N *is a Poisson random measure with mean measure* $\mu(x, \infty) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$.

Proof. Following Resnick [102, Theorem 5.3], (3.15) and (3.10) are equivalent. Moreover, a continuous mapping argument implies that, if $N_n \xrightarrow{\mathrm{d}} N$, then

$$\begin{aligned} \mathbb{P}(N_n(x, \infty) = 0) &= \mathbb{P}\Big(d_p \max_{i=1,\ldots,p} (S_n^{(i)}/\sqrt{n} - d_p) \leq x\Big) \\ &\to \mathbb{P}(N(x, \infty) = 0) = \mathbb{P}(-\log \Gamma_1 \leq x) = \exp(-\mathrm{e}^{-x})\,. \end{aligned} \tag{3.16}$$

Moreover, if (3.14) holds a Taylor expansion argument shows that

$$\begin{aligned} \mathbb{P}\Big(d_p \max_{i=1,\ldots,p} (S_n^{(i)}/\sqrt{n} - d_p) \leq x\Big) &= \Big(1 - \frac{p\,\mathbb{P}(S_n/\sqrt{n} > d_p + x/d_p)}{p}\Big)^p \\ &\to \exp(-\mathrm{e}^{-x})\,, \qquad n \to \infty\,, \end{aligned}$$

holds if and only if (3.10) does. □

This means that in case of i.i.d. points $(S_n^{(i)})$ the convergence of the maximum is equivalent to the convergence of the point processes (N_n). In general, the latter is a stronger statement. If (N_n) converges, the distribution of the maximum can always be recovered using (3.16).

3.3 Main results

3.3.1 Point process convergence of a sample covariance matrix

We consider the sample covariance matrix $\mathbf{S} = (S_{ij})_{i,j=1,\ldots,p}$ introduced in Section 3.1.2. The problem of showing limit theory for the associated point process is similar to Theorem 3.3 for i.i.d. random walks $(S_n^{(i)})$. In contrast to the i.i.d. copies $(S_n^{(i)})$ in Section 3.2 here we deal with $p(p-1)/2$ *dependent* off–diagonal entries of $\mathbf{S}$. Nevertheless, Theorem 3.1 will again be a main tool for proving these results.

Since the summands of S_{ij} are i.i.d. products $X_{it}X_{jt}$ we need to adjust the conditions (C2) and (C3) to this situation while (C1) remains unchanged.

(C2') There exists an increasing differentiable function g on $(0, \infty)$ such that $\mathbb{E}[\exp(g(|X_{11}X_{12}|))] < \infty$, $g'(x) \leq \tau g(x)/x$ for sufficiently large x and some $\tau < 1$, and $\lim_{x\to\infty} g(x)/\log x = \infty$.

(C3') There exists a constant $h > 0$ such that $\mathbb{E}[\exp(h\,|X_{11}X_{12}|)] < \infty$.

Remark 3.4. By Lemma 3.20, (C3') implies (C3). The reverse implication is not true. For example, if X is standard exponential, which satisfies (C3), then $X_{11}X_{21}$ has Weibull-type tail with parameter 1/2; see [7]; which does not satisfy (C3'). By Lemma 3.20, (C2') implies $\mathbb{E}[\exp(g(|X|))] < \infty$.

Theorem 3.5. *Assume the standard conditions on* (X_{it}) *and that* $p = p_n \to \infty$ *satisfies:*

- $p = O(n^{(s-2)/4})$ *if* (C1) *holds.*
- $p = \exp(o(g_n^2 \wedge n^{1/3}))$, *where* g_n *is the solution of the equation* $g_n^2 = g(g_n\sqrt{n})$, *if* (C2') *holds.*
- $p = \exp(o(n^{1/3}))$ *if* (C3') *holds.*

Define $\widetilde{d}_p = d_{p(p-1)/2}$. *Then the following point process convergence holds:*

$$N_n^S := \sum_{1\le i<j\le p} \varepsilon_{\widetilde{d}_p(S_{ij}/\sqrt{n}-\widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,,$$

where N *is the Poisson random measure defined in* (3.15).

The proof is given in Section 3.6.1.

Some comments

- The point process convergence in Theorem 3.5 remains valid if the standard conditions on (X_{it}) are relaxed to the following two conditions:
  - The columns $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of the matrix $(X_{it})_{i=1,\ldots,p;t=1\ldots,n}$ are i.i.d..
  - The random variables $X_{11}, \ldots, X_{p1}$ are independent, with mean zero and unit variance, but they are not necessarily identically distributed.

  The proof is the same as that of Theorem 3.5. All results in Section 3.3 hold under these relaxed conditions. For clarity of presentation and proof, all statements are presented under the standard conditions.
- Theorem 3.5 can be extended by introducing additional time stamps:

$$\sum_{1\le i<j\le p} \varepsilon_{\left(\frac{(i,j)}{p},\widetilde{d}_p(S_{ij}/\sqrt{n}-\widetilde{d}_p)\right)} \xrightarrow{\mathrm{d}} \widetilde{N}\,, \qquad n\to\infty\,,$$

  on $\{(x_1,x_2) : 0 \le x_1 \le 1\,, x_1 \le x_2\} \times \mathbb{R}$ where $\widetilde{N}$ is a Poisson random measure with mean measure $\mathbb{LEB} \times \mu$. This follows for example by using the techniques of [103, Proposition 3.21].
- Under any of the moment conditions (C2'),(C3') one can choose $p \sim \gamma n$ for $\gamma > 0$ in Theorem 3.5. Under (C1), one needs the condition $\mathbb{E}[|X|^6] < \infty$ in order to guarantee $p = O(n)$. This is in agreement with the minimal moment requirement for the results on W_n (see (3.3)).

Next we consider the order statistics of S_{ij}, $1 \le i < j \le p$:

$$\min_{1\le i<j\le p} S_{ij} =: S_{(p(p-1)/2)} \le \cdots \le S_{(1)} := \max_{1\le i<j\le p} S_{ij}\,.$$

Theorem 3.5 implies the convergence of the largest and smallest off–diagonal entries of $\mathbf{S}$.

Corollary 3.6. *Under the conditions of Theorem 3.5 we have joint convergence of the upper and lower order statistics: for any $k \geq 1$,*

$$\widetilde{d}_p\big(S_{(i)}/\sqrt{n} - \widetilde{d}_p\big)_{i=1,\dots,k} \xrightarrow{\mathrm{d}} (-\log \Gamma_i)_{i=1,\dots,k}\,, \tag{3.17}$$

$$\widetilde{d}_p\big(S_{(i)}/\sqrt{n} + \widetilde{d}_p\big)_{i=p(p-1)/2,\dots,p(p-1)/2-k+1} \xrightarrow{\mathrm{d}} (\log \Gamma_i)_{i=1,\dots,k}\,. \tag{3.18}$$

Moreover, the properly normalized maxima and minima are asymptotically independent, that is for any $x, y \in \mathbb{R}$ we have as $n \to \infty$,

$$\mathbb{P}\Big(\widetilde{d}_p(S_{(1)}/\sqrt{n} - \widetilde{d}_p) \leq x\,, \widetilde{d}_p(S_{(p(p-1)/2)}/\sqrt{n} + \widetilde{d}_p) \leq y\Big) \to \Lambda(x)(1 - \Lambda(-y))\,. \tag{3.19}$$

Proof. Relation (3.17) is immediate from $N_n^S \xrightarrow{\mathrm{d}} N$ and the continuous mapping theorem. The same argument works for (3.18) if one observes that

$$\widetilde{d}_p\big(S_{(i)}/\sqrt{n} + \widetilde{d}_p\big)_{i=p(p-1)/2,\dots,p(p-1)/2-k+1} = -\widetilde{d}_p\big((-S)_{(i)}/\sqrt{n} - \widetilde{d}_p\big)_{i=1,\dots,k}$$

where $(-S)_{(i)}$ is the ordered sample of $(-S_{ij})$. An application of (3.17) with (S_{ij}) replaced by $(-S_{ij})$ then yields (3.18).

Now we consider joint convergence of the maximum and the minimum: for $x, y \in \mathbb{R}$,

$$\begin{aligned} G_n(x,y) &= \mathbb{P}\Big(\widetilde{d}_p(S_{(1)}/\sqrt{n} - \widetilde{d}_p) \leq x\,, \widetilde{d}_p(S_{(p(p-1)/2)}/\sqrt{n} + \widetilde{d}_p) > y\Big) \\ &= \mathbb{P}\Big(-\widetilde{d}_p + y/\widetilde{d}_p < S_{ij}/\sqrt{n} \leq \widetilde{d}_p + x/\widetilde{d}_p \text{ for all } 1 \leq i < j \leq p\Big) \\ &= 1 - \mathbb{P}\Big(\bigcup_{1 \leq i < j \leq p} \{S_{ij}/\sqrt{n} > \widetilde{d}_p + x/\widetilde{d}_p\} \cup \{-S_{ij}/\sqrt{n} \geq \widetilde{d}_p - y/\widetilde{d}_p\}\Big)\,. \end{aligned}$$

Writing

$$A_{ij} = \{S_{ij}/\sqrt{n} > \widetilde{d}_p + x/\widetilde{d}_p\} \cup \{-S_{ij}/\sqrt{n} \geq \widetilde{d}_p - y/\widetilde{d}_p\}\,,$$

one can use the same arguments used for establishing $\mathbb{P}(N_n^S(B) = 0) \to \mathbb{P}(N(B) = 0)$ in the proof of Theorem 3.5 to show that

$$G_n(x,y) \to \exp\big(-(\mathrm{e}^{\,y} + \mathrm{e}^{\,-x})\big) = \Lambda(x)\Lambda(-y)\,, \qquad n \to \infty\,.$$

Hence

$$\begin{aligned} &\mathbb{P}\Big(\widetilde{d}_p(S_{(1)}/\sqrt{n} - \widetilde{d}_p) \leq x\,, \widetilde{d}_p(S_{(p(p-1)/2)}/\sqrt{n} + \widetilde{d}_p) \leq y\Big) \\ &= \mathbb{P}\Big(\widetilde{d}_p(S_{(1)}/\sqrt{n} - \widetilde{d}_p) \leq x\Big) - G_n(x,y) \\ &\to \Lambda(x) - \Lambda(x)\Lambda(-y) = \Lambda(x)(1 - \Lambda(-y))\,, \qquad n \to \infty\,. \end{aligned}$$

□

Remark 3.7. An immediate consequence is

$$\frac{S_{(1)}}{\sqrt{n \log p}} \xrightarrow{\mathbb{P}} 2 \quad \text{and} \quad \frac{S_{(p(p-1)/2)}}{\sqrt{n \log p}} \xrightarrow{\mathbb{P}} -2\,.$$

Remark 3.8. If $\mathbb{E}[|X|^s] < \infty$ for some $s > 4$ and $\mathrm{var}(X^2) > 0$, we conclude from Theorem 3.3 that for $p = O(n^{(s-4)/4})$,

$$\sum_{i=1}^{p} \varepsilon_{d_p\left((S_{ii}-n)/\sqrt{n\mathrm{var}(X^2)}-d_p\right)} \xrightarrow{\mathrm{d}} N\,. \tag{3.20}$$

In particular, $\big(\max_{i=1,\ldots,p} d_p\big((S_{ii}-n)/\sqrt{n\mathrm{var}(X^2)}-d_p\big)\big)$ converges to a Gumbel distribution. We notice that $d_p \sim \sqrt{2\log p}$ while the normalizing and centering constants for $(S_{ij}/\sqrt{n})_{i\neq j}$, in (3.19) are $\widetilde{d}_p \sim 2\sqrt{\log p}$.

Moreover, while we still have Gumbel convergence for the maxima of the off–diagonal elements S_{ij} for suitable (p_n) if $\mathbb{E}[|X|^s] < \infty$ for some $s \in (2,4)$, the point process convergence in (3.20) cannot hold. Indeed, then an appeal to Nagaev's large deviation result (3.13) shows that, under the regular variation condition (3.12) on X with $\alpha \in (2,4)$,

$$\sum_{i=1}^{p} \varepsilon_{a_{np}^{-2}(S_{ii}-n)} \xrightarrow{\mathrm{d}} N\,,$$

where N is Poisson random measure on the state space $(0,\infty)$ with mean measure $\mu_\alpha(x,\infty) = x^{-\alpha/2}$, $x > 0$, and a_k satisfies $k\mathbb{P}(|X| > a_k) \to 1$ as $k \to \infty$. In particular, the maxima of (S_{ii}) converge toward a standard Fréchet distribution:

$$\mathbb{P}\Big(a_{np}^{-2} \max_{i=1,\ldots,p} (S_{ii}-n) \le x\Big) \to \Phi_{\alpha/2}(x) = \exp(-x^{-\alpha/2})\,, \qquad x > 0\,.$$

Assume (3.12) on X with $\alpha \in (2,4)$. If we construct a point process by choosing the normalization a_{np}^2 for the diagonal and off–diagonal entries, the contribution of the (S_{ij}) vanishes in the limit:

$$\sum_{i=1}^{p} \varepsilon_{a_{np}^{-2}(S_{ii}-n)} + \sum_{1\le i<j\le p} \varepsilon_{a_{np}^{-2}S_{ij}} \xrightarrow{\mathrm{d}} N\,.$$

It is also proved in Heiny and Mikosch [60] that the diagonal entries (S_{ii}) of the sample covariance matrix dominate the off–diagonal terms in operator norm, that is $\|\mathbf{S} - \mathrm{diag}(\mathbf{S})\|/\|\mathrm{diag}(\mathbf{S})\| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. In turn, the asymptotic behavior of the largest eigenvalues of the sample covariance matrix are determined by the corresponding largest values of (S_{ii}).

The techniques in this paper straightforwardly extend to other transformations of the points (S_{ij}). As an example, we provide one such result for the squares (S_{ij}^2).

Corollary 3.9. *Assume the conditions of Theorem 3.5. Then*

$$N_n^{S^2} = \sum_{1\le i<j\le p} \varepsilon_{0.5S_{ij}^2/n - 0.5\,\widetilde{d}_p^2 - \log 2}$$

converges to the Poisson random measure N described in Theorem 3.5.

Proof. One can follow the arguments in the proof of Theorem 3.5. In order to show condition (i), observe that for $x \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\big[N_n^{S^2}(x,\infty)\big] &= \frac{p(p-1)}{2}\,\mathbb{P}\Big(\frac{S_{12}^2}{2n} - \frac{\widetilde{d}_p^2}{2} - \log 2 > x\Big) \\
&= \frac{p(p-1)}{2}\,\mathbb{P}\Big(\Big|\frac{S_{12}}{\sqrt{n}}\Big| > \sqrt{2(x+\log 2) + \widetilde{d}_p^2}\Big) \\
&\sim p^2\,\overline{\Phi}\Big(\sqrt{2(x+\log 2) + \widetilde{d}_p^2}\Big) \to \mathrm{e}^{-x}\,, \qquad n \to \infty\,.
\end{aligned}$$

□

3.3.2 Point process convergence of a sample correlation matrix

Based on Theorem 3.5 we can also derive point process convergence for the sample correlation matrix $\boldsymbol{R} = (R_{ij})_{i,j=1,\dots,p}$ defined in (3.1) and (3.2).

Theorem 3.10. *Assume the standard conditions on (X_{it}) and that $p = p_n \to \infty$ satisfies:*

- $p = O(n^{(s-2)/4})$ *if* (C1) *holds.*
- $p = \exp(o(g_n^2 \wedge n^{1/3}))$ *where g_n is the solution of the equation $g_n^2 = g(g_n\sqrt{n})$ if* (C2') *holds.*
- $p = \exp(o(n^{1/3}))$, *if* (C3') holds.

Then the following point process convergence holds,

$$N_n^R := \sum_{1\le i<j\le p} \varepsilon_{\widetilde{d}_p(\sqrt{n}R_{ij}-\widetilde{d}_p)} \xrightarrow{\mathrm{d}} N\,,$$

where N is the Poisson random measure defined in (3.15).

The proof is given in Section 3.6.3.
The results for the order statistics of R_{ij}, $1 \le i < j \le p$:

$$\min_{1\le i<j\le p} R_{ij} =: R_{(p(p-1)/2)} \le \cdots \le R_{(1)} := \max_{1\le i<j\le p} R_{ij}\,,$$

carry over from those for the order statistics of (S_{ij}).

Corollary 3.11. *Under the conditions of Theorem 3.10 we have joint convergence of the upper and lower order statistics: for any $k \ge 1$,*

$$\begin{aligned} \widetilde{d}_p\big(\sqrt{n}R_{(i)} - \widetilde{d}_p\big)_{i=1,\dots,k} &\xrightarrow{\mathrm{d}} (-\log\Gamma_i)_{i=1,\dots,k}\,,\\ \widetilde{d}_p\big(\sqrt{n}R_{(i)} + \widetilde{d}_p\big)_{i=p(p-1)/2,\dots,p(p-1)/2-k+1} &\xrightarrow{\mathrm{d}} (\log\Gamma_i)_{i=1,\dots,k}\,. \end{aligned}$$

Moreover, for any $x, y \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\widetilde{d}_p(\sqrt{n}R_{(1)} - \widetilde{d}_p) \le x\,, \widetilde{d}_p(\sqrt{n}R_{(p(p-1)/2)} + \widetilde{d}_p) \le y\Big) = \Lambda(x)(1-\Lambda(-y))\,.$$

and

$$\sqrt{\frac{n}{\log p}}R_{(1)} \xrightarrow{\mathbb{P}} 2 \quad \textit{and} \quad \sqrt{\frac{n}{\log p}}R_{(p(p-1)/2)} \xrightarrow{\mathbb{P}} -2\,.$$

3.4 Extensions and applications

3.4.1 Extensions

In this section, we extend our results for the point processes constructed from the off–diagonal entries of the sample covariance matrices $\mathbf{S}_n = \sum_{t=1}^n \mathbf{x}_t\mathbf{x}_t^\top$, where $\mathbf{x}_t$ are the

p–dimensional columns of the data matrix $\mathbf{X}$. We introduce the hypercubic random matrices (or tensors) of order m:

$$\mathbf{S}^{(m)} = \mathbf{S}_n^{(m)} = \sum_{t=1}^{n} \underbrace{\mathbf{x}_t \otimes \cdots \otimes \mathbf{x}_t}_{m \text{ times}}, \qquad m \in \mathbb{N}, \qquad n \geq 1. \tag{3.21}$$

with entries

$$S^{(m)}_{i_1,\ldots,i_m} = \Big(\sum_{t=1}^{n} X_{i_1 t} X_{i_2 t} \cdots X_{i_m t} \Big), \qquad 1 \leq i_1, \ldots, i_m \leq p.$$

It is easy to see that $\mathbf{S}^{(2)} = \mathbf{S}$ arises as a special case.

Next, we generalize the moment conditions (C2') and (C3') to the m–fold product $X_{11} \cdots X_{1m}$.

(C2(m)) There exists an increasing differentiable function g on $(0, \infty)$ such that $\mathbb{E}[\exp(g(|X_{11} \cdots X_{1m}|))] < \infty$, $g'(x) \leq \tau g(x)/x$ for sufficiently large x and some $\tau < 1$, and $\lim_{x\to\infty} g(x)/\log x = \infty$.

(C3(m)) There exists a constant $h > 0$ such that $\mathbb{E}[\exp(h\,|X_{11} \cdots X_{1m}|)] < \infty$.

The following result extends Theorem 3.5 to hypercubic matrices of order m.

Theorem 3.12. *Let $m \in \mathbb{N}$ and define $d_{p,m} = d_{\binom{p}{m}}$. Assume the standard conditions on (X_{it}) and that $p = p_n \to \infty$ satisfies:*

- *$p = O(n^{(s-2)/4})$ if (C1) holds.*
- *$p = \exp(o(g_n^2 \wedge n^{1/3}))$, where g_n is the solution of the equation $g_n^2 = g(g_n\sqrt{n})$, if (C2(m)) holds.*
- *$p = \exp(o(n^{1/3}))$ if (C3(m)) holds.*

Then the following point process convergence holds:

$$N_n^{(m)} = \sum_{1 \leq i_1 < \cdots < i_m \leq p} \varepsilon_{d_{p,m}(S^{(m)}_{i_1,\ldots,i_m}/\sqrt{n} - d_{p,m})} \xrightarrow{\mathrm{d}} N,$$

where N is the Poisson random measure defined in (3.15).

The proof is given in Section 3.6.2. Since $d_{p,m} \sim \sqrt{2m \log p}$, Theorem 3.12 implies the convergence of the largest and smallest off–diagonal entries of $\mathbf{S}^{(m)}$.

Corollary 3.13. *Under the assumptions of Theorem 3.12, we have, as $n \to \infty$,*

$$\max_{1 \leq i_1 < \cdots < i_m \leq p} \frac{S^{(m)}_{i_1,\ldots,i_m}}{\sqrt{n \log p}} \xrightarrow{\mathbb{P}} \sqrt{2m} \quad \text{and} \quad \min_{1 \leq i_1 < \cdots < i_m \leq p} \frac{S^{(m)}_{i_1,\ldots,i_m}}{\sqrt{n \log p}} \xrightarrow{\mathbb{P}} -\sqrt{2m}.$$

Analogously to Corollary 3.6, Theorem 3.12 yields the joint weak convergence of the off–diagonal entries of $\mathbf{S}^{(m)}$, thus extending Theorems 1 and 2 in [67] on the asymptotic Gumbel property of the largest off–diagonal entry of $\mathbf{S}^{(m)}$.

3.4.2 An application to threshold based estimators

A fundamental task in statistics is the estimation of the population covariance or correlation matrix of a multivariate distribution. If the dimension p becomes large, the sample versions $n^{-1}\mathbf{S}$ and $\boldsymbol{R}$ cease to be suitable estimators. Even for our simple model in Section 3.1.2, i.e., when the population covariance and correlation matrices are the p–dimensional identity matrix $\mathbf{I}_p$, the estimators $n^{-1}\mathbf{S}$ and $\boldsymbol{R}$ are not asymptotically consistent for $\mathbf{I}_p$. This phenomenon was explored in [59] among many other papers. Assuming $\mathbb{E}[X^4] < \infty$ and $p/n \to \gamma \in [0, \infty)$, [59] shows that, as $n \to \infty$,

$$\sqrt{n/p}\,\|n^{-1}\mathbf{S} - \mathbf{I}_p\| \xrightarrow{\mathbb{P}} 2 + \sqrt{\gamma} \quad \text{and} \quad \sqrt{n/p}\,\|\boldsymbol{R} - \mathbf{I}_p\| \xrightarrow{\mathbb{P}} 2 + \sqrt{\gamma}\,.$$

Note that p is allowed to grow at a slower rate than n. It was also observed in [59] that

$$\sqrt{n/p}\,\|n^{-1}\operatorname{diag}(\mathbf{S}) - \mathbf{I}_p\| \xrightarrow{\mathbb{P}} 0\,. \tag{3.22}$$

We would like to construct estimators $\widehat{\mathbf{S}}$, $\widehat{\boldsymbol{R}}$ based on $\mathbf{S}$ and $\boldsymbol{R}$, respectively, such that as $n \to \infty$,

$$\sqrt{n/p}\,\|n^{-1}\widehat{\mathbf{S}} - \mathbf{I}_p\| \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \sqrt{n/p}\,\|\widehat{\boldsymbol{R}} - \mathbf{I}_p\| \xrightarrow{\mathbb{P}} 0. \tag{3.23}$$

In view of (3.22), we know that we are able to deal with the diagonal. A natural approach is to eliminate the smallest off–diagonal entries by thresholding. Bickel and Levina [18, 19] considered estimators of the form

$$\widehat{\mathbf{S}} = \big(S_{ij}\mathbf{1}(|S_{ij}| > n\,t_n)\big) \quad \text{and} \quad \widehat{\boldsymbol{R}} = \big(R_{ij}\mathbf{1}(|R_{ij}| > t_n)\big)\,, \tag{3.24}$$

for some threshold sequence $t_n \to 0$. Choosing $t_n = t_n(C) = C\sqrt{(\log p)/n}$ with a sufficiently large constant C, [18, Theorem 1] shows (3.23) for standard normal X. In view of Remark 3.7, the order of the threshold perfectly matches the order of the largest off–diagonal entries. Based on our results, we provide a simple proof of (3.23) for a more general class of distributions.

Corollary 3.14. *Assume $p/n \to \gamma \in [0, \infty)$ and the conditions of Theorem 3.10. Then the estimators $\widehat{\boldsymbol{R}}$, $\widehat{\mathbf{S}}$ in* (3.24) *specified for $t_n(C)$, $C > 2$, satisfy relation* (3.23).

Proof. The diagonal part is taken care of by (3.22) and the fact that $\operatorname{diag}(\boldsymbol{R}) = \mathbf{I}_p$. The off–diagonal entries of $\widehat{\boldsymbol{R}}$ and $\widehat{\mathbf{S}}$ asymptotically vanish in view of Remark 3.7 and Corollary 3.11, respectively. □

Corollary 3.14 shows that the order of the threshold $t_n(C)$ is not affected by the distributional assumption. Under (C1) we thus allow for $p = O(n^{(s-2)/4})$ provided $\mathbb{E}[|X|^s] < \infty$. For comparison, Bickel and Levina [18, p. 2585] showed the first limit relation in (3.23) for the bigger threshold $t_n(C) = Cp^{4/s}/\sqrt{n}$ and dimension $p = o(n^{s/8})$.

3.4.3 An independence test

If the data (X_{it}) is centered Gaussian with identical distribution the null hypothesis of independence is equivalent to $H_0 : n^{-1}\mathbb{E}[\mathbf{S}] = \mathbf{I}_p$. Based on (3.3), Jiang [66] proposed the following test of H_0 with significance level $\alpha \in (0, 1)$:

$$\Psi_\alpha = \mathbf{1}(nW_n^2 - 4\log p + \log\log p \geq q_\alpha)\,,$$

where

$$q_\alpha = -\log(8\pi) - 2\log\log(1-\alpha)^{-1}$$

is the $(1-\alpha)$–quantile of the limiting non–standard Gumbel distribution. If $\Psi_\alpha = 1$, we reject H_0. Properties of this test are studied in [32].

In view of Corollary 3.6 we can propose a multitude of alternative tests based on the joint asymptotic distribution of the k largest or smallest off–diagonal entries of $\mathbf{S}$ and $\boldsymbol{R}$, respectively. Under the conditions of Theorem 3.5 we have as $n \to \infty$,

$$\widetilde{d}_p\Big(\frac{(S_{(1)},\ldots,S_{(k)})}{\sqrt{n}} - \widetilde{d}_p\Big) \stackrel{\mathrm{d}}{\to} (-\log\Gamma_1,\ldots,-\log\Gamma_k)$$

and $\Gamma_i = E_1 + \cdots + E_i$ for i.i.d. standard exponential random variables (E_j). For $k \geq 1$ and $\alpha \in (0,1)$, consider a set $\mathcal{A}_k^\alpha \subset \mathbb{R}^k$ such that

$$\mathbb{P}\big((-\log\Gamma_1,\ldots,-\log\Gamma_k) \in \mathcal{A}_k^\alpha\big) = 1-\alpha$$

and define the test $T(\mathcal{A}_k^\alpha)$ by

$$T(\mathcal{A}_k^\alpha) = \mathbf{1}\Big(\widetilde{d}_p\Big(\frac{(S_{(1)},\ldots,S_{(k)})}{\sqrt{n}} - \widetilde{d}_p\Big) \notin \mathcal{A}_k^\alpha\Big)\,.$$

If $T(\mathcal{A}_k^\alpha) = 1$, we reject H_0. Then $T(\mathcal{A}_k^\alpha)$ is an asymptotic independence test with significance level α.

Convenient univariate test statistics can be constructed from spacings of $S_{(1)},\ldots,S_{(k)}$. An advantage of using spacings is that one avoids centering by $\widetilde{d}_p$. For example, consider for some $k \geq 2$,

$$\begin{aligned}
T_k^{(1)} &= \widetilde{d}_p\,(S_{(1)} - S_{(k)})/\sqrt{n}\,,\\
T_k^{(2)} &= \widetilde{d}_p \max_{i=1,\ldots,k-1}(S_{(i)} - S_{(i+1)})/\sqrt{n}\,,\\
T_k^{(3)} &= \widetilde{d}_p^2 \frac{1}{n}\sum_{i=1}^{k-1}(S_{(i)} - S_{(i+1)})^2\,.
\end{aligned}$$

Recall the well–known fact that

$$\Big(\frac{\Gamma_1}{\Gamma_{k+1}},\ldots,\frac{\Gamma_k}{\Gamma_{k+1}}\Big) \stackrel{d}{=} \big(U_{(k)},\ldots,U_{(1)}\big)\,,$$

where the right–hand vector consists of the order statistics of k i.i.d. uniform random variables on $(0,1)$. Then we have

$$\begin{aligned}
T_k^{(1)} &\stackrel{\mathrm{d}}{\to} \log\big(\Gamma_k/\Gamma_1\big) = \log\frac{\Gamma_k/\Gamma_{k+1}}{\Gamma_1/\Gamma_{k+1}} \stackrel{d}{=} \log\big(U_{(1)}/U_{(k)}\big)\,,\\
T_k^{(2)} &\stackrel{\mathrm{d}}{\to} \max_{i=1,\ldots,k-1}\log(\Gamma_{i+1}/\Gamma_i) \stackrel{d}{=} \max_{i=1,\ldots,k-1}\log(U_{(k-i)}/U_{(k-i+1)})\,,\\
T_k^{(3)} &\stackrel{\mathrm{d}}{\to} \sum_{i=1}^{k-1}(\log(\Gamma_{i+1}/\Gamma_i))^2 \stackrel{d}{=} \sum_{i=1}^{k-1}(\log(U_{(k-i)}/U_{(k-i+1)}))^2\,.
\end{aligned}$$

Now, choosing q_α as the $(1-\alpha)$–quantiles of the limiting random variables we have $T(\mathcal{A}_k^\alpha) = \mathbf{1}(T_k^{(i)} > q_\alpha)$, $i = 1,2,3$.

3.5 Proof of Theorem 3.1

In view of Remark 3.2 it suffices to prove the theorem under (C1). Throughout this proof we assume the standard conditions on (X_{it}).

We start with a useful auxiliary result due to Einmahl [44] (Corollary 1(b), p. 31, in combination with Remark on p. 32).

Lemma 3.15. *Consider independent $\mathbb{R}^d$–valued random vectors $\xi_1, \ldots, \xi_n$ with mean zero. Assume that ξ_i, $i = 1, \ldots, n$, has finite moment generating function in some neighborhood of the origin and that the covariance matrix $\mathrm{var}(\xi_1 + \cdots + \xi_n) = B_n \mathbf{I}_d$ where $B_n > 0$ and $\mathbf{I}_d$ denotes the identity matrix. Let η_k be independent $N(0, \sigma^2 \mathrm{var}(\xi_k))$ random vectors independent of (ξ_k), and $\sigma^2 \in (0, 1]$. Let $\xi_k^* = \xi_k + \eta_k$, $k = 1, \ldots, n$, and write p_n^* for the density of $B_n^{-1/2}(\xi_1^* + \cdots + \xi_n^*)$. Choose $\alpha \in (0, 0.5)$ such that*

$$\alpha \sum_{k=1}^{n} \mathbb{E}[|\xi_k|^3 \exp(\alpha|\xi_k|)] \le B_n \,, \tag{3.25}$$

and write

$$\beta_n = \beta_n(\alpha) = B_n^{-3/2} \sum_{k=1}^{n} \mathbb{E}[|\xi_k|^3 \exp(\alpha|\xi_k|)] \,,$$

where $|x|$ denotes the Euclidean norm. If

$$|x| \le c_1 \alpha \, B_n^{1/2} \,, \qquad \sigma^2 \ge -c_2 \beta_n^2 \log \beta_n \,, \qquad B_n \ge c_3 \alpha^{-2} \,, \tag{3.26}$$

where c_1, c_2, c_3 are constants only depending on d, then

$$p_n^*(x) = \varphi_{(1+\sigma^2)\mathbf{I}_d}(x) \exp(\overline{T}_n(x)) \quad \text{with} \quad |\overline{T}_n(x)| \le c_4 \beta_n \left(|x|^3 + 1\right), \tag{3.27}$$

where $\varphi_{\mathbf{\Sigma}}$ is the density of a $N(0, \mathbf{\Sigma})$ random vector and c_4 is a constant only depending on d.

Proof under **(C1)**. We proceed by formulating and proving various auxiliary results. We will use the following notation: c denotes any positive constant whose value is not of interest, sometimes we write $c_0, c_1, c_2, \ldots$ for positive constants whose value or size is relevant in the proof,

$$\begin{aligned} \overline{X}_i &= X_i \mathbf{1}(|X_i| \le n^{1/s}) - \mathbb{E}[X \mathbf{1}(|X| \le n^{1/s})] \,, \qquad \underline{X}_i = X_i - \overline{X}_i \,, \\ \overline{S}_n &= \sum_{i=1}^{n} \overline{X}_i \,, \qquad \underline{S}_n = S_n - \overline{S}_n \,. \end{aligned}$$

Next we consider an approximation of the distribution of $\overline{S}_n$.

Lemma 3.16. *Let $\widetilde{p}_n$ be the density of*

$$n^{-1/2} \sum_{i=1}^{n} (\overline{X}_i + \sigma_n N_i)$$

where (N_i) is i.i.d. $N(0, 1)$, independent of (X_i) and $\sigma_n^2 = \mathrm{var}(\overline{X}) s_n^2$. If $n^{-2c_6} \log n \le s_n^2 \le 1$ with $c_6 = 0.5 - (1 - \delta)/s$ for arbitrarily small $\delta > 0$, then the relation

$$\widetilde{p}_n(x) = \varphi_{1+\sigma_n^2}(x)(1 + o(1)) \,, \qquad n \to \infty \,,$$

holds uniformly for $|x| = o(n^{1/6 - 1/(3s)})$.

Proof. We apply Lemma 3.15 to the i.i.d. random variables $\xi_i = \overline{X}_i$, $i = 1, \dots, n$. Notice that $\mathbb{E}[\overline{X}] = 0$ and $B_n = \operatorname{var}(\overline{S}_n) = n \operatorname{var}(\overline{X})$. Choose $\widetilde{\alpha} = c_5 n^{-1/s}$. Then

$$\begin{aligned} \widetilde{\alpha} \sum_{i=1}^{n} \mathbb{E}[|\overline{X}_i|^3 \exp(\widetilde{\alpha}|\overline{X}_i|)] &= \widetilde{\alpha}\, n\, \mathbb{E}[|\overline{X}|^3 \exp(\widetilde{\alpha}|\overline{X}|)] \\ &\le c_5\, n^{1-1/s}\, \mathbb{E}[|\overline{X}|^3]\, \exp(2c_5) \\ &\le 8c_5 \exp(2c_5) n^{1-\delta/s}\, \mathbb{E}[|X|^{2+\delta}]\,, \end{aligned}$$

where $\delta \in (0,1)$ is chosen such that $\mathbb{E}[|X|^{2+\delta}] < \infty$. Hence (3.25) is satisfied for $\alpha = \widetilde{\alpha}$ and sufficiently small c_5.

Next choose

$$\begin{aligned} \widetilde{\beta}_n &= B_n^{-3/2} \sum_{i=1}^{n} \mathbb{E}\big[|\overline{X}_i|^3\, \exp(\widetilde{\alpha}|\overline{X}_i|)\big] = B_n^{-3/2}\, n\, \mathbb{E}[|\overline{X}|^3 \exp(\widetilde{\alpha}|\overline{X}|)] \\ &\le c\, B_n^{-3/2} n^{1+(1-\delta)/s}\, \mathbb{E}[|\overline{X}|^{2+\delta}] \le c\, n^{-c_6}\,, \end{aligned} \tag{3.28}$$

where δ is chosen as above and $c_6 = 0.5 - (1-\delta)/s$.

Next we consider (3.26). We can choose x according to the restriction

$$|x| \le c_1\, \widetilde{\alpha}\, B_n^{1/2} \sim c\, n^{1/2-1/s}\,. \tag{3.29}$$

By (3.26) and (3.29) we can choose $\sigma^2 = \sigma_n^2$ according as

$$1 \ge \sigma_n^2 \ge c\, \log n\, n^{-2c_6}\,. \tag{3.30}$$

Moreover, $B_n \ge c_3\, \widetilde{\alpha}^{-2}$. An application of (3.27) yields

$$\widetilde{p}_n(x) = \varphi_{1+\sigma_n^2}(x) \exp(\overline{T}_n(x)) \quad \text{for} \quad |\overline{T}_n(x)| \le c_4 \widetilde{\beta}_n(|x|^3 + 1)\,,$$

but in view of (3.28) and (3.29), $\widetilde{\beta}_n(|x|^3+1) = o(1)$ uniformly for $|x|^3 = o(\min(n^{0.5-1/s}, n^{c_6})) = o(n^{0.5-1/s})$ for arbitrarily small $\delta > 0$. That is, the remainder term $|\overline{T}_n(x)|$ converges to zero, uniformly for the x considered. This proves the lemma. □

We add another auxiliary result.

Lemma 3.17. *Assume that $p = p_n \to \infty$ and $p = O(n^{(s-2)/2})$. Then for $x \in \mathbb{R}$, c_6 as in Lemma 3.16, an i.i.d. $N(0,1)$ sequence (N_i) and $\sigma_n^2 = c \log n\, n^{-2c_6}$, we have*

$$p\, \mathbb{P}\Big(n^{-1/2} \sum_{i=1}^{n} (\overline{X}_i + \sigma_n N_i) > d_p + x/d_p\Big) \to \mathrm{e}^{-x}\,, \qquad n \to \infty\,.$$

Proof. Write $y_n = \sqrt{(s-2)\log n}$. By virtue of Lemma 3.16 we observe that for any $C > 1$,

$$\begin{aligned} P_1 &= \mathbb{P}\Big(d_p + x/d_p < n^{-1/2} \sum_{i=1}^{n} (\overline{X}_i + \sigma_n N_i) \le y_n\Big) \sim \int_{d_p + x/d_p}^{y_n} \varphi_{1+\sigma_n^2}(y)\, dy\,, \\ P_2 &= \mathbb{P}\Big(y_n < n^{-1/2} \sum_{i=1}^{n} (\overline{X}_i + \sigma_n N_i) \le C\, y_n\Big) \sim \int_{y_n}^{C y_n} \varphi_{1+\sigma_n^2}(y)\, dy\,. \end{aligned}$$

However, using Mill's ratio and the definition of d_p, we have that

$$\begin{aligned} p\,P_1 &\sim \mathrm{e}^{-x}\,\frac{\overline{\Phi}\Big(\dfrac{d_p + x/d_p}{\sqrt{1+\sigma_n^2}}\Big)}{\overline{\Phi}(d_p + x/d_p)} - p\,\overline{\Phi}\Big(\frac{y_n}{\sqrt{1+\sigma_n^2}}\Big) \\ &\sim \mathrm{e}^{-x}\,\exp\Big(0.5(d_p + x/d_p)^2\frac{\sigma_n^2}{1+\sigma_n^2}\Big) - \frac{1}{\sqrt{2\pi}\sqrt{(s-2)\log n}} p\,n^{-(s-2)/2}\,, \end{aligned}$$

but the right-hand side converges to e^{-x} since $(d_p + x/d_p)^2\sigma_n^2 \sim d_p^2\sigma_n^2 = o(1)$, $p\,n^{-(s-2)/2} = O(1)$ and $(\log n)^2 n^{-2c_6} = o(1)$. A similar argument shows $pP_2 \to 0$.

We also have

$$\begin{aligned} \mathbb{P}\Big(n^{-1/2}\sum_{i=1}^n (\overline{X}_i + \sigma_n N_i) > C\,y_n\Big) &\le \mathbb{P}\Big(n^{-1/2}\overline{S}_n > 0.5\,C\,y_n\Big) + \overline{\Phi}\big(0.5\,C y_n/\sigma_n\big) \\ &= P_3 + P_4\,. \end{aligned}$$

It is easy to see that $pP_4 \to 0$. We observe that

$$\begin{aligned} |\overline{X}_i| &\le n^{1/s}\big(1 + o(n^{-1/s})\big) = c_n\,, \qquad a.s. \\ \mathrm{var}(\overline{X}) &\le \mathbb{E}[X^2\mathbf{1}(|X| \le n^{1/s})] \le \mathrm{var}(X) = 1\,. \end{aligned}$$

We apply Prokhorov's inequality (Petrov [98, Chapter III.5]) for any $C > 1$,

$$\begin{aligned} p\,\mathbb{P}\big(\overline{S}_n > C\sqrt{n}\,y_n\big) &\le p\,\exp\Big(-\frac{C\sqrt{(s-2)n\log n}}{2c_n}\,\log\big(1 + \frac{C\sqrt{(s-2)n\log n}\,c_n}{2n\mathrm{var}(\overline{X})}\big)\Big) \\ &\le p\,\exp\big(-\frac{C^2}{2}\frac{(s-2)\,\log n}{4}\big) \\ &= p\,n^{-C^2(s-2)/8}. \end{aligned}$$

The right-hand side converges to zero for sufficiently large C. This proves the lemma. □

Write $(X_{it})_{t\ge1}$ for the i.i.d. sequence of the summands constituting $S_n^{(i)}$ and

$$\overline{S}_n^{(i,N)} = \sum_{t=1}^n (\overline{X}_{it} + \sigma_n N_i) =: \overline{S}_n^{(i)} + \sigma_n\sqrt{n}\,\widetilde{N}_i\,,$$

where $(\widetilde{N}_i)$ are i.i.d. standard normal random variables independent of everything else. Then by Lemma 3.17,

$$\mathbb{P}\Big(\max_{i=1,\dots,p} d_p\,(\overline{S}_n^{(i,N)}/\sqrt{n} - d_p) \le x\Big) \to \Lambda(x)\,, \qquad x \in \mathbb{R}\,, \qquad n \to \infty\,.$$

We have

$$\begin{aligned} d_p n^{-1/2}\max_{i=1,\dots,p}|\overline{S}_n^{(i)} - \overline{S}_n^{(i,N)}| &\le d_p \max_{i=1,\dots,p}|\sigma_n(\widetilde{N}_i - d_p)| + \sigma_n\,d_p^2 \\ &\le d_p\sigma_n \max_{i=1,\dots,p}|\widetilde{N}_i - d_p| + \sigma_n\,d_p^2 \\ &= O_{\mathbb{P}}(d_p^2\sigma_n) = o_{\mathbb{P}}(1)\,, \qquad n\to\infty\,. \end{aligned}$$

Therefore

$$\mathbb{P}\Big(d_p\big(\max_{i=1,\dots,p}(\overline{S}_n^{(i)}/\sqrt{n}-d_p)\le x\Big)\to\Lambda(x)\,,\qquad x\in\mathbb{R}\,,\qquad n\to\infty\,,\tag{3.31}$$

and the latter relation is equivalent to

$$p\,\mathbb{P}(\overline{S}_n/\sqrt{n}>d_p+x/d_p)\to \mathrm{e}^{-x}\,,\qquad x\in\mathbb{R}\,,\qquad n\to\infty\,.\tag{3.32}$$

Our next goal is to prove that we can replace $\overline{S}_n$ by S_n in the latter relation. In view of the equivalence between (3.31) and (3.32) it suffices to show (3.31) with $\overline{S}_n^{(i)}$ replaced by $S_n^{(i)}$. Therefore we will show that

$$\frac{d_p}{\sqrt{n}}\max_{i=1,\dots,p}\big|\underline{S}_n^{(i)}\big|\stackrel{\mathbb{P}}{\to}0\,.$$

We have by the Fuk–Nagaev inequality [99, p. 78] for $y>0$ and suitable constants $c_0,c_1>0$,

$$\begin{aligned}\mathbb{P}\Big(d_p\max_{i=1,\dots,p}|\underline{S}_n^{(i)}/\sqrt{n}|>y\Big)\;&\le\;p\,\mathbb{P}\big(|\underline{S}_n|>\sqrt{n}y/d_p\big)\\&\le\;c_0 n\,\mathbb{E}[|\underline{X}|^s]\big(\frac{\sqrt{n}y}{d_p}\big)^{-s}+\exp\Big(-c_1\frac{y^2}{d_p^2\,\mathrm{var}(\underline{X})}\Big)\end{aligned}\tag{3.33}$$

Using partial integration and Markov's inequality of order s, we find that $\mathrm{var}(\underline{X})\le c\,n^{-0.5+1/(2s)}$ holds if $\mathbb{E}[|X|^s]<\infty$. Combining this bound with the rate $p=O(n^{-(s-2)/2})$, we see that $d_p^2\,\mathrm{var}(\underline{X})\to0$ and therefore the exponential term in (3.33) vanishes. The polynomial term in (3.33) converges to zero for the same reason. This proves (3.31) with $\overline{S}_n^{(i)}$ replaced by $S_n^{(i)}$ and finishes the proof of the theorem. □

3.6 Proofs of sample covariance and correlation results

3.6.1 Proof of Theorem 3.5

By Kallenberg's criterion for the convergence of simple point processes (see for instance [46, p. 233, Theorem 5.2.2]) it suffices to verify the following conditions:

(i) For any $-\infty<a<b<\infty$, one has $\mathbb{E}[N_n^S(a,b]]\to\mathbb{E}[N(a,b]]=\mu(a,b]$ as $n\to\infty$.

(ii) For $B=\cup_{i=1}^{\ell}(b_i,c_i]\subset(-\infty,\infty)$ with $-\infty<b_1<c_1<\cdots<b_\ell<c_\ell<\infty$, one has $\mathbb{P}(N_n^S(B)=0)\to\mathbb{P}(N(B)=0)=\mathrm{e}^{-\mu(B)}$ as $n\to\infty$.

We start with (i). Note that $\mu(a,b]=\mathrm{e}^{-a}-\mathrm{e}^{-b}$. Since the assumptions of Theorem 3.1 hold it follows from (3.10) (with p replaced by $p(p-1)/2$), that as $n\to\infty$

$$\mathbb{E}[N_n^S(a,b]]=\frac{p(p-1)}{2}\mathbb{P}\big(\widetilde{d}_p+a/\widetilde{d}_p<S_{12}/\sqrt{n}<\widetilde{d}_p+b/\widetilde{d}_p\big)\to\mu(a,b]\,.$$

To show (ii), we consider

$$1-\mathbb{P}(N_n^S(B)=0)=\mathbb{P}\Big(\bigcup_{1\le i<j\le p}A_{ij}\Big)\,,\qquad A_{ij}=\{\widetilde{d}_p(S_{ij}/\sqrt{n}-\widetilde{d}_p)\in B\}\,.$$

By an inclusion–exclusion argument we get for $k \geq 1$,

$$\sum_{d=1}^{2k}(-1)^{d-1}W_d \leq \mathbb{P}\Big(\bigcup_{1\leq i<j\leq p} A_{ij}\Big) \leq \sum_{d=1}^{2k-1}(-1)^{d-1}W_d\,, \tag{3.34}$$

where

$$W_d = \textstyle\sum_{(I,J)\in I_d}\mathbb{P}(A_{i_1j_1}\cap\cdots\cap A_{i_dj_d}) =: \sum_{(I,J)\in I_d} q_{(I,J)}$$

and the summation runs over the set

$$\begin{aligned} I_d = \{(I,J) = ((i_1,j_1),\ldots,(i_d,j_d)) \text{ such that } 1 \leq i_t < j_t \leq p, t = 1,\ldots,d\,,\\ \text{and } (i_1,j_1) < (i_2,j_2) < \cdots < (i_d,j_d)\}\,. \end{aligned}$$

In the definition of I_d, we use the lexicographic ordering of pairs $(i_s,j_s),(i_t,j_t)$:

$$(i_s,j_s) < (i_t,j_t) \quad \text{ if and only if } \quad i_s < i_t \text{ or } (i_s = i_t \text{ and } j_s < j_t)\,.$$

A combinatorial argument yields

$$|I_d| = \binom{\frac{p(p-1)}{2}}{d} \sim \frac{1}{d!}\Big(\frac{p^2}{2}\Big)^d, \qquad n\to\infty\,. \tag{3.35}$$

Proof of (ii) under (C1). Consider the set $\widehat{I}_d$ consisting of all elements $(I,J)\in I_d$ such that all $i_t, j_t, t = 1,\ldots,d$ are mutually distinct. For $(I,J)\in\widehat{I}_d$ the random variables $S_{i_t,j_t}, t = 1,\ldots d$, are i.i.d. and therefore

$$q_{(I,J)} = \big(\mathbb{P}(A_{12})\big)^d\,. \tag{3.36}$$

For $(I,J)\in I_d\backslash\widehat{I}_d$ we write

$$\mathbf{S}_n^{(I,J)} = \big(S_{i_1j_1},\ldots,S_{i_d,j_d}\big)^\top = \sum_{t=1}^n \big(X_{i_1t}X_{j_1t},\ldots,X_{i_dt}X_{j_dt}\big)^\top =: \sum_{t=1}^n \xi_t\,,$$

and also $\mathbf{1} = (1,\ldots,1)^\top\in\mathbb{R}^d$. The i.i.d. $\mathbb{R}^d$–valued summands ξ_t with generic element ξ have mean zero and covariance matrix $\mathbf{I}_d$. We have

$$q_{(I,J)} \quad = \quad \mathbb{P}\Big(n^{-1/2}\mathbf{S}_n^{(I,J)} \in \widetilde{d}_p\,\mathbf{1} + B^d/\widetilde{d}_p\Big)$$

We will apply Lemma 3.15 to (ξ_t). We will prove it under (C1); the proof under (C2') and (C3') is analogous; we will indicate some necessary changes. In this case, $\mathbb{E}[|\xi_i|^s] < \infty$ for some $s > 2$. Write

$$\begin{aligned} \overline{\xi}_t &= \Big(\xi_t^{(l)}\mathbf{1}\big(|\xi_t^{(l)}|\leq n^{1/s}\big) - \mathbb{E}[\xi^{(l)}\mathbf{1}(|\xi^{(l)}|\leq n^{1/s})]\Big)^\top_{l=1,\ldots,d}\,,\\ \underline{\xi}_t &= \xi_t - \overline{\xi}_t\,,\\ \overline{\mathbf{S}}_n^{(I,J)} &= \sum_{t=1}^n \overline{\xi}_t\,, \qquad \underline{\mathbf{S}}_n^{(I,J)} = \mathbf{S}_n^{(I,J)} - \overline{\mathbf{S}}_n^{(I,J)} = \sum_{t=1}^n \underline{\xi}_t\,. \end{aligned}$$

Proceeding as in the proof of Lemma 3.16, we obtain the following result.

Lemma 3.18. *Let $\widetilde{p}_n$ be the density of*

$$n^{-1/2}\sum_{i=1}^{n}(\overline{\xi}_i+\sigma_n N_i)$$

where (N_i) is i.i.d. $N(0,\mathbf{I}_d)$, independent of (ξ_i) and $\sigma_n^2=\mathrm{var}(\overline{\xi}^{(1)})s_n^2$. If $n^{-2c_6}\log n\le s_n^2\le 1$ with $c_6=0.5-(1-\delta)/s$ for arbitrarily small $\delta>0$, then the relation

$$\widetilde{p}_n(x)=\varphi_{(1+\sigma_n^2)\mathbf{I}_d}(x)(1+o(1))\,,\qquad n\to\infty\,,$$

holds uniformly for $|x|=o(n^{1/6-1/(3s)})$.

Following the lines of the proof of Lemma 3.17, we obtain the following result:

Lemma 3.19. *Assume that $p=p_n\to\infty$ and $p^2=O(n^{(s-2)/2})$. Then for $\sigma_n^2=c\log n\,n^{-2c_6}$ and an i.i.d. $N(0,1)$ sequence $(\widetilde{N}_i)$, uniformly for (I,J) in I_d,*

$$\Big(\frac{p^2}{2}\Big)^d q_{(I,J)}\sim\Big(\frac{p^2}{2}\mathbb{P}\Big(n^{-1/2}\sum_{t=1}^{n}\big(\overline{\xi}_t^{(1)}+\sigma_n\widetilde{N}_i\big)\in B\Big)\Big)^d\sim(\mu(B))^d\,. \tag{3.37}$$

Finally, we need to prove that $\overline{\mathbf{S}}_n^{(I,J)}$ in (3.37) can be replaced by $\mathbf{S}_n^{(I,J)}$. However, this follows in the same way as the corresponding steps in the proof of Theorem 3.3. Indeed, since we need to show that $n^{-1/2}\underline{\mathbf{S}}_n^{(I,J)}$ does not contribute asymptotically to $n^{-1/2}\mathbf{S}_n^{(I,J)}$ it suffices to prove this fact for each of the components of $n^{-1/2}\underline{\mathbf{S}}_n^{(I,J)}$.

We conclude that as $n\to\infty$

$$W_d=\Big(\sum_{(I,J)\in I_d\setminus\widehat{I}_d}+\sum_{(I,J)\in\widehat{I}_d}\Big)q_{(I,J)}\sim\frac{1}{d!}\Big(\frac{p^2}{2}\Big)^d\big(\mathbb{P}(A_{12})\big)^d\sim\frac{(\mu(B))^d}{d!}. \tag{3.38}$$

We recall that (3.34) provides an upper and lower bound for $\mathbb{P}(N_n(B)=0)$. Letting first $n\to\infty$ and then $k\to\infty$, thanks to (3.38) we see that both bounds converge to the same limit. More precisely, we have

$$\lim_{n\to\infty}P(N_n(B)=0)=1-\sum_{d=1}^{\infty}(-1)^{d-1}\frac{\big(\mu(B)\big)^d}{d!}=\sum_{d=0}^{\infty}\frac{\big(-\mu(B)\big)^d}{d!}=\mathrm{e}^{-\mu(B)}\,.$$

The proof of (ii) is complete.

Proof of (ii) under (C2'), (C3'). Write $b_0=\min_{1\le q\le\ell}b_q$, $c_0=\max_{1\le q\le\ell}c_q$ and for $(I,J)\in I_d\setminus\widehat{I}_d$,

$$\widetilde{S}_n=S_{i_1j_1}+\cdots+S_{i_dj_d}=\sum_{t=1}^{n}(X_{i_1t}X_{j_1t}+\cdots+X_{i_dt}X_{j_dt})\,.$$

We have

$$\begin{aligned}q_{(I,J)}&\le\mathbb{P}\Big(\big(\frac{\widetilde{S}_n}{d\sqrt{n}}-\widetilde{d}_p\big)\,\widetilde{d}_p\in(b_0,c_0]\Big)\\&=\mathbb{P}\Big(\sqrt{d}\big(b_0/\widetilde{d}_p+\widetilde{d}_p\big)<\frac{\widetilde{S}_n}{\sqrt{d\,n}}<\sqrt{d}\big(c_0/\widetilde{d}_p+\widetilde{d}_p\big)\Big)\end{aligned} \tag{3.39}$$

Note that $\widetilde{S}_n/\sqrt{d}$ has i.i.d. summands with mean zero and unit variance. Since $\sqrt{d}(c_0/\widetilde{d}_p + \widetilde{d}_p) = o(n^{1/6})$ under (C3') and $\sqrt{d}(c_0/\widetilde{d}_p + \widetilde{d}_p) = o(n^{1/6} \wedge g_n)$ under (C2') applications of [98, Theorem 1 in Section VIII.2] and [87, Theorem 4], respectively, yield

$$q_{(I,J)} \quad \leq \quad c\Big(\overline{\Phi}\big(\sqrt{d}(b_0/\widetilde{d}_p + \widetilde{d}_p)\big) - \overline{\Phi}\big(\sqrt{d}(c_0/\widetilde{d}_p + \widetilde{d}_p)\big)\Big) = O(p^{-2d+\varepsilon})\,.$$

for an arbitrarily small $\varepsilon > 0$. This shows that (3.38) holds. Now one can proceed as under condition (C1).

3.6.2 Proof of Theorem 3.12

We proceed as in the proof Theorem 3.5 and show $(i), (ii)$ therein. For $-\infty < a < b < \infty$, it follows from (3.10) that as $n \to \infty$

$$\begin{aligned}\mathbb{E}[N_n^{(m)}(a,b]] &= \binom{p}{m}\mathbb{P}\big(d_{p,m} + a/d_{p,m} < S_{12}/\sqrt{n} < d_{p,m} + b/d_{p,m}\big)\\ &\to \mathrm{e}^{-a} - \mathrm{e}^{-b}\,.\end{aligned}$$

This proves condition (i). The proof of (ii) is completely analogous to the proof of Theorem 3.5. The main difference is that $i < j$ needs to be replaced with $i_1 < i_2 < \cdots < i_m$. For example, instead of the index set I_d whose elements are d distinct m–tuples, with $|I_d| = \binom{\binom{p}{2}}{d}$; see (3.35); one would get an index set $I_d^{(m)}$ of d distinct m–tuples satisfying $|I_d^{(m)}| = \binom{\binom{p}{m}}{d}$. We omit details.

3.6.3 Proof of Theorem 3.10

First, assume $\mathrm{var}(X^2) = 0$. Then $S_{ii} = n$ a.s. for all i and hence $\sqrt{n}R_{ij} = S_{ij}/\sqrt{n}$ so that the claim follows immediately from Theorem 3.5.

In the remainder of this proof, we therefore assume $\mathrm{var}(X^2) > 0$. By Theorem 3.5, we already know that the point processes N_n^S converge to a Poisson random measure with mean measure $\mu(x,\infty] = \mathrm{e}^{-x}$, $x > 0$. Our idea is to transfer the convergence of N_n^S onto N_n^R. To this end, it suffices to show that (see [70, Theorem 4.2]) $N_n^R - N_n^S \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$, or equivalently that for any continuous function f on $\mathbb{R}$ with compact support,

$$\int f\,\mathrm{d}N_n^R - \int f\,\mathrm{d}N_n^S \xrightarrow{\mathbb{P}} 0\,, \quad n \to \infty\,.$$

Suppose the compact support of f is contained in $[K+\gamma_0, \infty)$ for some $\gamma_0 > 0$ and $K \in \mathbb{R}$. Since f is uniformly continuous, $\omega(\gamma) := \sup\{|f(x) - f(y)| : x, y \in \mathbb{R}, |x-y| \leq \gamma\}$ tends to zero as $\gamma \to 0$. We have to show that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\sum_{1\leq i<j\leq p}\Big(f\big((\sqrt{n}R_{ij} - \widetilde{d}_p)\widetilde{d}_p\big) - f\big((S_{ij}/\sqrt{n} - \widetilde{d}_p)\widetilde{d}_p\big)\Big)\Big| > \varepsilon\Big) = 0\,. \tag{3.40}$$

On the sets

$$A_{n,\gamma} = \Big\{\max_{1\leq i<j\leq p}\big|\widetilde{d}_p(\sqrt{n}R_{ij} - S_{ij}/\sqrt{n})\big| \leq \gamma\Big\}\,, \quad \gamma \in (0, \gamma_0)\,, \tag{3.41}$$

we have

$$\big|f\big((\sqrt{n}R_{ij} - \widetilde{d}_p)\widetilde{d}_p\big) - f\big((S_{ij}/\sqrt{n} - \widetilde{d}_p)\widetilde{d}_p\big)\big| \leq \omega(\gamma)\,\mathbf{1}((S_{ij}/\sqrt{n} - \widetilde{d}_p)\widetilde{d}_p > K)\,.$$

Therefore, we see that, for $\gamma \in (0, \gamma_0)$,

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{1\le i<j\le p}\Big(f\big((\sqrt{n}R_{ij}-\widetilde{d}_p)\widetilde{d}_p\big)-f\big((S_{ij}/\sqrt{n}-\widetilde{d}_p)\widetilde{d}_p\big)\Big)\Big|>\varepsilon, A_{n,\gamma}\Big)\\
&\le \mathbb{P}\Big(\omega(\gamma)\,\#\{1\le i<j\le p : (S_{ij}/\sqrt{n}-\widetilde{d}_p)\widetilde{d}_p>K\}>\varepsilon\Big)\\
&\le \frac{\omega(\gamma)}{\varepsilon}\mathbb{E}\big[\#\{1\le i<j\le p : (S_{ij}/\sqrt{n}-\widetilde{d}_p)\widetilde{d}_p>K\}\big]\\
&= \frac{\omega(\gamma)}{\varepsilon}\underbrace{\frac{p(p-1)}{2}\mathbb{P}((S_{12}/\sqrt{n}-\widetilde{d}_p)\widetilde{d}_p>K)}_{\to \mathrm{e}^{-K} \text{ by Theorem 3.1}}.
\end{aligned} \tag{3.42}$$

Moreover, we have

$$\begin{aligned}
\mathbb{P}(A^c_{n,\gamma}) &= \mathbb{P}\Big(\max_{1\le i<j\le p}\big|\widetilde{d}_p(\sqrt{n}R_{ij}-S_{ij}/\sqrt{n})\big|>\gamma\Big)\\
&= \mathbb{P}\Big(\max_{1\le i<j\le p}\widetilde{d}_p\frac{|S_{ij}|}{\sqrt{n}}\Big|\frac{n}{\sqrt{S_{ii}S_{jj}}}-1\Big|>\gamma\Big).
\end{aligned}$$

Since $\max_{1\le i<j\le p}(S_{ij}/\sqrt{n}-\widetilde{d}_p)\widetilde{d}_p \to \Lambda$, we get that $\max_{1\le i<j\le p}\widetilde{d}_p\frac{|S_{ij}|}{\sqrt{n}} = O_{\mathbb{P}}(\widetilde{d}_p^2)$. Thus,

$$\lim_{n\to\infty}\mathbb{P}(A^c_{n,\gamma})=0 \tag{3.43}$$

is implied by

$$\lim_{n\to\infty}\mathbb{P}\Big(\widetilde{d}_p^2\max_{1\le i<j\le p}\Big|\frac{n}{\sqrt{S_{ii}S_{jj}}}-1\Big|>\beta\Big)=0\,,\quad \beta>0\,. \tag{3.44}$$

Then taking the limits $n\to\infty$ followed by $\gamma\to 0^+$ in (3.42) and (3.43) establishes (3.40).

It remains to prove (3.44). By the law of large numbers, $|S_{ii}/n| \overset{\text{a.s.}}{\to} 1$ as $n\to\infty$. We have

$$\begin{aligned}
\max_{1\le i<j\le p}\Big|\frac{n}{\sqrt{S_{ii}S_{jj}}}-1\Big| &= \Big(\frac{n}{\min_{1\le i<j\le p}\sqrt{S_{ii}S_{jj}}}-1\Big)\vee\Big(1-\frac{n}{\max_{1\le i<j\le p}\sqrt{S_{ii}S_{jj}}}\Big)\\
&\le \max_{1\le i\le p}\Big|\frac{n}{S_{ii}}-1\Big|
\end{aligned}$$

so that (3.44) follows from

$$\lim_{n\to\infty}\mathbb{P}\Big(\widetilde{d}_p^2\max_{1\le i\le p}\Big|\frac{S_{ii}}{n}-1\Big|>\beta\Big)=0\,,\quad \beta>0\,. \tag{3.45}$$

We have

$$\mathbb{P}\Big(\widetilde{d}_p^2\max_{1\le i\le p}\Big|\frac{S_{ii}}{n}-1\Big|>\beta\Big)\le p\,\mathbb{P}\Big(\frac{1}{\sqrt{n}}\Big|\sum_{t=1}^n(X_{1t}^2-1)\Big|>\frac{\beta\sqrt{n}}{\widetilde{d}_p^2}\Big):=\Psi_n\,.$$

It remains to prove that $\Psi_n\to 0$ under each of the conditions (C1), (C2'), (C3').

First, assume (C1). Thus we have $\mathbb{E}[|X|^s]<\infty$ and $p=O(n^{(s-2)/4})$ for some $s>2$. An application of Markov's inequality yields

$$\Psi_n \;\le\; c\,\widetilde{d}_p^{\,s}\,n^{-(s+2)/4}\,\mathbb{E}\big[|S_{11}-n|^{s/2}\big]\,.$$

By [39, Lemma A.4] one has

$$\mathbb{E}\big[|S_{11}-n|^{s/2}\big] \le c\, n^{\max(1,s/4)}$$

and therefore it is easy to conclude that $\Psi_n = O((\log n)^{s/2} n^{-(1/4)\min(s-2,2)}) \to 0$ as $n \to \infty$.

Next, assume condition (C3'). By [98, Section VIII.4, No. 8], we have for $0 \le x \le n^{\alpha}/\rho(n)$ with $0 < \alpha \le 1/6$ and $\rho(n) \to \infty$ arbitrarily slowly that

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}}\Big|\sum_{t=1}^{n}(X_{1t}^2-1)\Big| > x\Big) \sim 2\overline{\Phi}(x/\sqrt{\operatorname{var}(X^2)})\,, \quad n\to\infty\,, \tag{3.46}$$

if $\mathbb{E}\big[\exp\big(|X_{11}^2-1|^{4\alpha/(2\alpha+1)}\big)\big] < \infty$. We apply this result with $\alpha = 1/6$. Then the latter moment requirement reads $\mathbb{E}\big[\exp\big(|X_{11}^2-1|^{1/2}\big)\big] < \infty$ which in view of Lemma 3.20 is implied by (C3'). By definition of $\widetilde{d}_p$ and $p = \exp(o(n^{1/3}))$, we have

$$\frac{\sqrt{n}}{\widetilde{d}_p^2} \sim \frac{\sqrt{n}}{4\log p} > \frac{n^{1/6}}{\rho(n)} \tag{3.47}$$

for any $\rho(n) \to \infty$. Using (3.47), applying Mill's ratio and (3.46) yield for a sequence $\rho(n) \to \infty$ sufficiently slowly that as $n \to \infty$

$$\Psi_n \le p\,\mathbb{P}\Big(\frac{1}{\sqrt{n}}\Big|\sum_{t=1}^{n}(X_{1t}^2-1)\Big| > \frac{n^{1/6}}{\rho(n)}\Big) \sim 2p\,\overline{\Phi}\Big(\frac{n^{1/6}}{\rho(n)\sqrt{\operatorname{var}(X^2)}}\Big) \to 0\,.$$

Finally, assume (C2') and $p = \exp(o(n^{1/3} \wedge g_n^2))$. We can proceed in the same way as under $(C1)$. By Lemma 3.20, we have $\mathbb{E}[\exp(g(|X|))] < \infty$. For any $\rho(n) \to \infty$ we have

$$\frac{\sqrt{n}}{\widetilde{d}_p^2} > \frac{n^{1/6}}{\rho(n)} \ge \frac{n^{1/6} \wedge g_n'}{\rho(n)}\,.$$

An application of [87, Theorem 3] shows that $\Psi_n \to 0$. The proof is complete.

Lemma 3.20. *Let $Z, Z' \ge 0$ be i.i.d. random variables, h a positive constant and g an increasing function on $(0,\infty)$ such that $\mathbb{E}[\exp(g(hZZ'))] < \infty$. Then we have $\mathbb{E}[\exp(g(Z))] < \infty$.*

Proof. If Z is bounded, the claim is trivial. Otherwise there exists $\alpha > 1/h$ such that $\mathbb{P}(Z \le \alpha) < 1$. Writing F for the distribution function of Z, we have

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{g(Z)}](1-F(\alpha)) &= \int_{\alpha}^{\infty} \mathbb{E}[\mathrm{e}^{g(Z)}]\,\mathrm{d}F(t)\\ &\le \int_{\alpha}^{\infty} \mathbb{E}[\mathrm{e}^{g(Z\,ht)}]\,\mathrm{d}F(t) \le \mathbb{E}[\mathrm{e}^{g(hZZ')}].\end{aligned}$$

This implies $\mathbb{E}[\mathrm{e}^{g(Z)}] \le \mathbb{E}[\mathrm{e}^{g(hZZ')}]/(1-F(\alpha)) < \infty$. □

Chapter 4

Fitting inhomogeneous phase–type distributions to data: the univariate and the multivariate case

Hansjörg Albrecher, Mogens Bladt & Jorge Yslas
Submitted for publication

Abstract

The class of inhomogeneous phase-type distributions was recently introduced in [3] as a dense extension of classical phase-type distributions that leads to more parsimonious models in the presence of heavy tails. In this paper we propose a fitting procedure for this class to given data. We furthermore consider an analogous extension of Kulkarni's multivariate phase-type class [73] to the inhomogeneous framework and study parameter estimation for the resulting new and flexible class of multivariate distributions. As a by-product, we amend a previously suggested fitting procedure for the homogeneous multivariate phase-type case and provide appropriate adaptations for censored data. The performance of the algorithms is illustrated in several numerical examples, both for simulated and real-life insurance data.

Keywords: heavy tails; inhomogeneous phase-type; matrix Pareto distribution; matrix Weibull distribution; multivariate phase-type; parameter estimation

4.1 Introduction

The development, study and fitting of flexible distributions for random phenomena is an important branch of applied probability and statistics. Some respective approaches are based on a nice blend of theory and practice, among which the class of phase–type (PH) distributions is a prominent example. Originally initiated by Neuts [91], the realization of a (univariate) phase–type distributed random variable is interpreted as the time until absorption of a time–homogeneous, finite state–space Markov jump process with one absorbing state and the rest being transient. The explicit description through matrix exponentials makes the resulting class of distributions at the same time versatile and analytically tractable (see e.g. [24] for a recent survey). The class of phase–type distributions is known to be dense (in the sense of weak convergence) among all distributions on the positive halfline, but for distributions whose shape is very different from combinations of exponential components (which are the building blocks of the probabilistic Markov jump process construction), a suitable phase–type approximation will need a large dimension of the involved matrix (representing the number of phases of the underlying Markov process) and – in addition to computational challenges – may then be seen unnatural. This is particularly the case for heavy–tailed distributions, where the focus in modelling often lies on the tail of the distribution, and the latter is not well captured by the combination of exponential components of the PH construction. After some first amendment involving infinite–dimensional matrices was suggested in Bladt et al. [26, 27], recently a new way to circumvent this problem was proposed in [3]. Concretely, when the Markov jump process is allowed to be time–inhomogeneous, one gains a lot of flexibility in terms of the structure of the individual components entering the matrix framework, which can reduce the complexity of appropriate fitting distributions drastically, in particular for distributions with heavy tails. The intensity matrices of the Markov jump process are then a function of time. In the general case, they may not commute at different time epochs, which complicates their statistical estimation due to a lack of appropriate sufficient statistics. However, there is an important sub–class for which the intensity matrices can be written as a constant matrix scaled by some real function. In this class all matrices commute, and it was shown in [3] that along this way one in fact obtains, for instance, Pareto, Weibull and Generalized Extreme Value (GEV) distributions with matrix-valued parameters. These distribution classes are all dense in the class of distributions on the positive halfline and inherit the computational advantages of the PH-type class, but also provide excellent fits for heavy–tailed data already for small dimensions, something that the original PH class could not achieve. In particular, if by some preliminary exploratory analysis one has a good guess for an appropriate scaling function (typically suggested by the empirical tail behavior), the resulting matrix distributions can be very parsimoneous yet effective model improvements of the respective base distributions with a genuinely heavy tail. However, while parameter estimation for univariate PH distributions by a standard maximum likelihood procedure based on an EM algorithm has been studied in the seminal paper of Asmussen et al. [10] (see also the later extension of Olsson [95] dealing with censored observations and Bladt et al. [22] for an MCMC approach), parameter estimation for the time–inhomogeneous case has not yet been addressed.

Motivated by the flexibility of the approach, in this paper we will also consider an inhomogeneous extension of the multivariate version of the PH distribution. The multivariate phase–type distribution (of MPH* type) was originally introduced by Kulkarni [73] and is constructed as the joint distribution of certain state-dependent accumulated

rewards earned on the same underlying Markov jump process. It has PH-distributed marginals and also enjoys a denseness property in the class of all distributions on the respective positive orthant. Multivariate phase–type distributions have found applications in diverse areas. For instance, Cai et al. [30] consider them for determining conditional tail expectations in risk management, Cai et al. [31] studied several types of ruin probabilities for a multivariate compound Poisson risk model when the claim size vector follows an MPH* distribution, and Herbertsson [62] used this class to model default contagion in credit risk. More recently, Bladt et al. [25] applied MPH* distributions for the calculation of Parisian type ruin probabilities. In terms of fitting of the (time–homogeneous) multivariate MPH* distribution, Ahlström et al. [1] introduced an algorithm for a bivariate subclass of MPH*, and an EM algorithm for parameter estimation in the general case was proposed in [28]. However, the latter was not actually implemented and contains an inconsistency in the maximum likelihood estimator (which we amend in this paper).

The inhomogeneous extension of the MPH* to be proposed in this paper will then again serve the purpose of keeping the dimension of the involved matrices low when one faces a non-exponential behavior in the marginals and the joint multivariate behaviour. We would like to point out that an alternative analytically tractable deviation from exponential behavior utilizing Mittag-Leffler distributions in both the univariate and multivariate case can be found in [4, 5, 6]. A number of commonly used heavy-tailed multivariate distributions are in fact transformed multivariate exponential distributions. For instance, Mardia [81] was the first to systematically study multivariate Pareto distributions, which he introduced by transforming a Wicksell–Kibble–type multivariate exponential distribution (see [71]). He also noticed that estimation methods for the multivariate exponential can then be translated directly towards the estimation of the multivariate Pareto distribution. Arnold [8] presents some approaches to extend Mardia's analysis to obtain other multivariate distributions with Pareto marginals. Likewise, multivariate versions of the Weibull distribution have been obtained as power transforms of multivariate exponential distributions, see e.g. Lee [74]. The inhomogeneous MPH* extension that we propose in this paper can to some extent be seen as a generalization and unification of these above models.

The main purpose of this paper is to provide algorithms for the statistical fitting of all these flexible classes of distributions and illustrate and discuss their implementation. We will present a unified maximum–likelihood based approach to fitting phase–type distributions (PH), inhomogeneous phase–type distributions (IPH), multivariate phase–type distributions (of MPH* type) and its newly introduced inhomogeneous extension. These classes contain a large number of mathematically tractable distributions that are sufficiently general to fit any non–negative data set, in the body and for both light or heavy tails. We will also consider extensions of the procedures to adapt for censored data and to the fitting of theoretically known joint distributions.

The structure of the paper is as follows. In Section 4.2 we provide an overview of the class of IPH distributions and present a new fitting procedure, which we then exemplify on two particular cases, one on a simulated data set and the other on actual data for lifetimes of the Danish population. In Section 4.3 we shortly recollect some facts about the MPH* class, review existing methods for parameter estimation and provide a substantiation and correction of an algorithm that was previously proposed in the literature. We then extend the algorithm to the case of censored observations, and give more details on an important particular bivariate subclass with explicit density. The section finishes with illustrations of the algorithms for a simulated bivariate sample

as well as a phase–type approximation to a known bivariate exponential distribution. In Section 4.4 we introduce some multivariate extensions to distributions in the IPH class, derive basic properties, provide an EM algorithm for its parameter estimation and again illustrate its use in several examples, including multivariate matrix–Pareto models, multivariate matrix–Weibull models as well as a real data application to a bivariate Danish fire insurance data set. Section 4.5 concludes.

4.2 Inhomogeneous phase–type distributions

4.2.1 Preliminaries

Let $(J_t)_{t\geq 0}$ denote a time–inhomogeneous Markov jump process on a state space $\{1,\ldots,p,p+1\}$, where states $1,\ldots,p$ are transient and state $p+1$ is absorbing. Then $(J_t)_{t\geq 0}$ has an intensity matrix of the form

$$\boldsymbol{\Lambda}(t) = \begin{pmatrix} \boldsymbol{T}(t) & \boldsymbol{t}(t) \\ \boldsymbol{0} & 0 \end{pmatrix}, \quad t \geq 0\,,$$

where $\boldsymbol{T}(t)$ is a $p \times p$ matrix and $\boldsymbol{t}(t)$ is a p–dimensional column vector. Here, for any time $t \geq 0$, $\boldsymbol{t}(t) = -\boldsymbol{T}(t)\,\boldsymbol{e}$, where $\boldsymbol{e}$ is the p–dimensional column vector of ones. Let $\pi_k = \mathbb{P}(J_0 = k)$, $k = 1,\ldots,p$, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_p)$ and assume that $\mathbb{P}(J_0 = p+1) = 0$. Then we say that the time until absorption

$$\tau = \inf\{t \geq 0 \mid J_t = p+1\}$$

has an inhomogeneous phase–type distribution with representation $(\boldsymbol{\pi}, \boldsymbol{T}(t))$ and we write $\tau \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}(t))$. If $\boldsymbol{T}(t) = \lambda(t)\,\boldsymbol{T}$, where $\lambda(t)$ is some known non–negative real function and $\boldsymbol{T}$ is a sub–intensity matrix, then we write $\tau \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$. If $X \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$, then there exists a function g such that

$$X \sim g(Y)\,, \tag{4.1}$$

where $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. Specifically, g is defined by

$$g^{-1}(x) = \int_0^x \lambda(t)dt$$

or, equivalently,

$$\lambda(t) = \frac{d}{dt} g^{-1}(t)\,.$$

The density f_X and distribution function F_X for $X \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$ are given by

$$\begin{aligned} f_X(x) &= \lambda(x)\,\boldsymbol{\pi} \exp\left(\int_0^x \lambda(t)dt\ \boldsymbol{T}\right)\boldsymbol{t}\,, \\ F_X(x) &= 1 - \boldsymbol{\pi} \exp\left(\int_0^x \lambda(t)dt\ \boldsymbol{T}\right)\boldsymbol{e}\,. \end{aligned}$$

For further reading on inhomogeneous phase–type distributions and motivations for their use in modelling we refer to Albrecher & Bladt [3]. Note that for $\lambda(t) \equiv 1$ one returns to the time-homogeneous case, which corresponds to the conventional phase–type distribution with notation $\mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ (a comprehensive account of phase–type distributions can

be found in Bladt & Nielsen [24]).

As illustrated in [3], a number of IPH distributions can be expressed as classical distributions with matrix-valued parameter. For the representation of such distributions we make use of functional calculus. If h is an analytic function and $\boldsymbol{A}$ is a matrix, define

$$h(\boldsymbol{A}) = \frac{1}{2\pi i}\oint_\gamma h(z)(z\boldsymbol{I}-\boldsymbol{A})^{-1}dz\,,$$

where γ is a simple path enclosing the eigenvalues of $\boldsymbol{A}$ (cf. [24, Sec. 3.4] for details). Important examples include the transformation $g(y) = \beta\,(e^y - 1)$ for $\beta > 0$ in (4.1) leading to a matrix–Pareto distribution with density function and survival function

$$f_X(x) = \boldsymbol{\pi}\left(\frac{x}{\beta}+1\right)^{\boldsymbol{T}-\boldsymbol{I}}\boldsymbol{t}\,\frac{1}{\beta}\,,\quad \bar{F}_X(x) = 1-F_X(x) = \boldsymbol{\pi}\left(\frac{x}{\beta}+1\right)^{\boldsymbol{T}}\boldsymbol{e}\,, \tag{4.2}$$

respectively, as well as the matrix–Weibull distribution with density and survival function

$$f_X(x) = \boldsymbol{\pi}e^{\boldsymbol{T}x^\beta}\boldsymbol{t}\beta x^{\beta-1}\,,\quad \bar{F}_X(x) = \boldsymbol{\pi}e^{\boldsymbol{T}x^\beta}\boldsymbol{e}\,,$$

obtained from $g(y) = y^{1/\beta}$ ($\beta > 0$), see [3] for further details.

4.2.2 Parameter estimation

For the matrix–Pareto distribution (4.2) and $\beta = 1$, the transform is parameter-independent, so that the distribution can be fitted to i.i.d. data $x_1,\ldots,x_N$ by fitting a phase–type distribution $\mathrm{PH}(\boldsymbol{\pi},\boldsymbol{T})$ to the transformed data $\log(1+x_1),\ldots,\log(1+x_N)$ using an EM algorithm [10]. This was the procedure employed in [3] for the numerical illustration there. The general case – where the transform does depend on parameters – is more subtle and shall be dealt with here. The key will be to apply a parameter-dependent transformation in each step of the EM algorithm.

Let $x_1,\ldots,x_N$ be an i.i.d. sample of an inhomogeneous phase–type distribution with representation $X \sim \mathrm{IPH}(\boldsymbol{\pi},\boldsymbol{T},\lambda(\,\cdot\,;\boldsymbol{\beta}))$, where $\lambda(\,\cdot\,;\boldsymbol{\beta})$ is a parametric non–negative function depending on the vector $\boldsymbol{\beta}$. We then know that $X \overset{d}{=} g(Y;\boldsymbol{\beta})$ with $Y \sim \mathrm{PH}(\boldsymbol{\pi},\boldsymbol{T})$ and g is defined in terms of its inverse function $g^{-1}(x;\boldsymbol{\beta}) = \int_0^x \lambda(t;\boldsymbol{\beta})dt$. In particular $g^{-1}(X;\boldsymbol{\beta}) \overset{d}{=} Y \sim \mathrm{PH}(\boldsymbol{\pi},\boldsymbol{T})$. The EM algorithm for fitting $\mathrm{IPH}(\boldsymbol{\pi},\boldsymbol{T},\lambda(\,\cdot\,;\boldsymbol{\beta}))$ then works as follows.

Algorithm 4.1 (EM algorithm for transformed phase–type distributions)**.**

0. Initialize with some "arbitrary" $(\boldsymbol{\pi},\boldsymbol{T},\boldsymbol{\beta})$.

1. Transform the data into $y_i = g^{-1}(x_i;\boldsymbol{\beta})$, $i = 1,\ldots,N$, *and apply the E– and M–steps of the conventional EM algorithm of Asmussen [10] by which we obtain the estimators* $(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{T}})$.

2. Compute

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^N \log(f_X(x_i;\hat{\boldsymbol{\pi}},\hat{\boldsymbol{T}},\boldsymbol{\beta}))\\ &= \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^N \log\left(\lambda(x_i;\boldsymbol{\beta})\hat{\boldsymbol{\pi}}\exp\left(\int_0^{x_i}\lambda(t;\boldsymbol{\beta})dt\;\hat{\boldsymbol{T}}\right)\hat{\boldsymbol{t}}\right)\,.\end{aligned}$$

3. Assign $(\boldsymbol{\pi},\boldsymbol{T},\boldsymbol{\beta}) = (\hat{\boldsymbol{\pi}},\hat{\boldsymbol{T}},\hat{\boldsymbol{\beta}})$ *and GOTO 1.*

Then the likelihood function increases for each iteration, and hence converges to a (possibly local) maximum.

Proof. Since the data points x_i are assumed to be i.i.d. realisations from the unknown distribution $\mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$, there exists a function g such that $y_i = g^{-1}(x_i; \boldsymbol{\beta})$ are i.i.d. realisations of phase–type distributed random variables $\mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. That function g is assumed to be known up to the value of $\boldsymbol{\beta}$. In turn, $x_i = g(y_i; \boldsymbol{\beta})$, so a data point x_i can be interpreted as the absorption time of the Markov jump process corresponding to $\mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$, which is y_i, but with the scale of the time axis for the y_i–data converted (stretched) into $g(\,\cdot\,; \boldsymbol{\beta})$–coordinates instead. The full data likelihood is then given by

$$L(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}; \boldsymbol{y}) \;=\; \prod_{k=1}^{p} \pi_k^{B_k} \prod_{k=1}^{p} \prod_{l \neq k} t_{kl}^{N_{kl}} e^{-t_{kl} Z_k(\boldsymbol{\beta})} \prod_{k=1}^{p} t_k^{N_k} e^{-t_k Z_k(\boldsymbol{\beta})} \,,$$

where B_k is the number of times the Markov process underlying the phase–type distribution initiates in state k, N_{kl} denotes the total number of transitions from state k to l, N_k denotes the number of times an exit to the absorbing state was caused by a jump from state k, and $Z_k(\boldsymbol{\beta})$ is the total time the Markov process has spent in state k. We notice that $Z_k(\boldsymbol{\beta})$ is the only sufficient statistic which depends on the transformation of the time axis for the y–data and hence on $\boldsymbol{\beta}$. Consequently, for any given $\boldsymbol{\beta}$, the E–step is simply the one as in [10], and so is the M–step for $(\boldsymbol{\pi}, \boldsymbol{T})$.

The $\boldsymbol{\beta}$ update in 2. requires a general, usually numerical, maximization of the incomplete data likelihood. Each iteration of the algorithm increases the likelihood. Indeed, let L^I denote the incomplete data likelihood and consider parameter values $(\boldsymbol{\pi}_n, \boldsymbol{T}_n, \boldsymbol{\beta}_n)$ after the n-th iteration. In the $(n+1)$-th iteration, we first obtain $(\boldsymbol{\pi}_{n+1}, \boldsymbol{T}_{n+1})$ in 1. so that

$$L^I(\boldsymbol{\pi}_n, \boldsymbol{T}_n, \boldsymbol{\beta}_n; g^{-1}(\boldsymbol{y}; \boldsymbol{\beta}_n)) \leq L^I(\boldsymbol{\pi}_{n+1}, \boldsymbol{T}_{n+1}, \boldsymbol{\beta}_n; g^{-1}(\boldsymbol{y}; \boldsymbol{\beta}_n)) \,.$$

By monotonicity of g and the transformation theorem,

$$L^I(\boldsymbol{\pi}_n, \boldsymbol{T}_n, \boldsymbol{\beta}_n; \boldsymbol{y}) \leq L^I(\boldsymbol{\pi}_{n+1}, \boldsymbol{T}_{n+1}, \boldsymbol{\beta}_n; \boldsymbol{y})$$

and hence

$$\begin{aligned} L^I(\boldsymbol{\pi}_n, \boldsymbol{T}_n, \boldsymbol{\beta}_n; \boldsymbol{y}) &\leq L^I(\boldsymbol{\pi}_{n+1}, \boldsymbol{T}_{n+1}, \boldsymbol{\beta}_n; \boldsymbol{y}) \\ &\leq \sup_{\boldsymbol{\beta}} L^I(\boldsymbol{\pi}_{n+1}, \boldsymbol{T}_{n+1}, \boldsymbol{\beta}; \boldsymbol{y}) = L^I(\boldsymbol{\pi}_{n+1}, \boldsymbol{T}_{n+1}, \boldsymbol{\beta}_{n+1}; \boldsymbol{y}) \,. \end{aligned}$$

□

Example 4.2. (Matrix–Gompertz) Let $X = \log(\beta Y + 1)/\beta$, where $Y \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ and $\beta > 0$. Then

$$\bar{F}_X(x) = \boldsymbol{\pi} e^{\boldsymbol{T}(e^{\beta x} - 1)/\beta} \boldsymbol{e} \quad \text{and} \quad f_X(x) = \boldsymbol{\pi} e^{\boldsymbol{T}(e^{\beta x} - 1)/\beta} \boldsymbol{t} e^{\beta x} \,. \tag{4.3}$$

We refer to the distribution of X as a *matrix–Gompertz distribution*, since the scale parameter of the usual Gompertz distribution is now replaced by a matrix. Note that the resulting distribution has a lighter tail than a conventional phase–type distribution. The Gompertz distribution is used in a number of applications, most notably it is historically used for the modelling of human lifetimes [55]. Its matrix version (4.3) provides a natural flexible extension. As an illustration, we fitted a matrix–Gompertz distribution with 3 phases using Algorithm 4.1 with 2 500 iterations to the lifetime of the Danish population

that died in the year 2 000 at ages 50 to 100 (data obtained from the Human Mortality Database (HMD) and available in the R-package *MortalitySmooth* [35]). Here and in later examples, the number of iterations in the algorithm is chosen in such a way that the changes in the successive log–likelihoods become negligible. Concerning running times, our implementation makes use of the gradient ascent method for the maximization part of the algorithm, in which the running times highly depend on the step–length and the actually chosen stopping criterion. In the present example we employed a step–length of 10^{-8} and run gradient ascent until the absolute value of the derivative is less than 0.001 leading to a running time of about 35 seconds on a usual PC (with 2.9 GHz Dual–Core Intel Core i5 processor 5287U) for the 2 500 iterations of the EM algorithm. Note that this choice of stopping criterion is to prioritize precision over speed, and an improvement on running times can be attained by using a different maximization procedure. The obtained parameters are as follows:

$$\hat{\boldsymbol{\pi}} = (0.0450,\, 0.1303,\, 0.8246)\,,$$
$$\hat{\boldsymbol{T}} = \begin{pmatrix} -0.1357 & 0.1214 & 0 \\ 0.0130 & -0.0421 & 0.0288 \\ 0.1415 & 0.0184 & -0.1620 \end{pmatrix},$$
$$\hat{\beta} = 0.1019\,.$$

Figure 4.1 shows that the fitted density recovers the structure of the data quite well. Note that conventional phase–type distributions have been used to model the distribution of lifespans (see for instance Asmussen at al. [9]). However, the number of phases required to capture the tail behavior of the data with the latter is rather large, due to the lighter than exponential tail. In contrast, the matrix–Gompertz distribution provides an excellent fit with comparably few parameters (phases).

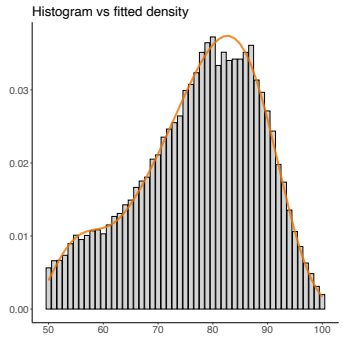

Figure 4.1: Histogram of lifetimes of the Danish population that died in the year 2 000 at ages 50 to 100 versus the density of the fitted matrix–Gompertz distribution.

Example 4.3. (Matrix–GEV) Algorithm 4.1 can also be applied to estimate distributions that are not IPH in a strict sense, but that are defined as a transformation of a PH distribution. This is for instance the case for $g(y) = \mu - \sigma(y^{-\xi} - 1)/\xi$ with $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$. Recall from [3] that

$$F_X(x) = \begin{cases} \boldsymbol{\pi} \exp\left(\boldsymbol{T}\left(1 + \xi\dfrac{x-\mu}{\sigma}\right)^{-1/\xi}\right)\boldsymbol{e}\,, & \xi \neq 0\,, \\ \boldsymbol{\pi} \exp\left(\boldsymbol{T}\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right)\boldsymbol{e}\,, & \xi = 0\,, \end{cases}$$

$$f_X(x) = \begin{cases} \frac{1}{\sigma}\boldsymbol{\pi} \exp\left(\boldsymbol{T}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right)\boldsymbol{t}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-(1+\xi)/\xi}, & \xi \neq 0\,, \\ \frac{1}{\sigma}\boldsymbol{\pi} \exp\left(\boldsymbol{T}\exp\left(-\frac{x-\mu}{\sigma}\right)\right)\boldsymbol{t}\exp\left(-\frac{x-\mu}{\sigma}\right), & \xi = 0\,, \end{cases}$$

from which it becomes clear that this distribution can be interpreted as a matrix version of the generalized extreme value (GEV) distribution, see e.g. [16]. As an illustration, we generated an i.i.d. sample of size 5 000 from such a distribution of 3 phases with parameters

$$\begin{gathered} \boldsymbol{\pi} = (1,\, 0,\, 0)\,, \\ \boldsymbol{T} = \begin{pmatrix} -1 & 0.5 & 0 \\ 0.2 & -2 & 0.8 \\ 1 & 1 & -5 \end{pmatrix}, \\ \mu = 2\,, \quad \sigma = 0.5\,, \quad \xi = 0.4\,, \end{gathered}$$

which has theoretical moments $\mathbb{E}(X) = 2.2524$ and $SD(X) = 1.4423$. The generated sample has moments $\hat{\mathbb{E}}(X) = 2.2607$ and $\hat{SD}(X) = 1.3307$. We then fitted such a matrix–GEV distribution with the same number phases using Algorithm 4.1 with 1 500 steps, obtaining the following parameters:

$$\begin{gathered} \hat{\boldsymbol{\pi}} = (0.0772,\, 0.1268,\, 0.7960)\,, \\ \hat{\boldsymbol{T}} = \begin{pmatrix} -8.9772 & 0.0964 & 0.0001 \\ 0.2891 & -2.8439 & 0.3542 \\ 3.2353 & 0.0137 & -5.7731 \end{pmatrix}, \\ \hat{\mu} = 1.3852\,, \quad \hat{\sigma} = 0.2285\,, \quad \hat{\xi} = 0.4251\,. \end{gathered}$$

We observe that the algorithm estimates pretty well the shape parameter ξ, which determines the heaviness of the tail. Moreover, the fitted distribution has moments $\mathbb{E}(X) = 2.2640$ and $SD(X) = 1.6587$, which resemble the ones of the sample, and Figure 4.2 shows that the algorithm recovers both body and tail of the data. Note also that the log–likelihood of the fitted matrix–GEV is $-4\,104.541$, while the log–likelihood using the original matrix–GEV distribution is $-4\,107.005$. Such a comparison of the log–likelihoods works as an additional evidence of the performance of the algorithm. One can observe that the parameters estimated for $\boldsymbol{\pi}$, $\boldsymbol{T}$, μ and σ do not resemble the original parameter values, but this is linked with the well-known identifiability issue for phase–type distributions (namely that other parameter combinations may lead to a very similar density shape). In fact, the algorithm finds the parameters that maximize the likelihood for the given sample, and as the concrete numbers above show, the present parameters even outperform the original model underlying the sample(!), see also the convincing QQ–plot in Figure 4.2. Here, the step–length is 10^{-5} and the gradient ascent is run until the norm of the derivative is less than 0.1 leading to a running time of 2 609 seconds for the 1 500 iterations of Algorithm 4.1.

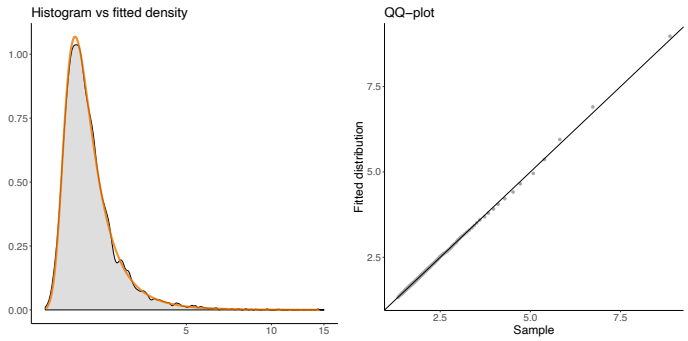

Figure 4.2: Histogram of simulated sample versus density of the fitted matrix–GEV distribution in log–scale (left) as well as QQ–plot of simulated sample versus fit (right).

4.3 Multivariate phase-type distributions

4.3.1 Preliminaries

Let $\tau \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$ be a (conventional) p–dimensional phase–type distributed random variable with underlying time–homogeneous Markov jump process $(J_t)_{t\geq 0}$. Let $\boldsymbol{r}_j = (r_j(1), \ldots, r_j(p))'$ be non–negative p–dimensional column vectors, $j = 1, \ldots, d$, and let

$$\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_d)$$

be a $p \times d$–dimensional *reward matrix*. Now define

$$Y^{(j)} = \int_0^\tau r_j(J_t)\, dt$$

for all $j = 1, \ldots, d$. If we interpret $r_j(k)$ as the rate at which a reward is obtained while J_t is in state k, then $Y^{(j)}$ is the total reward for component j obtained prior to absorption. We then say that the random vector $\boldsymbol{Y} = \left(Y^{(1)}, \ldots, Y^{(d)}\right)'$ has a multivariate phase–type distribution of the MPH* type (as defined in [73], see also [24]) and we write $\boldsymbol{Y} \sim \mathrm{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$.

While each member of the MPH* class has an explicit expression for the (joint) Laplace transform and the joint moments of any order (see Section 8.1.1 of [24]), there are no general explicit expressions for the density and distribution functions. However, for certain structures and sub–classes explicit expressions for the latter do exist (like Example 8.1.13 of [24]).

If $\boldsymbol{Y} = \left(Y^{(1)}, \ldots, Y^{(d)}\right)' \sim \mathrm{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$, then each marginal $Y^{(j)}$ has a phase–type distribution, $\mathrm{PH}(\boldsymbol{\pi}_j, \boldsymbol{T}_j)$ say. First we decompose

$$\boldsymbol{r}_j = \begin{pmatrix} \boldsymbol{r}_j^+ \\ \boldsymbol{r}_j^0 \end{pmatrix}, \quad \boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}^+ & \boldsymbol{\pi}^0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}^{++} & \boldsymbol{T}^{+0} \\ \boldsymbol{T}^{0+} & \boldsymbol{T}^{00} \end{pmatrix},$$

where we have reordered the state space such that the $+$ terms correspond to the states k for which the rewards $r_k^{(j)}$ are strictly positive, and the 0 terms to the states with zero rewards. E.g., $\boldsymbol{T}_{+0}$ corresponds to the intensities by which the underlying Markov jump process $(J_t)_{t\geq 0}$ jumps from a state with positive reward to a state with zero reward.

Then the phase–type distribution of $Y^{(j)}$ is given by an atom at zero of size $\boldsymbol{\pi}_0(\boldsymbol{I} - (-\boldsymbol{T}_{00})^{-1}\boldsymbol{S}_{0+})\,\boldsymbol{e}$, where $\boldsymbol{e}$ is the column vector of ones of appropriate dimension, and

$$\boldsymbol{\pi}_j = \boldsymbol{\pi}^{+} + \boldsymbol{\pi}^{0}(-\boldsymbol{T}^{00})^{-1}\boldsymbol{T}^{0+} \quad \text{and} \quad \boldsymbol{T}_j = \boldsymbol{\Delta}(\boldsymbol{r}_j^{+})^{-1}\left(\boldsymbol{T}^{++} + \boldsymbol{T}^{+0}(-\boldsymbol{T}_{00})^{-1}\boldsymbol{T}^{0+}\right), \quad (4.4)$$

where $\Delta(\boldsymbol{a})$ denotes the $d' \times d'$ diagonal matrix with entries $a^{(m)}$, $m = 1, \ldots, d'$, from a d'–dimensional vector $\boldsymbol{a}$. The atom appears in case there is a positive probability of starting in a non–reward–earning state (0) and the underlying Markov process gets absorbed before visiting a reward earning state (+). The Markov jump process generating $Y^{(j)}$ starts in the same state as J_t if the reward is positive (hence $\boldsymbol{\pi}_+$) or it starts in the first state with positive rewards that J_t enters after starting in a zero reward state (hence the term $\boldsymbol{\pi}^0(-\boldsymbol{T}^{00})^{-1}\boldsymbol{T}^{0+}$). Similar arguments apply to the generator $\boldsymbol{T}_j$, where only reward-earning terms will form part of the state space for $Y^{(j)}$. We refer to [24] for further details.

Summarizing, each marginal $Y^{(j)}$ has a phase–type distribution, which is based on the original Markov process $(J_t)_{t\geq 0}$, but with a possibly smaller state space and with rescaled parameters.

4.3.2 Parameter estimation

We next provide an algorithm for estimating MPH* distributed data. The data consist of a d–dimensional multivariate sample of N i.i.d. observations

$$\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})', \quad i = 1, \ldots, N\,.$$

That is, we only observe the times to absorption, $y_i^{(j)}$, of each phase–type distributed marginal. Hence we are clearly in an incomplete data set–up and we shall employ the EM algorithm for fitting $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$.

The EM algorithm works by replacing unavailable sufficient statistics by their conditional expectations given data under given parameters, and thereby updating the parameters by using known formulas for the maximum likelihood estimator in the complete data domain. Iteration of the procedure then produces a sequence of parameter values which increases the likelihood in each step.

For the present situation, we define the complete data as both the trajectories of the underlying Markov process which generates the phase–type distribution from which the marginals of the multivariate vector are constructed, and the Markov jump processes representing the rewards in all marginal distributions. It is not sufficient with complete knowledge of the marginal trajectories only. Indeed, one can easily construct examples where the underlying processes cannot be reconstructed from the marginals only. The complete knowledge of both marginals and the underlying Markov process which generates the marginals creates another problem in relation to the incomplete data since we do not have observations for the absorption times of the underlying Markov process. We can get around this problem by assuming that the rows of the reward matrix $\boldsymbol{R}$ sum to one, i.e. $\boldsymbol{Re} = \boldsymbol{e}$. This assumption is not restrictive and can be imposed without losing generality due to the great ambiguity of (multivariate) phase–type representations. Hence our data consists of marginals $\boldsymbol{y}^{(j)} = (y_1^{(j)}, \ldots, y_N^{(j)})'$, $j = 1, \ldots, d$, and their sums $\boldsymbol{y}^{(S)} = \sum_{l=1}^{d} \boldsymbol{y}^{(l)}$.

In the complete data domain, the estimation is straightforward and works as follows. The complete data MLE for $(\boldsymbol{\pi}, \boldsymbol{T})$ is given by

$$\hat{\pi}_k = \frac{B_k}{N}\,, \quad \hat{t}_{kl} = \frac{N_{kl}}{Z_k}\,, \quad \hat{t}_k = \frac{N_k}{Z_k}\,, \quad \hat{t}_{kk} = -\sum_{l\neq k} \hat{t}_{kl} - \hat{t}_k\,.$$

The rewards of the marginals are then given by

$$\hat{r}_j(k) = \frac{Z_k^{(j)}}{Z_k} = \frac{Z_k^{(j)}}{\sum_{l=1}^d Z_k^{(l)}},$$

where $Z_k^{(j)}$ is the over-all amount of time the j'th component has spent in state k.

In the EM algorithm, we now must replace all aforementioned sufficient statistics by their conditional expectations given data. Concerning Z_k, N_{kl}, N_k and B_k, these only depend on the underlying Markov jump process and are computed conditionally on $\boldsymbol{y}^{(S)}$ only. Their formulas are then as stated in the algorithm below (see [10]).

Concerning the conditional expectation of $Z_k^{(j)}$, we must calculate the expected reward (under $(\boldsymbol{\pi}, \boldsymbol{T})$) given all data of marginal j, which amounts to calculating the conditional expected time given data for the corresponding phase–type representation of the j-th marginal, $(\boldsymbol{\pi}_j, \boldsymbol{T}_j)$. These are readily given by (again using [10])

$$\mathbb{E}\left(Z_k^{(j)} \mid \boldsymbol{Y}^{(j)} = \boldsymbol{y}^{(j)}\right) = \sum_{i=1}^N \frac{\int_0^{y_i^{(j)}} \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}_j (y_i^{(j)} - u)} \boldsymbol{t}_j \boldsymbol{\pi}_j \mathrm{e}^{\boldsymbol{T}_j u} \boldsymbol{e}_k du}{\boldsymbol{\pi}_j \mathrm{e}^{\boldsymbol{T}_j y_i^{(j)}} \boldsymbol{t}_j}.$$

Then

$$\hat{r}_j(k) = \frac{\mathbb{E}\left(Z_k^{(j)} \mid \boldsymbol{Y}^{(j)} = \boldsymbol{y}^{(j)}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)}.$$

Iterating the above finally provides a (single) full EM algorithm for the estimation of $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$. We summarize the results in the following.

Algorithm 4.4 (EM algorithm for MPH* distributions)**.**

0. Initialize with some "arbitrary" $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ *with* $\boldsymbol{Re} = \boldsymbol{e}$*, and compute* $\boldsymbol{\pi}_j$ *and* $\boldsymbol{T}_j$*,* $j = 1, \ldots, d$*, using* (4.4)*.*

1. (E–step) Calculate

$$\mathbb{E}\left(B_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^N \frac{\pi_k \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T} y_i^{(S)}} \boldsymbol{t}}{\boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} y_i^{(S)}} \boldsymbol{t}}$$

$$\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^N \frac{\int_0^{y_i^{(S)}} \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}(y_i^{(S)} - u)} \boldsymbol{t} \boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} u} \boldsymbol{e}_k du}{\boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} y_i^{(S)}} \boldsymbol{t}}$$

$$\mathbb{E}\left(N_{kl} \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^N t_{kl} \frac{\int_0^{y_i^{(S)}} \boldsymbol{e}_l' \mathrm{e}^{\boldsymbol{T}(y_i^{(S)} - u)} \boldsymbol{t} \boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} u} \boldsymbol{e}_k du}{\boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} y_i^{(S)}} \boldsymbol{t}}$$

$$\mathbb{E}\left(N_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^N t_k \frac{\boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} y_i^{(S)}} \boldsymbol{e}_k}{\boldsymbol{\pi} \mathrm{e}^{\boldsymbol{T} y_i^{(S)}} \boldsymbol{t}}$$

$$\mathbb{E}\left(Z_k^{(j)} \mid \boldsymbol{Y}^{(j)} = \boldsymbol{y}^{(j)}\right) = \sum_{i=1}^N \frac{\int_0^{y_i^{(j)}} \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}_j (y_i^{(j)} - u)} \boldsymbol{t}_j \boldsymbol{\pi}_j \mathrm{e}^{\boldsymbol{T}_j u} \boldsymbol{e}_k du}{\boldsymbol{\pi}_j \mathrm{e}^{\boldsymbol{T}_j y_i^{(j)}} \boldsymbol{t}_j}.$$

2. (M–step) Let

$$\hat{\alpha}_k = \frac{1}{N}\mathbb{E}\left(B_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right),\ \hat{t}_{kl} = \frac{\mathbb{E}\left(N_{kl} \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)},\ \hat{t}_k = \frac{\mathbb{E}\left(N_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)},$$

$$\hat{t}_{kk} = -\sum_{l\neq k}\hat{t}_{kl} - \hat{t}_k\,,\quad \hat{\boldsymbol{\pi}} = (\hat{\pi}_1,\ldots,\hat{\pi}_p)\,,\quad \hat{\boldsymbol{T}} = \{\hat{t}_{kl}\}_{k,l=1,\ldots,p} \text{ and } \hat{\boldsymbol{t}} = (\hat{t}_1,\ldots,\hat{t}_p)'.$$

and

$$\hat{r}_j(k) := \frac{\mathbb{E}\left(Z_k^{(j)} \mid \boldsymbol{Y}^{(j)} = \boldsymbol{y}^{(j)}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)} \quad \text{and} \quad \hat{\boldsymbol{R}} = \{\hat{r}_j(k)\}_{k=1,\ldots,p,j=1,\ldots,d}\,.$$

3. Assign $\boldsymbol{\pi} := \hat{\boldsymbol{\pi}}$, $\boldsymbol{T} := \hat{\boldsymbol{T}}$, $\boldsymbol{t} := \hat{\boldsymbol{t}}$, $\boldsymbol{R} := \hat{\boldsymbol{R}}$ *and compute* $\boldsymbol{\pi}_j$, $\boldsymbol{T}_j$, $j = 1,\ldots,d$, *using* (4.4). *GOTO 1.*

Remark 4.5. Algorithm 4.4 was originally proposed in [28] as two consecutive EM algorithms and its original statement contained a minor error in the M–step update for the reward matrix. To see why Algorithm 4.4 can be decomposed into the two consecutive EM algorithms, we argue as follows. Running the EM Algorithm 4.4, $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}})$ will eventually converge (without input from the part involving the reward components). For constant $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}})$, Algorithm 4.4 is indeed equivalent to the second EM algorithm in [28]. More specifically, the algorithm takes the following form.

First EM.

0. Initialize with some "arbitrary" $(\boldsymbol{\pi}, \boldsymbol{T})$.
1. (E–step) Calculate

$$\mathbb{E}\left(B_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^{N}\frac{\pi_k \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}y_i^{(S)}}\boldsymbol{t}}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}y_i^{(S)}}\boldsymbol{t}}$$

$$\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^{N}\frac{\displaystyle\int_0^{y_i^{(S)}} \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}(y_i^{(S)}-u)}\boldsymbol{t}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}y_i^{(S)}}\boldsymbol{t}}$$

$$\mathbb{E}\left(N_{kl} \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^{N} t_{kl}\frac{\displaystyle\int_0^{y_i^{(S)}} \boldsymbol{e}_l' \mathrm{e}^{\boldsymbol{T}(y_i^{(S)}-u)}\boldsymbol{t}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}y_i^{(S)}}\boldsymbol{t}}$$

$$\mathbb{E}\left(N_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right) = \sum_{i=1}^{N} t_k\frac{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}y_i^{(S)}}\boldsymbol{e}_k}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}y_i^{(S)}}\boldsymbol{t}}\,.$$

2. (M–step) Let

$$\hat{\alpha}_k = \frac{1}{N}\mathbb{E}\left(B_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right),\ \hat{t}_{kl} = \frac{\mathbb{E}\left(N_{kl} \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)},\ \hat{t}_k = \frac{\mathbb{E}\left(N_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y}^{(S)} = \boldsymbol{y}^{(S)}\right)},$$

$$\hat{t}_{kk} = -\sum_{l\neq k}\hat{t}_{kl} - \hat{t}_k\,,\quad \hat{\boldsymbol{\pi}} = (\hat{\pi}_1,\ldots,\hat{\pi}_p)\,,\quad \hat{\boldsymbol{T}} = \{\hat{t}_{kl}\}_{k,l=1,\ldots,p} \quad \text{and} \quad \hat{\boldsymbol{t}} = (\hat{t}_1,\ldots,\hat{t}_p)'.$$

3. Assign $\boldsymbol{\pi} := \hat{\boldsymbol{\pi}}$, $\boldsymbol{T} := \hat{\boldsymbol{T}}$, $\boldsymbol{t} := \hat{\boldsymbol{t}}$ and GOTO 1.

Second EM. Use the estimated $(\boldsymbol{\pi}, \boldsymbol{T})$ of the first EM.

0. Initialize with some "arbitrary" $\boldsymbol{R}$ with $\boldsymbol{Re} = \boldsymbol{e}$, and compute $\boldsymbol{\pi}_j$ and $\boldsymbol{T}_j$, $j = 1, \ldots, d$, using (4.4).

1. (E–step) Calculate

$$\mathbb{E}\left(Z_k^{(j)} \mid \boldsymbol{Y}^{(j)} = \boldsymbol{y}^{(j)}\right) = \sum_{i=1}^{N} \frac{\int_0^{y_i^{(j)}} \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}_j (y_i^{(j)} - u)} \boldsymbol{t}_j \boldsymbol{\pi}_j \mathrm{e}^{\boldsymbol{T}_j u} \boldsymbol{e}_k du}{\boldsymbol{\pi}_j \mathrm{e}^{\boldsymbol{T}_j y_i^{(j)}} \boldsymbol{t}_j} .$$

2. (M–step) Let

$$\hat{r}_j(k) := \frac{\mathbb{E}\left(Z_k^{(j)} \mid \boldsymbol{Y}^{(j)} = \boldsymbol{y}^{(j)}\right)}{\sum_{l=1}^{d} \mathbb{E}\left(Z_k^{(l)} \mid \boldsymbol{Y}^{(l)} = \boldsymbol{y}^{(l)}\right)} \quad \text{and} \quad \hat{\boldsymbol{R}} = \{\hat{r}_j(k)\}_{k=1,\ldots,p, j=1,\ldots,d} .$$

3. Assign $\boldsymbol{R} := \hat{\boldsymbol{R}}$ and compute $\boldsymbol{\pi}_j$ and $\boldsymbol{T}_j$, $j = 1, \ldots, d$, using (4.4). GOTO 1.

Remark 4.6. The main computational burden lies in the E–steps, where matrix exponentials and integrals thereof must be evaluated. In Asmussen et al. [10] this is done by converting the problem into a system of ODEs, which are then solved via a Runge–Kutta method of fourth order (a C implementation, called EMpht, is available online [96]). While this approach is adequate for fitting univariate phase–type distributions, the Runge–Kutta method fails to work in some cases in the multivariate setting, in particular for the second EM, when an element in the reward matrix approaches zero. The reason is that the sub-intensity matrix of (at least) one of the marginals will adjust to this change by increasing some of the entries of the matrix in each iteration, and thus requiring an increasingly smaller step–size in the Runge–Kutta method to accurately approximate the solution to the system. Our implementation includes an approach for the computation of matrix exponentials based on uniformization, and it is a slight variation of the method in Neuts [93, p.232]. We explain briefly the method. By taking $\phi = \max(-t_{kk})_{k=1,\ldots,p}$ and defining $\boldsymbol{P} := \phi^{-1}(\phi \boldsymbol{I} + \boldsymbol{T})$, which is in fact a transition matrix, we have that

$$\exp(\boldsymbol{T}y) = \sum_{n=0}^{\infty} \frac{(\phi y)^n}{n!} e^{-\phi y} \boldsymbol{P}^n .$$

Then

$$\begin{aligned} \left| e^{\boldsymbol{T}y} - \sum_{n=0}^{M} \frac{(\phi y)^n}{n!} e^{-\phi y} \boldsymbol{P}^n \right| &\leq \sum_{n=M+1}^{\infty} \frac{(\phi y)^n}{n!} e^{-\phi y} |\boldsymbol{P}^n| \\ &\leq \sum_{n=M+1}^{\infty} \frac{(\phi y)^n}{n!} e^{-\phi y} = \mathbb{P}(N_{\phi y} > M) , \end{aligned}$$

where $N_{\phi y}$ is Poisson distributed with mean ϕy. Hence, we can find M such that the difference of the matrix exponential with a finite sum is less than or equal to a given error $\epsilon > 0$. Of course, larger values of ϕy give bigger values of M, dismissing any computational improvement for large observations. A way to circumvent this problem is to observe that $e^{\boldsymbol{T}y} = (e^{\boldsymbol{T}y/2^m})^{2^m}$, thus we can find m such that $\phi y/2^m < 1$, compute $e^{\boldsymbol{T}y/2^m}$ by a finite sum and then retrieve $e^{\boldsymbol{T}y}$ by squaring.

To compute the integrals involving matrix exponentials, we observe that by defining

$$\boldsymbol{G}(y;\boldsymbol{\pi},\boldsymbol{T}) := \int_0^y \mathrm{e}^{\boldsymbol{T}(y-u)}\boldsymbol{t}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}du\,,$$

we have that (see Van Loan [109])

$$\exp\left(\begin{pmatrix} \boldsymbol{T} & \boldsymbol{t}\boldsymbol{\pi} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} y\right) = \begin{pmatrix} \mathrm{e}^{\boldsymbol{T}y} & \boldsymbol{G}(y;\boldsymbol{\pi},\boldsymbol{T}) \\ \boldsymbol{0} & \mathrm{e}^{\boldsymbol{T}y} \end{pmatrix}.$$

Correspondingly, a simple (and efficient) way to compute $\boldsymbol{G}(y;\boldsymbol{\pi},\boldsymbol{T})$ is by calculating the matrix exponential of the left hand side.

Approaches to improve the speed of the EM algorithm in the univariate case exist in the literature; for instance, Okamura et al. [94] proposed a method also based on uniformization. □

4.3.3 Parameter estimation for censored data

In certain applications, some or all of the data may be censored. We call a data point *right–censored* at v if it takes an unknown value above v, *left–censored* at w if it takes an unknown value below w, and *interval–censored* if it is contained in the interval $(v, w]$, but its exact value is unknown. Left–censoring is a special case of interval–censoring with $v = 0$, while right–censoring can be obtained by fixing v and letting $w \to \infty$.

The EM Algorithm 4.4 works much in the same way as for uncensored data, with the only difference that we are no longer observing exact data points $Y^{(j)} = y^{(j)}$, but only $Y^{(j)} \in (v^{(j)}, w^{(j)}]$. This will only change the E–steps, where the conditional expectations can be calculated using the formulas in Olsson [95]. We now explain in detail how to adapt Algorithm 4.4 to censored data.

First EM

It is possible that a data point consists of a combination of marginals with both censored (not necessarily in the same intervals) and uncensored data (this is relevant in the first EM algorithm when considering data of the sum of the marginals). Table 4.1 contains all possible combinations one might have in the data and the way of treating them. Note that for $d > 2$, one simply repeats the same rules iteratively.

$Y^{(1)}$	$Y^{(2)}$	$Y^{(S)} = Y^{(1)} + Y^{(2)}$
Uncensored with value $y^{(1)}$	Uncensored with value $y^{(2)}$	Uncensored with value $y^{(1)} + y^{(2)}$
Right–censored at $v^{(1)}$	Uncensored with value $y^{(1)}$	Right–censored at $v^{(1)} + y^{(2)}$
Right–censored at $v^{(1)}$	Right–censored at $v^{(2)}$	Right–censored at $v^{(1)} + v^{(2)}$
Right–censored at $v^{(1)}$	Interval–censored $(v^{(2)}, w^{(2)}]$	Right–censored at $v^{(1)} + v^{(2)}$
Interval–censored $(v^{(1)}, w^{(1)}]$	Uncensored with value $y^{(2)}$	Interval–censored $(v^{(1)} + y^{(2)}, w^{(1)} + y^{(2)}]$
Interval–censored $(v^{(1)}, w^{(1)}]$	Interval–censored $(v^{(2)}, w^{(2)}]$	Interval–censored $(v^{(1)} + v^{(2)}, w^{(1)} + w^{(2)}]$

Table 4.1: Rules for censored data.

For completeness, we include here the conditional expectations needed (see also [95]).

$$\mathbb{E}\left(B_k \mid Y^{(S)} \in (v,w]\right) = \frac{\pi_k \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}v}\boldsymbol{e} - \pi_k \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}w}\boldsymbol{e}}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}v}\boldsymbol{e} - \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}w}\boldsymbol{e}},$$

$$\begin{aligned}\mathbb{E}&\left(Z_k \mid Y^{(S)} \in (v,w]\right)\\ &= \frac{\displaystyle\int_v^w \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du - \left(\int_0^w \boldsymbol{e}_k'\mathrm{e}^{\boldsymbol{T}(w-u)}\boldsymbol{e}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du - \int_0^v \boldsymbol{e}_k'\mathrm{e}^{\boldsymbol{T}(v-u)}\boldsymbol{e}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du\right)}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}v}\boldsymbol{e} - \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}w}\boldsymbol{e}},\end{aligned}$$

$$\begin{aligned}\mathbb{E}&\left(N_{kl} \mid Y^{(S)} \in (v,w]\right)\\ &= t_{kl}\frac{\displaystyle\int_v^w \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du - \left(\int_0^w \boldsymbol{e}_l'\mathrm{e}^{\boldsymbol{T}(w-u)}\boldsymbol{e}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du - \int_0^v \boldsymbol{e}_l'\mathrm{e}^{\boldsymbol{T}(v-u)}\boldsymbol{e}\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du\right)}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}v}\boldsymbol{e} - \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}w}\boldsymbol{e}},\end{aligned}$$

$$\mathbb{E}\left(N_k \mid Y^{(S)} \in (v,w]\right) = t_k \frac{\displaystyle\int_v^w \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}u}\boldsymbol{e}_k du}{\boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}v}\boldsymbol{e} - \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{T}w}\boldsymbol{e}}.$$

Second EM

The second EM algorithm works as above, with the only difference that for marginals with censored data the corresponding conditional expectation is calculated as

$$\begin{aligned}\mathbb{E}&\left(Z_k^{(j)} \mid Y^{(j)} \in (v^{(j)}, w^{(j)}]\right)\\ &= \frac{\displaystyle\int_{v^{(j)}}^{w^{(j)}} \boldsymbol{\pi}_j\mathrm{e}^{\boldsymbol{T}_j u}\boldsymbol{e}_k du - \left(\int_0^{w^{(j)}} \boldsymbol{e}_k'\mathrm{e}^{\boldsymbol{T}_j(w^{(j)}-u)}\boldsymbol{e}\boldsymbol{\pi}_j\mathrm{e}^{\boldsymbol{T}_j u}\boldsymbol{e}_k du - \int_0^{v^{(j)}} \boldsymbol{e}_k'\mathrm{e}^{\boldsymbol{T}_j(v^{(j)}-u)}\boldsymbol{e}\boldsymbol{\pi}_j\mathrm{e}^{\boldsymbol{T}_j u}\boldsymbol{e}_k du\right)}{\boldsymbol{\pi}_j\mathrm{e}^{\boldsymbol{T}_j v^{(j)}}\boldsymbol{e} - \boldsymbol{\pi}_j\mathrm{e}^{\boldsymbol{T}_j w^{(j)}}\boldsymbol{e}}.\end{aligned}$$

4.3.4 A bivariate phase–type distribution with explicit density

For a general MPH* distribution an explicit density is not available. Kulkarni [73] characterized the density by a system of partial differential equations, and in Breuer [28] a semi–explicit form is deduced. The following type of bivariate phase–type distributions does lead to an explicit density:
Let $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)})' \sim \mathrm{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ with

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{0} & \boldsymbol{T}_{22} \end{pmatrix}, \quad \boldsymbol{\pi} = (\boldsymbol{\alpha}, \boldsymbol{0}) \quad \text{and} \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{e} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{e} \end{pmatrix}, \tag{4.5}$$

where $\boldsymbol{T}_{11}$ and $\boldsymbol{T}_{22}$ are sub–intensity matrices of dimensions p_1 and p_2 ($p = p_1 + p_2$), respectively, and $\boldsymbol{T}_{11}\,\boldsymbol{e} + \boldsymbol{T}_{12}\,\boldsymbol{e} = \boldsymbol{0}$. Then the joint density of $\boldsymbol{Y}$ is given by

$$f_{\boldsymbol{Y}}\left(y^{(1)}, y^{(2)}\right) = \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}_{11}y^{(1)}}\boldsymbol{T}_{12}\mathrm{e}^{\boldsymbol{T}_{22}y^{(2)}}(-\boldsymbol{T}_{22})\,\boldsymbol{e}, \tag{4.6}$$

with marginals $Y^{(1)} \sim \mathrm{PH}(\boldsymbol{\alpha}, \boldsymbol{T}_{11})$ and $Y^{(2)} \sim \mathrm{PH}(\boldsymbol{\alpha}(-\boldsymbol{T}_{11})^{-1}\boldsymbol{T}_{12}, \boldsymbol{T}_{22})$. Note that the Baker–type bivariate distributions introduced in Bladt et al. [25] are a particular case. The latter have some remarkable properties: one can construct a distribution of this type with specific given marginals and a given Pearson correlation coefficient; this class is also dense within the set of bivariate distributions with support in $\mathbb{R}_+^2$ (this follows from the fact that the class of Bernstein copulas can be used to approximate arbitrarily well any copula (see Sencetta and Satchell [105]) and that the class of phase–type distributions can approximate arbitrarily well any distribution with support on $\mathbb{R}_+$), making the bigger class of bivariate distributions also dense.

4.3.4.1 Tail independence

The existence of an explicit form of the density allows us to compute the upper tail dependence coefficient λ_U. Recall that the latter is defined as

$$\lambda_U = \lim_{q\to 1^-} \mathbb{P}\left(Y^{(1)} > F^{-1}_{Y^{(1)}}(q) \mid Y^{(2)} > F^{-1}_{Y^{(2)}}(q)\right).$$

It is a classical measure of dependence in the tail and of considerable interest in applications in insurance and finance, where the modelling of tail events is crucial. From (4.6) we have

$$\bar{F}_{\boldsymbol{Y}}(y^{(1)}, y^{(2)}) = \mathbb{P}\left(Y^{(1)} > y^{(1)}, Y^{(2)} > y^{(2)}\right) = \boldsymbol{\alpha}\left(-\boldsymbol{T}_{11}\right)^{-1} \mathrm{e}^{\boldsymbol{T}_{11}y^{(1)}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}y^{(2)}} \boldsymbol{e}\,.$$

Then, if $-\lambda_j$ is the real part of the eigenvalue of $\boldsymbol{T}_{jj}$ with largest real part and k_j is the dimension of the Jordan block of λ_j for $j = 1, 2$, it is easy to see that

$$\bar{F}_{\boldsymbol{Y}}(y^{(1)}, y^{(2)}) \sim b(y^{(1)})^{k_1-1}\mathrm{e}^{-\lambda_1 y^{(1)}}(y^{(2)})^{k_2-1}\mathrm{e}^{-\lambda_2 y^{(2)}}, \quad \text{as} \quad y^{(1)}, y^{(2)} \to \infty\,,$$

where b is a positive constant. Hence

$$\lambda_U = \lim_{q\to 1^-} \frac{b(F^{-1}_{Y^{(1)}}(q))^{k_1-1}\mathrm{e}^{-\lambda_1(F^{-1}_{Y^{(1)}}(q))}(F^{-1}_{Y^{(2)}}(q))^{k_2-1}\mathrm{e}^{-\lambda_2(F^{-1}_{Y^{(2)}}(q))}}{c(F^{-1}_{Y^{(2)}}(q))^{k_2-1}\mathrm{e}^{-\lambda_2(F^{-1}_{Y^{(2)}}(q))}} = 0\,,$$

with c positive constant. In other words, $\boldsymbol{Y}$ is upper-tail-independent.

4.3.4.2 Estimation

The density (4.6) allows for a special form of EM algorithm. Such an algorithm was introduced in Ahlström et al. [1] and we include it for completeness, subsequent use and comparison purposes.

Algorithm 4.7.

0. Initialize with some "arbitrary" $(\boldsymbol{\alpha}, \boldsymbol{T})$.

1. (E–step) Calculate

$$\mathbb{E}\left(B_k \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \sum_{i=1}^{N} \frac{\alpha_k \boldsymbol{e}_k' \mathrm{e}^{\boldsymbol{T}_{11}y_i^{(1)}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}y_i^{(2)}} \left(-\boldsymbol{T}_{22}\right)\boldsymbol{e}}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, \quad k = 1, \ldots, p_1$$

$$\mathbb{E}\left(Z_k \mid \boldsymbol{Y} = \boldsymbol{y}\right)$$

$$= \begin{cases} \displaystyle\sum_{i=1}^{N} \frac{\int_0^{y_i^{(1)}} \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}_{11}u}\boldsymbol{e}_k\boldsymbol{e}_k'\mathrm{e}^{\boldsymbol{T}_{11}(y_i^{(1)}-u)}\boldsymbol{T}_{12}\mathrm{e}^{\boldsymbol{T}_{22}y_i^{(2)}}\left(-\boldsymbol{T}_{22}\right)\boldsymbol{e}du}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, & k = 1, \ldots, p_1 \\ \displaystyle\sum_{i=1}^{N} \frac{\int_0^{y_i^{(2)}} \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}_{11}y_i^{(1)}}\boldsymbol{T}_{12}\mathrm{e}^{\boldsymbol{T}_{22}u}\boldsymbol{e}_{(k-p_1)}\boldsymbol{e}_{(k-p_1)}'\mathrm{e}^{\boldsymbol{T}_{22}(y_i^{(2)}-u)}\left(-\boldsymbol{T}_{22}\right)\boldsymbol{e}du}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, & k = p_1+1, \ldots, p \end{cases}$$

$$\mathbb{E}\left(N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}\right)$$

$$= \begin{cases} \displaystyle\sum_{i=1}^{N} t_{kl} \frac{\displaystyle\int_0^{y_i^{(1)}} \boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11}u} \boldsymbol{e}_k \boldsymbol{e}_l' \mathrm{e}^{\boldsymbol{T}_{11}(y_i^{(1)}-u)} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}y_i^{(2)}} (-\boldsymbol{T}_{22}) \boldsymbol{e} du}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, & k,l = 1,\ldots,p_1,\ k \neq l \\ \displaystyle\sum_{i=1}^{N} t_{kl} \frac{\boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11}y_i^{(1)}} \boldsymbol{e}_k \boldsymbol{e}_{l-p_1}' \mathrm{e}^{\boldsymbol{T}_{22}y_i^{(2)}} (-\boldsymbol{T}_{22}) \boldsymbol{e}}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, & k = 1,\ldots,p_1,\ l = p_1+1,\ldots,p \\ \displaystyle\sum_{i=1}^{N} t_{kl} \frac{\displaystyle\int_0^{y_i^{(2)}} \boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11}y_i^{(1)}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}u} \boldsymbol{e}_{k-p_1} \boldsymbol{e}_{l-p_1}' \mathrm{e}^{\boldsymbol{T}_{22}(y_i^{(2)}-u)} (-\boldsymbol{T}_{22}) \boldsymbol{e} du}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, & k,l = p_1+1,\ldots,p,\ k \neq l \end{cases}$$

$$\mathbb{E}\left(N_k \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11}y_i^{(1)}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}y_i^{(2)}} \boldsymbol{e}_{k-p_1}}{f_{\boldsymbol{Y}}(y_i^{(1)}, y_i^{(2)}; \boldsymbol{\alpha}, \boldsymbol{T})}, \quad k = p_1+1,\ldots,p\,.$$

2. (M–step) Let

$$\hat{\alpha}_k = \frac{1}{N}\mathbb{E}\left(B_k \mid \boldsymbol{Y} = \boldsymbol{y}\right), \quad \hat{t}_{kl} = \frac{\mathbb{E}\left(N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y} = \boldsymbol{y}\right)}, \quad \hat{t}_k = \frac{\mathbb{E}\left(N_k \mid \boldsymbol{Y} = \boldsymbol{y}\right)}{\mathbb{E}\left(Z_k \mid \boldsymbol{Y} = \boldsymbol{y}\right)},$$
$$\hat{t}_{kk} = -\sum_{l \neq k} \hat{t}_{kl} - \hat{t}_k, \quad \hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{p_1}), \quad \hat{\boldsymbol{T}} = \{\hat{t}_{kl}\}_{k,l=1,\ldots,p} \quad \textit{and} \quad \hat{\boldsymbol{t}} = (\hat{t}_{p_1+1}, \ldots, \hat{t}_p)'.$$

3. Assign $\boldsymbol{\alpha} := \hat{\boldsymbol{\alpha}}$, $\boldsymbol{T} := \hat{\boldsymbol{T}}$, $\boldsymbol{t} := \hat{\boldsymbol{t}}$ and GOTO 1.

We now provide two detailed illustrations. When Algorithm 4.4 is employed, given that an explicit form of the joint density is not available, we choose the number of iterations in such a way that the changes in the successive log–likelihoods in the first EM become negligible and the changes in the successive parameter estimates become negligible in the second EM. For Algorithm 4.7 we used a criterion similar to the univariate case.

Example 4.8 (Simulation study)**.** The objective of the present example is to compare the performance of Algorithm 4.4 and Algorithm 4.7. We will illustrate that the more general Algorithm 4.4 also provides reasonable results when dealing with a sample from a bivariate distribution with density (4.6), for which the more specific Algorithm 4.7 is particularly well-suited. We generated an i.i.d. sample of size 10 000 from a MPH* distribution with parameters

$$\boldsymbol{\pi} = (0.15,\, 0.85,\, 0,\, 0)\,, \quad \boldsymbol{T} = \begin{pmatrix} -2 & 0 & 2 & 0 \\ 9 & -11 & 0 & 2 \\ 0 & 0 & -1 & 0.5 \\ 0 & 0 & 0 & -5 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

which has theoretical mean $\mathbb{E}(\boldsymbol{Y}) = \left(\mathbb{E}(Y^{(1)}), \mathbb{E}(Y^{(2)})\right)' = (0.5, 0.9609)'$, and correlation coefficient $\rho(Y^{(1)}, Y^{(2)}) = 0.1148$. Moreover, we know that $\lambda_U = 0$. The simulated sample has numerical values $\hat{\mathbb{E}}(\boldsymbol{Y}) = (0.5046, 0.9650)'$, $\hat{\rho} = 0.1235$ and Kendall's tau $\hat{\rho}_\tau = 0.1613$. We now use Algorithms 4.4 and 4.7 to recover the underlying structure of the data, and we assess the quality of the estimation by comparing densities, QQ plots, numerical properties of the distributions and contour plots.

Using Algorithm 4.4 with the same number of phases $p = 4$, random initial values and 3 500 steps in each EM algorithm, we obtain the following parameters:

$$\hat{\boldsymbol{\pi}} = (0.8592,\ 0.0002,\ 0.0005,\ 0.1402)\ ,$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -9.2654 & 0.3805 & 7.2657 & 1.6192 \\ 0.0039 & -1.1038 & 0.0002 & 0.0118 \\ 0.7429 & 0.1640 & -7.1356 & 4.9516 \\ 0.3295 & 1.6162 & 0.5278 & -2.4740 \end{pmatrix},$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 0.5950 & 0.4050 \\ 0 & 1 \\ 0.4254 & 0.5746 \\ 0.8347 & 0.1653 \end{pmatrix}.$$

The fitted distribution has mean $\mathbb{E}(\boldsymbol{Y}) = (0.5205, 0.9491)'$ and $\rho = 0.2559$, which approximates reasonably well the mean of the original distribution and to a lesser degree well the correlation coefficient. We also approximated λ_U and ρ_τ via simulation obtaining $\hat{\lambda}_U = 0.0058$ and $\hat{\rho}_\tau = 0.2843$. Figures 4.3 and 4.4 show that the algorithm is able to recover the structure of the marginals and the sum of the marginals. Moreover, Figure 4.5 shows that the contour plot of the fitted distribution is similar to the one of the sample. Here, the running time of Algorithm 4.4 with its 3 500 iterations was 924 seconds.

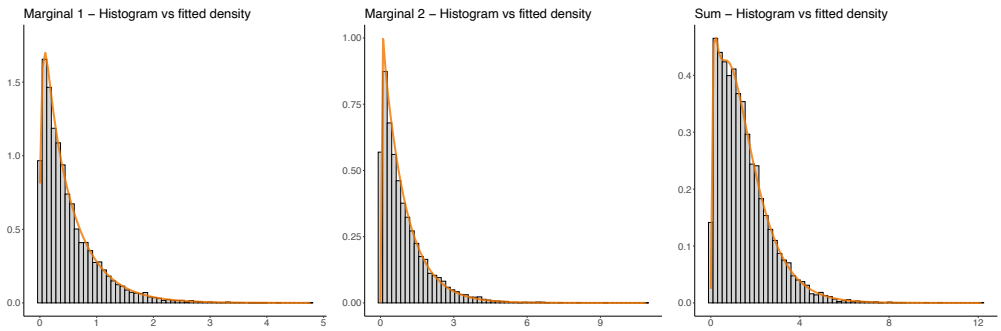

Figure 4.3: Histograms of simulated sample versus densities of the MPH* distribution fitted using Algorithm 4.4.

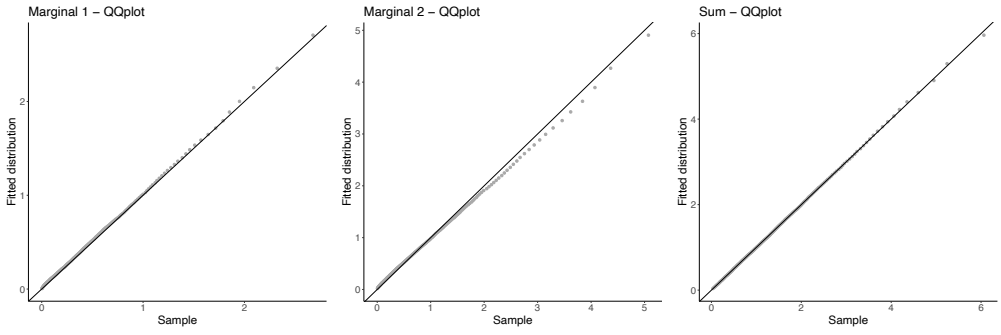

Figure 4.4: QQ plots of simulated sample versus fitted MPH* distribution using Algorithm 4.4.

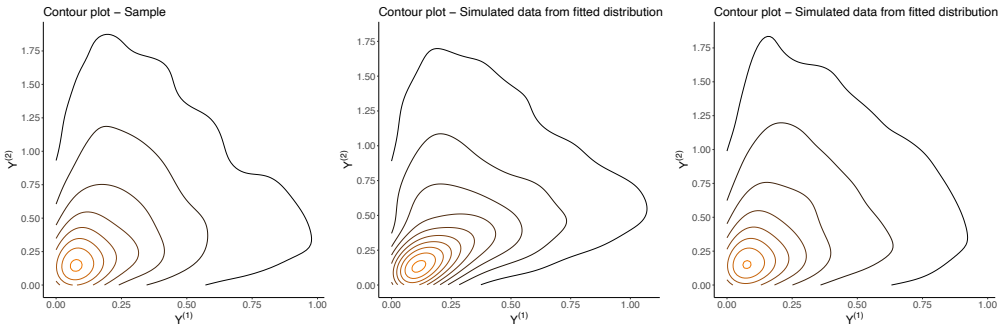

Figure 4.5: Contour plot of sample (left), contour plot of a simulated sample from the MPH* distribution fitted with Algorithm 4.4 (center) and contour plot of of a simulated sample from the distribution fitted with Algorithm 4.7 (right).

Next we use Algorithm 4.7 with 4 phases ($p_1 = 2$ and $p_2 = 2$), random initial values and 2 500 steps in the EM algorithm (leading to a running time of 623 seconds). The estimated parameters are:

$$\hat{\boldsymbol{\pi}} = (0.1218,\, 0.8782,\, 0,\, 0)\ ,$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -2.2656 & 0.2832 & 1.9823 & 0.0002 \\ 8.2508 & -10.2295 & 0.0257 & 1.9530 \\ 0 & 0 & -1.0115 & 0.6419 \\ 0 & 0 & 0.0992 & -4.7781 \end{pmatrix}\ ,$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}\ ,$$

with corresponding mean $\mathbb{E}(\boldsymbol{Y}) = (0.5046, 0.9650)'$ and $\rho = 0.1156$. ρ_τ for this fit can be estimated via simulation, giving $\hat{\rho}_\tau = 0.1448$. A contour plot of the fit is available in Figure 4.5 and we see that the algorithm recovers the original structure of the data even better. Note that again the log–likelihood of the fitted MPH* distribution ($-12\,329.82$) outperforms the log–likelihood using the original MPH* distribution ($-12\,330.96$).

Finally we would like to remark that Algorithm 4.7 already starts with a more specific structure on its parameters which resembles the one of the distribution from which the data come from. On the other hand, Algorithm 4.4 does not require any prior assumption on the initial structure of its parameters. Thus, a better fit from Algorithm 4.7 is expected, since Algorithm 4.4 needs to find a distribution in a larger set. In line with the non-identifiability issue, one sees that one can obtain a quite reasonable fit in that larger class that captures some main features of the original distribution, whereas the more specific Algorithm 4.7 finds a fit that even exhibits nicely the original correlation pattern. Yet, the flexibility of Algorithm 4.4 is a considerable advantage when dealing with data sets without the additional knowledge about the underlying distribution.

Example 4.9 (Known distribution – Marshall-Olkin exponential)**.** We now would like to illustrate that Algorithms 4.4 and 4.7 can be modified to fit a MPH* model to a theoretically given joint distribution H. The idea is along the lines of [10] and consists of considering sequences of empirical distributions with increasing sample size. We ex-

emplify this by considering a bivariate Marshall–Olkin exponential distribution, whose joint survival function is of the form

$$\overline{F}\left(y^{(1)}, y^{(2)}\right) = \exp\left(-\lambda_1 y^{(1)} - \lambda_2 y^{(2)} - \lambda_{12}\max(y^{(1)}, y^{(2)})\right) .$$

We take $\lambda_1 = 1$, $\lambda_2 = 3$ and $\lambda_{12} = 1$, then the distribution has theoretical moments $\mathbb{E}(\boldsymbol{Y}) = (0.5, 0.25)$ and $\rho = 0.2$. It is easy to see that the Marshall–Olkin bivariate exponential is upper-tail-independent, i.e., $\lambda_U = 0$. Moreover, we approximate ρ_τ via simulation, obtaining $\hat{\rho}_\tau = 0.2012$. Then we fit a MPH* distribution using the Algorithm 4.4. With 3 phases and random initial values together with 2 500 steps in each EM algorithm (overall running time about 120 seconds), we obtain the parameters

$$\hat{\boldsymbol{\pi}} = (0.8233,\ 0.1633,\ 0.0134) ,$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} 2.5894 & 1.6637 & 0.7601 \\ 0.0087 & -2.0699 & 0.2102 \\ 0.1032 & 0.3465 & -4.2765 \end{pmatrix} ,$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 0.4412 & 0.5588 \\ 0.9514 & 0.0486 \\ 0.2348 & 0.7652 \end{pmatrix} ,$$

which has corresponding moments $\mathbb{E}(\boldsymbol{Y}) = (0.4942, 0.2565)'$ and $\rho = 0.3260$. λ_U and ρ_τ for the resulting model can be approximated by simulation to be $\hat{\lambda}_U = 0.0547$ and $\hat{\rho}_\tau = 0.2745$. Together with the densities (Figure 4.6) and QQ plots (Figure 4.7), one sees that this algorithm recovers rather well the structure of the original joint distribution.

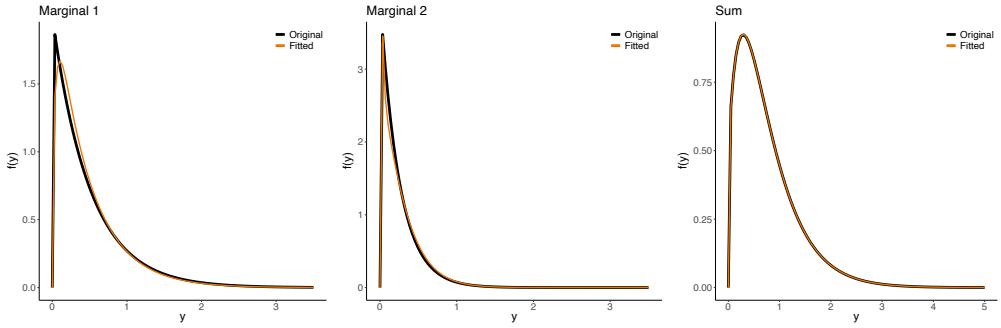

Figure 4.6: Densities of the original Marshall–Olkin distribution versus densities of the MPH* distribution fitted using Algorithm 4.4.

4.4 Multivariate inhomogeneous phase-type distributions

There are various possibilities for extending inhomogeneous phase-type distributions to more than one dimension. In the following we suggest one particular approach and provide an algorithm for the parameter estimation.

4.4.1 Definition and Properties

Let $\boldsymbol{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ and define $\boldsymbol{X} := (g_1(Y^{(1)}), \ldots, g_d(Y^{(d)}))'$, where $g_j : \mathbb{R}_+ \to \mathbb{R}_+$ are increasing and differentiable functions for $j = 1, \ldots, d$, then we say that $\boldsymbol{X}$ has an *inhomogeneous MPH** distribution.

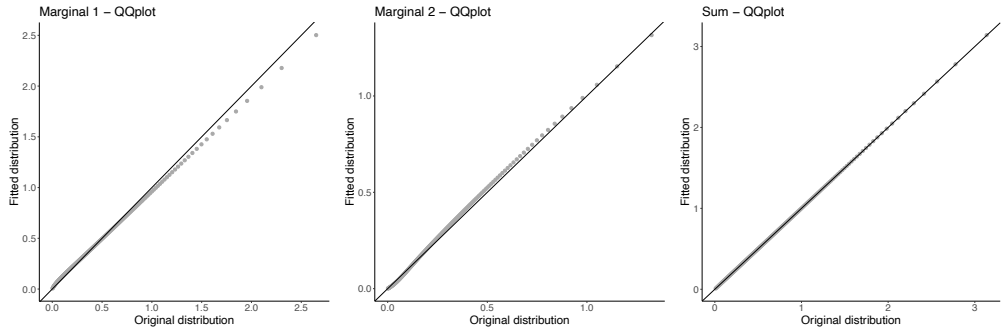

Figure 4.7: QQ plots of original Marshall–Olkin distribution versus fitted MPH* distribution using Algorithm 4.4.

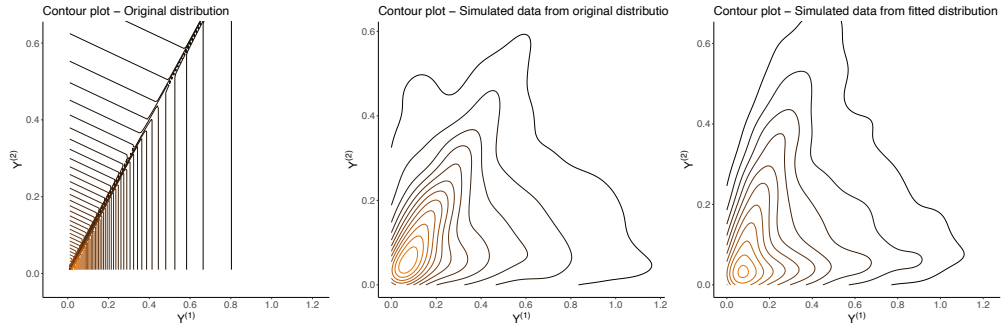

Figure 4.8: Contour plot of the original Marshall–Olkin distribution (left) contour plot of simulated sample from Marshall–Olkin distribution (middle) and contour plot of simulated sample from the distribution fitted with Algorithm 4.4 (right).

Several of its properties follow directly from the definition:

1. The marginals $X^{(j)} = g_j(Y^{(j)})$ are IPH distributed, since each $Y^{(j)}$ is phase–type distributed, $j = 1, \ldots, d$.

2. Since g_j is increasing for all $j = 1, \ldots, d$, the copula of $\boldsymbol{X}$ is the same as the copula of $\boldsymbol{Y}$ (see e.g. [82, Prop.7.7]).

3. For fixed $g_j(\cdot)$, $j = 1, \ldots, d$, this new class is dense in $\mathbb{R}_+^d$ (by the denseness of the MPH* class).

In the sequel we will provide an algorithm for parameter estimation for which an explicit expressions of the bivariate density is needed. We therefore restrict it to the bivariate case.

4.4.2 Parameter estimation in the bivariate case

In the bivariate case, for any $\boldsymbol{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ with parameters (4.5), it is easy to see that the density of $\boldsymbol{X} = (g_1(Y^{(1)}), g_2(Y^{(2)}))'$ is given by

$$f_{\boldsymbol{X}}\left(x^{(1)}, x^{(2)}\right) = \boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11} g_1^{-1}(x^{(1)})} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22} g_2^{-1}(x^{(2)})} (-\boldsymbol{T}_{22}) \boldsymbol{e} \, \frac{1}{g_1'(g_1^{-1}(x^{(1)})) g_2'(g_2^{-1}(x^{(2)}))} \,. \tag{4.7}$$

If we assume that $g_j(\,\cdot\,;\boldsymbol{\beta}_j)$ is a parametric non–negative function depending on the vector $\boldsymbol{\beta}_j$, $j = 1, 2$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Then, we can formulate an algorithm analogous to Algorithm 4.1:

Algorithm 4.10 (EM algorithm for bivariate inhomogeneous MPH* distributions)**.**

0. Initialize with some "arbitrary" $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\beta})$.

1. Transform the data into $y_i^{(j)} := g_j^{-1}(x_i^{(j)}; \boldsymbol{\beta}_j)$, $i = 1, \ldots, N$, $j = 1, 2$, *and apply the E– and M–steps of Algorithm 4.7 by which we obtain the estimators* $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}})$.

2. Compute

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log(f_{\boldsymbol{X}}(x_i^{(1)}, x_i^{(2)}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}}, \boldsymbol{\beta}))$$

3. Assign $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\beta}})$ *and GOTO 1.*

We now consider particular multivariate distributions obtained through such a transformation of an MPH* random vector.

4.4.3 Multivariate matrix–Pareto models

Let $\boldsymbol{X} = (g_1(Y^{(1)}), \ldots, g_d(Y^{(d)}))'$, where $\boldsymbol{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ and

$$g_j(y) = \beta_j(e^y - 1), \quad \beta_j > 0, \; j = 1, \ldots, d. \tag{4.8}$$

Then we say that $\boldsymbol{X}$ follows a *multivariate matrix–Pareto distribution*. Some special properties of this class of distributions are:

1. Marginal distributions are matrix–Pareto distributed.
2. Moments and cross–moments of $\boldsymbol{X}$ can be obtained from the moment generating function of $\boldsymbol{Y}$ (see [24, Theorem 8.1.2]), provided that they exist.
3. Products of the type $\prod_{i=1}^{M} \left(\frac{X^{(j_i)}}{\beta_{j_i}} + 1\right)^{a_i}$ are matrix–Pareto distributed, $a_i > 0$, $j_i \in \{1, \ldots, d\}$, $i = 1, \ldots, M$, $M \leq N$, since linear combinations of $\boldsymbol{Y}$ are PH distributed.

In the bivariate case, (4.8) and (4.7), lead to

$$f_{\boldsymbol{X}}(x^{(1)}, x^{(2)}) = \boldsymbol{\alpha} \left(\frac{x^{(1)}}{\beta_1} + 1\right)^{\boldsymbol{T}_{11} - \boldsymbol{I}} \boldsymbol{T}_{12} \left(\frac{x^{(2)}}{\beta_2} + 1\right)^{\boldsymbol{T}_{22} - \boldsymbol{I}} (-\boldsymbol{T}_{22}) \, \boldsymbol{e} \, \frac{1}{\beta_1 \beta_2} \,.$$

and

$$\bar{F}_{\boldsymbol{X}}(x^{(1)}, x^{(2)}) = \boldsymbol{\alpha} \, (-\boldsymbol{T}_{11})^{-1} \left(\frac{x^{(1)}}{\beta_1} + 1\right)^{\boldsymbol{T}_{11}} \boldsymbol{T}_{12} \left(\frac{x^{(2)}}{\beta_2} + 1\right)^{\boldsymbol{T}_{22}} \boldsymbol{e}$$

Moreover, in this bivariate case, linear combinations of $X^{(i)}$ are regularly varying, and the respective index is the real part of the eigenvalue with largest real part of the sub–intensity matrices of the marginals, which follows from [40, Lem. 2.1] and asymptotic independence of $\boldsymbol{Y}$. The general case is not clear since the condition of asymptotic independence does not hold.

Remark 4.11. The Marshall–Olkin Pareto distribution (see Hanagal [57]) is a particular case of this class of distributions.

4.4.3.1 Parameter estimation

As in the univariate case, if we assume the simpler transformation $g_j(y) = e^y - 1$, $j = 1, \ldots, d$, then we can use fitting methods of the MPH* class by taking the logarithm of the marginal observations. I.e., we can apply Algorithms 4.4 and 4.7 to the transformed data $y_i^{(j)} := \log(x_i^{(j)}+1)$, $i = 1, \ldots, N$, $j = 1, \ldots, d$, to estimate the parameters $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$. We exemplify the use of this method in two examples.

Example 4.12. (Mardia type I) We generated an i.i.d. sample of size 10 000 from a (translated) Mardia type I Pareto distribution (see [81]) with parameters $\sigma_1 = \sigma_2 = 1$ and $\alpha = 2$. This distribution has theoretical numerical values $\mathbb{E}(\boldsymbol{X}) = (1, 1)'$, $\lambda_U = 0$ and $\rho_\tau = 0.2$. The simulated sample has numerical values $\hat{\mathbb{E}}(\boldsymbol{X}) = (0.9812, 0.9712)'$ and $\hat{\rho}_\tau = 0.2049$.

We fitted a bivariate matrix–Pareto distribution using Algorithm 4.7 with $p_1 = p_2 = 2$ (i.e., $p = 4$) and 2 000 steps on the transformed data (with a running time of 530 seconds), getting the following parameters:

$$\hat{\boldsymbol{\alpha}} = (0.1666,\ 0.8334,\ 0,\ 0)\ ,$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -2.0022 & 0.0238 & 1.9784 & 0 \\ 9.2506 & -11.2967 & 0.1204 & 1.9256 \\ 0 & 0 & -2.0175 & 1.8247 \\ 0 & 0 & 0.0834 & -12.8829 \end{pmatrix},$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

The tails of the marginals of the fitted distribution are determined by the real part of the eigenvalues with largest real part of the sub–intensity matrices of the marginal distributions, which are $\lambda_1^{(\max)} = -1.9785$ and $\lambda_2^{(\max)} = -2.0035$. These resemble well the ones of the original distribution. The fitted distribution has first moment $\mathbb{E}(\boldsymbol{X}) = (1.0164, 0.9963)'$. Moreover, we estimated ρ_τ via simulation, obtaining $\hat{\rho}_\tau = 0.1582$. The QQ plots are available in Figure 4.9 and contour plots are depicted in Figure 4.10, from where it becomes clear that the algorithm recovers the structure of the data well. Again, the log–likelihood of the fitted bivariate matrix–Pareto ($-15\,550.52$) exceeds the log–likelihood using the original Mardia distribution ($-16\,146.65$).

The concrete structure of the intensity and reward matrix underlying Algorithm 4.7 restricts its application to tail-independent models. While in the previous example this was justified, in situations with tail-dependent data one should rather look for fits in the general MPH* class by using Algorithm 4.4 on the transformed data. The following example illustrates such an approach.

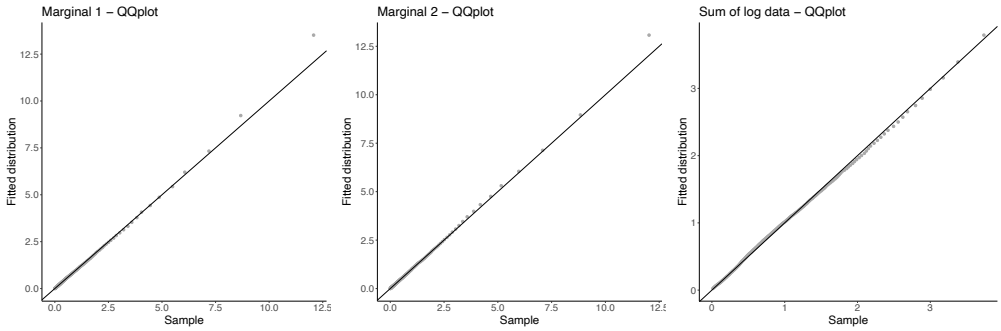

Figure 4.9: QQ plots of simulated Mardia Pareto sample versus fitted multivariate matrix–Pareto distribution using Algorithm 4.7.

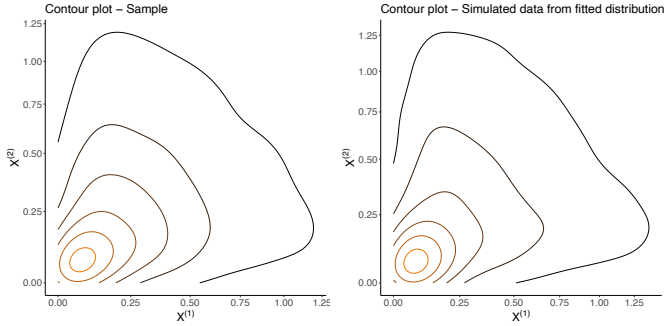

Figure 4.10: Contour plot of simulated Mardia Pareto sample (left) and contour plot of a simulated sample from the Matrix–Pareto distribution fitted using Algorithm 4.7 (right).

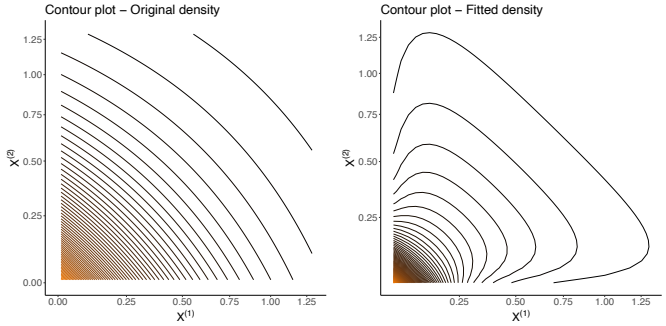

Figure 4.11: Contour plot of original Mardia Pareto (left) and contour plot of Matrix–Pareto distribution fitted using Algorithm 4.7 (right).

Example 4.13. (Gumbel copula with matrix–Pareto marginals) We generated an i.i.d. sample of size 5 000 from a three-dimensional random vector with first marginal being a matrix–Pareto with parameters

$$\boldsymbol{\pi}_1 = (0.2,\, 0.8)\,,$$
$$\boldsymbol{T}_1 = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix},$$

$$\beta_1 = 1\,,$$

second marginal being a matrix–Pareto with parameters

$$\boldsymbol{\pi}_2 = (0.5, 0.5)\,,$$
$$\boldsymbol{T}_2 = \begin{pmatrix} -2 & 0 \\ 0 & -1.5 \end{pmatrix},$$
$$\beta_2 = 1\,,$$

and the third marginal being a conventional Pareto with shape parameter 2.5, and a Gumbel copula with parameter $\theta = 4$. The choice of a Gumbel copula instead of a multivariate matrix–Pareto model based on an MPH* construction is to show that the algorithm can be used to model any type of dependence structure. The Gumbel copula is known to have positive tail dependence. This distribution has theoretical numerical values $\mathbb{E}(\boldsymbol{X}) = (0.6, 1.5, 0.6667)'$, $\lambda_U = 2 - 2^{1/4} \approx 0.8108$ and $\rho_\tau = 0.75$, and the real part of the eigenvalues that determine the heaviness of the tails are $\lambda_1^{(\max)} = -2$, $\lambda_2^{(\max)} = -1.5$ and $\lambda_3^{(\max)} = -2.5$. The simulated sample has numerical values $\hat{\mathbb{E}}(\boldsymbol{X}) = (0.6035, 1.4721, 0.6732)'$, $\hat{\lambda}_U(X_1, X_2) = 0.8142$, $\hat{\lambda}_U(X_1, X_3) = 0.7429$, $\hat{\lambda}_U(X_2, X_3) = 0.7857$, $\hat{\rho}_\tau(X_1, X_2) = 0.7513$, $\hat{\rho}_\tau(X_1, X_3) = 0.7526$ and $\hat{\rho}_\tau(X_2, X_3) = 0.7544$. Then, we fitted a multivariate matrix–Pareto distribution using Algorithm 4.4 with $p = 6$ and 5 000 steps on the transformed data (running time 6 501 seconds), obtaining the following parameters:

$$\hat{\boldsymbol{\alpha}} = (0.0062\,, 0.0606\,, 0.4394\,, 0.0600\,, 0.0608\,, 0.3731)\,,$$
$$\hat{\boldsymbol{T}} = \begin{pmatrix} -16.9867 & 1.7532 & 1.3902 & 3.2835 & 1.0736 & 2.1224 \\ 0.2710 & -1.3282 & 0.2835 & 0.0757 & 0.0562 & 0.2987 \\ 0.3714 & 0.0161 & -2.8735 & 0.6658 & 0.8798 & 0.9208 \\ 1.1742 & 0.5387 & 0.2421 & -4.0067 & 0.5156 & 0.5571 \\ 0.0786 & 0.5130 & 0.2916 & 1.1454 & -4.2894 & 1.5148 \\ 2.0244 & 0.2912 & 0.7007 & 0.7777 & 2.0237 & -6.2507 \end{pmatrix},$$
$$\hat{\boldsymbol{R}} = \begin{pmatrix} 0.9414 & 0.0558 & 0.0028 \\ 0.2856 & 0.4594 & 0.2549 \\ 0.2934 & 0.5233 & 0.1833 \\ 0.0906 & 0.5548 & 0.3546 \\ 0.4230 & 0.0750 & 0.5020 \\ 0.0748 & 0.5992 & 0.3260 \end{pmatrix}.$$

The real part of the eigenvalues with largest real part of the sub–intensity matrices of the marginal distributions are $\lambda_1^{(\max)} = -2.5370$, $\lambda_2^{(\max)} = -1.5941$ and $\lambda_3^{(\max)} = -2.4167$, which are close to the ones of the original distribution. The fitted distribution has first moment $\mathbb{E}(\boldsymbol{X}) = (0.6006, 1.5344, 0.6850)'$. Moreover, we estimated λ_U and ρ_τ via simulation, obtaining $\hat{\lambda}_U(X_1, X_2) = 0.70403$, $\hat{\lambda}_U(X_1, X_3) = 0.736438$, $\hat{\lambda}_U(X_2, X_3) = 0.672417$, $\hat{\rho}_\tau(X_1, X_2) = 0.6924$, $\hat{\rho}_\tau(X_1, X_3) = 0.7470$ and $\hat{\rho}_\tau(X_2, X_3) = 0.7728$. Comparing all numerical properties, together with the QQ plots (see Figure 4.12) and contour plots (see Figure 4.13), we see that this algorithm also recovers relatively well the structure of the data.

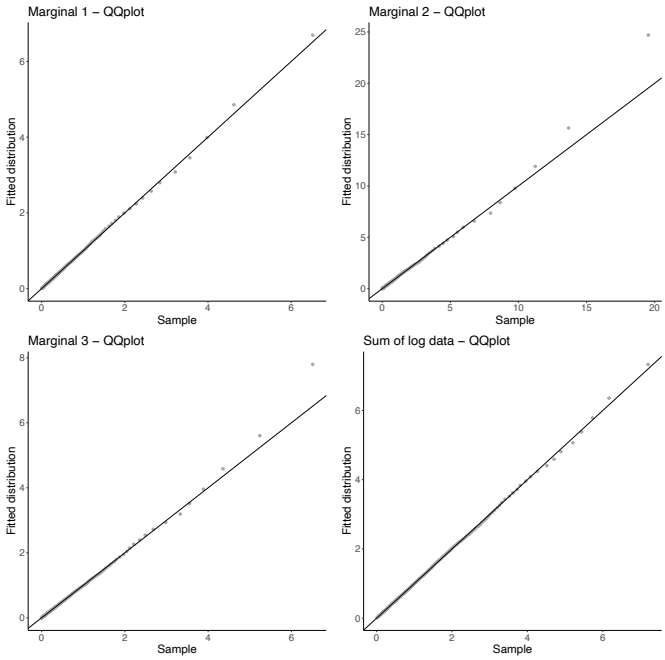

Figure 4.12: QQ plots of simulated sample versus fitted multivariate matrix–Pareto distribution using Algorithm 4.4.

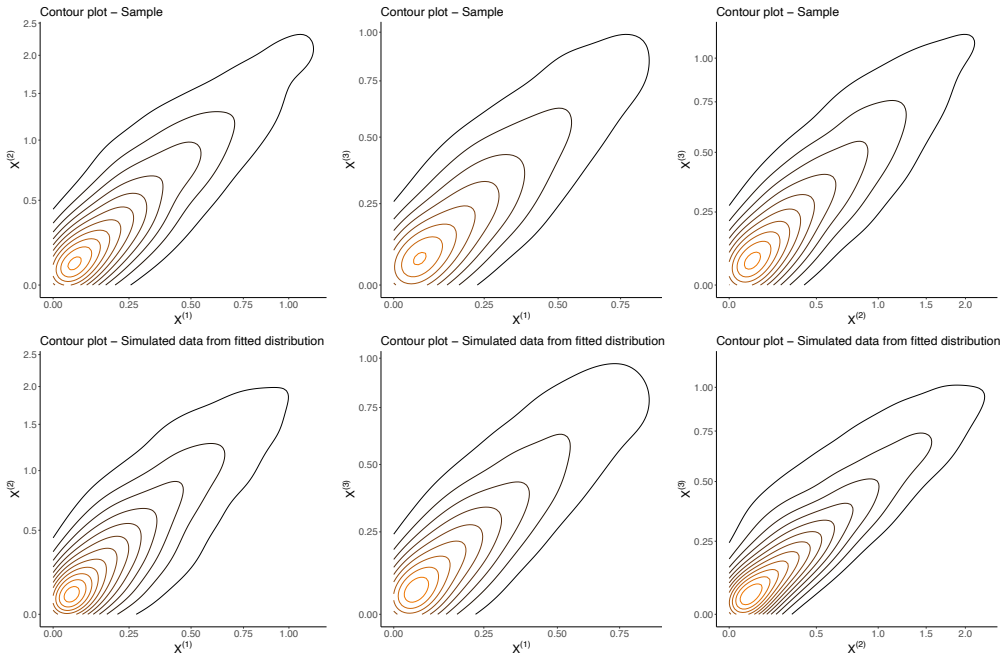

Figure 4.13: Contour plots of simulated sample from Gumbel copula with matrix-Pareto marginals (top) and contour plots of a simulated sample from the Matrix–Pareto distribution fitted using Algorithm 4.4 (bottom).

We would like to comment on the relative inaccuracy of the obtained estimate for $\lambda_1^{(\max)}$, especially when compared to the convincing estimates for the other marginals. The first marginal here is in fact a mixture of Pareto distributions with shape parameters 2 and 3, and mixing probabilities 0.2 and 0.8, respectively. That is, a major proportion of the data for the first marginal are to be expected to stem from the Pareto distribution with shape parameter 3. Since Algorithm 4.4 fits body and tail at the same time, one cannot expect the resulting estimate for the tail to be as good as techniques of classical extreme value theory, which focusses on the tail fit only. To illustrate this point, we also applied the algorithm to data simulated from changed initial probabilities in the first marginal according to $\boldsymbol{\pi}_1 = (0.8,\, 0.2)$ (and otherwise identical parameters). Now most of the data points will stem from the heavier distribution in the mixture. Indeed, the estimate for $\lambda_1^{(\max)}$ in that case turns out to be -2.0152, while the tail estimate for the other marginals remains almost at the same value (-1.5966 and -2.4154).

Next we present a parameter–dependent example with real data, employing Algorithm 4.10.

Example 4.14. (Danish fire insurance data) Consider the famous Danish fire insurance claim data set (see e.g. [56]). We propose here a bivariate matrix–Pareto distribution as a model for the components building and content with observations in the set $(1,\infty) \times (1,\infty)$. To that end, we first translate the sample to the origin, thus the sample has numerical values $\hat{\mathbb{E}}(\boldsymbol{X}) = (3.2476, 4.6856)'$ and $\hat{\rho}_\tau = 0.2143$. Then, we fit a bivariate matrix–Pareto distribution with $p_1 = p_2 = 2$ using Algorithm 4.10 with 2 000 steps (with a step–length of 0.05 and gradient ascent until the norm of the derivative is less than 0.1, the running time is 438 seconds), obtaining the following parameters:

$$\hat{\boldsymbol{\alpha}} = (0,\, 1,\, 0,\, 0)\,,$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -2.6333 & 0 & 0.0005 & 2.6328 \\ 1.2788 & -3.8336 & 2.5548 & 0 \\ 0 & 0 & -10.9822 & 0 \\ 0 & 0 & 2.4131 & -2.4732 \end{pmatrix},$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

$$\hat{\beta}_1 = 5.0377\,, \quad \hat{\beta}_2 = 13.4538\,.$$

The real part of the eigenvalues that determines the heaviness of the tails are $\lambda_1^{(\max)} = -2.6333$ and $\lambda_2^{(\max)} = -2.4732$, respectively. The fitted distribution has mean $\mathbb{E}(\boldsymbol{X}) = (3.1698, 4.6803)'$. Moreover, estimating ρ_τ via simulation gives $\hat{\rho}_\tau = 0.2190$. From the QQ plots (Figure 4.14) and contour plots (Figure 4.15), we see that the fitted distribution is a reasonable model for the data.

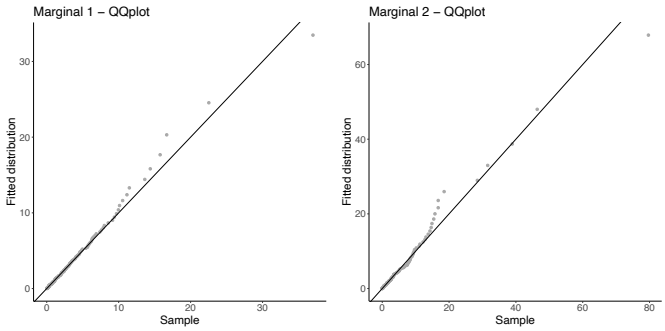

Figure 4.14: QQ plots of Danish fire insurance claim size sample versus fitted bivariate matrix–Pareto distribution using Algorithm 4.10.

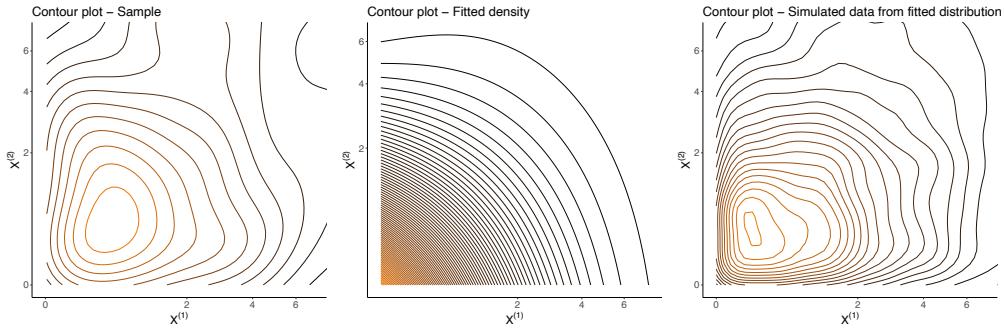

Figure 4.15: Contour plot of Danish fire insurance claim size sample (left), contour plot of Matrix–Pareto distribution fitted using Algorithm 4.10 (middle) and contour plot of simulated sample from fitted distribution (right).

This bivariate data set was recently also studied in Albrecher et al. [2, Sec. 4.5.2], where a splicing model with a bivariate mixed Erlang for the body and a bivariate Generalized Pareto distribution (GPD) for the tail was proposed. That approach required a threshold selection for the fitting of the tails, and univariate extreme value analysis led there to values of regular variation of around 2 for the building component and 1.67 for the contents component. Even though our estimates are further away from these values than their bivariate model (1.75 and 1.54, respectively), we would like to emphasize that the fitting of a matrix–Pareto distribution does not require any threshold selection. Furthermore, if we were to use more phases and a general form of the sub–intensity and reward matrices, the fit would quickly improve and for about 6 phases reach the accuracy of the bivariate GPD model, but then the overall number of parameters compared to the size of the present data set may not be considered commensurate, which is why we stick to the above choice. Note that our proposed procedure is fully automatic and the respective implementation can easily be applied to any other data set as well.

4.4.4 Multivariate Matrix–Weibull models

Let $\boldsymbol{X} = (g_1(Y^{(1)}), \ldots, g_d(Y^{(d)}))'$, where $\boldsymbol{Y} \sim \mathrm{MPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$ and $g_j(y) = y^{1/\beta_j}$, $\beta_j > 0$, $j = 1, \ldots, d$, then we say that $\boldsymbol{X}$ has a *multivariate matrix–Weibull distribution*. Some special properties of this type of distribution are:

1. Marginal distributions are matrix–Weibull distributed.
2. For a vector $\boldsymbol{a} = (a^{(1)}, \ldots, a^{(d)})$, with $a^{(j)} > 0$, $j = 1, \ldots, d$, $\Delta(\boldsymbol{a})\boldsymbol{X}$ is multivariate matrix–Weibull distributed.

For the bivariate case we get

$$f_{\boldsymbol{X}}\left(x^{(1)}, x^{(2)}\right) = \boldsymbol{\alpha} \mathrm{e}^{\boldsymbol{T}_{11}(x^{(1)})^{\beta_1}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}(x^{(2)})^{\beta_2}} (-\boldsymbol{T}_{22})\boldsymbol{e}\, \beta_1 (x^{(1)})^{\beta_1 - 1} \beta_2 (x^{(2)})^{\beta_2 - 1} .$$

and

$$\bar{F}_{\boldsymbol{X}}(x^{(1)}, x^{(2)}) = \boldsymbol{\alpha} \left(-\boldsymbol{T}_{11}\right)^{-1} \mathrm{e}^{\boldsymbol{T}_{11}(x^{(1)})^{\beta_1}} \boldsymbol{T}_{12} \mathrm{e}^{\boldsymbol{T}_{22}(x^{(2)})^{\beta_2}} \boldsymbol{e}$$

Remark 4.15. The Marshall–Olkin Weibull distribution (see [58]) is a particular case of this distribution.

4.4.4.1 Parameter estimation

In contrast to Section 4.4.3.1, all transformations are parameter–dependent, and the fitting procedures of the previous subsection are not applicable. However, we can apply Algorithm 4.10 in the bivariate case.

Example 4.16. (Bivariate Matrix–Weibull) We generate an i.i.d. sample of size 5 000 of a bivariate random vector with matrix–Weibull marginals with parameters

$$\begin{gathered} \boldsymbol{\pi}_1 = (0.8,\ 0.2) , \\ \boldsymbol{T}_1 = \begin{pmatrix} -1 & 0.5 \\ 0 & -0.5 \end{pmatrix} , \\ \beta_1 = 0.4 \end{gathered}$$

for the first marginal and

$$\begin{gathered} \boldsymbol{\pi}_2 = (0.5,\ 0.25,\ 0.25) , \\ \boldsymbol{T}_2 = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -0.5 & 0.5 \\ 0 & 0 & -0.1 \end{pmatrix} , \\ \beta_2 = 0.6 \end{gathered}$$

for the second marginal, and a Gaussian copula with parameter $\rho = 0.5$. While any copula, or also simply a bivariate matrix–Weibull based on a MPH* construction could be used, we choose the Gaussian copula here to illustrate that the algorithm is able to work with any type of dependence structure. This distribution has theoretical mean $\mathbb{E}(\boldsymbol{X}) = (18.7997, 86.3711)'$. The sample has numerical values $\hat{\mathbb{E}}(\boldsymbol{X}) = (18.7690, 88.1637)'$ and $\hat{\rho}_\tau = 0.3431$.

We fit a bivariate matrix–Weibull distribution with $p_1 = p_2 = 3$ using Algorithm 4.10 with 1 500 steps (with a running time of 3 930 seconds for a step–length of 10^{-5}), getting the following parameters:

$$\hat{\boldsymbol{\alpha}} = (0.2101,\ 0,\ 0.7899,\ 0,\ 0,\ 0) ,$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -4.2507 & 0.5527 & 1.2916 & 0 & 0 & 2.4064 \\ 0 & -0.3069 & 0 & 0 & 0.3069 & 0 \\ 0.0089 & 0.2575 & -0.6903 & 0.4238 & 0 & 0 \\ 0 & 0 & 0 & -0.0946 & 0 & 0.0946 \\ 0 & 0 & 0 & 0.0360 & -0.0360 & 0 \\ 0 & 0 & 0 & 0.0026 & 0 & -0.2542 \end{pmatrix},$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

$$\beta_1 = 0.4689\,, \quad \beta_2 = 0.7340\,.$$

One sees that the algorithm estimates the shape parameters of the matrix–Weibull marginals reasonably well. The fitted distribution has mean $\mathbb{E}(\boldsymbol{X}) = (18.6988, 88.1166)'$, and from simulated data we get $\hat{\rho}_\tau = 0.3236$. The QQ and contour plots are given in Figure 4.16 and Figure 4.18, respectively. The log–likelihood of the fitted bivariate matrix–Weibull is $-39\,748$, which is to be compared with the log–likelihood $-39\,687.19$ using the original distribution.

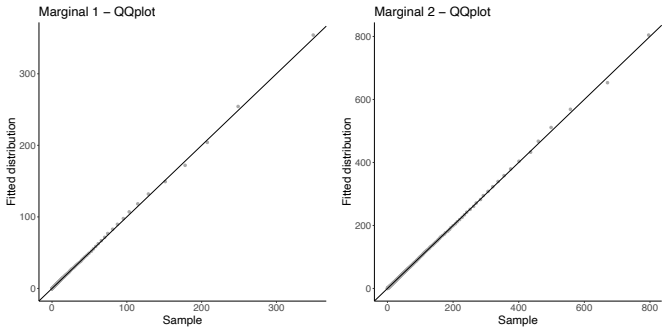

Figure 4.16: QQ plots of sample versus fitted bivariate matrix–Weibull distribution using Algorithm 4.10.

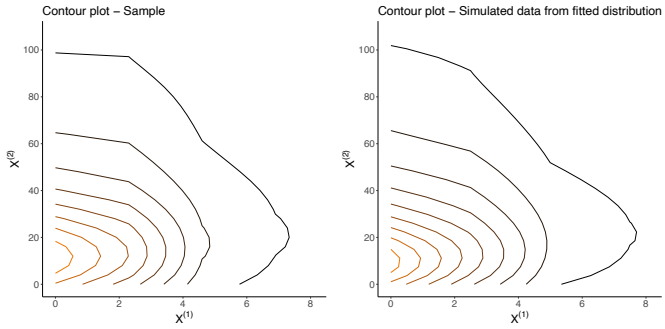

Figure 4.17: Contour plot of simulated sample (left) and contour plot of a simulated sample from the bivariate Matrix–Weibull distribution fitted using Algorithm 4.10 (right).

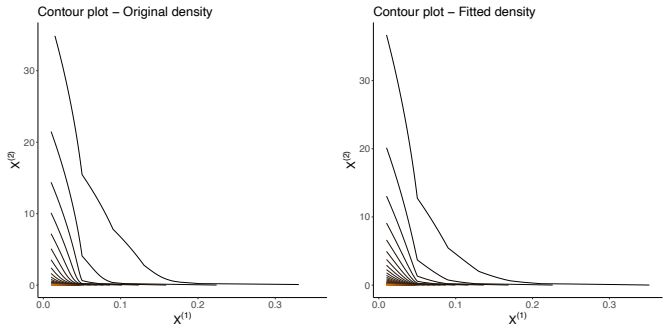

Figure 4.18: Contour plot of original bivarite matrix-Weibull (left) and contour plot of bivariate Matrix–Weibull distribution fitted using Algorithm 4.10 (right).

Remark 4.17. In all examples of this section the marginals were assumed to be of the same type (both matrix–Pareto or both matrix–Weibull). We would like to mention that the generality of Algorithm 4.10 also allows to fit models with marginals of different types (e.g. one marginal matrix–Pareto and the other matrix–Weibull).

4.5 Conclusion

In this paper we provided a guide for the statistical fitting of homogeneous and inhomogeneous phase–type distributions to data, both for the univariate and multivariate case. For that purpose, we derived a new EM algorithm for IPH distributions that are obtained through parameter–dependent transformations. In addition, we introduced new classes of multivariate distributions with IPH marginals and some attractive properties. As a by-product, we amended the estimation method proposed by Breuer [28] for the homogeneous MPH* case and illustrated its usefulness and flexibility. We furthermore discussed extensions for censored data and the fitting of the phase–type classes to given continuous joint distribution functions. The performance of the proposed algorithms was exemplified in various numerical examples, both on simulated and real data. In order to facilitate the implementation of the proposed algorithms for fitting this general class of distributions to given data, a respective R package is in preparation and will be made available on Cran.

Bibliography

[1] Ahlström, L., Olsson, M., and Nerman, O. A parametric estimation procedure for relapse time distributions. *Lifetime Data Analysis 5*, 2 (1999), 113–132.

[2] Albrecher, H., Beirlant, J., and Teugels, J. L. *Reinsurance: Actuarial and Statistical Aspects*. John Wiley & Sons, Chichester, 2017.

[3] Albrecher, H., and Bladt, M. Inhomogeneous phase-type distributions and heavy tails. *Journal of Applied Probability 56*, 4 (2019), 1044–1064.

[4] Albrecher, H., Bladt, M., and Bladt, M. Matrix Mittag-Leffler distributions and modeling heavy-tailed risks. *Extremes* (2020). To appear.

[5] Albrecher, H., Bladt, M., and Bladt, M. Multivariate fractional phase–type distributions. *arXiv preprint arXiv:2003.11122* (2020).

[6] Albrecher, H., Bladt, M., and Bladt, M. Multivariate matrix Mittag-Leffler distributions. *Annals of the Institute of Statistical Mathematics* (2020), 1–26.

[7] Arendarczyk, M., and Dębicki, K. Asymptotics of supremum distribution of a Gaussian process over a Weibullian time. *Bernoulli 17*, 1 (2011), 194–210.

[8] Arnold, B. C. *Pareto Distributions*. Chapman and Hall/CRC, 2015.

[9] Asmussen, S., Laub, P. J., and Yang, H. Phase-type models in life insurance: fitting and valuation of equity-linked benefits. *Risks 7*, 1 (2019), 17.

[10] Asmussen, S., Nerman, O., and Olsson, M. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* (1996), 419–441.

[11] Assaf, D., Langberg, N. A., Savits, T. H., and Shaked, M. Multivariate phase-type distributions. *Operations Research 32*, 3 (1984), 688–702.

[12] Auffinger, A., Ben Arous, G., and Péché, S. Poisson convergence for the largest eigenvalues of heavy tailed random matrices. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques 45*, 3 (2009), 589–610.

[13] Bai, Z., and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices*, second ed. Springer Series in Statistics. Springer, New York, 2010.

[14] Balkema, A. A., and De Haan, L. Residual life time at great age. *The Annals of Probability* (1974), 792–804.

[15] Basrak, B., and Segers, J. Regularly varying multivariate time series. *Stochastic Processes and their Applications 119*, 4 (2009), 1055–1080.

[16] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, Chichester, 2004.

[17] Bhatia, R. *Matrix Analysis*, vol. 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[18] BICKEL, P. J., AND LEVINA, E. Covariance regularization by thresholding. *The Annals of Statistics 36*, 6 (2008), 2577–2604.

[19] BICKEL, P. J., AND LEVINA, E. Regularized estimation of large covariance matrices. *The Annals of Statistics 36*, 1 (2008), 199–227.

[20] BINGHAM, N. H., GOLDIE, C. M., AND TEUGELS, J. L. *Regular Variation*, vol. 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987.

[21] BINGHAM, N. H., GOLDIE, C. M., AND TEUGELS, J. L. *Regular Variation*, vol. 27. Cambridge University Press, 1989.

[22] BLADT, M., GONZALEZ, A., AND LAURITZEN, S. L. The estimation of phase-type related functionals using Markov Chain Monte Carlo methods. *Scandinavian Actuarial Journal 2003*, 4 (2003), 280–300.

[23] BLADT, M., AND NIELSEN, B. F. Multivariate matrix-exponential distributions. *Stochastic Models 26*, 1 (2010), 1–26.

[24] BLADT, M., AND NIELSEN, B. F. *Matrix-Exponential Distributions in Applied Probability*. Springer, 2017.

[25] BLADT, M., NIELSEN, B. F., AND PERALTA, O. Parisian types of ruin probabilities for a class of dependent risk-reserve processes. *Scandinavian Actuarial Journal 2019*, 1 (2019), 32–61.

[26] BLADT, M., NIELSEN, B. F., AND SAMORODNITSKY, G. Calculation of ruin probabilities for a dense class of heavy tailed distributions. *Scandinavian Actuarial Journal 2015*, 7 (2015), 573–591.

[27] BLADT, M., AND ROJAS-NANDAYAPA, L. Fitting phase–type scale mixtures to heavy–tailed data and distributions. *Extremes* (2017), 1–29.

[28] BREUER, L. A semi-explicit density function for Kulkarni's bivariate phase-type distribution. *Stochastic Models 32*, 4 (2016), 632–642.

[29] BUN, J., BOUCHAUD, J.-P., AND POTTERS, M. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports 666* (2017), 1–109.

[30] CAI, J., AND LI, H. Conditional tail expectations for multivariate phase-type distributions. *Journal of Applied Probability 42*, 3 (2005), 810–825.

[31] CAI, J., AND LI, H. Multivariate risk model of phase type. *Insurance: Mathematics and Economics 36*, 2 (2005), 137–152.

[32] CAI, T., LIU, W., AND XIA, Y. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*, 501 (2013), 265–277.

[33] CAI, T. T., AND JIANG, T. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics 39*, 3 (2011), 1496–1525.

[34] CAI, T. T., AND JIANG, T. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis 107* (2012), 24–39.

[35] CAMARDA, C. G. Mortalitysmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software 50* (2012), 1–24.

[36] CLINE, D. B. H., AND HSING, T. Large deviation probabilities for sums of random variables with heavy or subexponential tails. *Technical report. Statistics Dept., Texas A&M University.* (1998).

[37] DAVIS, R. A., HEINY, J., MIKOSCH, T., AND XIE, X. Extreme value analysis for the sample autocovariance matrices of heavy-tailed multivariate time series. *Extremes 19*, 3 (2016), 517–547.

[38] DAVIS, R. A., AND HSING, T. Point process and partial sum convergence for weakly dependent random variables with infinite variance. *The Annals of Probability* (1995), 879–917.

[39] DAVIS, R. A., MIKOSCH, T., AND PFAFFEL, O. Asymptotic theory for the sample covariance matrix of a heavy-tailed multivariate time series. *Stochastic Processes and their Applications 126*, 3 (2016), 767–799.

[40] DAVIS, R. A., AND RESNICK, S. I. Limit theory for bilinear processes with heavy-tailed noise. *The Annals of Applied Probability 6*, 4 (1996), 1191–1210.

[41] DENISOV, D., DIEKER, A. B., AND SHNEER, V. Large deviations for random walks under subexponentiality: the big-jump domain. *The Annals of Probability 36*, 5 (2008), 1946–1991.

[42] DONOHO, D. High-dimensional data analysis: the curses and blessings of dimensionality. *Technical Report, Stanford University* (2000).

[43] DURRETT, R. Maxima of branching random walks vs. independent random walks. *Stochastic Processes and their Applications 9*, 2 (1979), 117–135.

[44] EINMAHL, U. Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *Journal of Multivariate Analysis 28*, 1 (1989), 20–68.

[45] EL KAROUI, N. Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability 19*, 6 (2009), 2362–2405.

[46] EMBRECHTS, P., KLÜPPELBERG, C., AND MIKOSCH, T. *Modelling Extremal Events for Insurance and Finance*, vol. 33 of *Applications of Mathematics (New York)*. Springer, Berlin, 1997.

[47] ERLANG, A. K. Sandsynlighedsregning og telefonsamtaler. *Nyt tidsskrift for Matematik 20* (1909), 33–39.

[48] FELLER, W. *An Introduction to Probability Theory and its Applications. Vol. II.* John Wiley & Sons, Inc., New York-London-Sydney, 1966.

[49] FISHER, R. A., AND TIPPETT, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society* (1928), vol. 24, Cambridge University Press, pp. 180–190.

[50] FOSS, S., KORSHUNOV, D., AND ZACHARY, S. *An Introduction to Heavy-Tailed and Subexponential Distributions*, vol. 6. Springer, 2011.

[51] FYODOROV, Y. V., AND BOUCHAUD, J.-P. Freezing and extreme-value statistics in a random energy model with logarithmically correlated potential. *Journal of Physics A: Mathematical and Theoretical 41*, 37 (2008), 372001.

[52] GANTERT, N., AND HÖFELSAUER, T. Large deviations for the maximum of a branching random walk. *Electronic Communications in Probability 23* (2018), Paper No. 34, 12.

[53] GEMAN, S. A limit theorem for the norm of random matrices. *The Annals of Probability 8*, 2 (1980), 252–261.

[54] GNEDENKO, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics* (1943), 423–453.

[55] GOMPERTZ, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London 115* (1825), 513–583.

[56] GRÜN, B., AND MILJKOVIC, T. Extending composite loss models using a general framework of advanced computational tools. *Scandinavian Actuarial Journal 2019*, 8 (2019), 642–660.

[57] HANAGAL, D. D. A multivariate Pareto distribution. *Communications in Statistics-Theory and Methods 25*, 7 (1996), 1471–1488.

[58] HANAGAL, D. D. A multivariate Weibull distribution. *Economic Quality Control 11* (1996), 193–200.

[59] HEINY, J. Large correlation matrices: a comparison theorem and its applications. *Submitted.* (2019).

[60] HEINY, J., AND MIKOSCH, T. Eigenvalues and eigenvectors of heavy-tailed sample covariance matrices with general growth rates: the iid case. *Stochastic Processes and their Applications 127*, 7 (2017), 2179–2207.

[61] HEINY, JOHANNES, M. T., AND JORGE, Y. Point process convergence for the off-diagonal entries of sample covariance matrices. *Annals of Applied Probability*. To appear.

[62] HERBERTSSON, A. Modelling default contagion using multivariate phase-type distributions. *Review of Derivatives Research 14*, 1 (2011), 1–36.

[63] HULT, H., LINDSKOG, F., MIKOSCH, T., SAMORODNITSKY, G., ET AL. Functional large deviations for multivariate regularly varying random walks. *The Annals of Applied Probability 15*, 4 (2005), 2651–2680.

[64] IBRAGIMOV, I., AND LINNIK, Y. V. Independent and Stationary Sequences of Random Variables. *Wolters-Noordhoff.* (1971).

[65] JENSEN, A. *A Distribution Model, Applicable to Economics.* Munksgaard, 1954.

[66] JIANG, T. The asymptotic distributions of the largest entries of sample correlation matrices. *The Annals of Applied Probability 14*, 2 (2004), 865–880.

[67] JIANG, T., AND XIE, J. Limiting behavior of largest entry of random tensor constructed by high-dimensional data. *arXiv:1910.12701v1* (2019).

[68] JOHNSTONE, I. M. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics 29*, 2 (2001), 295–327.

[69] JOHNSTONE, I. M., AND TITTERINGTON, D. M. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences 367*, 1906 (2009), 4237–4253.

[70] KALLENBERG, O. *Random Measures*, third ed. Akademie-Verlag, Berlin; Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London, 1983.

[71] KIBBLE, W. A two-variate gamma type distribution. *Sankhyā: The Indian Journal of Statistics* (1941), 137–150.

[72] KLÜPPELBERG, C., AND MIKOSCH, T. Large deviations of heavy-tailed random sums with applications in insurance and finance. *Journal of Applied Probability 34*, 2 (1997), 293–308.

[73] KULKARNI, V. G. A new class of multivariate phase type distributions. *Operations Research 37*, 1 (1989), 151–158.

[74] LEE, L. Multivariate distributions having Weibull properties. *Journal of Multivariate Analysis 9*, 2 (1979), 267–277.

[75] LI, D., LIU, W.-D., AND ROSALSKY, A. Necessary and sufficient conditions for the asymptotic distribution of the largest entry of a sample correlation matrix. *Probability Theory and Related Fields 148*, 1-2 (2010), 5–35.

[76] LI, D., QI, Y., AND ROSALSKY, A. On Jiang's asymptotic distribution of the largest entry of a sample correlation matrix. *Journal of Multivariate Analysis 111* (2012), 256–270.

[77] LI, D., AND ROSALSKY, A. Some strong limit theorems for the largest entries of sample correlation matrices. *The Annals of Applied Probability 16*, 1 (2006), 423–447.

[78] LINNIK, Y. V. Limit theorems allowing large deviations for the sums of independent variables I, II. *Theory of Probability and its Applications 6* (1961), 145–161.

[79] LIU, W.-D., LIN, Z., AND SHAO, Q.-M. The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *The Annals of Applied Probability 18*, 6 (2008), 2337–2366.

[80] MARČENKO, V. A., AND PASTUR, L. A. Distribution of eigenvalues in certain sets of random matrices. *Matematičeski Sbornik 72 (114)* (1967), 507–536.

[81] MARDIA, K. V., ET AL. Multivariate Pareto distributions. *The Annals of Mathematical Statistics 33*, 3 (1962), 1008–1015.

[82] MCNEIL, A. J., FREY, R., AND EMBRECHTS, P. *Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition*. Princeton University Press, 2015.

[83] MICHEL, R. Results on probabilities of moderate deviations. *The Annals of Probability* (1974), 349–353.

[84] MIKOSCH, T., AND NAGAEV, A. V. Large deviations of heavy-tailed sums with applications in insurance. *Extremes 1*, 1 (1998), 81–110.

[85] MIKOSCH, T., AND WINTENBERGER, O. A large deviations approach to limit theory for heavy-tailed time series. *Probability Theory and Related Fields 166*, 1-2 (2016), 233–269.

[86] NAGAEV, A. Limit theorems for large deviations where Cramér's conditions are violated. *Transactions of the Academy of Sciences of the Uzbek Republic, Series Mathematics and Physics 7* (1969), 17–22.

[87] NAGAEV, A. V. Integral limit theorems for large deviations when Cramer's condition is not fulfilled 1, 2. *Theory of Probability and its Applications 14* (1969), 51–64.

[88] NAGAEV, A. V. On a property of sums of independent random variables. *Theory of Probability and its Applications 22*, 2 (1978), 326–338.

[89] NAGAEV, S. V. Some limit theorems for large deviations. *Theory of Probability and its Applications 10*, 2 (1965), 214–235.

[90] NAGAEV, S. V. Large deviations of sums of independent random variables. *The Annals of Probability 7*, 5 (1979), 745–789.

[91] NEUTS, M. Probability distributions of phase type. In *Liber Amicorum Professor Emeritus H. Florin*. Department of Mathematics, University of Louvain, Belgium, 1975, pp. 173–206.

[92] NEUTS, M. F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Courier Corporation, 1994.

[93] NEUTS, M. F. *Algorithmic Probability: A Collection of Problems*, vol. 3. CRC Press, 1995.

[94] OKAMURA, H., DOHI, T., AND TRIVEDI, K. S. A refined EM algorithm for PH distributions. *Performance Evaluation 68*, 10 (2011), 938–954.

[95] OLSSON, M. Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics* (1996), 443–460.

[96] OLSSON, M. The EMpht programme. Manual. Chalmers University of Technology and Gothenburg University, 1998.

[97] PAUL, D., AND AUE, A. Random matrix theory in statistics: a review. *Journal of Statistical Planning and Inference 150* (2014), 1–29.

[98] PETROV, V. V. *Sums of Independent Random Variables (in Russian)*. Nauka, Moscow, 1972.

[99] PETROV, V. V. *Limit Theorems of Probability Theory*, vol. 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

[100] PICKANDS III, J., ET AL. Statistical inference using extreme order statistics. *The Annals of Statistics 3*, 1 (1975), 119–131.

[101] RESNICK, S., AND STĂRICĂ, C. Consistency of Hill's estimator for dependent data. *Journal of Applied Probability 32*, 1 (1995), 139–167.

[102] RESNICK, S. I. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2007.

[103] RESNICK, S. I. *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2008. Reprint of the 1987 original.

[104] ROZOVSKIĬ, L. V. Probabilities of large deviations on the whole axis. *Theory of Probability and its Applications 38*, 1 (1993), 79–109.

[105] SANCETTA, A., AND SATCHELL, S. The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory 20*, 3 (2004), 535–562.

[106] SLAVÍK, A. *Product Integration, Its History and Applications*. Matfyzpress Prague, 2007.

[107] SOSHNIKOV, A. Poisson statistics for the largest eigenvalues of Wigner random matrices with heavy tails. *Electronic Communications in Probability 9* (2004), 82–91.

[108] SOSHNIKOV, A. Poisson statistics for the largest eigenvalues in random matrix ensembles. In *Mathematical Physics of Quantum Mechanics*. Springer, 2006, pp. 351–364.

[109] VAN LOAN, C. Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control 23*, 3 (1978), 395–404.

[110] YAO, J., ZHENG, S., AND BAI, Z. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, 2015.

[111] ZHOU, W. Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Transactions of the American Mathematical Society 359*, 11 (2007), 5345–5363.