Modelling of Extremes

Adrien Hitz

Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

5 October 2016

Acknowledgements

I would like to start by thanking very much my supervisor Robin Evans for his guidance and trust through these three years of doctorate. I also thank Anthony Davison for suggesting graphical modeling of extreme value distributions, Richard Davis for proposing the notion of asymptotic conditional independence, Thomas Mikosch and Gennady Samorodnitsky for stimulating lectures on regular variation and long range dependence respectively, as well as David Steinsaltz, Steffen Lauritzen and Sebastian Engelke for valuable discussions. I am very grateful to the Marquise and Marquis de Amodio for their funding. I also benefited from inspiring conversations on my research with colleagues and friends Francois Lavergne, Thibaut Lienart, Thomas Lugrin, Charles Savoie, Owen Thomas, Alexander Vervuurt, Phyllis Wan, Matthieu Wilhelm, Nikola Yolov and Zhe Zhou. Finally, I owe much to Aurélia, Karen, Michael, Mireille, Marc and the rest of my family for their presence and love.

Abstract

This work focuses on statistical methods to understand how frequently rare events occur and what the magnitude of extreme values such as large losses is. It lies in a field called extreme value analysis whose scope is to provide support for scientific decision making when extreme observations are of particular importance such as in environmental applications, insurance and finance.

In the univariate case, I propose new techniques to model tails of discrete distributions and illustrate them in an application on word frequency and multiple birth data. Suitably rescaled, the limiting tails of some discrete distributions are shown to converge to a discrete generalized Pareto distribution and generalized Zipf distribution respectively.

In the multivariate high-dimensional case, I suggest modeling tail dependence between random variables by a graph such that its nodes correspond to the variables and shocks propagate through the edges. Relying on the ideas of graphical models, I prove that if the variables satisfy a new notion called asymptotic conditional independence, then the density of the joint distribution can be simplified and expressed in terms of lower dimensional functions. This generalizes the Hammersley–Clifford theorem and enables us to infer tail distributions from observations in reduced dimension. As an illustration, extreme river flows are modeled by a tree graphical model whose structure appears to recover almost exactly the actual river network.

A fundamental concept when studying limiting tail distributions is regular variation. I propose a new notion in the multivariate case called one-component regular variation, of which Karamata's and the representation theorem, two important results in the univariate case, are generalizations.

Eventually, I turn my attention to website visit data and fit a censored copula Gaussian graphical model allowing the visualization of users' behavior by a graph.

Contents

List of Figures

List of Abbreviations

cdf . . . . . Cumulative distribution function

w.p. . . . . . With probability

MDA . . . . . Maximum domain of attraction

RV . . . . . Regular variation

GEV . . . . . Generalized extreme value distribution

GPD . . . . . Generalized Pareto distribution

D-GPD . . . . Discrete Generalized Pareto distribution

GZD . . . . . Generalized Zipf distribution

BIC . . . . . Bayesian information criterion

QQ-plot . . . . Quantile-quantile plot

1 Introduction

1.1 Preliminary

The table below should help the reader find his way through the thesis as it indicates whether a chapter is rather theoretical or applied and if it is in the univariate or multivariate case.

	Univariate	*Multivariate*
Theoretical	Discrete Extremes (Ch. 2)	One-Component Regular Variation (Ch. 4) Graphical Modeling of Extremes (Ch. 5)
Applied		River Flows (Ch. 3) Website Visits (Ch. 6)

Depending on the reader's interest, we recommend the following chapters:

- univariate discrete extremes: 1.3 and 2,
- graphical modeling of extremes: 1.4, 3 and 5,
- regular variation: 1.5 and 4,
- applications in multivariate analysis: 3 and 6.

Before starting, I would like to add that the thesis is my own writing. Chapter 4 and 5 are based on Hitz and Evans [2016] and, together with Chapter 6 and 3, they

were the product of my work under the supervision of Robin Evans. For Chapter 2, I chose the direction of research and received advice and correction from Gennady Samorodnitsky and Richard Davis. The latter proposed the notion of asymptotic conditional independence and Anthony Davison suggested to combine graphical models and extreme value distributions as a subject for my Master's thesis.

1.2 Notation

Boldface type is used for vectors, $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to $x \in \mathbb{R}$ and $\lceil x \rceil$ the smallest integer larger than or equal to x. Unless stated otherwise, functions are from $\mathbb{R}$ to $\mathbb{R}$. We use the notation $\mathbb{R}_+ = (0, \infty)$, $\mathbb{N}_+ = \{1, 2, 3, \ldots\}$, $\mathbb{N}_0 = \{0\} \cup \mathbb{N}_+$ and p-norms on $\mathbb{R}^d$ are written $||\boldsymbol{x}||_p = (|x_1|^p + \ldots + |x_d|^p)^{1/p}$ for $p \geq 1$ and $||\boldsymbol{x}||_\infty = \max_{i=1,\ldots,d} |x_i|$. Moreover, $|A|$ denotes the cardinality of a set $A \subset \mathbb{R}^d$, $\lambda A = \{\lambda \boldsymbol{x} : \boldsymbol{x} \in A\}$, $\lambda \in \mathbb{R}$ and $1_A = 1_A(x)$ is 1 if $x \in A$ and 0 otherwise. The arrow "$\to$" denotes pointwise limit in t. Unless specified, convergence is on the whole domain of definition of the sequence. The relation $f \sim g$ means $f(t)/g(t) \to 1$. In equalities or pointwise convergences involving probability densities, we omit mentioning "a.e." which stands for almost everywhere. Random vectors $\boldsymbol{X} : \mathbb{R}^d \to \mathbb{R}$ are measurable functions from a probability space $(\Omega, \Sigma, \Pr)$ to the measurable space $\{\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\}$, where $\mathcal{B}(\cdot)$ is the Borel σ-algebra, see Appendix A.1 for details about measure theory. The cumulative distribution function of a random variable X is $F(x) = \Pr(X \leq x)$, its survival function is $\bar{F}(x) = \Pr(X > x)$ and the expectation of X is denoted by $\mathbb{E}(X)$. Weak convergence between measures is written "$\overset{w}{\Rightarrow}$" and vague convergence "$\overset{v}{\Rightarrow}$", see A.1 for a definition.

1.3 Modeling Discrete Extremes

Let $\boldsymbol{x} = \{x^{(i)}\}_{i=1}^n$ be i.i.d. realizations of a random variable X. We are interested in estimating the probability $\Pr(X \geq y)$ for a large $y \in \mathbb{R}$. This can be done using the empirical estimate

$$\hat{p} = |\{i : x^{(i)} \geq y\}|/n. \tag{1.1}$$

However, when y is too large, few observations fall above y resulting in a noisy estimator. Moreover, if y is larger than the maximal observation, then $\hat{p} = 0$ which can underestimate important risks if $\Pr(X \geq y) > 0$. This illustrates why extreme quantile estimation should sometimes not be performed naively but requires special techniques. As we will now explain, extreme value theory motivates a parametric family distribution which can be used to approximate the tails of a broad class of distributions.

Let X be a continuous or discrete random variable taking values in $[0, x_F)$ with survival function $\bar{F}$, where $x_F \in \mathbb{R}_+ \cup \{\infty\}$. One of the most famous results in probability theory is the central limit theorem. It states that for any i.i.d. copies $\{X^{(i)}\}_{i=1}^n$ of a random variable X whose variance is finite, there exist sequences c_n and d_n such that

$$d_n^{-1}\left(\sum_{i=1}^{n} X^{(i)} - c_n\right) \xrightarrow{d} Y,$$

as $n \to \infty$, where Y follows a normal distribution; one can choose $c_n = \mathbb{E}(X)$ and $d_n = \sqrt{n}$ (see e.g. Billingsley [1995]). Instead of being interested in the sum of the $X^{(i)}$, let us consider the maximum of them. The counterpart of the central limit theorem, known as the Fisher—Tippett-–Gnedenko theorem, states that if there exist sequences c_n and d_n such that

$$d_n^{-1}\left(\max_{i=1,\ldots,n} X^{(i)} - c_n\right) \xrightarrow{d} Z, \tag{1.2}$$

and Z is non-degenerate (i.e., it is not constant with probability 1), then Z follows a generalized extreme value distribution (GEV) defined by its cdf

$$G(x) = \exp\left\{-\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\} 1_{\{d_0<x<d_1\}}, \tag{1.3}$$

for $d_1 = \infty$ if $\xi \geq 0$ and $d_1 = \mu + \sigma/|\xi|$ if $\xi < 0$, $d_0 = -\infty$ if $\xi \leq 0$ and $d_0 = \mu - \sigma/\xi$ if $\xi > 0$ (see Embrechts et al. [1997]). If (1.2) holds, one says that X or its distribution belongs to the maximum domain of attraction of an extreme value distribution, written $X \in \mathrm{MDA}_\xi$. Most common continuous distributions such as normal, log-normal, Student, exponential, Weibull, Fréchet, Beta, Gamma, uniform belong to the maximum domain of attraction of an extreme value distribution.

Interestingly, when $X \in \mathrm{MDA}_\xi$, the behavior of $X \mid X \geq t$ as $t \to x_F$ is characterized as follows: $X \in \mathrm{MDA}_\xi$ is equivalent to saying that there exists a strictly positive sequence a_t such that

$$a_t^{-1}(X - t) \mid X \geq t \xrightarrow{d} Z, \tag{1.4}$$

as $t \to x_F$, for some Z following a non-degenerate probability distribution on $[0, \infty)$. In this case, one can take $d_n = a_t$ and $c_n = t$ for $n \equiv n(t) = 1/\bar{F}(t)$ in (1.2).

Under (1.4), Z follows a generalized Pareto distribution (GPD), defined by its cdf

$$F_{\mathrm{GPD}}(x; \sigma, \xi) = 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi} 1_{\{x < e_1\}}, \quad x \geq 0, \tag{1.5}$$

with $e_1 = \infty$ if $\xi \geq 0$, $e_1 = \sigma/|\xi|$ if $\xi < 0$, and $F_{\mathrm{GPD}}(x) = 1 - \exp(-x/\sigma)$ if $\xi = 0$ [Pickands, 1975]. In addition, the sequence of cdfs of $a_t^{-1}(X - t) \mid X \geq t$ converges uniformly to F_{GPD} on $[0, \infty)$. When $\xi > 0$, (1.4) is equivalent to

$$t^{-1}X \mid X \geq t \xrightarrow{d} Y, \tag{1.6}$$

on $[1, \infty)$ for some non-degenerate distribution Y; in this case, Y is Pareto distributed of order $1/\xi$ (and X is called regularly varying) (see Resnick [1987]).

Let us come back to the problem of estimating $\Pr(X \geq y)$ and assume that $X \in \mathrm{MDA}_\xi$. Let t be a threshold above which there are sufficiently many observations and write

$$\Pr(X \geq y) = \Pr(X - t \geq y_t \mid X \geq t) \Pr(X \geq t),$$

where $y_t = y - t$. At present, the empirical estimator in (1.1) can be used for $\Pr(X \geq t)$ and, if t is sufficiently large,

$$\Pr(X - t \geq y_t \mid X \geq t) = \Pr\{a_t^{-1}(X - t) \geq a_t^{-1} y_t \mid X \geq t) \approx 1 - F_{\mathrm{GPD}}(y_t; a_t\sigma, \xi). \tag{1.7}$$

This motivates the GPD method which consists in approximating the distribution of exceedances of X above a large threshold t by a GPD [Davison and Smith, 1990]. The normalizing sequence a_t that is unknown in practice is absorbed in the scale parameter σ which can be estimated by maximum likelihood jointly with the shape

parameter ξ. There is an inherent trade-off between bias and variance in the choice of t: when t is larger, the approximation above becomes more accurate but there are fewer observations above the threshold which increases the variance of the estimate. In the experimental parts of this thesis, the threshold levels are chosen by standard diagnostic methods, and thus the results are relatively stable over a range of local threshold choices.

Most common continuous distributions belong to MDA_ξ, and the GPD method often works well in practice when X follows a continuous distribution, although the approximation can be poor when u is not large enough (for instance if X is normal and the threshold t is around the 90th percentile of the distribution).

In the discrete case, two issues are apparent with applying the GPD method. First, a necessary condition for a discrete distribution F to be in MDA_ξ is that F is long-tailed, i.e.,

$$\bar{F}(t+1)/\bar{F}(t) \longrightarrow 1 \tag{1.8}$$

[Shimura, 2012, Anderson, 1970, 1980]. However, many common discrete distributions, including geometric, Poisson and negative binomial distributions, are not long-tailed. The second issue is that one approximates the tail of a distribution that is discrete by the GPD, a continuous distribution, which can be rough when many ties occur in the sample.

To overcome these issues, in Chapter 2, we suggest two alternative methods of modeling the tails of discrete observations, each one relying on a specific assumption on the underlying distribution.

The first one is to approximate $\Pr(X - u = k \mid X \geq u)$ for a large integer threshold u by

$$p_{\text{D-GPD}}(k; \sigma, \xi) = \bar{F}_{\text{GPD}}(k; \sigma, \xi) - \bar{F}_{\text{GPD}}(k+1; \sigma, \xi), \quad k \in \mathbb{N}_0, \tag{1.9}$$

where $\bar{F}_{\text{GPD}}$ is the survival function of the GPD. The probability mass function $p_{\text{D-GPD}}$ defines the discrete generalized Pareto distribution (D-GPD), which has been used by Prieto et al. [2014] to model road accidents; various aspects of the discrete

Pareto distribution (a special case) were studied in Krishna and Singh Pundir [2009], Buddana and Kozubowski [2014], and Kozubowski et al. [2015].

The second one approximates $\Pr(X - u = k \mid X \geq u)$ by

$$p_{\text{GZD}}(k; \sigma, \xi) = \frac{\left(1 + \xi \frac{k}{\sigma}\right)^{-1/\xi - 1} 1_{\{k < e_1\}}}{\sum_{i=0}^{\infty} \left(1 + \xi \frac{i}{\sigma}\right)^{-1/\xi - 1} 1_{\{i < e_1\}}}, \quad k \in \mathbb{N}_0, \tag{1.10}$$

called generalized Zipf distribution (GZD). The endpoint e_1 is defined as in (1.5). In the case $\xi > 0$, the GZD corresponds to a Zipf–Mandelbrot distribution. The probability mass function of the latter is usually written in the form $p(k) = (k + q)^{-s}/H_{s,q}$, for $k = 0, 1, 2, \ldots$ for $s > 1$ and $q > 0$, where $H_{s,q}$ is the Hurwitz-Zeta function [Mandelbrot, 1953]. It coincides with the GZD when $s = 1 + 1/\xi$ and $q = \sigma/\xi$. When $q = 1$ the distribution is a Zipf law which is sometimes presented as the counterpart of the Pareto distribution because its probability mass function, after a shift, can be written in a homogeneous form (Arnold [2015]). Zipf-type families of discrete distributions have been fitted to various data sets such as word frequencies (Booth [1967]), city sizes (Gabaix [1999]), company sizes (Axtell [2001]) and website hits (Clauset et al. [2009]). In the case $\xi < 0$, the GZD has a finite endpoint $\lceil e_1 - 1 \rceil$. Finally, when $\xi = 0$, we use the convention $(1 + \xi x)^{1/\xi} = e^x$ and the GZD (as well as the D-GPD) is simply a geometric distribution with probability of success $p = 1 - e^{-1/\sigma}$.

In the first part of Chapter 2, we justify theoretically the GPD, D-GPD and GZD methods in the case $\xi > 0$ by showing that under some conditions,

$$\frac{\Pr(X = k + u \mid X \geq u)}{q(k; \xi u, \xi)} \to 1, \quad k \in \mathbb{N}_0,$$

where $q \equiv f_{\text{GPD}}$, $p_{\text{D-GPD}}$ or p_{GZD}. A similar justification is provided for $q \equiv p_{\text{D-GPD}}$ and p_{GZD} in the case $\xi = 0$ under more restrictive assumptions shown to be satisfied by the geometric and negative binomial distributions.

In the second part, we simulate data from several discrete distributions and compare the ability of the three methods to estimate extreme quantiles. The D-GPD and GZD methods outperform the GPD when there are many tied observations,

otherwise the results are similar. The D-GPD having analytical survival and probability mass functions, it allows for more efficient inference than the GZD.

Finally, we study data sets counting the occurrence of words in English and French corpus and show that the three methods are appropriate to describe the tail distribution of these word frequencies. We also use the D-GPD and GZD methods to model the number of births in the United States and in France over 20 years; they appear to be useful for estimating the probability of regions far from the origin despite the observations being very small integer values.

1.4 Graphical Modeling of Extremes

In the multivariate case, we are interested in inferring the distribution of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ in regions far from the origin. Suppose that we observe i.i.d. realizations $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$ of $\boldsymbol{X}$ and want to estimate the probability that X_1 is large given the rest of the vector, that is $\Pr(X_1 \geq y \mid \boldsymbol{X}_{-1} = \boldsymbol{x}_{-1})$ for $\boldsymbol{x}_{-1} \in \mathbb{R}^{d-1}$ and some large $y > 0$. This is a difficult problem in general, especially in high dimensions.

Let us start again by an analogy with the central limit theorem. Suppose that $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$ are i.i.d. copies of $\boldsymbol{X}$. The multivariate central limit theorem states that if $\boldsymbol{X}$ has finite variance, then there exist sequences $\boldsymbol{c}_n$ and $\boldsymbol{d}_n$ in $\mathbb{R}^d$ such that

$$\boldsymbol{d}_n^{-1}\left(\sum_{i=1}^n \boldsymbol{X}^{(i)} - \boldsymbol{c}_n\right) \xrightarrow{d} \boldsymbol{Y},$$

(operations are here performed componentwise) where $\boldsymbol{Y}$ follows a multivariate normal distribution; one can take $\boldsymbol{c}_n = \mathbb{E}(\boldsymbol{X})$ and $\boldsymbol{d}_n = \sqrt{n}\boldsymbol{1}$. The counterpart result for maxima states that if there exist sequences $\boldsymbol{c}_n$ and $\boldsymbol{d}_n$ such that

$$\boldsymbol{d}_n^{-1}\left(\max_{i=1,\ldots,n} \boldsymbol{X}^{(i)} - \boldsymbol{c}_n\right) \xrightarrow{d} \boldsymbol{Z}, \tag{1.11}$$

and the univariate marginals of $\boldsymbol{Z}$ are non-degenerate, then $\boldsymbol{Z}$ follows a distribution with cdf of the form

$$G(\boldsymbol{x}) = \exp\{-\nu([-\boldsymbol{\infty}, \boldsymbol{x}]^c\}, \tag{1.12}$$

for some measure ν on $[-\infty, \infty)$ called the exponent measure, see e.g. Beirlant et al. [2004]. The distributions determined by (1.12) are called multivariate extreme value distributions and, in contrast to the central limit theorem, form a nonparametric class. In this case, the univariate marginals follow the GEV distribution defined in (1.3). Suppose further that the latter are all Fréchet distributed of order α, i.e., $\mu = 1$, $\sigma = 1/\alpha$ and $\xi = 1/\alpha$, and that ν is concentrated on $[\mathbf{0}, \boldsymbol{\infty}) \setminus \{\mathbf{0}\}$. It can then be shown that ν is finite on every Borel set A bounded away from $\mathbf{0}$ and is homogeneous of order $-\alpha$, i.e., $\nu(\lambda A) = \lambda^{-\alpha}\nu(A)$, $\forall \lambda > 0$; in addition, one can take $\boldsymbol{c}_n = \mathbf{0}$ and $\boldsymbol{d}_n = n\mathbf{1}$ in (1.11).

As in the univariate case, there is a relation between the limit of rescaled maxima and the behavior of $\boldsymbol{X}$ in regions far away from the origin: (1.11) and common Fréchet marginals is equivalent to

$$t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t \xrightarrow{d} \boldsymbol{Y}, \tag{1.13}$$

on $C_{||\cdot||_\infty} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_\infty \geq 1\}$ and $||\boldsymbol{Y}||_\infty \not\sim \delta_1$, the Dirac measure at 1. In this case, $\boldsymbol{Y} \sim \nu(\cdot)/\nu(C_{||\cdot||_\infty})$ and $Y_i \mid Y_i \geq 1$ is Pareto distributed of order $-\alpha$, i.e., its cdf is $1 - x^{-\alpha}$, for all i such that $\Pr(Y_i \geq 1) > 0$. The convergence above is usually written as a vague convergence to a non-null Radon measure on $[\mathbf{0}, \boldsymbol{\infty}] \setminus \{\mathbf{0}\}$ as in Basrak [2000], but for simplicity we stick to weak convergence. Condition (1.13) is appealing because it describes the limiting distribution of a random vector given that at least one of its marginals is extreme; in addition, it is not only restricted to vectors with positive values. (If $\boldsymbol{X}$ satisfies (1.13) we say that it is regularly varying, a notion discussed in the following section.) A stronger condition is

$$f_{t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}|| \geq t} \to f_{\boldsymbol{Y}}, \tag{1.14}$$

for a probability density $f_{\boldsymbol{Y}}$ on $C_{||\cdot||_\infty}$. It follows from the homogeneity property of ν that $f_{\boldsymbol{Y}}$ is homogeneous of order $-\alpha - d$. The following implication can be useful: (1.14) holds if

$$\frac{f_{\boldsymbol{X}}(t\boldsymbol{x})}{f_{\boldsymbol{X}}(t\mathbf{1})} \to v(\boldsymbol{x}), \tag{1.15}$$

on $C_{||\cdot||_\infty}$ for $v \not\equiv 0$ and if the sequence above is dominated by an integrable function. In this case, v is proportional to $f_{\boldsymbol{Y}}$. (A density satisfying (1.14) is called regularly varying.) For instance, if $\boldsymbol{X}$ follows a multivariate Student distribution with degrees of freedom ν, mean $\mathbf{0}$ and covariance matrix $\frac{\nu}{\nu-2}\Sigma$, then

$$f_{t^{-1}\boldsymbol{X}\,|\,||\boldsymbol{X}||_\infty \geq t}(\boldsymbol{x}) \to f_{\boldsymbol{Y}}(\boldsymbol{x}) = c^{-1}(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x})^{-(\nu+d)/2}, \tag{1.16}$$

for some normalizing constant $c > 0$. Up to a constant and a transformation of the univariate marginals, $f_{\boldsymbol{Y}}$ coincides with the density of the exponent measure of the extremal t distribution derived in Ribatet [2013]. By contrast, if $\boldsymbol{X}$ follows a multivariate Gaussian distribution with marginals transformed to unit Fréchet, one can show that

$$\frac{f_{\boldsymbol{X}}(t\boldsymbol{x})}{f_{\boldsymbol{X}}(t\mathbf{1})} \to x_1^{-d-1}\cdots x_d^{-d-1}, \tag{1.17}$$

(except if $|X_i| \overset{d}{=} |X_j|$ for some $i \neq j$), but the limit is not integrable on $C_{||\cdot||_\infty}$ so (1.14) does not hold. However, (1.13) is still true; in the bivariate case for instance, the limit is

$$(Y_1, Y_2) = \begin{cases} (W_1, 0) & \text{w.p. } \frac{1}{2}, \\ (0, W_2) & \text{w.p. } \frac{1}{2} \end{cases} \tag{1.18}$$

where W_1, W_2 are independent and Pareto distributed random variables of index 1. Why do we fall on this strange dependence pattern where extremes of Y_1 and Y_2 never occur jointly? Importantly, independence — and conditional independence — between marginals is only meaningful on a product space [Dawid, 2001]. Here, $\boldsymbol{Y}$ has values in the truncated cone $C_{||\cdot||_\infty}$, thus the range of Y_i depends on the other marginals and independence cannot be expressed, leading to the degenerate case above. In this case, one says that the marginals of $\boldsymbol{X}$ are asymptotically independent (or tail independent), a notion that we formally define in the next paragraph.

Let $\boldsymbol{X}$ be a random vector such that X_i has cumulative distribution function F for all $i = 1, \ldots, d$ and suppose for simplicity that $0 < F < 1$ on $\mathbb{R}$. Two marginals X_i and X_j are called asymptotically independent if

$$\Pr(\sigma X_i \geq t \mid \tilde{\sigma} X_j \geq t) \to 0, \quad \forall \sigma, \tilde{\sigma} \in \{-1, 1\}, \tag{1.19}$$

see e.g. de Haan and Resnick [1977] and Ledford and Tawn [1996]. If the limits in (1.19) exist and their values belong to $(0, 1]$ for all $\sigma, \tilde{\sigma} \in \{-1, 1\}$, one says that X_i and X_j are asymptotically dependent. For example, the marginals of a Student distribution are asymptotically dependent, whereas the ones of a Gaussian distribution are asymptotically independent (unless two marginals are fully dependent). As illustrated in (1.18), if the limit of $t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t$ exists and if $\boldsymbol{X}$ has marginals that are asymptotically independent, then the limiting distribution only allocates mass on the axis; this is problematic when modeling large observations of $\boldsymbol{X}$ and different techniques are required in this case, see e.g. Bortot et al. [2000], Maulik et al. [2002] and Wadsworth et al. [2017].

To summarize, we have introduced conditions under which the probability density of $t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t$ converges to a probability density $f_{\boldsymbol{Y}}$ on $C_{||\cdot||_\infty}$ and have seen that in this case $f_{\boldsymbol{Y}}$ is homogeneous. This motivates the use of a homogeneous density for modeling observations from $\boldsymbol{X}$ that are far from the origin in the case of asymptotic dependence. Modeling multivariate maxima is essentially the same problem: it requires the density of the cdf G defined in (1.12); derivating G leads to 2^d terms, hence, when d is modestly large, one typically performs a Taylor expansion of the exponential function around 0 (assuming all components of $\boldsymbol{x}$ are large) to approximate the density of G by the density of the exponent measure ν, which is, up to a constant, $f_{\boldsymbol{Y}}$.

In low dimensions, there is a rich class of models for G and thus for the homogeneous density $f_{\boldsymbol{Y}}$, such as Gumbel copula, Hüsler–Reiss model, extremal t copula as well as nonparametric procedures [Gudendorf and Segers, 2010]. A common strategy is to write $f_{\boldsymbol{Y}}(\boldsymbol{y}) = ||\boldsymbol{y}||^{-d-\alpha}\tilde{h}(\boldsymbol{y}/||\boldsymbol{y}||)$, where $\tilde{h}$ is an angular density on the sphere $\{\boldsymbol{w} \in \mathbb{R}^d : ||\boldsymbol{w}|| = 1\}$ for any norm $||\cdot||$ such that $\tilde{h}$ satisfies moment constraints to ensure that $\sigma Y_i \mid \sigma Y_i \geq 1$ is Pareto distributed for $\sigma \in \{-1, 1\}$. Modeling $f_{\boldsymbol{Y}}$ in larger dimensions has also received much attention in multivariate analysis. Several parametric models for h have been suggested such as the asymmetric logistic model [Tawn, 1990], the pairwise beta distribution [Cooley et al., 2010], its generalization [Ballani and Schlather, 2011], and the angular density

of the Hüsler–Reiss exponent measure [Engelke et al., 2015]. Semi-parametric approaches include mixtures of Dirichlet distributions [Boldi and Davison, 2007]. These models suffer limitations in high dimensions as the number of parameters explodes or they may lack flexibility to describe multivariate tails.

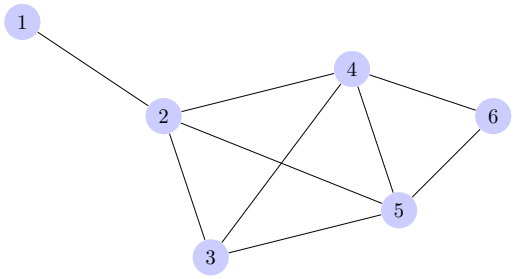

Figure 1.1: A decomposable graph with set of nodes $V = \{1, \dots, 6\}$. The maximal cliques of the graph are $\{1, 2\}$, $\{2, 3, 4, 5\}$ and $\{4, 5, 6\}$. The separator sets are $\{2\}$ and $\{4, 5\}$. If $\boldsymbol{Z}$ satisfies the pairwise Markov property according to this graph, the following conditional independence relations between pairs of its marginals hold: $Z_1 \perp\!\!\!\perp Z_i \mid \boldsymbol{Z}_{V\setminus\{1,i\}}$ for $i = 3, 4, 5, 6$, and $Z_6 \perp\!\!\!\perp Z_i \mid \boldsymbol{Z}_{V\setminus\{6,i\}}$ for $i = 1, 2, 3$.

We will try to apply the ideas of graphical models which have been successful in reducing dimensionality by imposing conditional independence relations between variables, see for instance Lauritzen [1996] and Wainwright and Jordan [2008]. We start by briefly introducing the definition of conditional independence and stating an important result. Let $\boldsymbol{Z}$ be a random vector with values in a product space $E \subseteq \mathbb{R}^d$ and with an a.e. continuous probability density $f_{\boldsymbol{Z}}$ w.r.t. some measure μ_0 and suppose that $\boldsymbol{Z}_A \perp\!\!\!\perp \boldsymbol{Z}_C \mid \boldsymbol{Z}_B$, which means that, knowing $\boldsymbol{Z}_B$, all information contained in $\boldsymbol{Z}_C$ is irrelevant to predict $\boldsymbol{Z}_A$. Formally,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{f_{AB}(\boldsymbol{z}_{AB}) f_{BC}(\boldsymbol{z}_{BC})}{f_B(\boldsymbol{z}_B)}, \quad \forall \boldsymbol{z} \in E.$$

replacing subscripts of the form $\boldsymbol{Z}_D$ by D for clarity. In other words, the joint density is determined by lower dimensional densities. There is a very convenient way of visualizing conditional independence between marginals by a graph. Consider the graph $\mathcal{G}$ with set of nodes $V = \{1, \dots, d\}$ (i.e., each node i corresponds to a marginal Z_i) and set of edges E defined as follows:

$$(i, j) \notin E \quad \text{implies} \quad Z_i \perp\!\!\!\perp Z_j \mid \boldsymbol{Z}_{V\setminus\{i,j\}},$$

that is, if there is no edge between i and j, then Z_i and Z_j are conditionally independent given the rest of the vector. In this case, one says that $\boldsymbol{Z}$ satisfies the pairwise Markov property according to $\mathcal{G}$. An example is provided in Figure 1.1. A clique of the graph is a subset of E where each node is connected to each other node in the subset, and the clique is called maximal if it cannot be extended. An example is provided in Figure 1.1.

One of the main results in graphical models is the Hammersley–Clifford theorem [Hammersley and Clifford, 1971]. It states that if $f_{\boldsymbol{Z}}$ is positive, then $\boldsymbol{Z}$ satisfies the pairwise Markov property according to a graph $\mathcal{G}$ if and only if its density factorizes w.r.t. the graph, that is,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = c^{-1} \prod_{C \in \mathcal{C}} \phi(\boldsymbol{z}_C),$$

where $\mathcal{C}$ is the set of maximal cliques, c is a normalizing constant and ϕ are some functions. If one assumes further that the graph is decomposable, then the joint density can be written in terms of some of the marginals of $\boldsymbol{Z}$:

$$f_{\boldsymbol{Z}} = \frac{\prod_{C \in \mathcal{C}} f_{\boldsymbol{Z}_C}}{\prod_{D \in \mathcal{D}} f_{\boldsymbol{Z}_D}},$$

where $\mathcal{D}$ is the multiset containing the separator sets. We refer to Appendix A.1.3 for the definitions of decomposability and separator sets. For instance, if $f_{\boldsymbol{Z}} > 0$, then $\boldsymbol{Z}$ satisfies the pairwise Markov property according to the decomposable graph in Figure 1.1 if and only if

$$f_{\boldsymbol{Z}} = f_{12}\, f_{345|2}\, f_{6|45} = \frac{f_{12} f_{2345} f_{456}}{f_2 f_{45}}.$$

Coming back to our context, the first complication is that the limiting tail density $f_{\boldsymbol{Y}}$ lies on the truncated cone $C_{||\cdot||_\infty}$ and conditional independence is only meaningful on a product space as mentioned earlier. To overcome this issue, we will express the distribution on the following sets:

$$C_k := \{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\}, \quad k = 1, \ldots, d.$$

Note that a distribution is determined on $C_{||\cdot||_\infty}$ exactly when it is determined on C_k for all k; moreover, each C_k is a product space. Let $\mathcal{G}$ be a graph and consider the following assumption:

$$f_{\boldsymbol{Y}\,|\,|Y_k|\geq 1} \text{ factorizes w.r.t. } \mathcal{G} \text{ on } C_k,\ \forall k = 1, \ldots, d. \tag{1.20}$$

Alternatively, one can assume that $f_{\boldsymbol{Y}\,|\,|Y_k|\geq 1}$ factorizes on C_k for $k = 1$ only, as discussed by Segers [2016] in the case of a tree graphical model.

Recall that there is no homogeneous probability density on $C_{||\cdot||_\infty}$ for which two marginals are independent (as illustrated in (1.18)). There is therefore doubt that conditional independence can be achieved under the homogeneity constraint. In fact, probability densities on $C_{||\cdot||_\infty}$ satisfying (1.20) are easily constructed. Consider for instance

$$f_{\boldsymbol{Y}}(y_1, y_2, y_3) = c^{-1} \frac{(y_1 + y_2)^{-3}(y_2 + y_3)^{-3}}{y_2^{-2}},$$

for some normalizing constant $c > 0$, which is homogeneous of order -4 and clearly satisfies

$$Y_1 \perp\!\!\!\perp Y_3 \mid \boldsymbol{Y}_2,\ |Y_k| \geq 1, \quad \forall k = 1, 2, 3.$$

Other instances will be presented such as a factorization w.r.t. a tree in Example 5.1.3 and a factorization of the Hüsler–Reiss exponent measure density in Example 5.1.4.

Suppose now that the probability density of $t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t$ converges to a probability density $f_{\boldsymbol{Y}} > 0$ on $C_{||\cdot||_\infty}$ known to be homogeneous. Two crucial questions are addressed in Chapter 5: first, what assumptions on $\boldsymbol{X}$ are necessary to enforce a factorization of $f_{\boldsymbol{Y}}$? Second, if $f_{\boldsymbol{Y}}$ can be expressed as a product of low-dimensional functions, how would we infer these from low-dimensional marginals of $\boldsymbol{X}$?

As an answer, we say that two marginals of $\boldsymbol{X}$ are *asymptotically conditionally independent* given the rest, written

$$X_i \,\tilde{\perp\!\!\!\perp}\, X_j \mid \boldsymbol{X}_D,$$

for $D = \{1, \dots, d\} \setminus \{i, j\}$, if

$$f_{t,\, D\cup\{i,j\}}\, f_{t,\, D} - f_{t,\, D\cup\{i\}}\, f_{t,\, D\cup\{j\}} \to 0, \quad \text{on } C_k,\ \forall k,$$

as $t \to \infty$, where f_t is the density of $\boldsymbol{X}_t := t^{-1}\boldsymbol{X} \mid |X_k| \geq t$. In Proposition 5.2.1, we generalize Hammersley–Clifford theorem for sequences, which states in this context that $\boldsymbol{X}$ satisfies the *asymptotic* pairwise Markov property according to a graph $\mathcal{G}$ if and only if (1.20) holds, i.e.,

$$f_{\boldsymbol{Y}\,|\,|Y_k|\geq 1} \text{ factorizes w.r.t. } \mathcal{G} \text{ on } C_k,\ \forall k = 1, \dots, d.$$

Under the additional assumption that $\mathcal{G}$ is decomposable, we demonstrate in Theorem 5.1.2 that (1.20) is equivalent to the existence of a class $\{h_S\}_{S\in\mathcal{C}\cup\mathcal{D}}$ of homogeneous, positive and measurable functions on $\mathbb{R}^{|S|}$, integrable on $C^S_{||\cdot||_\infty}$, satisfying

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = c^{-1}\frac{\prod_{C\in\mathcal{C}} h_C(\boldsymbol{y}_C)}{\prod_{D\in\mathcal{D}} h_D(\boldsymbol{y}_D)}, \quad \boldsymbol{y} \in C_{||\cdot||_\infty}, \tag{1.21}$$

for some $c > 0$, and such that $h_D = \int_{\mathbb{R}^{|C\setminus D|}} h_C\, d\boldsymbol{y}_{C\setminus D}$, $D \subset C$, $D \in \mathcal{D}$, $C \in \mathcal{C}$ (recall that $\mathcal{C}$ and $\mathcal{D}$ are defined in Appendix A.1.3). In this case, h_S is proportional to the probability density of $\boldsymbol{Y}_S \mid ||\boldsymbol{Y}_S||_\infty \geq 1$ on $C^S_{||\cdot||_\infty}$ for all $S \in \mathcal{C}\cup\mathcal{D}$. This allows us to perform inference in lower dimensions, for instance by relying upon the approximation

$$t^{-1}\boldsymbol{X}_S \mid ||\boldsymbol{X}_S||_\infty \geq t \ \approx\ \boldsymbol{Y}_S \mid ||\boldsymbol{Y}_S||_\infty \geq 1 \tag{1.22}$$

for large t. An alternative approach based on censoring non-extremal observations is often preferred to (1.22) (see e.g. Coles [2001]); we motivate it in the next paragraph.

Let $\boldsymbol{X}^C_t$ be a sequence of censored versions of $\boldsymbol{X}$ defined as

$$\boldsymbol{X}^C_t := \begin{cases} (t^{-1}X_1\, 1_{|X_1|\geq t}, \dots, t^{-1}X_d\, 1_{|X_d|\geq t}) \mid ||\boldsymbol{X}||_\infty \geq t & \text{w.p. } p, \\ \boldsymbol{0} & \text{w.p. } 1-p, \end{cases} \tag{1.23}$$

for $p \in (0, 1)$, and taking values in

$$F^d := \{(-\infty, -1] \cup \{0\} \cup [1, \infty)\}^d.$$

If $f_{t^{-1}\boldsymbol{X}\,|\,||\boldsymbol{X}||_\infty \geq t} \to f_{\boldsymbol{Y}} > 0$ on $C_{||\cdot||_\infty}$, then the density of $\boldsymbol{X}_t^C$ converges to the one of $\boldsymbol{Y}^C$, defined as the censored version of $\boldsymbol{Y}$, that is,

$$\boldsymbol{Y}^C := \begin{cases} (Y_1 1_{|Y_1|\geq 1}, \ldots, Y_d 1_{|Y_d|\geq 1}) & \text{w.p. } p, \\ \mathbf{0} & \text{w.p. } 1-p. \end{cases} \tag{1.24}$$

It is helpful to think of Y_i^C as a variable taking value zero when there is no extreme and values above and below 1 when there is a positive and negative extreme respectively. Since $f_{\boldsymbol{Y}}$ is homogeneous, $\sigma Y_i^C \mid \sigma Y_i^C \geq 1$ is Pareto distributed for $\sigma \in \{-1, 1\}$. The probability density of $\boldsymbol{Y}^C$ is defined w.r.t. the mixed measure in (3.2) and we refer to it as a *censored homogeneous density*. For example, the censored homogeneous limiting density in (1.16) is

$$f_{\boldsymbol{Y}^C}(\boldsymbol{x}_A, \mathbf{0}_{A^c}) = p\, c^{-1} \int_{|\boldsymbol{x}_{A^c}|<\mathbf{1}} (\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x})^{-(\nu+d)/2} d\boldsymbol{x}_{A^c}, \tag{1.25}$$

for $A = \{i : |x_i| \geq 1\}$. One can show that $f_{\boldsymbol{X}_t^C} \to f_{\boldsymbol{Y}^C}$ implies $f_{\boldsymbol{X}_{t,S}^C} \to f_{\boldsymbol{Y}_S^C}$ for any subset S, where $\boldsymbol{Y}_S^C$ is the censored version of the marginal $\boldsymbol{Y}_S$. This motivates a different approximation than (1.22) to infer the limiting density h_S:

$$\boldsymbol{X}_{t,S}^C \approx \boldsymbol{Y}_S^C, \tag{1.26}$$

for large t. In other words, h_S is inferred by using information about the occurrence of extremes and their magnitude only, but the value taken by non extremal marginals is assumed to be irrelevant.

To sum up, if $f_{t^{-1}\boldsymbol{X}\,|\,||\boldsymbol{X}||_\infty \geq t} \to f_{\boldsymbol{Y}} > 0$ for a probability density $f_{\boldsymbol{Y}}$ on $C_{||\cdot||_\infty}$, then the following equivalence holds: $\boldsymbol{X}$ satisfies the asymptotic pairwise Markov property according to a decomposable graph $\mathcal{G}$ if and only if $f_{\boldsymbol{Y}}$ can be written as in (1.21), i.e., expressed in terms of low-dimensional homogeneous densities h_S that are proportional to $f_{\boldsymbol{Y}_S\,|\,||\boldsymbol{Y}_S||_\infty \geq 1}$ on $C^S_{||\cdot||_\infty}$ for $S \in \mathcal{C} \cup \mathcal{D}$. Moreover, $\boldsymbol{Y}_S \mid ||\boldsymbol{Y}_S||_\infty \geq 1$ corresponds to the limit of $t^{-1}\boldsymbol{X}_S \mid ||\boldsymbol{X}_S||_\infty \geq t$, and its censored version $\boldsymbol{Y}_S^C$ is the limit of $\boldsymbol{X}_{t,S}^C$. This motivates the following procedure for inferring tails of multivariate distributions in high dimensions.

i. Assume that the data are i.i.d. realizations of a random vector $\tilde{\boldsymbol{X}}$.

ii. For every i and $\sigma \in \{-1, 1\}$, select a large threshold u_i^σ for $\sigma\tilde{X}_i$, fit a generalized Pareto distribution to $\sigma\tilde{X}_i \mid \sigma\tilde{X}_i \geq u_i^\sigma$, transform it to unit Pareto on $\sigma[1, \infty)$ and set observations between the thresholds to 0. This results in a vector $\boldsymbol{X}^C$ with values in $F^d = \{(-\infty, -1] \cup \{0\} \cup [1, \infty)\}^d$.

iii. Test asymptotic independence and asymptotic conditional independence between the marginals of $\boldsymbol{X}$ and select a decomposable graph $\mathcal{G}$.

iv. For each $S \in \mathcal{C} \cup \mathcal{D}$, choose a parametric censored homogeneous density such as (1.25) and estimate it using observations from $\boldsymbol{X}_S^C$. Check that t is sufficiently large and that it makes sense to use a homogeneous density.

v. This determines the homogeneous densities h_S and the full joint limiting density

$$f_{\boldsymbol{Y}} = c^{-1} \frac{\prod_{C \in \mathcal{C}} h_C}{\prod_{D \in \mathcal{D}} h_D}.$$

Remark that it is standard in extreme value analysis to transform the marginals of $\tilde{\boldsymbol{X}} \mid ||\tilde{\boldsymbol{X}}||_\infty \geq u$ to a common distribution such as unit Fréchet, see [Coles and Tawn, 1991]; the approach here slightly differs as it only transforms observations of the marginals of $\tilde{\boldsymbol{X}}$ whose absolute values fall above large thresholds.

In this thesis, we will not focus on testing asymptotic conditional independence, which is left for future work. Several methods exist for testing asymptotic independence, see e.g. Falk and Michel [2006].

We have seen how to construct valid high-dimensional homogeneous densities on $C_{||\cdot||_\infty}$ by enforcing a factorization of $f_{\boldsymbol{Y}}$ on the sets $C_k = \{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\}$ for all k. In this paragraph, we discuss a few alternatives: the first is to impose a factorization of $f_{\boldsymbol{Z}}$, the probability density of rescaled maxima in (1.11). Papastathopoulos and Strokorb [2016] show that conditional independence between two marginals of $\boldsymbol{Z}$ given the rest of the vector implies their independence, that is, the density of an extreme value distribution can only factorize trivially. A second alternative is to consider a factorization of $f_{\boldsymbol{Y}^C}$, the density of the censored vector, which lies on the product space $F^d = \{(-\infty, -1] \cup \{0\} \cup [1, \infty)\}^d$ [Hitz

and Evans, 2016]. This approach seems appealing because it is directly in line with the inference procedure using censored observations as described in (1.26). Nevertheless, we later found that the homogeneity constraint imposes prohibitive restrictions on the low-dimensional densities h_S, complicating the construction of a valid factorization (see Example A.3.3). Eventually, we mention the approach followed by Gissibl and Klüppelberg [2015] which consists in propagating extremes through a directed acyclic graph using max-linear functions.

In practice, we should be careful when modeling extremes using a homogeneous multivariate distribution because it excludes fundamental dependence structures such as the asymptotic independence regime. To avoid this limitation, a possibility is to assume a factorization of $f_{\boldsymbol{X}_t^C}$ w.r.t. $\mathcal{G}$ for all t sufficiently large, that is

$$f_{\boldsymbol{X}_t^C} = \frac{\prod_{C \in \mathcal{C}} f_{\boldsymbol{X}_{t,C}^C}}{\prod_{D \in \mathcal{D}} f_{\boldsymbol{X}_{t,D}^C}}. \tag{1.27}$$

One can then model $\boldsymbol{X}_{t,S}^C$ for $S \in \mathcal{C} \cup \mathcal{D}$ using a probability density that does not necessarily fulfill the homogeneity constraint, see e.g. a class of bivariate models based on censored copulas in Example 3.2.2. Modeling $\boldsymbol{X}_{t,S}^C$ using a censored homogeneous density can still be appropriate if $f_{t^{-1}\boldsymbol{X}_S \,|\, ||\boldsymbol{X}_S||_\infty \geq t} \to f_{\boldsymbol{Y}^{(S)}}$ for some random vector $\boldsymbol{Y}^{(S)}$. For instance, if the latter condition holds for $S = \{i\}$ $\forall i = 1, \ldots, d$ and if the univariate marginals of $\boldsymbol{X}_t^C$ are mutually independent for t large enough (i.e., $\mathcal{G}$ has no edges), then (1.27) can be approximated as follows:

$$f_{\boldsymbol{X}_t^C}(\boldsymbol{x}_A, \boldsymbol{0}_{A^c}) \approx \prod_{i \in A} \Pr(\sigma_{x_i} Y^{(i)} \geq 1) |x_i|^{-2} \prod_{i \in A^c} \Pr(|Y^{(i)}| \leq 1), \tag{1.28}$$

for large t, where $A = \{i : |x_i| \geq 1\}$ and σ_x denotes the sign of x — compare (1.28) to the degenerate limit in (1.18). The approach described in this paragraph is not established on a proper limiting distribution anymore, but allows us to capture a broader class of dependence structure. It is illustrated in an application on extreme river flows in Chapter 3.

Let us come back to the introductory problem of estimating the probability $\Pr(\tilde{X}_1 \geq y_1 \mid \tilde{\boldsymbol{X}}_{-1} = \tilde{\boldsymbol{x}}_{-1})$ for some large y_1. After having transformed the vector

into $\boldsymbol{X}^C$ as in (ii.), we rely on (1.27) to find

$$\begin{aligned}\Pr(X_1 \geq y_1 \mid \boldsymbol{X}_{-1} = \boldsymbol{x}_{-1}) &\approx \Pr(X_1^C \geq y_1 \mid \boldsymbol{X}_A^C = \boldsymbol{x}_A, \boldsymbol{X}_{A^c}^C = \boldsymbol{0}) \\ &\approx \Pr(X_1^C \geq y_1 \mid \boldsymbol{X}_{A\cap N}^C = \boldsymbol{x}_{A\cap N}, \boldsymbol{X}_{A^c\cap N}^C = \boldsymbol{0}), \quad (1.29)\end{aligned}$$

where $A = \{i \neq 1 : |x_i| \geq 1\}$ and $N = \{i : (1, i) \in E\}$, the set of neighbors of 1 in the graph $\mathcal{G}$. In other words, the conditional distribution in (1.29) is obtained by ignoring information about small observations and irrelevant marginals. If N is a clique, (1.29) corresponds to the conditional distribution of the model fitted to $\boldsymbol{X}_N^C$. If N is not a clique, the expression is not analytical in general but can be computed using techniques such as Markov chain Monte Carlo (see e.g. Brooks et al. [2011]).

The multivariate Gaussian distribution is a graphical model with two remarkable properties: its conditional distributions are known, and conditional independence corresponds to the following constraint: if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, then $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{\{1,\ldots,d\}\setminus\{i,j\}}$ if and only if $\Sigma_{ij}^{-1} = 0$. The benefit is that graph selection and parameter estimation can be performed at once, which led to successful methods such as the graphical lasso [Friedman et al., 2008]. The conditional distributions of the multivariate Student are also known, and in this case a zero in entry (i, j) of the inverse covariance matrix implies that X_i and X_j are conditionally uncorrelated given the rest of the vector [Finegold and Drton, 2011]. In Chapter 6, we describe a multivariate distribution on F^d that is based on the Gaussian and Student graphical models and can readily be used for modeling multivariate extremes. The model is illustrated on website visits and, in Chapter 3, we apply it to the extreme river flows and compare it to a tree graphical model.

1.5 One-Component Regular Variation

Regular variation for a random variable X is defined as

$$t^{-1}X \mid X \geq t \xrightarrow{d} Y, \quad (1.30)$$

on $[1, \infty)$ as $t \to \infty$, where $Y \not\sim \delta_1$, the Dirac measure at 1. This notion is important as it describes the limiting behavior of extreme values of X as their

magnitude grows. Denoting by $\bar{F}$ the survival function of X and $\bar{G}$ the one of Y, (1.30) can equivalently be written as

$$\frac{\bar{F}(tx)}{\bar{F}(t)} \xrightarrow{d} \bar{G}(x).$$

Regular variation for functions is defined similarly: a measurable function $u : (0, \infty) \to (0, \infty)$ is said to be regularly varying if there exists some function $v > 0$ such that

$$\frac{u(tx)}{u(t)} \to v(x), \quad x > 0, \tag{1.31}$$

(Karamata [1933], see also Seneta [1973]). A striking result is that in this case there exists $\alpha \in \mathbb{R}$ such that $v(x) = x^\alpha$, which is shown as follows: $\forall x, y > 0$,

$$v(xy) \leftarrow \frac{u(txy)}{u(t)} = \frac{u(txy)}{u(tx)} \frac{u(tx)}{u(t)} \to v(y)v(x),$$

and from Cauchy's functional equation, any measurable and positive function satisfying $v(xy) = v(x)v(y)$, $\forall x, y > 0$, has the form $v(x) = x^\alpha$ (see e.g. Bingham et al. [1989]). For this reason, regular variation is denoted by $u \in \mathrm{RV}_\alpha$.

We now give a few examples of regularly varying functions. The Fréchet distribution of index $\alpha > 0$ has survival function $\bar{F}(x) = 1 - \exp(-x^{-\alpha})$ which satisfies $\bar{F} \in \mathrm{RV}_{-\alpha}$ and its probability density belongs to $\mathrm{RV}_{-\alpha-1}$. The survival function $\bar{F}(x) = e^{-\lambda x}$ of the exponential distribution of index $\lambda > 0$ is not regularly varying because its tail decays too rapidly to zero. A possible misunderstanding is to believe that $u \in \mathrm{RV}_\alpha$ implies $u(t) \sim t^\alpha$, which is incorrect: the probability density $f(x) = x^{-2} \log x$ on $[1, \infty)$ satisfies $f \in \mathrm{RV}_{-2}$, nevertheless, $f(t)/t^{-2} \to \infty$. Regular variation of the survival function $\bar{F}$ does not guarantee that the limit is also a survival function: $\bar{F}(x) = (1+\log x)^{-1}$ satisfies $\bar{F} \in \mathrm{RV}_0$. Similarly, regular variation of the probability density does not ensure that the limit is also a probability density.

Regularly varying functions are known to have interesting properties. First, if $u \in \mathrm{RV}_{-\alpha}$ for $\alpha > 0$, then the convergence in (1.31) is actually uniform in x on $[1, \infty)$. Second, the representation theorem states that $u \in \mathrm{RV}_{-\alpha}$ for $\alpha \in \mathbb{R}$ if and only if

$$u(x) = c(x) \exp\left\{ - \int_{x_0}^{x} \alpha(z) z^{-1} dz \right\},$$

for x_0 sufficiently large and some measurable functions $c(\cdot)$, $\alpha(\cdot)$ such that $c(t) \to c > 0$, $\alpha(t) \to \alpha$, as $t \to \infty$. A direct consequence is that $\forall \epsilon > 0$, $\exists\, c_1, c_2 > 0$ such that

$$c_1 x^{-\alpha-\epsilon} < \frac{u(tx)}{u(t)} < c_2 x^{-\alpha+\epsilon},$$

for x sufficiently large. This relation is useful to bound u by the integrable function $c_2 x^{-\alpha+\epsilon}$ when $\alpha > 0$ before applying dominated convergence.

At present, suppose that $\bar{U}(x) = \int_x^\infty u(z)dz$ exists for x sufficiently large and let $\alpha > 0$. A third result known as Karamata's theorem states that $u \in \mathrm{RV}_{-\alpha-1}$ if and only if

$$\frac{tu(t)}{\bar{U}(t)} \to \alpha, \tag{1.32}$$

and in this case $\bar{U} \in \mathrm{RV}_{-\alpha}$.

These properties of regularly varying functions can directly be translated for distributions. If the survival function of X satisfies $\bar{F} \in \mathrm{RV}_{-\alpha}$ for $\alpha > 0$, then it can be shown that (1.30) holds, i.e., $t^{-1}X \mid X \geq t \xrightarrow{d} Y$ on $[1, \infty)$ for some random variable $Y \not\sim \delta_1$, and thus in this case, Y follows a Pareto distribution of order α. We denote condition (1.30) by $X \in \mathrm{RV}_{-\alpha}$.

According to Karamata's theorem, if X admits a probability density f_X and $\alpha > 0$, then $f_X \in \mathrm{RV}_{-\alpha-1}$ if and only if

$$\frac{tf_X(t)}{\bar{F}(t)} \to \alpha. \tag{1.33}$$

and in this case $\bar{F} \in \mathrm{RV}_{-\alpha}$ and thus $X \in \mathrm{RV}_{-\alpha}$. It is quite remarkable that the convergence in (1.33) in only one point is strong enough to ensure regular variation of X, that is, convergence of $t^{-1}X \mid X \geq t$ on the whole space. Checking $f_X \in \mathrm{RV}_{-\alpha-1}$ is relatively easy if one has access to the density as it only requires to compute the limit of $f_X(tx)/f_X(t)$.

We now discuss the multivariate case and see if similar properties can be derived. A measurable function u on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ is regularly varying with limit v if $u(\lambda \mathbf{1})$ and $v(\lambda \mathbf{1})$ are positive for all $\lambda > 0$ and if

$$\frac{u(t\boldsymbol{x})}{u(t\mathbf{1})} \to v(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}, \tag{1.34}$$

see Resnick [1987]. In this case, the limit is homogeneous, i.e., there exists $\alpha \in \mathbb{R}$ such that $v(\lambda \boldsymbol{x}) = \lambda^{\alpha} v(\boldsymbol{x})$, $\forall \lambda > 0$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$, and we write $u \in \mathrm{RV}_{\alpha}$.

The counterpart notion for distributions is defined as follows: a random vector $\boldsymbol{X}$ is regularly varying if

$$t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_{\infty} \geq t \xrightarrow{d} \boldsymbol{Y}, \tag{1.35}$$

on $C_{||\cdot||_{\infty}} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_{\infty} \geq 1\}$ such that $||\boldsymbol{Y}||_{\infty} \not\sim \delta_1$. Similarly to the univariate case, it is possible to show that the distribution ν of the limit $\boldsymbol{Y}$ on $C_{||\cdot||}$ is homogeneous of order $-\alpha$ for some $\alpha > 0$, that is, $\nu(\lambda A) = \lambda^{-\alpha}\nu(A)$, for all Borel sets A and $\lambda \geq 1$ and we denote in this case regular variation by $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}$. If $\boldsymbol{X}$ has a probability density satisfying $f_{\boldsymbol{X}} \in \mathrm{RV}_{-\alpha-1}$ for $\alpha > 0$ such that the sequence is dominated by an integrable function, then $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}$.

From (1.34), it is not clear if the representation and Karamata's theorem can be extended into the multivariate setting. In Chapter 4, we suggest a notion for functions called *one-component regular variation* and show that

- it can express the multivariate regular variation in (1.34),
- the representation and Karamata's theorem find a generalization for such functions.

We call u *regularly varying w.r.t. its first component* if

$$\frac{u(tx, \boldsymbol{y})}{u(t, \boldsymbol{1})} \to v(x, \boldsymbol{y}),$$

where u and v are measurable non-negative functions on $(0, \infty) \times \mathbb{R}^{d-1}$ such that $u(\cdot, \boldsymbol{1}) > 0$ and $v(\cdot, \boldsymbol{1}) > 0$. In this case, $v(x, \boldsymbol{y}) = x^{\alpha}\, h(\boldsymbol{y})$ for some non-negative function h on $\mathbb{R}^{d-1}$ such that $h(\boldsymbol{1}) = 1$ and we write $u \in \mathrm{RV}^{x}_{-\alpha}(h)$.

One-component regular variation includes regular variation: for instance, $\tilde{u}(\boldsymbol{x})$ is regularly varying on $\mathbb{R}^d_+ \setminus \{\boldsymbol{0}\}$ if and only if $u(r, \boldsymbol{\theta}) := \tilde{u}(r\boldsymbol{\theta})$ is regularly varying in its first component on $\mathbb{R}_+ \times S_{d-1}$, where $S_{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}|| = 1\}$.

The representation theorem becomes: $u \in \mathrm{RV}^{x}_{\alpha}(h)$ if and only if

$$u(x, \boldsymbol{y}) = c(x) \exp\left\{\int_{x_0}^{x} \alpha(z) z^{-1} dz\right\} q(x, \boldsymbol{y}),$$

for measurable functions s.t. $c(t) \to c > 0$, $\alpha(t) \to \alpha$, $q(t, \boldsymbol{y}) \to h(\boldsymbol{y})$ as $t \to \infty$.

The extended version of Karamata's theorem states that if $\alpha > 0$ and $\bar{U}(x, \boldsymbol{y}) := \int_x^\infty u(z, \boldsymbol{y})dz$ exists $\forall x, \boldsymbol{y}$, then $u(x, \boldsymbol{y}) \in \mathrm{RV}^x_{-\alpha-1}(h)$ if and only if

$$\frac{tu(t, \boldsymbol{y})}{\bar{U}(t, \mathbf{1})} \to \alpha\, h(\boldsymbol{y}). \tag{1.36}$$

In this case, $\bar{U} \in \mathrm{RV}^x_{-\alpha}(h)$.

One-component regular variation also finds a natural formulation for distributions: we say that $(X, \boldsymbol{Y})$ is *regularly varying w.r.t. its first component*,

$$(t^{-1}X, \boldsymbol{Y}) \mid X \geq t \xrightarrow{d} \boldsymbol{Z}, \tag{1.37}$$

on $[1, \infty) \times \mathbb{R}^{d-1}$, for a random vector $\boldsymbol{Z}$ such that $Z_1 \not\sim \delta_1$. Limits of random vectors with an extreme component have already been studied [Heffernan and Resnick, 2007, Resnick and Zeber, 2014] and it is known that $\boldsymbol{Z} \sim P_\alpha \times H$, where P_α is the Pareto distribution of order $\alpha > 0$ and H is a probability distribution on $\mathbb{R}^{d-1}$. We denote (1.37) by $(X, \boldsymbol{Y}) \in \mathrm{RV}^x_{-\alpha}(H)$. Maulik et al. [2002] introduce (1.37) as a stronger condition than asymptotic conditional independence; in particular, they show that if $(X, \boldsymbol{Y}) \in \mathrm{RV}^x_{-\alpha}$ and $(X, \boldsymbol{Y}) \in \mathrm{RV}_{-\alpha}$, then $\boldsymbol{Z}$ only takes values on the axis. We use this notion in a different context here: interpreting X as the radius and $\boldsymbol{Y}$ as the angle of a vector $\boldsymbol{V}$, we will see that regular variation of $\boldsymbol{V}$ is equivalent to one-component regular variation of $\varphi(\boldsymbol{V}) = (X, \boldsymbol{Y})$, where φ is a change of variable to radial coordinates defined in (4.7). For instance, a non-negative random vector $\boldsymbol{V}$ is regularly varying if and only $(||\boldsymbol{V}||, \boldsymbol{V}/||\boldsymbol{V}||) \in \mathrm{RV}^x_{-\alpha}(H)$. This is a well-known result and in this case H is called the angular measure [Beirlant et al., 2004]. If H admits a density, the latter corresponds to the angular density $\tilde{h}$ introduced in Section 1.4.

As a last point, let us come back to the definition of a regularly varying function u. We have seen that if $u(tx)/u(t) \to v(x) > 0$, then x^α for some $\alpha \in \mathbb{R}$. Multiplying x by t is an arbitrary choice of operation that can be replaced by a more general scaling [Bingham and Ostaszewski, 2010]. Let us choose $t \star x \mapsto T^{-1}\{T(t)T(x)\}$ for

a diffeomorphism $T : [e_0, e_1) \to [1, \infty)$ for $e_0 \in \mathbb{R}$, $e_1 \in \mathbb{R} \cup \{\infty\}$. A diffeomorphism is a differentiable bijection whose inverse is differentiable as well. If

$$\frac{u(t \star x)}{u(t \star e_0)} \to v(x) > 0,$$

then $v(x) = T(x)^\alpha$. For instance, if $e_1 = \infty$ and $\star$ is addition, $v(x) = e^{\alpha x}$. If $x \star y = ||x, y||_p$ for $p > 0$, $v(x) = e^{\alpha x^p}$. If $e_1 < \infty$ and $x \star y = x + y - e_1 xy$, then $v(x) = (1 - e_1^{-1} x)^\alpha$. This extends regular variation to other decays than the power law. It is not just a trivial transformation, but enriches the representation and Karamata's theorem that we extend to more general decays in Chapter 4.

2
Discrete Extremes

Contents

This chapter attempts to widen the scope of extreme value analysis applied to discrete-valued data containing many tied observations. Extreme values of a random variable X are commonly modeled using the generalized Pareto distribution as explained in 1.7, a method that often gives good results in practice. When X is discrete, we propose two other methods using a discrete generalized Pareto and a generalized Zipf distribution respectively. The latter are not only motivated for a more general class of discrete distributions on theoretical grounds, but they perfom well for estimating rare events in several simulated data. The chapter ends with an illustration of the methods on two discrete-valued data sets consisting of word frequency and multiple birth data.

2.1 Maximal Domain of Attraction for Discrete Distributions

Recall that a random variable X with survival function $\bar{F}_X(x) = \Pr(X > x)$ is in the maximum domain of attraction (MDA) of an extreme value distribution if there exists a strictly positive sequence a_t such that

$$\frac{\bar{F}_X(a_t x + t)}{\bar{F}_X(t)} \xrightarrow{d} \bar{G}(x),$$

as $t \to x_F$, where $\bar{G}$ is the survival function of a non-degenerate probability distribution and x_F is the right endpoint of $\bar{F}_X$. We have seen in (1.5) that in this case the limit is a generalized Pareto distribution (GPD), i.e., $\bar{G}(x) = \bar{F}_{\text{GPD}}(x; \sigma, \xi)$, for $\sigma > 0$ and $\xi \in \mathbb{R}$ and we write $X \in \text{MDA}_\xi$ or $\bar{F}_X \in \text{MDA}_\xi$. We now focus on deriving asymptotic results for the tails of some discrete distributions in the case $x_F = \infty$; these results are obviously not applicable for discrete distributions having a finite endpoint.

As explained in Chapter 1.3, common discrete distributions such as geometric, Poisson and negative binomial do not belong to the maximum domain of attraction of an extreme value distribution. This leads us to define a maximal domain of attraction specific for discrete distributions.

We say that a discrete random variable X with non-negative values and infinite right endpoint satisfies the discrete extreme value condition, which we write as $X \in \text{D-MDA}_\xi$, if there exists a survival function $\bar{F}_Y$ of a random variable $Y \in \text{MDA}_\xi$ with $\xi \geq 0$ such that $\bar{F}_X \equiv \bar{F}_Y$ on $\mathbb{N}_0$, i.e., the equality in distribution, $X \overset{d}{=} \lfloor Y \rfloor$, holds. We call Y an extension of X and such an extension is not unique. Shimura [2012] showed that $X \in \text{MDA}_\xi$ for some $\xi \geq 0$ if and only if $X \in \text{D-MDA}_\xi$ and X is long-tailed, that is, $\bar{F}_X(x+1)/\bar{F}_X(x) \to 1$ as $x \to \infty$. It was also shown by Shimura [2012] that geometric, Poisson and negative binomial distributions are in D-MDA. This set is, therefore, strictly larger than the set of all discrete distributions in MDA_ξ.

Let $X \in \text{D-MDA}_\xi$ and $Y \in \text{MDA}_\xi$ be the corresponding extension satisfying $X \overset{d}{=} \lfloor Y \rfloor$. Then, for large integers u,

Shimura [2012] proves that $X \in \text{D-MDA}_\xi$ and X is long-tailed if and only if $X \in \text{MDA}_\xi$, for $\xi \geq 0$. (X is long-tailed if $\bar{F}_X(t+1)/\bar{F}_X(t) \to 1$ as defined in (1.8)). In addition, he also shows that such discrete distributions as geometric, Poisson and negative binomial distributions are in the set D-MDA. This set is, therefore, strictly larger than the set of all discrete distributions in MDA. Let $X \in \text{D-MDA}_\xi$, and let $Y \in \text{MDA}_\xi$ be a random variable such that $X \stackrel{d}{=} \lfloor Y \rfloor$. If u is an integer, then $\{X \geq u\} \equiv \{Y \geq u\}$. Thus for any large integers u,

$$\begin{aligned} \Pr(X - u = k \mid X \geq u) &= \Pr(Y - u \geq k \mid Y \geq u) - \Pr(Y - u \geq k+1 \mid Y \geq u) \\ &\approx \bar{F}_{\text{GPD}}(k; \sigma a_u, \xi) - \bar{F}_{\text{GPD}}(k+1; \sigma a_u, \xi) \\ &= p_{\text{D-GPD}}(k; \sigma a_u, \xi), \end{aligned} \tag{2.1}$$

using (1.7). Note that the extension Y of X is not unique and thus the scale parameter in (2.1) remains undetermined.

We now discuss an alternative assumption on the distribution of the discrete random variable X that will allow us to construct another approximation of its tail. Let p_X be the probability mass function of X and suppose that there exists a non-negative random variable $Y \in \text{MDA}_{\xi/(1+\xi)}$ with $\xi \geq 0$ such that $p_X \equiv c\bar{F}_Y$ on $\mathbb{N}_0 \cap [d, \infty)$ for some $c > 0$ and $d \geq 0$, a condition that we denote by $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$. For any large integers u,

$$\begin{aligned} \Pr(X - u = k \mid X \geq u) &= \frac{p_X(u+k)/p_X(u)}{\sum_{i=0}^{\infty} p_X(u+i)/p_X(u)} \\ &= \frac{P(Y > u+k)/P(Y > u)}{\sum_{i=0}^{\infty} P(Y > u+i)/P(Y > u)} \\ &\approx \frac{\left(1 + \xi \frac{k}{a_u \sigma}\right)^{-1/\xi - 1} 1_{\{k < e_1\}}}{\sum_{i=0}^{\infty} \left(1 + \xi \frac{i}{a_u \sigma}\right)^{-1/\xi - 1} 1_{\{i < e_1\}}} \\ &= p_{\text{GZD}}(k; \sigma a_u, \xi). \end{aligned} \tag{2.2}$$

relying again on (1.7). We will show that geometric, Poisson and negative binomial probability mass functions belong to D-MDA_0 so the tail approximation (2.2) makes sense in these cases. In addition, we will see that $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ for $\xi \geq 0$ implies $X \in \text{MDA}_\xi$.

Equations (2.1) and (2.2) motivate the approximation of the exceedances of some discrete random variables above a large threshold by a D-GPD and GZD respectively, which we refer to as the D-GPD and GZD method. If $\xi \geq 0$, then

$$\sup_{k\in\mathbb{N}_0}\left|\frac{f_{\mathrm{GPD}}(k;\sigma,\xi)}{q(k;\sigma,\xi)}-1\right| \underset{\sigma\to\infty}{\longrightarrow} 0,$$

for either $q = f_{\mathrm{GZD}}$ or $q = f_{\text{D-GPD}}$. One thus expects the GPD, D-GPD and GZD methods to give similar results when the scale parameter σ is large.

In what follows, we provide theoretical justification for the D-GPD and GZD methods, followed by experimental results assessing their performance. We end by two applications on word frequency and multiple birth data.

2.2 Theoretical Results

We now describe a number of properties of the approximation procedures introduced earlier.

Proposition 2.2.1. *If $X \in$ D-MDA$_\xi$, then there exists a positive sequence a_u such that*

$$\lim_{u\in\mathbb{N}_0,\, u\to\infty}\sup_{k\in\mathbb{N}_0}\Big|P(X=u+k\mid X\geq u)-p_{\textit{D-GPD}}(k;a_u,\xi)\Big|=0. \tag{2.3}$$

Proof. By assumption, there exists a random variable Y and a positive sequence $\tilde{a}_u$ for $u>0$ such that $X \overset{d}{=} \lfloor Y \rfloor$ and the sequence of functions $P\{\tilde{a}_u^{-1}(Y-u)\geq x \mid Y\geq u\}$, $x\geq 0$, converges uniformly, as $u\to\infty$, to the function $\bar{F}_{\mathrm{GPD}}(x;\sigma,\xi)$, $x\geq 0$, for some $\sigma>0$ and $\xi\geq 0$. For a positive integer u we let $a_u=\tilde{a}_u/\sigma$. Since $u\in\mathbb{N}_0$, it holds $\{X\geq u\}\equiv\{Y\geq u\}$. Then

$$\begin{aligned}
&\sup_{k\in\mathbb{N}_0}\mid P(X=u+k\mid X\geq u)-p_{\text{D-GPD}}(k;a_u,\xi)\mid\\
&=\sup_{k\in\mathbb{N}_0}\Big|P\Big(\tilde{a}_u^{-1}(Y-u)\geq\tilde{a}_u^{-1}k\mid Y\geq u\Big)-P\Big(\tilde{a}_u^{-1}(Y-u)\geq\tilde{a}_u^{-1}(k+1)\mid Y\geq u\Big)\\
&\qquad\qquad-\bar{F}_{\mathrm{GPD}}(k;a_u,\xi)+\bar{F}_{\mathrm{GPD}}(k+1;a_u,\xi)\Big|\\
&\leq 2\sup_{x\geq 0}\Big|P\left(\tilde{a}_u^{-1}(Y-u)\geq x\mid Y\geq u\right)-\bar{F}_{\mathrm{GPD}}(x;\sigma,\xi)\Big|\to 0
\end{aligned}$$

as $u\to\infty$ over the integers. □

We will need the following auxiliary result.

Lemma 2.2.2. *If $\xi > 0$, then, as $u \to \infty$,*

$$u^{1/\xi} H_{1+1/\xi,u} \to \xi,$$

where $H_{s,q} = \sum_{i=0}^{\infty}(q+i)^{-s}$ is the Hurwitz-Zeta function.

Proof.

$$\begin{aligned} u^{1/\xi} H_{1+1/\xi,u} &= u^{-1} \sum_{i=0}^{\infty} \left(1 + \frac{i}{u}\right)^{-1-1/\xi} = u^{-1} \int_0^{\infty} \left(1 + \frac{\lfloor x \rfloor}{u}\right)^{-1-1/\xi} dx \\ &= \int_0^{\infty} \left(1 + \frac{\lfloor uy \rfloor}{u}\right)^{-1-1/\xi} dy \ \to \ \xi, \end{aligned}$$

because $(1 + \lfloor uy \rfloor / u)^{-1-1/\xi} \to (1+y)^{-1-1/\xi}$; exchange limit and integral by dominated convergence since the integrand is bounded by $y^{-1-1/\xi}$. □

Recall that a positive and measurable function f on $[1, \infty)$ is regularly varying if there exists a positive function ℓ such that

$$\lim_{u \to \infty} \frac{f(ux)}{f(u)} \to \ell(x), \quad x \geq 1.$$

In this case, there exists $\alpha \in \mathbb{R}$ such that $\ell(x) = x^{\alpha}$ and we write $f \in \mathrm{RV}_{\alpha}$ (see Bingham et al. [1989]). If $f \in \mathrm{RV}_{-\alpha}$ for $\alpha \geq 0$, then

$$\lim_{u \to \infty} \sup_{x \in [1,b]} \left| \frac{f(ux)}{f(u)} - x^{-\alpha} \right| \to 0, \tag{2.4}$$

for $b = \infty$ if $\alpha > 0$ and $b < \infty$ if $\alpha = 0$. If $f \in \mathrm{RV}_{-\alpha}$ for $\alpha > 0$, then for any $\epsilon > 0$ there is $u_{\epsilon} \in (0, \infty)$ such that

$$e^{-\epsilon} x^{-\alpha-\epsilon} \leq \frac{f(ux)}{f(u)} \leq e^{\epsilon} x^{-\alpha+\epsilon}, \quad x \geq 1, \tag{2.5}$$

for $u \geq u_{\epsilon}$; these bounds are called the Potter bounds. We say that X is regularly varying if $\bar{F}_X \in \mathrm{RV}_{-\alpha}$ for some $\alpha > 0$, a necessary and sufficient condition for $X \in \mathrm{MDA}_{1/\alpha}$.

The following result justifies more rigorously the approximation suggested in (2.2).

Theorem 2.2.3. *If $p_X \in$ D-MDA$_{\xi/(1+\xi)}$ for $\xi > 0$, then $X \in$ MDA$_\xi$ and for any sequence of nonnegative integers k_u such that $\sup_u k_u/u < \infty$,*

$$\lim_{u \in \mathbb{N}_0,\, u \to \infty} \frac{P(X = k_u + u \mid X \geq u)}{p_{GZD}(k_u; \xi u, \xi)} = 1. \tag{2.6}$$

Proof. By assumption, there exists a survival function $\bar{F}$ such that $\bar{F}(k) = p_X(k)$ for k large enough and $\bar{F} \in \mathrm{RV}_{-1/\xi-1}$. The last condition is equivalent to $\bar{F}(\lfloor \cdot \rfloor) \in \mathrm{RV}_{-1/\xi-1}$ (Shimura [2012]). Therefore,

$$\begin{aligned}\frac{\bar{F}_X(ux)}{\bar{F}_X(u)} &= \frac{\sum_{i=\lceil ux \rceil}^{\infty} p_X(i)}{\sum_{i=u} p_X(i)} = \frac{\int_{ux}^{\infty} \bar{F}(\lfloor y \rfloor) dy - \int_{ux}^{\lceil ux \rceil} \bar{F}(\lfloor y \rfloor) dy}{\int_u^{\infty} \bar{F}(\lfloor y \rfloor) dy} \\ &= \frac{u \int_x^{\infty} \bar{F}(\lfloor uz \rfloor)/\bar{F}(\lfloor u \rfloor) dz - (\lceil ux \rceil - ux)\bar{F}(\lfloor ux \rfloor)/\bar{F}(\lfloor u \rfloor)}{u \int_1^{\infty} \bar{F}(\lfloor uz \rfloor)/\bar{F}(\lfloor u \rfloor) dz} \to x^{-1/\xi}, \quad x \geq 1,\end{aligned}$$

applying (2.5) and dominated convergence. This shows $\bar{F}_X \in \mathrm{RV}_{-1/\xi}$, that is, $X \in \mathrm{MDA}_\xi$.

For the second part of the theorem, we have

$$\frac{P(X = k_u + u \mid X \geq u)}{p_{\mathrm{GZD}}(k_u; \xi u, \xi)} = \frac{\bar{F}(u + k_u)/\bar{F}(u)}{(1 + k_u/u)^{-1/\xi-1}} \frac{\sum_{i=0}^{\infty} (1 + i/u)^{-1/\xi-1}}{\sum_{i=0}^{\infty} \bar{F}(u+i)/\bar{F}(u)}.$$

By the uniform convergence (2.4) and the fact that k_u grows at most linearly fast, we conclude that

$$\frac{\bar{F}(u + k_u)/\bar{F}(u)}{(1 + k_u/u)^{-1/\xi}} \to 1,$$

as $u \to \infty$ over the integers. Second, Lemma 2.2.2 yields

$$u^{-1} \sum_{i=0}^{\infty} (1 + i/u)^{-1/\xi-1} \to \xi.$$

Third, it follows from (2.5) that for $\epsilon \in (0, 1/\xi)$, there exists $u_\epsilon > 0$ such that for $u \geq u_\epsilon$,

$$u^{-1} \sum_{i=0}^{\infty} \bar{F}(u+i)/\bar{F}(u) \leq u^{-1} e^{\epsilon} \sum_{i=0}^{\infty} \left(1 + \frac{i}{u}\right)^{-1-1/\xi+\epsilon} \to \frac{\xi e^{\epsilon}}{1 - \xi\epsilon}, \tag{2.7}$$

using again Lemma 2.2.2. A lower bound is found in the same manner and we let $\epsilon \to 0$ to conclude. □

We now present a tail equivalence property between the GZD and D-GPD probability mass functions and the GPD density function. A direct consequence is that the denominator p_{GZD} in (2.6) can be replaced either by $p_{\text{D-GPD}}$ or by f_{GPD}.

Proposition 2.2.4. *If $\xi \geq 0$, then*

$$\lim_{\sigma\to\infty} \sup_{k\in\mathbb{N}_0} \left| \frac{p_{D\text{-}GPD}(k;\sigma,\xi)}{p_{GZD}(k;\sigma,\xi)} - 1 \right| = \lim_{\sigma\to\infty} \sup_{k\in\mathbb{N}_0} \left| \frac{p_{D\text{-}GPD}(k;\sigma,\xi)}{f_{GPD}(k;\sigma,\xi)} - 1 \right| = 0.$$

Proof. Suppose first that $\xi > 0$,. Then

$$\begin{aligned} \frac{p_{\text{D-GPD}}(k;\sigma,\xi)}{f_{\text{GPD}}(k;\sigma,\xi)} &= \frac{\left(1+\xi\frac{k}{\sigma}\right)^{-1/\xi} - \left(1+\xi\frac{k+1}{\sigma}\right)^{-1/\xi}}{\frac{1}{\sigma}\left(1+\xi\frac{k}{\sigma}\right)^{-1/\xi-1}} \\ &= \left\{1 - \left(1+\frac{\xi}{\sigma+\xi k}\right)^{-1/\xi}\right\}(\sigma+\xi k) \to 1, \end{aligned} \tag{2.8}$$

uniformly in $k \in \mathbb{N}_0$ as $\sigma \to \infty$. Furthermore,

$$\sup_{k\in\mathbb{N}_0} \frac{f_{\text{GPD}}(k;\sigma,\xi)}{p_{\text{GZD}}(k;\sigma,\xi)} = \sigma^{-1} \sum_{i=0}^{\infty} (1+\xi i/\sigma)^{-1/\xi-1} \to 1 \tag{2.9}$$

by Lemma 2.2.2. In the case $\xi = 0$,

$$p_{\text{D-GPD}}(k;\sigma,0)/f_{\text{GPD}}(k;\sigma,0) = p_{\text{GZD}}(k;\sigma,0)/f_{\text{GPD}}(k;\sigma,0) = \sigma(1-e^{-1/\sigma}) \to 1$$

as $\sigma \to \infty$. □

Next we extend Theorem 2.2.3 to the case $\xi = 0$. Recall that a distribution F is in MDA_0 if and only if the survival function has a representation

$$\bar{F}(x) = c(x) \exp\left\{-\int_0^x \frac{1}{a(y)} dy\right\}, \quad -\infty < x < x_F, \tag{2.10}$$

where $c(\cdot)$ is a positive function with $c(x) \to c > 0$ as $x \uparrow x_F$, and $a(\cdot)$ is a positive, differentiable function $a(\cdot)$ with $\lim_{x\uparrow x_F} a'(x) = 0$. If $c(x) = c$ on (d, x_F) for some $d < x_F$, then we say that the distribution F satisfies the von Mises condition. The function $a(\cdot)$ in (2.10) is sometimes called the *auxiliary function*. Note, however, that it is only uniquely defined (on (d, x_F)) under the von Mises condition; see Embrechts et al. [1997]. Recall that we only consider the case of unbounded support, i.e. $\bar{F}_X > 0$.

Proposition 2.2.5. *Suppose that $p_X \in MDA_0$ and, moreover, that a distribution F such that $p_X(k) = \bar{F}(k)$ has the property that an auxiliary function of $\bar{F}$ satisfies $\lim_{x\to\infty} a(x) = \sigma \in [0,\infty]$. Then*

$$\lim_{u\in\mathbb{N}_0,\, u\to\infty} P(X = k+u \mid X \geq u) = p_{Ge}(k;\sigma), \quad k \in \mathbb{N}_0\,, \tag{2.11}$$

where $p_{Ge}(k;\sigma) = (1 - e^{-1/\sigma})\, e^{-k/\sigma}$ *is the probability mass function of a geometric distribution if* $0 < \sigma < \infty$, *and* $p_{Ge}(k;\infty) = p_{Ge}(k;0) = 0$. *Furthermore, if* $\sigma \in [0,\infty)$, *then* $X \in$ *D-MDA*$_0$.

Proof. Note that for large integers u,

$$P(X = k + u \mid X \geq u) = \frac{\bar{F}(k+u)/\bar{F}(u)}{\sum_{i=0}^{\infty} \bar{F}(i+u)/\bar{F}(u)}\,. \tag{2.12}$$

We have for every $i = 0, 1, 2, \ldots,$

$$\bar{F}(i+u)/\bar{F}(u) = \frac{c(i+u)}{c(u)} \exp\left\{-\int_0^i 1/a(u+y)dy\right\} \to e^{-i/\sigma}$$

as $u \to \infty$. If $0 < \sigma < \infty$, then the dominated convergence theorem gives us

$$\sum_{i=0}^{\infty} \bar{F}(i+u)/\bar{F}(u) \to \sum_{i=0}^{\infty} e^{-i/\sigma} = 1/(1 - e^{-1/\sigma})\,,$$

and (2.11) follows. If $\sigma = \infty$,

$$\sum_{i=0}^{\infty} \bar{F}(i+u)/\bar{F}(u) \to \infty$$

by Fatou's lemma, and (2.11), once again, follows. If $\sigma = 0$, the claim follows from the fact that the denominator in (2.12) cannot be smaller than 1.

For the second part of the proposition, it follows from

$$p_X(n) = c(n) \exp\left\{-\int_0^n \frac{1}{a(y)} dy\right\}$$

for all n and $a(y) \to \sigma \in [0,\infty)$ that

$$\lim_{n\to\infty} \frac{p_X(n)}{P(X \geq n)} = 1 - e^{-1/\sigma} \in (0,\infty)\,,$$

which immediately implies that $X \in$ D-MDA$_0$ as well. □

The condition $p_X \in$ D-MDA is satisfied, among others, by the Zipf–Mandelbrot, geometric, Poisson and negative binomial distributions, as shown in the next example.

Example 2.2.6. The probability mass function of a Zipf–Mandelbrot distribution with parameters $s > 1$ and $\sigma > 0$ is in D-MDA$_{1/s}$ because it is regularly varying of order $-s$.

The probability mass function of a geometric distribution belongs to D-MDA$_0$ as it coincides up to a constant with the survival function of an exponential distribution. The latter distribution clearly satisfies the von Mises condition and thus is a member of MDA$_0$. The auxiliary function is, in fact, equal (eventually) to $1/\lambda$, where λ is the rate of the exponential distribution.

The probability mass function p of a Poisson distribution with rate $\lambda > 0$ coincides on $k \in \mathbb{N}_0$ with the function

$$g(x) = \frac{\lambda^x e^{-\lambda}}{\Gamma(x+1)},$$

a continuous function on $\mathbb{R}_+$ satisfying $\lim_{x\to\infty} g(x) = 0$. Moreover,

$$\frac{d}{dx} \log g(x) = -\psi_0(x+1) + \log \lambda \,,$$

where ψ_0 is the polygamma function of order 0. Since $\psi_0(x) \to \infty$ as $x \to \infty$, we see that $g'(x) < 0$ for x sufficiently large. Therefore, $\bar{F}_Y(x) = g(x)/g(d)$ is a survival function on $[d, \infty)$ for some $d \geq 0$. Furthermore,

$$\frac{d}{dx}\left(-\frac{1}{g'(x)}\right) = -\frac{\psi_1(x+1)}{(\psi_0(x+1) - \log\lambda)^2} \,,$$

where $\psi_1 = \psi_0'$ is is the polygamma function of order 1. Since $\psi_1(x) \to 0$ as $x \to \infty$, we conclude that F satisfies the von Mises condition, with the auxiliary function $a(x) = (\psi_0(x+1) - \log\lambda)^{-1} \to 0$ as $x \to \infty$. Therefore, the Poisson probability mass function is in D-MDA$_0$.

Similarly, the probability mass function of the negative binomial distribution with probability of success $p \in (0,1)$ and number of successes $r > 0$ is also in D-MDA$_0$ because it coincides on $\{0, 1, 2, \ldots\}$ with the function

$$g(x) = \frac{p^r}{\Gamma(r)} \frac{\Gamma(x+r)}{\Gamma(x+1)} (1-p)^x \,,$$

a continuous function on $\mathbb{R}_+$. It is simple to check that $\lim_{x\to\infty} g(x) = 0$, and $g'(x) < 0$ for x large enough, so that $\bar{F}_Y(x) = g(x)/g(d)$ is a survival function on

$[d, \infty)$ for some $d \geq 0$. Furthermore, $g(x) \sim cx^{r-1}(1-p)^x$ for large x, where c is a positive constant. Therefore, $\bar{F}_Y$ is of the form (2.10) with the auxiliary function

$$a(x) = \frac{1}{-\log(1-p) - (r-1)/x}, \quad x \text{ large},$$

and so it converges to $-1/\log(1-p)$ as $x \to \infty$.

To summarize, for a discrete random variable X, the conditions $X \in \text{MDA}$, $X \in \text{D-MDA}$ and $p_X \in \text{D-MDA}$ are related to each other as follows. If $\xi \geq 0$, then $X \in \text{MDA}_\xi$ if and only if $X \in \text{D-MDA}_\xi$ and X is long-tailed.

We end by providing some intuition on how the approximation methods suggested above differ. First of all, one would expect the D-GPD and GZD methods to perform similarly when ξ is close to zero because both approximating distributions coincide with a geometric distribution when $\xi = 0$. Second, Proposition 2.2.4 suggests that if one uses either $p_{\text{D-GPD}}(k; \sigma, \xi)$, $f_{\text{GPD}}(k; \sigma, \xi)$ or $p_{\text{GZD}}(k; \sigma, \xi)$ as an approximation to $P(X - u = k \mid X \geq u)$, one should not expect major differences as long as one chooses the scale parameter σ to be large. This would always be the case if X is long-tailed (see (1.8)), since the scale parameter is chosen to be proportional to the normalization sequence a_u defined in (1.4), which grows to infinity if and only if X is long-tailed.

When using a continuous distribution, such as the generalized Pareto distribution, to approximate a discrete distribution, it is common to use a continuity correction and shift the argument in the continuous approximation by some $\delta \in [0, 1)$. In our situation this corresponds to replacing $f_{\text{GPD}}(k; \sigma, \xi)$ by $f_{\text{GPD}}(k + \delta; \sigma, \xi)$, some $\delta \in [0, 1)$. When $\xi > 0$, we can check that

$$\frac{p_{\text{D-GPD}(\sigma,\xi)}(k)}{f_{\text{GPD}(\sigma,\xi)}(k+\delta)} = 1 + \frac{(1+\xi)(2\delta - 1)}{2\sigma} + O(\sigma^{-2}),$$

as $\sigma \to \infty$, for every $k \in \mathbb{N}_0$. Therefore, the approximations by $p_{\text{D-GPD}(\sigma,\xi)}(k)$ and $f_{\text{GPD}}(k + \delta; \sigma, \xi)$ with large scale σ are most similar when $\delta = 1/2$, a property that will be illustrated in the experimental part. Similarly, in the case $\xi = 0$, as $\sigma \to \infty$,

$$\frac{p_{\text{GPD}(\sigma,\xi)}(k)}{f_{\text{GPD}(\sigma,\xi)}(k+\delta)} = \sigma e^{\delta/\sigma}(1 - e^{-1/\sigma}) = 1 + \frac{2\delta - 1}{2\sigma} + O(\sigma^{-2}), \tag{2.13}$$

for every $k \in \mathbb{N}_0$, and the fastest convergence is again found when $\delta = 1/2$.

Eventually, we present a final result related to the well-known invariance property of the generalized Pareto distribution: its residual lifetime is again generalized Pareto distributed. More precisely, if $Y \sim$ GPD with scale parameter σ and shape parameter $\xi \in \mathbb{R}$, then the exceedance $Y - u \mid Y > u \sim$ GPD with scale parameter $\sigma + \xi u$ and shape parameter ξ, for all $u \in [0, e_1)$. This invariance property is important when approximating the exceedance distribution using generalized Pareto distributions because changing the threshold does not alter the distributional assumptions used in the approximation. The proposition below shows that the D-GPD has an analogous property. Moreover, it also shows that discretizing a GPD using different types of rounding does not affect the fact that a D-GPD is obtained and the shape parameter remains invariant.

Proposition 2.2.7. *Let Y have the generalized Pareto distribution with scale parameter $\sigma > 0$ and shape parameter $\xi \geq 0$. Let $0 < h \leq 1$. If $X = \lfloor \lambda Y + 1 - h \rfloor$, then for any integer $u \geq 1 - h$,*

$$X - u \mid X \geq u \sim \textit{D-GPD}(\sigma', \xi),$$

where $\sigma' = \lambda\sigma + \xi(u + h - 1)$.

Proof. For all $k \in \mathbb{N}$,

$$\Pr(X - u = k \mid X \geq u) = \Pr(k + u + h - 1 \leq \lambda Y < k + u + h \mid \lambda Y \geq u + h - 1)$$
$$= \left\{ \frac{\lambda\sigma + \xi(k + u + h - 1)}{\lambda\sigma + \xi(u + h - 1)} \right\}^{-1/\xi} - \left\{ \frac{\lambda\sigma + \xi(k + u + h)}{\lambda\sigma + \xi(u + h - 1)} \right\}^{-1/\xi} = p_{\text{D-GPD}}\left(k; \sigma', \xi\right).$$

□

2.3 Experimental Results

We now assess the performance of the generalized Pareto distribution (GPD), the discrete generalized Pareto distribution (D-GPD) and the generalized Zipf distribution (GZD) methods in estimating extreme quantiles from several simulated and real data sets. We start by a simple example to illustrate the methods.

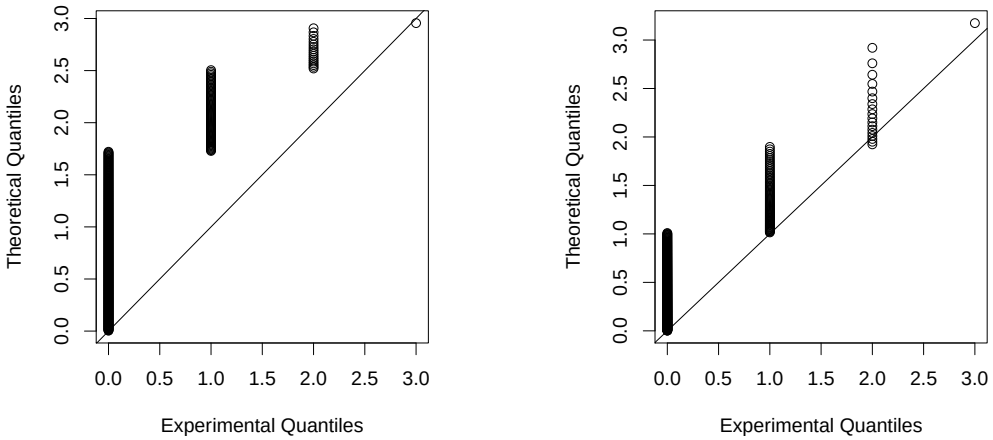

Figure 2.1: Quantile-quantile plots for the fit of a GPD (left) and D-GPD (right) to $X - 3 \mid X \geq 3$, where X is a sample of size 5000 drawn from a Poisson distribution with rate 1. A good fit occurs on the right because the lowest part of each accumulation of points falls close to the line, but not on the left.

2.3.1 Simulated Data

Let us draw 5000 i.i.d. realizations of X, a Poisson distributed random variable with rate $\lambda = 1$. We are interested in inferring the distribution of the exceedances $X - u \mid X \geq u$ above a large threshold u, say the 95th percentile of the data, which is 3 in this case. Since $X \in \text{D-MDA}_0$, we can approximate the distribution of $X - u \mid X \geq u$ by a D-GPD as explained in Section 1.3, which we refer to as the D-GPD approximation. Moreover, p_X, the probability mass function of X, also belongs to D-MDA_0 as shown in Example 2.2.6, motivating the GZD approximation. However, X is not in MDA and thus the GPD approximation does not necessarily apply. In order to compare these three approximations, we fit a GPD, D-GPD and GZD to the observations above $u = 3$, estimating their two parameters σ and ξ by maximum likelihood. Figure 2.1 shows quantile-quantile plots (QQ-plots) for the GPD and D-GPD fits, plotting $F^{-1}_{\text{GPD}}\{i/(n+1); \sigma, \xi\}$ for $i = 1, \ldots, n$ against the corresponding empirical quantiles. A good fit occurs if the minimum of points sharing the same x-coordinate falls close to the diagonal line, which is the case for the D-GPD but not the GPD. As we can see, the GPD approximation is clearly

outperformed by the D-GPD one in this case. We mention that the D-GPD and GZD approximations produce very similar estimates in this case and the QQ-plot for the GZD fit, although not displayed, is almost visually identical as the one on the right-hand side in Figure 2.1. The estimated scale parameters σ for the GPD, D-GPD and GZD are 1.91, 0.71 and 0.69 respectively; these are small numbers and thus Proposition 2.2.4 does not apply. As expected, increasing λ would increase the estimate for σ and render the three methods indistinguishable.

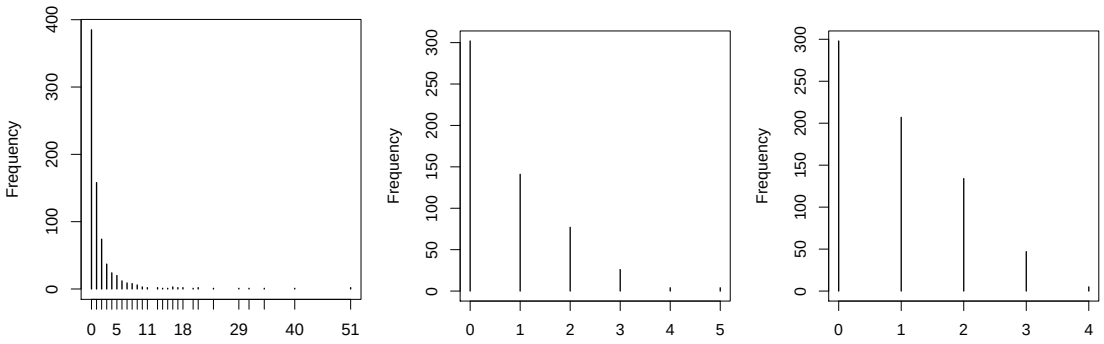

Figure 2.2: Frequency tables of a sample of $X_i - u \mid X_i \geq u$ in experiments $i = 1, 2, 3$ (from left to right).

	$P(X \geq 70) \times 10^3$		ξ	σ
	mean (cov, len)	true length	mean (cov, len)	mean (len)
truth	**0.10**	0.50		
$Y - u \mid Y \geq u$				
GPD	0.10 (87%, 0.16)	0.16	0.49 (95%, 0.22)	1.19 (0.30)
$X - u \mid X \geq u$				
D-GPD	**0.10** (86%, 0.17)	0.16	0.49 (95%, 0.23)	1.19 (0.33)
GZD	**0.11** (88%, 0.17)	0.17	0.50 (95%, 0.24)	1.39 (0.29)
$\text{GPD}_{\delta=-\frac{1}{2}}$	0.04 (20%, 0.07)	0.08	0.37 (22%, 0.18)	1.43 (0.32)
$\text{GPD}_{\delta=0}$	7.93 (83%, 23.97)	11.25	8.27 (0%, 1.24)	0.00 (0.00)[1]

Table 2.1: Performance of several methods in estimating the probability of a region far from the origin in experiment 1. The following experiment was repeated 500 times: a sample of size 8000 is drawn from $X = \lfloor Y \rfloor$, where Y is Student distributed with degree of freedom $\nu = 4$; a discrete generalized Pareto distribution (D-GPD), a generalized Zipf distribution (GZD) and a generalized Pareto distribution (GPD) with location parameter $-\delta$ are then fitted to $X - u \mid X \geq u$, where u is the 95th percentile of the observations. In addition, a GPD is fitted to $Y - u \mid Y \geq u$. The table displays average maximum likelihood estimators for the parameters σ and ξ and the probability p_e of having an observation above the rounded 99.99 percentile of Y. Between brackets, it shows average length of 90% confidence intervals based on asymptotic normality of the estimators, the length of 90% confidence intervals for the estimates across the experiments (true length) and the percentage of time the interval contains the true value (coverage). The average threshold over the experiments was $u = 2$ and there were about 700 exceedances.

	$P(X \geq 11) \times 10^3$		ξ	σ
	mean (cov, len)	true length	mean (cov, len)	mean (len)
truth	**0.23**		0.00	
$Y - u' \mid Y \geq u'$				
GPD	0.21 (78%, 0.36)	0.38	−0.14 (6%, 0.15)	1.50 (0.33)
$X - u \mid X \geq u$				
D-GPD	**0.21** (78%, 0.37)	0.40	−0.14 (10%, 0.16)	1.49 (0.35)
GZD	**0.21** (77%, 0.37)	0.40	−0.14 (10%, 0.15)	1.43 (0.30)
$\text{GPD}_{\delta=-\frac{1}{2}}$	0.08 (29%, 0.17)	0.20	−0.17 (0%, 0.13)	1.60 (0.33)
$\text{GPD}_{\delta=0}$	9.24 (100%, 27.88)	2.54	9.80 (0%, 1.75)	0.00 (0.00)

Remark: 1 and 6 N/A's in confidence interval length for the GZD and $\text{GPD}_{\delta=-\frac{1}{2}}$ respectively.

Table 2.2: Experiment 2. Here, $X = \text{round}(Y)$ for $Y \sim \mathcal{N}(0, 9)$. The average threshold is 5 and about 500 observations exceed it.

[1]The values rounded off to two decimals are 0.00.

	$P(X \geq 15) \times 10^3$,		ξ	σ
truth	mean (cov, len) **2.27**	true length	mean (cov, len) -0.50	mean (len)
$Y - u' \mid Y \geq u'$				
GPD	2.20 (87%, 1.14)	1.16	-0.51 (94%, 0.13)	1.86 (0.37)
$X - u \mid X \geq u$				
D-GPD	**2.25** (89%, 1.65)	1.60	-0.50 (95%, 0.21)	1.85 (0.45)
GZD	**2.26** (89%, 1.69)	1.62	-0.49 (98%, 0.20)	1.61 (0.36)
$\text{GPD}_{\delta=-\frac{1}{2}}$	0.44 (0%, 0.62)	0.39	-0.50 (100%, 0.15)	1.91 (0.41)
$\text{GPD}_{\delta=0}$	7.25 (70%, 19.01)	11.57	7.77 (0%, 1.36)	0.00 (0.00)

Table 2.3: Experiment 3. Here, $X = \lceil Y \rceil$, where $Y = 14.7\,Z$ and $Z \sim \text{Beta}(1, 2)$. The average threshold is $u = 12$ and about 500 observations exceed it.

We now compare the performances of the GPD, D-GPD and GZD approximations in estimating the probability of a rare event in three different simulated cases. Let

$$Y_1 \sim t_4, \quad Y_2 \sim \mathcal{N}(0, 9), \quad Y_3 = 14.7\,Z, \quad Z \sim \text{Beta}(1, 2),$$

where t_ν denotes a centered Student distribution with degree of freedom ν. The random variables Y_1, Y_2 and Y_3 all belong to MDA_ξ for $\xi = 1/4$, $\xi = 0$ and $\xi = -1/2$ respectively. We define further three discrete random variables using various types of rounding to better illustrate the invariance property of the D-GPD in Proposition 2.2.7:

$$X_1 = \lfloor Y_1 \rfloor, \quad X_2 = \text{round}(Y_1), \quad X_3 = \lceil Y_3 \rceil.$$

The first random variables satisfies $X_1 \in \text{D-MDA}_{1/4}$ and $X_1 \in \text{MDA}_{1/4}$, the second $X_2 \in \text{D-MDA}_0$ but $X_2 \notin \text{MDA}$ since X_2 is not long-tailed.[2] The distribution of X_3 has a finite endpoint $x_F = 14.7$, a situation which we did not treat in the theoretical part.

For each $i = 1, 2, 3$, we repeat 500 times the following experiment. We simulate 8000 i.i.d. realizations of X_i and, from these observations, the goal is to estimate the probability of the extreme region

$$p_e = P(X_i \geq \lfloor q_e + 1 - h_i \rfloor),$$

[2]It is not immediate if the probability mass functions of X_1 and X_2 are in D-MDA.

where q_e is the 99.99 percentile of Y_i, i.e., the value exceeded once every 10 000 times on average, and $h_1 = 1$, $h_2 = \frac{1}{2}$, $h_3 = 0$. The strategy pursued is to select a threshold u as the 95th percentile of the sample, fit a parametric distribution to the exceedances $X - u \mid X > u$, and use it to extrapolate the tail and estimate p_e. In this context, we gave grounds to the choice of a GPD when $i = 1$ and D-GPD when $i = 1, 2$. In all other cases, the GPD, D-GPD and GZD approximations will be applied heuristically.

As discussed in Section 2.2, when approximating a discrete distribution by a continuous one, it is common to incorporate a continuity correction. We thus implement two variants of the GPD approximation: the first has no continuity correction ($\delta = 0$), the second shifts the observations by $\delta = \frac{1}{2}$.

Besides comparing the estimations of the GPD, D-GPD and GZD approximations for p_e from observations of X_i, we are also interested in comparing them to an estimation of p_e based on Y_i (as opposed to its discretization X_i). Perhaps discretizing continuous data provides superior estimators for p_e? We therefore fit a GPD to the exceedances $Y_i - u' \mid Y_i \geq u'$ for $u' = u - 1 + h_i$, where u' is chosen to obtain the same sample size as for $X_i - u \mid X_i \geq u$.

Frequency tables of a sample of $X_i - u \mid X_i \geq u$ are displayed in Figure 2.2 for $i = 1, 2, 3$. For each experiment and each model, we compute maximum likelihood estimators for σ and ξ by performing a two dimensional maximization using the function `optim` of R Core Team [2015] with starting values $(1, 1)$. We then compute p_e and 90% confidence intervals under asymptotic normality of the estimator — the assumption of normality leads to an approximation error. Table 2.1, 2.2 and 2.3 display: the average parameters p_e, ξ and σ over the 500 experiments, the average length of the confidence intervals, the true length and their coverage. True length is the length the intervals should have had to contain the estimates across the 500 experiments 90% of the time. Coverage indicates the proportion of time the truth lies in the confidence interval.

It appears that the D-GPD and GZD approximations accurately estimate p_e with a coverage close to the correct one of 90%. On the other hand, the GPD approximation with $\delta = \frac{1}{2}$ is inaccurate and the GPD with $\delta = 0$ performs very

poorly (misleadingly, the latter has a larger likelihood at the maximum likelihood estimate than the former). This illustrates why the D-GPD or GZD approximations should be preferred to the GPD approximation in situations similar to theses simulated cases.

Regarding the simulated case $i = 2$, we can see in Table 2.2 that the tail of the discretized normal distribution is not well approximated by any model — the estimate for ξ is significantly smaller than the truth. Improving the estimation of ξ would require a larger threshold and possibly more observations. Nevertheless, the D-GPD and GZD approximations still provide a relatively accurate estimation of p_e, underestimating a little the uncertainty.

We know that the D-GPD and GZD coincide with a geometric distribution when $\xi = 0$ and that they are asymptotically equivalent when $\xi > 0$ and $\sigma \to \infty$ (Proposition 2.2.4). In the simulated cases $i = 1, 2$, although σ is not particularly large, the D-GPD and GZD approximations deliver very similar results. In the case $i = 3$, some slight differences are apparent because the two fitted distributions have a small endpoint close to which they allocate mass in a distinct manner.

In the three simulation cases $i = 1, 2, 3$, we note few differences between the GPD fitted to exceedances of Y_i and the D-GPD fitted to exceedances of its discretization X_i, except that the former method yields slightly shorter confidence intervals. We emphasize that the two methods are always estimating the same true parameter ξ, no matter if the data were discretized using floor, rounding or ceiling function, a consequence of the invariance property of the D-GPD in Proposition 2.2.7.

In conclusion, we showed that the D-GPD and GZD approximations were able to estimate the probability of extreme regions in these examples. The D-GPD is more efficient as its probability mass, survival and quantile functions are analytical, but both are useful to describe tails of discrete distributions, which we will further illustrate on real data sets.

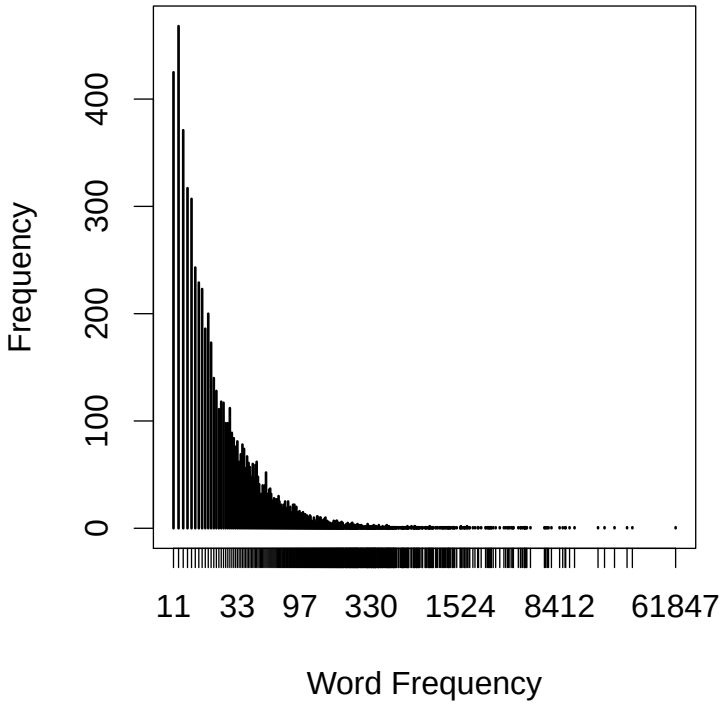

Figure 2.3: Word frequency table of the 7476 most frequent words in a British corpus (x axis on log-scale).

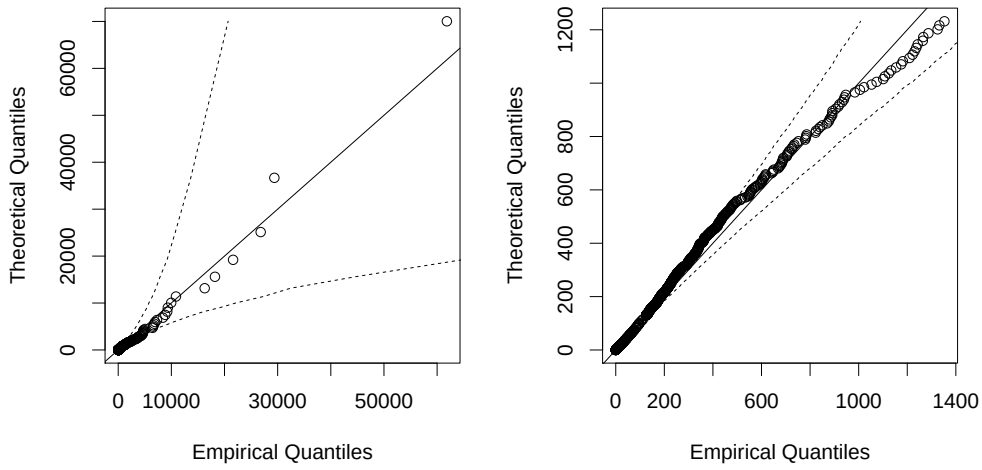

Figure 2.4: Quantile-quantile plots for a D-GPD fitted to the frequencies of the 7476 most frequent words in a British corpus. On the right, only percentiles below 99% are plotted. Dashed lines denote pointwise 90% confidence intervals for empirical quantile estimates.

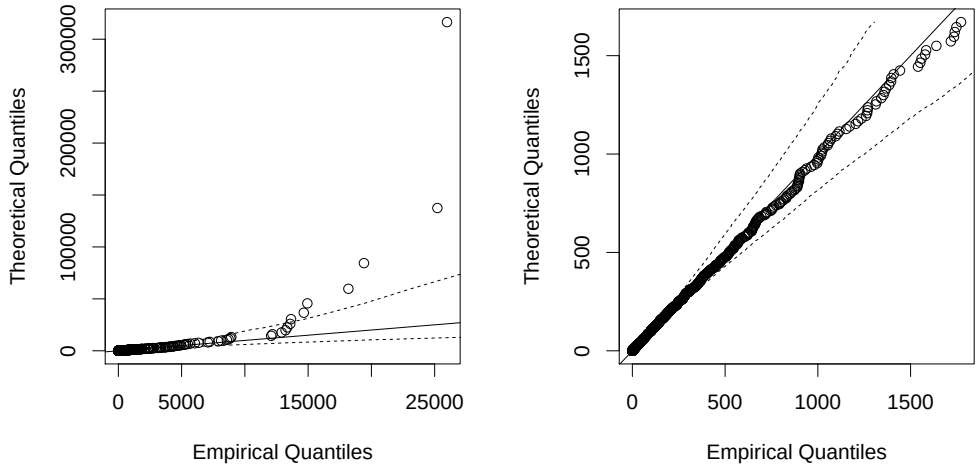

Figure 2.5: QQ-plots for a D-GPD fitted to frequencies of the 7668 most frequent words in a collection of French movie subtitles.

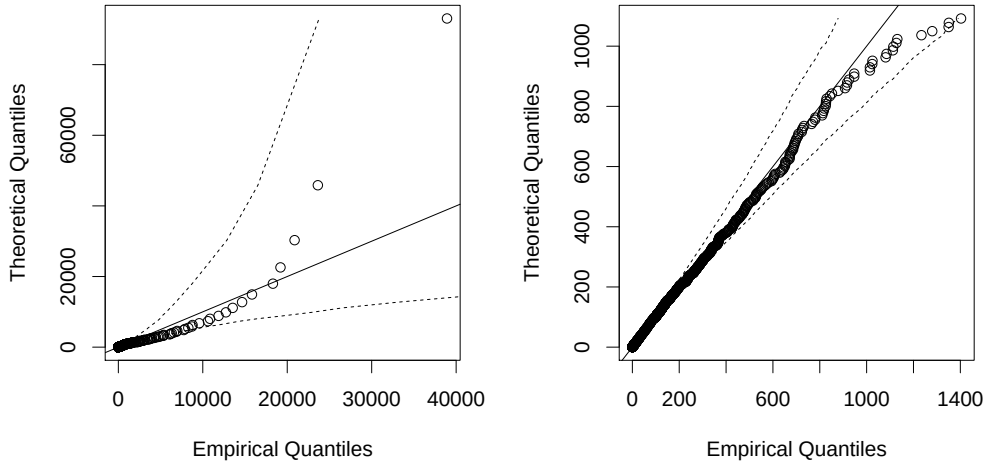

Figure 2.6: QQ-plots for a D-GPD fitted to frequencies of the 7468 most frequent words in a collection of French books.

2.3.2 Word Frequencies and Word Lengths

We consider three data sets counting word occurrences: the first is based on the British National corpus and counts written and spoken English words [Bri, 2007], the second and third consist of rounded relative frequencies of French words in a collection of books and movie subtitles respectively [New et al., 2004].

Let X denote the frequency of a word in one of these corpora. As we are interested in modeling the frequency of the most popular words, we select a large

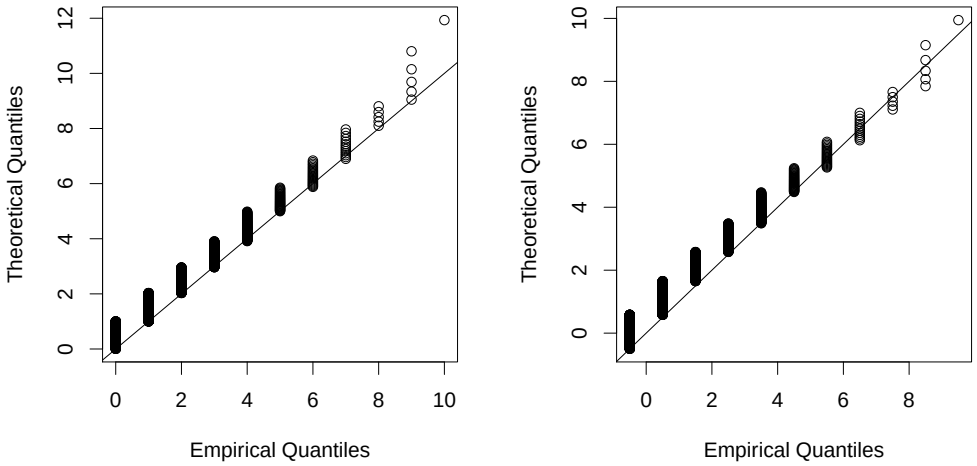

Figure 2.7: QQ-plots for a D-GPD (left) and GPD with $\delta = -\frac{1}{2}$ (right) fitted to the length of the 2875 French words consisting of 15 letters or more. A good fit occurs when the lowest part of each accumulation of points is close to the line.

threshold u and fit to the exceedances $X - u \mid X \geq u$ a generalized Pareto distribution (GPD), a discrete generalized Pareto distribution (D-GPD) and a generalized Zipf distribution (GZD) by maximum likelihood. Two variants are implemented for the GPD: shifting the observations by $\delta = \frac{1}{2}$ or by $\delta = 0$. We chose u by relying upon mean residual plots [Davison and Smith, 1990]; for the three data sets, about 7500 words satisfy $X \geq u$.

Figure 2.3 displays the frequency table of $X \mid X \geq u$ for the British corpus. The six most frequent words in this case are: “the”, “of”, “and”, “a”, “in” and “to.” The D-GPD delivers a good fit for the three data sets as revealed by the QQ-plots in Figure 2.4, 2.5 and 2.6 because most observations lie within the pointwise 90% confidence intervals. The latter are obtained by simulating 2000 times from the fitted model and computing empirical quantiles at each simulation. In Figure 2.5, the six most common words counted in French subtitles seem to behave as outliers; these are: “je, “de”, “est”, “pas”, “la”, “tu”.

Table 2.4 presents maximum likelihood estimators for σ and ξ. Except in the case of the GPD with $\delta = 0$, all models provide similar results as expected from Proposition 2.2.4 because σ is relatively large. A Kolmogorov–Smirnov test for discrete data [Arnold and Emerson, 2011] shows that the D-GPD and GZD only

Model	p-value	NLL	ξ	σ
Word Frequency				
British				
D-GPD	**0.09**	35148.3	$0.93_{[0.90,0.97]}$	$15.93_{[15.32,16.53]}$
GZD	**0.09**	35148.3	$0.93_{[0.90,0.97]}$	$16.37_{[15.78,16.97]}$
$\text{GPD}_{\delta=-\frac{1}{2}}$	**0.10**		$0.93_{[0.89,0.97]}$	$15.95_{[15.35,16.56]}$
$\text{GPD}_{\delta=0}$	0.00		$1.01_{[0.97,1.05]}$	$14.34_{[13.77,14.92]}$
French (in subtitles)				
D-GPD	**0.82**	33313.4	$1.20_{[1.16,1.24]}$	$7.94_{[7.63,8.25]}$
GZD	**0.80**	33313.4	$1.20_{[1.16,1.24]}$	$8.52_{[8.22,8.83]}$
$\text{GPD}_{\delta=-\frac{1}{2}}$	0.54		$1.19_{[1.15,1.23]}$	$8.03_{[7.72,8.35]}$
$\text{GPD}_{\delta=0}$	0.00		$1.38_{[1.33,1.43]}$	$6.20_{[5.92,6.48]}$
French (in books)				
D-GPD	**0.92**	33343.7	$1.02_{[0.98,1.06]}$	$10.25_{[9.86,10.63]}$
GZD	**0.92**	33343.7	$1.02_{[0.98,1.06]}$	$10.75_{[10.37,11.13]}$
$\text{GPD}_{\delta=-\frac{1}{2}}$	**0.90**		$1.02_{[0.98,1.05]}$	$10.30_{[9.92,10.68]}$
$\text{GPD}_{\delta=0}$	0.00		$1.13_{[1.09,1.17]}$	$8.71_{[8.35,9.06]}$
Word length				
French				
D-GPD	**1.00**	3894.0	$0.02_{[-0.01,0.06]}$	$1.36_{[1.30,1.43]}$
GZD	**1.00**	3894.0	$0.02_{[-0.01,0.06]}$	$1.37_{[1.32,1.43]}$
$\text{GPD}_{\delta=-\frac{1}{2}}$	0.02		$-0.04_{[-0.06,-0.01]}$	$1.51_{[1.45,1.57]}$
$\text{GPD}_{\delta=0}$	N/A			

Table 2.4: Discrete Kolmogorov-Smirnov test (p-value), negative log-likelihood (NLL) and maximum likelihood estimators for the GPD, D-GPD and GZD fitted to the frequencies of most frequent words in a British corpus, a collection of French subtitles, of French books and to the length of the longest words in French. The sample size of each word frequency data sets and length data sets is around 7500 and 3000 respectively.

outperform the GPD for the French subtitles. (In order to perform the test for the GPD, which is a continuous distribution, we assumed that data were rounded realizations of the fitted model). Notice that $\xi \geq 1$ for the French word data and thus, according to the fitted model, the mean of X is infinite; a discussion on distributions with infinite moments can be found in Berger and Mandelbrot [1963]. Since the GZD coincides with a Zipf–Mandelbrot distribution when $\xi > 0$, the above results are consistent with the common hypothesis in linguistic that word frequencies follow a Zipf-type law (see e.g. Booth [1967]).

We now consider a fourth data set where tied observations are more frequent. Let

us consider all 150 000 words in the French lexical [New et al., 2004] and denote by X the length of a word. For instance, the longest word is "anticonstitutionnellement", consisting of 25 letters. We fit the usual models to $X - u \mid X \geq u$ with threshold $u = 15$ corresponding to the 98% percentile of the data, which leaves about 3000 exceedances. The D-GPD and GZD deliver very similar estimations and, this time, clearly model the extremes more accurately than the GPD with $\delta = \frac{1}{2}$ as shown by QQ-plots in Figure 2.7 and discrete Kolmogorov–Smirnov tests at the bottom of Table 2.4.

In conclusion, we have studied the distribution of the frequencies of the most common and longest words from large corpora. The D-GPD and GZD were useful in accurately describing these data and summarizing them by a two parameter distribution. In addition, we have supported the conclusion drawn earlier from the simulated cases: the D-GPD and GZD are preferred over the GPD to model extremes of discrete data when tied observations are frequent.

2.3.3 Multiple Births

We now focus on a data set consisting of very small integer values to see if the discrete generalized Pareto distribution (D-GPD) and generalized Zipf distribution (GZD) can still describe tails in this context. We examine data counting multiple births in the United States and in metropolitan France from 1995 to 2014 (Hamilton et al., 2015 and Beaumel and Bellamy, 2015). The frequency table reads

	single	twin	triplet	quadruplet	quint. or more
France	15 036 159	240 402	4 286	110	3
US	78 178 588	2 500 340	117 603	8 108	1 353.

Let X be the number of children at birth. We fit a right-censored GPD, D-GPD and GZD to $X - u \mid X \geq u$ for $u = 1$ with censored point $C = 5$, and display maximum likelihood estimates and p-values of χ^2-tests for goodness-of-fit in the table below.

For French births, the χ^2-test assesses the adequacy of the D-GPD and GZD. According to the D-GPD model, the probability of having quintuplets is 7.3×10^{-6}

	NLL	ξ	σ	p-value
France				
D-GPD	22578.4	$0.02_{[0.01,0.04]}$	$0.24_{[0.23,0.24]}$	0.99
GZD	22578.4	$0.02_{[0.01,0.04]}$	$0.24_{[0.24,0.25]}$	0.99
US				
D-GPD	546490.2	$0.07_{[0.06,0.07]}$	$0.30_{[0.30,0.30]}$	0.00
GZD	546489.5	$0.07_{[0.06,0.07]}$	$0.32_{[0.31,0.32]}$	0.00

with 90% confidence interval $[6.2 \times 10^{-6}, 8.4 \times 10^{-6}]$. A naive empirical estimate for p_e counting the number of quintuplets or more would give 0.1×10^{-6}, which underestimates the probability of the region and is not useful to evaluate the uncertainty.

For American births, the test rejects both models. Nevertheless, the following experiment repeated 500 times indicates that the models remain relatively accurate in estimating the probability of a rare event: we draw with replacement 0.1% of all births in the US on the period of study. We then fit a D-GPD and GZD to observations in this subsample that are greater or equal to 1 and use the fitted model to compute p_e, the probability of having quintuplets or more. The next table displays average maximum likelihood estimators for p_e over the experiments as well as length, true length and coverage of 90% confidence intervals as defined in Table 2.1. Each subsample contained some quadruplets and, on average, 1.3 quintuplets or more.

	mean	coverage	length
truth	$\mathbf{1.7} \times 10^{-5}$		
D-GPD	$\mathbf{1.6} \times 10^{-5}$	76%	3.6×10^{-5}
GZD	$\mathbf{1.4} \times 10^{-5}$	93%	14.7×10^{-5}

Despite the few distinct values taken by X, the D-GPD and GZD tail approximations were able to estimate p_e relatively accurately and quantify the uncertainty. This exemplifies how extreme value analysis can be useful for estimating probabilities of extreme regions even when the data consist of particularly small integers.

3

Graphical Modeling of Extreme River Flows

We study the data set used by Asadi et al. [2015] which reports average daily discharges in m^3/s in 31 stations along the Danube during the summers from 1960 to 2009 in Bavaria, Germany. Some temporal dependence in the data is removed by considering only maxima of flows occuring at a station within 9 days windows; this gives $n = 428$ flow events $\tilde{\boldsymbol{x}} = \{\tilde{\boldsymbol{x}}^{(k)}\}_{k=1}^{n}$ assumed to be i.i.d. realizations of a random vector $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_d)$, for $d = 31$. Figure 3.1 displays a map of the river.

The authors model the spatial dependence between these flows using a Brown–Resnick process with a specific non-Euclidean distance. In our application, contrary to the spatial modeling framework, we do not include information on the location of each station. We are interested in estimating a graphical model for the tail distribution of $\tilde{\boldsymbol{X}}$. Graphical models are useful to simplify a joint density by assuming conditional independence between some of the marginals and visualize the dependence structure as a graph. We address here the following questions:

- Is conditional independence a relevant notion for understanding this data set?
- Does the selected graph recover some aspects of the network formed by the rivers?

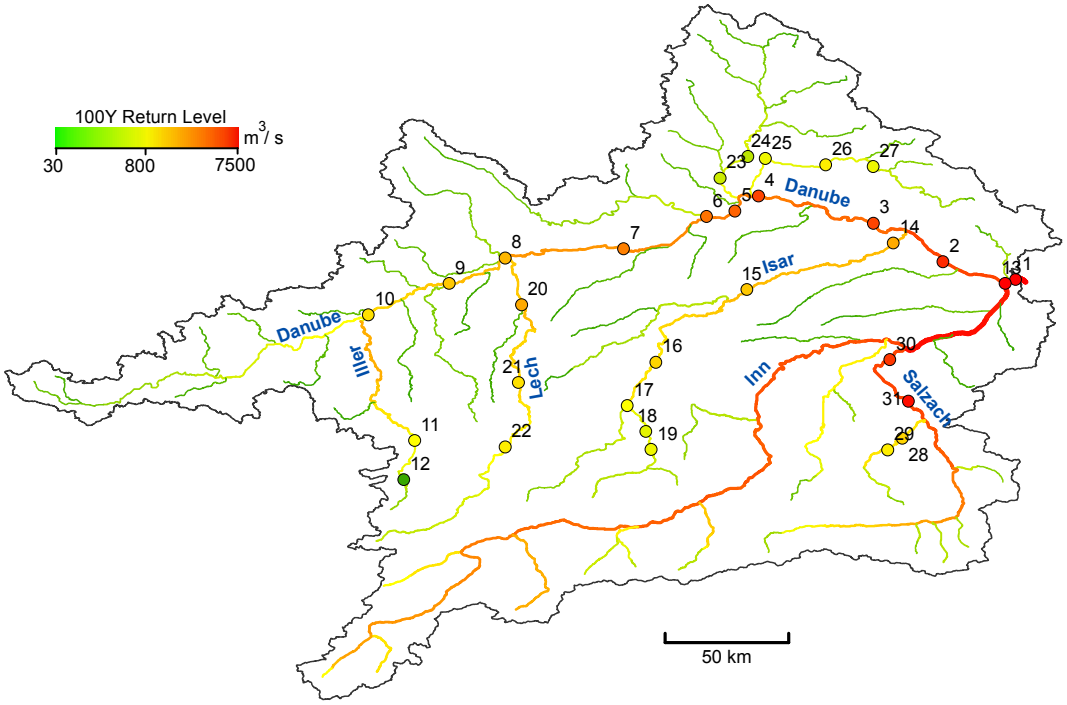

Figure 3.1: Map of the upper Danube basin showing the 100-year return level for river flow at the 31 stations [Asadi et al., 2015].

- Can we describe simply and accurately the dependence between the extreme flow events using graphical models?

3.1 Factorization of the Censored Tail Density

In this section, we explain how to estimate the tail distribution of $\tilde{\boldsymbol{X}}$ using graphical models. As we want to model accurately flows that are extreme, we consider a censored version of $\tilde{\boldsymbol{X}}$ that only includes information on the occurrence of extremes and their magnitude as in Ledford and Tawn [1996]. This censored vector is

$$\tilde{\boldsymbol{X}}^C \equiv \tilde{\boldsymbol{X}}^C(\boldsymbol{u}) = (\tilde{X}_1 1_{\tilde{X}_1 \geq u_1}, \ldots, \tilde{X}_d 1_{\tilde{X}_d \geq u_d}),$$

and we choose u_i to be the 80% quantile of $\tilde{\boldsymbol{x}}_i$.

We then fit a generalized Pareto distribution to $\tilde{X}_i \mid \tilde{X}_i \geq u_i$ and transform the latter to unit Pareto for every i, obtaining a censored vector $\boldsymbol{X}^C$ with values in $F_+^d = (\{0\} \cup [1, \infty))^d$. Transforming the univariate marginals of $\tilde{\boldsymbol{X}} \mid ||\tilde{\boldsymbol{X}}||_\infty \geq u$ to a common distribution such as unit Fréchet for a large u is recurrent in extreme

value analysis [Coles and Tawn, 1991]; our procedure slightly differs as it transforms the marginals of the censored vector $\tilde{\boldsymbol{X}}^C$.

We are interested in the dependence structure of $\boldsymbol{X}^C$, especially in knowing whether the following conditional independence relations hold for $i, j \in \{1, \ldots, d\}$:

$$X_i^C \perp\!\!\!\perp X_j^C \mid \boldsymbol{X}^C_{\{1,\ldots,d\}\setminus\{i,j\}}. \tag{3.1}$$

Suppose that $\boldsymbol{X}^C$ admits a probability density $f_{\boldsymbol{X}^C}$ w.r.t. the measure

$$\mu_0(A_1 \times \ldots \times A_d) := \sum_{D\subseteq\{1,\ldots,d\}} \lambda^{|D^c|}\left(\prod_{i\in D^c} A_i\right)\delta_{\mathbf{0}}^{|D|}\left(\prod_{i\in D} A_i\right), \tag{3.2}$$

on F^d, where λ^d is the d-dimensional Lebesgue measure and $\delta_{\mathbf{0}}^d$ the Dirac measure on $\mathbf{0} \in \mathbb{R}^d$. For instance, $f_{X_i^C}(x) = \frac{\partial}{\partial x}\Pr(X_i^C \le x, X_i^C > 0)$ for $x \ge 1$ and $f_{X_i^C}(0) = \Pr(X_i^C = 0)$.

We say that $\boldsymbol{X}^C$ satisfies the pairwise Markov property according to a graph $\mathcal{G}$ with set of nodes $V = \{1, \ldots, d\}$ if its set of edges E satisfies

$$(i,j) \notin E \quad \Longrightarrow \quad X_i^C \perp\!\!\!\perp X_j^C \mid \boldsymbol{X}^C_{\{1,\ldots,d\}\setminus\{i,j\}}.$$

If $f_{\boldsymbol{X}^C} > 0$, it follows from the Hammersley–Clifford theorem that $\boldsymbol{X}^C$ satisfies the pairwise Markov property according to a decomposable graph $\mathcal{G}$ if and only if $f_{\boldsymbol{X}^C}$ factorizes w.r.t. $\mathcal{G}$, i.e.,

$$f_{\boldsymbol{X}^C}(\boldsymbol{x}) = \frac{\prod_{C\in\mathcal{C}} f_{\boldsymbol{X}_C^C}(\boldsymbol{x}_C)}{\prod_{D\in\mathcal{D}} f_{\boldsymbol{X}_D^C}(\boldsymbol{x}_D)}, \quad \boldsymbol{x} \in F_+^d, \tag{3.3}$$

where $\mathcal{C}$ is the set of maximal cliques and $\mathcal{D}$ is the multiset containing the separator sets defined in (A.4). Essentially, assuming conditional independence between some marginals simplifies the joint density, which can then be expressed as a product of its lower dimensional marginal densities.

We now examine the probability density of $\boldsymbol{X}_S^C$ for $S \in \mathcal{C} \cup \mathcal{D}$ and provide assumptions under which it converges to a *censored homogeneous density* as the threshold grows. Suppose that $\boldsymbol{X}^C$ coincides for some t with a sequence $\boldsymbol{X}^C(t)$ defined as follows:

$$\boldsymbol{X}^C \equiv \boldsymbol{X}^C(t) = \begin{cases} (t^{-1}X_1\,1_{|X_1|\ge t}, \ldots, t^{-1}X_d\,1_{|X_d|\ge t}) & \text{w.p. } p, \\ \mathbf{0} & \text{w.p. } 1-p, \end{cases}$$

for $p = \Pr(||\boldsymbol{X}^C||_\infty \geq 1)$ and a random vector $\boldsymbol{X}$. Assume further that the latter satisfies

$$f_{t^{-1}\boldsymbol{X}_S \,|\, ||\boldsymbol{X}_S||_\infty \geq t} \to f_{\boldsymbol{Y}}, \tag{3.4}$$

for some random vector $\boldsymbol{Y}$, for every $S \in \mathcal{C} \cup \mathcal{D}$; (in particular this implies that the marginals of $\boldsymbol{X}_S$ are asymptotically dependent, see (1.19) for a definition). As a consequence, it can be shown that the marginal $\boldsymbol{X}_S^C$ has a probability density converging to the one of the censored version

$$\boldsymbol{Y}^C = \begin{cases} (Y_1 1_{|Y_1|\geq 1}, \ldots, Y_{|S|} 1_{|Y_{|S|}|\geq 1}) & \text{w.p. } p, \\ \mathbf{0} & \text{w.p. } 1-p. \end{cases}$$

Moreover, we will see in (4.12) that $f_{\boldsymbol{Y}}$ is necessarily homogeneous of order $-1-d$ (in particular $Y_i^C \mid Y_i^C \geq 1$ for $i \in S$ is unit Pareto distributed, hence the transformation of $\tilde{X}_i^C \mid \tilde{X}_i^C \geq u_i$ to unit Pareto). When t is large we can approximate $\boldsymbol{X}_S^C$ by $\boldsymbol{Y}^C$, which motivates the modeling of $f_{\boldsymbol{X}_S^C}$ by a censored homogeneous density.

The following example illustrates a multivariate distribution on F_+^d that factorizes and can be written as a product of low-dimensional censored homogeneous densities.

Example 3.1.1 (Factorization of a Censored Tail Density)**.** Let $\boldsymbol{X} = (X_1, X_2, X_3)$ be a non-negative random vector and suppose that its censored version $\boldsymbol{X}^C = (t^{-1}X_1 1_{X_1 \geq t}, \ldots, t^{-1}X_3 1_{X_3 \geq t})$ with values in $\mathbb{F}_+^3 = (\{0\} \cup [1, \infty))^3$ satisfies

$$X_1^C \perp\!\!\!\perp X_3^C \mid X_2^C,$$

for $t > 0$. From Hammersley–Clifford theorem, $X_1^C \perp\!\!\!\perp X_3^C \mid X_2^C$ if and only if $f_{\boldsymbol{X}^C} = f_{\boldsymbol{X}_{12}^C} f_{\boldsymbol{X}_{23}^C} / f_{\boldsymbol{X}_2^C}$. In addition, assume that the density of $t^{-1}\boldsymbol{X}_{12} \mid ||\boldsymbol{X}_{12}||_\infty \geq t$ converges to

$$f_{\boldsymbol{Y}}(x, y) = \frac{4}{3}(x+y)^{-3},$$

on $C_{||\cdot||_\infty}^{12}$, and that $\boldsymbol{X}_{23} \overset{d}{=} \boldsymbol{X}_{12}$. When t is large, this motivates the approximation

$$f_{\boldsymbol{X}^C}(x, y, z) \approx \frac{f_{\boldsymbol{Y}^C}(x, y) f_{\boldsymbol{Y}^C}(y, z)}{f_{Y_2^C}(y)}, \quad \forall (x, y, z) \in F_+^3,$$

where $\boldsymbol{Y}^C$ admits the censored homogeneous density

$$f_{\boldsymbol{Y}^C}(x,0) = p\int_0^1 f_{\boldsymbol{Y}}(x,y)dy = \frac{2}{3}p(1+2x)x^{-2}(1+x)^{-2},$$
$$f_{Y_2^C}(x) = p\int_0^\infty f_{\boldsymbol{Y}}(x,y)dy = \frac{2}{3}px^{-2},$$
$$f_{Y_2^C}(0) = (1-p) + p\int_1^\infty\int_0^1 f_{\boldsymbol{Y}}(x,y)dxdy = 1 - \frac{2}{3}p,$$

etc, for $x \geq 1$ and $p = \Pr(||\boldsymbol{X}_{12}^C||_\infty \geq 1) \in (0,1)$.

Let us come back to the problem of modeling the vector of extreme river flows $\boldsymbol{X}^C$. We have introduced assumptions under which its density $f_{\boldsymbol{X}^C}$ factorizes w.r.t. a graph and discussed the convergence of some of its marginals $\boldsymbol{X}_S^C$ to a censored homogeneous distribution. The rest of the chapter focuses on selecting the graph $\mathcal{G}$ and estimating appropriate models for $\boldsymbol{X}_S^C$.

First, we want to know if some marginals of $\boldsymbol{X}^C$ are mutually independent. We start by considering the binary vector

$$\boldsymbol{B}^C = (1_{X_1^C \geq 1}, \ldots, 1_{X_d^C \geq 1}),$$

and test independence between its pairs of univariate marginals. A χ^2 test with level 0.05 rejects the hypothesis $B_i \perp\!\!\!\perp B_j$ for all pairs i, j (see Appendix A.2 for a definition of several goodness-of-fit tests). The average correlation between the marginals of $\tilde{\boldsymbol{X}}$ is 0.73, indicating a strong extremal dependence as found in Asadi et al. [2015].

A mutual information test rejects $B_i \perp\!\!\!\perp B_j \mid \boldsymbol{B}_{\{1,\ldots,d\}\setminus\{i,j\}}$ for 27% of all pairs only, which could mean that there are many conditional independence relations to exploit between the marginals of $\boldsymbol{X}^C$.

Two different approaches are possible at this point. Either one forces the graph to have a specific structure, such as being decomposable and having cliques of relatively small size so that each clique can then be modeled parametrically or nonparametrically by a low-dimensional distribution. Or one imposes no restriction on the graph but choose a parametric multivariate distribution whose joint density factorizes when its parameters take certain values, such as the multivariate Gaussian distribution.

In this chapter, we focus mainly on the first approach when the graph is a tree and merely compare it to the second by fitting a censored Gaussian and Student copula graphical model to $\boldsymbol{X}^C$ as explained in detail in Chapter 6 in an application on website visits.

3.2 Tree Graphical Model Estimation

We now explain how to estimate a tree graphical model for $\boldsymbol{X}^C$, adapting the procedure described in Lafferty et al. [2012] to our context. We are not interested in testing conditional independence but in enforcing $f_{\boldsymbol{X}^C}$ to factorize w.r.t. to a tree $\mathcal{G} = (V, E)$. In this case,

$$f_{\boldsymbol{X}^C}(\boldsymbol{x}) = \prod_{(i,j)\in E} \frac{f(x_i, x_j)}{f(x_i)f(x_j)} \prod_{i=1}^{d} f(x_i), \tag{3.5}$$

dropping the subscript notation for simplicity. In other words, the density of $\boldsymbol{X}^C$ is expressed in terms of its univariate and bivariate marginals only. Since the univariate marginal distributions are assumed to be Pareto, we take $f_i(0) = \hat{p}_{i,0}$, $f_i(x) = (1 - \hat{p}_{i,0})x^{-2}$, where $\hat{p}_{i,0}$ is the empirical estimator for $\Pr(X_i^C = 0)$. Moreover, we suppose that $\boldsymbol{X}_{ij}^C$ has bivariate parametric density $f_{ij}(x, y; \boldsymbol{\theta}_{ij})$, denoting by $\boldsymbol{\theta} = \{\theta_{ij}\}_{(i,j)\in E}$ the set of disjoint parameters for the bivariate models. The log-likelihood reads

$$\sum_{k=1}^{n} \log f_{\boldsymbol{X}^C}(\boldsymbol{x}^{(k)}; \boldsymbol{\theta}) = \sum_{(i,j)\in E} \underbrace{\sum_{k=1}^{n} \log \frac{f(x_i^{(k)}, x_j^{(k)}; \boldsymbol{\theta}_{ij})}{f(x_i^{(k)})f(x_j^{(k)})}}_{w_{ij}} + \sum_{i=1}^{d}\sum_{k=1}^{n} \log f(x_i^{(k)}). \tag{3.6}$$

Consequently, the tree that maximizes the log-likelihood is solution of

$$\mathcal{G}^\star = \underset{\mathcal{G}=(V,E)}{\arg\max} \sum_{(i,j)\in E} w_{ij}, \tag{3.7}$$

i.e., it is the tree maximizing the sum of the weights w_{ij} associated to its edges and is called maximum spanning tree; it can be obtained by applying Kruskal's algorithm [Kruskal, 1956], see Appendix A.1.3. This means that a tree graphical model can be estimated by the following procedure: fit a bivariate model $f(x_i, x_j; \boldsymbol{\theta}_{ij})$ to each pair of $\boldsymbol{X}^C$ separately, compute the weights w_{ij} and solve (3.7). Similarly, one obtains the tree that minimizes, say, the Bayesian information criterion (BIC) by

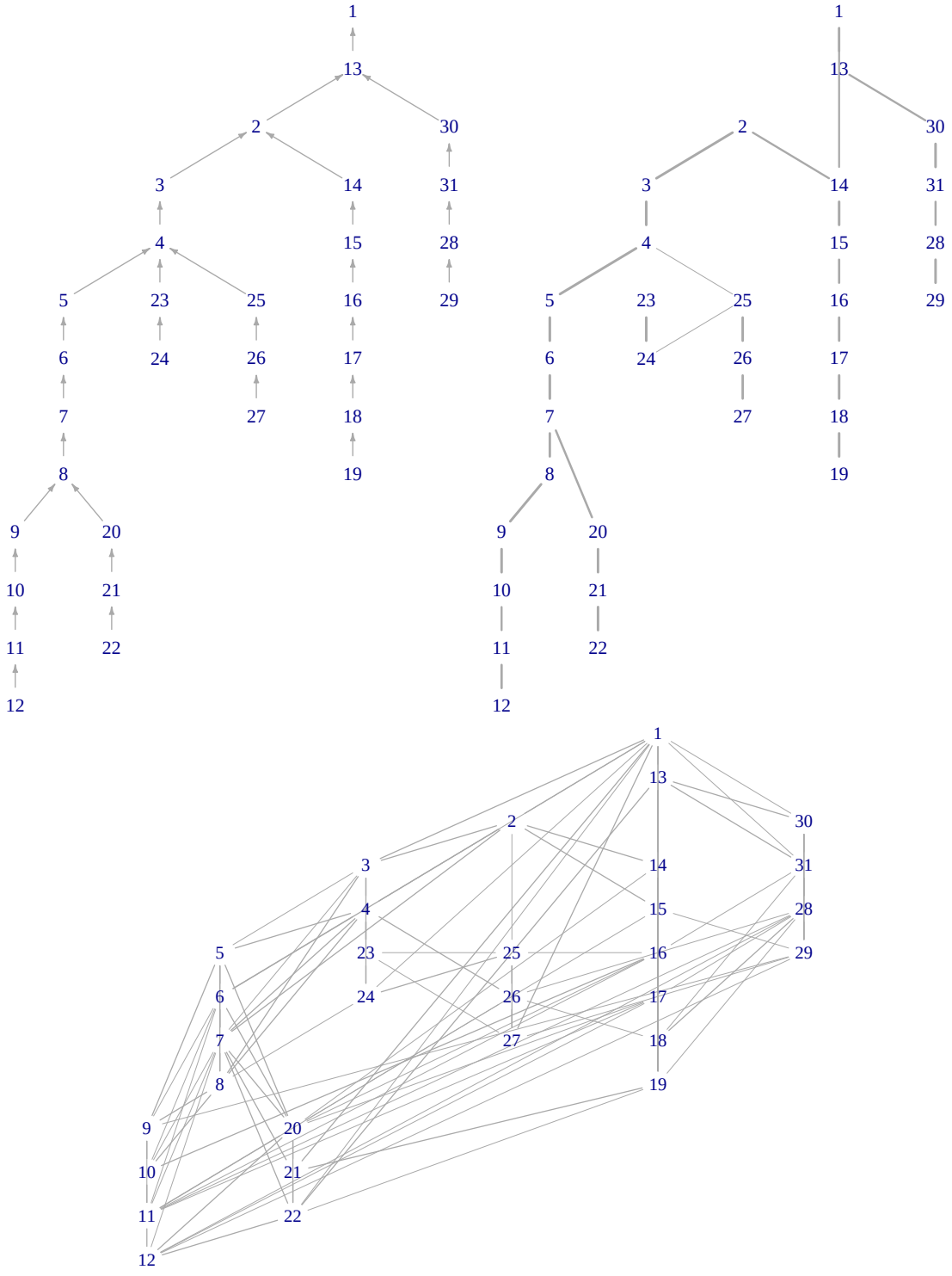

Figure 3.2: Upper left: the river network of the Danube through 31 stations in Bavaria. Upper right: the tree of a graphical model for extreme flows $\boldsymbol{X}^C = (X_1^C, \ldots, X_{31}^C)$ selected using pairwise distance correlation. Below: the graph of a censored Student copula graphical lasso fitted to $\boldsymbol{X}^C$. The size of the edges is proportional to the dependence between connected marginals on a log scale.

replacing the weights by $|\boldsymbol{\theta}_{ij}| \log n - 2w_{ij}$, where $|\boldsymbol{\theta}_{ij}|$ is the number of parameter in the bivariate model for $\boldsymbol{X}_{ij}^C$. Alternatively, the weights can be any other measure of dependence such as the mutual information between the pairs.

The upper part of Figure 3.2 compares the actual tree (formed by the rivers between the stations) to the maximal spanning tree obtained by choosing distance correlation between the pairs of $\boldsymbol{X}^C$ as weight [Székely et al., 2007]. Other weight choices selected similar trees. Up to three edges, the selected graph corresponds exactly to the river network. Recall that we forced the graph to be a tree, thus the graph according to which $\boldsymbol{X}^C$ satisfies the pairwise Markov property could have a much more complex structure. As a comparison, the censored Student copula graphical model fitted to $\boldsymbol{X}^C$, selects a graph with many more edges as seen in the bottom of Figure 3.2.

To fully determine the joint distribution of the tree graphical model, it remains to fit bivariate family distributions on F_+^2 to $\boldsymbol{X}_{ij}^C$ for every pairs (i, j) connected in the graph. In the previous section, we have motivated the approximation $\boldsymbol{X}_{ij}^C \approx \boldsymbol{Y}_{ij}^C$ for large u_{ij} such that $f_{\boldsymbol{Y}_{ij}^C}$ is a censored homogeneous density. The example below shows how to obtain such bivariate distributions on F_+^2 and gives two instances based on the Cauchy and Hüsler–Reiss distributions. In practice, imposing the homogeneity constraint is sometimes quite restrictive because $\boldsymbol{X}_{ij}$ is not necessarily regularly varying, or because the threshold is too low for the asymptotic approximation to be accurate. Instead, we might prefer working with a broader class of bivariate distributions encompassing asymptotic dependence and independence. For that purpose, we explain how any bivariate copula can be turned into a bivariate family distribution on F_+^2 with unit Pareto positive marginals and two parameters dictating how far in the tail the dependence structure of the copula should be considered.

Example 3.2.1 (Censored Homogeneous Bivariate Distribution)**.** Let $h : \mathbb{R}_+^2 \to [0, \infty)$ be a symmetric function which is homogeneous of order -3 and integrable on sets bounded away from $(0, 0)$. Consider a random vector (Y_1^C, Y_2^C) with values

in $F_+^2 = (\{0\} \cup [1, \infty))^2$ and a censored homogeneous probability density w.r.t. μ_0 defined by

$$\begin{aligned} \frac{\partial}{\partial x \, \partial y} \Pr(Y_1^C \leq x, Y_2^C \leq y) &= (1 - p_{00}) c^{-1} h(x, y), \\ \frac{\partial}{\partial x} \Pr(Y_1^C \leq x, Y_2^C = 0) &= (1 - p_{00}) c^{-1} \int_0^1 h(x, y) dy, \\ \Pr(Y_1^C = 0, Y_2^C = 0) &= p_{00}, \end{aligned} \tag{3.8}$$

where $p_{00} \in (0, 1)$, $c = \int_{C^+_{||\cdot||_\infty}} h(x, y) dx dy$ and $C^+_{||\cdot||_\infty} = \{(x, y) \in \mathbb{R}_+^2 : \max(x, y) \geq 1\}$. As a consequence of the homogeneity of h of order -3, $Y_1^C \mid Y_1^C \geq 1$ and $Y_2^C \mid Y_2^C \geq 1$ are unit Pareto. It also follows from the homogeneity of h that integrating it over $\{\max(x, y) \geq k\}$ for $k > 0$ determines the same distribution as in the case $k = 1$. All the quantities required to sample from (Y_1^C, Y_2^C) are provided in Example A.3.1.

We now present two concrete examples related to the extremal Cauchy and Hüsler–Reiss bivariate distributions respectively.

- Suppose that (Y_1^C, Y_2^C) is the censored version of the limit of $t^{-1}(Y_1, Y_2) \mid \max(Y_1, Y_2) \geq t$ when (Y_1, Y_2) is bivariate Cauchy; in other words, h is the density of the exponent measure of the extremal t distribution with degree of freedom $\nu = 1$ given by

  $$h(x, y) = (x^2 - 2\rho xy + y^2)^{-3/2},$$

  [Demarta and McNeil, 2005, Thibaud and Opitz, 2015]. In this case, Equation (3.8) defines a valid distribution on F_+^2 when $\rho \in (-\infty, 1)$. Taking ρ close to 1 makes Y_1^C and Y_2^C strongly positively dependent, whereas ρ close to $-\infty$ corresponds to strong negative dependence.

- In the bivariate case, the Hüsler–Reiss exponent measure defined in (5.4) has density

  $$h(x, y) = x^{-3/2} y^{-3/2} \exp\left\{ -\frac{\log^2 (y/x)}{2\rho} \right\}.$$

In this case, the distribution in (3.8) is well-defined for $\rho > 0$. Asymptotic independence occurs when $\rho \to \infty$, whereas strong positive dependence arises as $\rho \to 0$.

Example 3.2.2 (Censored Bivariate Copula)**.** Let

$$\tilde{Y}_1^C = \begin{cases} 0 & \text{if } U < s, \\ 1/\{1 - F_{U|U\geq s}(U)\} & \text{else,} \end{cases}$$

$$\tilde{Y}_2^C = \begin{cases} 0 & \text{if } V < t, \\ 1/\{1 - F_{V|V\geq t}(V)\} & \text{else,} \end{cases}$$

where $s, t \in [0, 1]$ and (U, V) is distributed according to a symmetric copula $C(u, v)$, i.e., any continuous bivariate distribution on $\mathbb{R}^2$ with marginals satisfying $U \overset{d}{=} V \sim \mathcal{U}(0, 1)$ and $(U, V) \overset{d}{=} (V, U)$. The model is defined by

$$(Y_1^C, Y_2^C) = \begin{cases} (\tilde{Y}_1^C, \tilde{Y}_2^C) | \max(\tilde{Y}_1^C, \tilde{Y}_2^C) \geq 1 & \text{w.p. } p_{00}, \\ (0, 0) & \text{w.p. } 1 - p_{00}, \end{cases}$$

which takes values on F_+^2. In particular, $Y_1^C | Y_1^C \geq 1$ and $Y_2^C | Y_2^C \geq 1$ are Pareto distributed of order 1. Equivalently, the model can be formulated as in (3.8):

$$h(x, y) = c(\tilde{u}, \tilde{v})(1 - s)x^{-2}(1 - t)y^{-2},$$
$$\int_0^1 h(x, y)dy = C_{Y_2|Y_1}(t \mid \tilde{u})(1 - s)x^{-2}, \quad c = 1 - C(s, t),$$

for $\tilde{u} = F^{-1}_{U|U\geq s}(1 - x^{-1})$, and $\tilde{v} = F^{-1}_{V|V\geq t}(1 - y^{-1})$. All necessary quantities for sampling are presented in Example A.3.1.

Contrary to the previous example, the density h is not necessarily homogeneous. However, suppose that the density of $n^{-1}(Y_1, Y_2) \mid \max(Y_1, Y_2) \geq n$ converges to a probability density as n grows, where (Y_1, Y_2) has joint cumulative distribution function $C(1-x^{-1}, 1-y^{-1})$. Then h converges to this homogeneous limiting density as $s = t \to \infty$. Thereby, parameters s and t decide how far in the tail the dependence structure of the given copula is taken for modeling (Y_1^C, Y_2^C).

Recall that we are estimating a tree graphical model for $\boldsymbol{X}^C$ and have already selected the tree. We now fit to every pair $\boldsymbol{X}_{ij}^C$ that is connected in the tree the

following bivariate distributions defined in Example 3.2.1 and 3.2.2: the extremal Cauchy and Hüsler–Reiss censored homogeneous distributions and the censored copula distributions using Joe, Gaussian and Student copula (with degrees of freedom $\nu = 1, 10, 30, 60$ for the latter) using the `R` package `copula` [Yan, 2007]. We then select the model that minimizes the BIC. The table below shows how many times the bivariate models listed above were preferred for modeling a connected pairs. For each of them, the symmetric case $s = t$ was optimal.

Gaussian	Hüsler–Reiss	$t_{\nu=10}$	ext. Cauchy	$t_{\nu=30}$	Joe	$t_{\nu=1}$	$t_{\nu=60}$
9	6	6	5	3	1	0	0

It may seem surprising at first that the Gaussian model is so often selected because it can only represent a regime of asymptotic independence whereas the extreme flows exhibit a particularly strong dependence. Recall however that we do not fit the model to the entire vector $\boldsymbol{X}_{ij}$ but only to large or censored observations, and this formulation allows the Gaussian to describe strong tail dependence accurately, at least in this case.

As explained in (3.5), the joint density $f_{\boldsymbol{X}^C}$ is now fully determined because we have selected the tree and estimated bivariate models for each connected pair. We now compute the BIC of the tree graphical model using (3.6) and compare it to the one of the censored Student copula graphical model fitted to $\boldsymbol{X}^C$. The latter has the following parameters: the degree of freedom ν, the covariance matrix $\frac{\nu}{\nu-2}\Sigma$, and a regularization parameter λ determining the sparsity of Σ^{-1}, that is, the percentage of zero entries in the upper triangular part and diagonal. We choose λ by minimizing the BIC computed from (6.4) for $\nu = 10, 20, 30, 60, \infty$ and reported in the table below together with the sparsity of Σ^{-1}.

	Tree	$t_{\nu=30}$	Gaussian
BIC	**25.3**	25.9	27.4
Sparsity		61%	53%
λ		0.0073	0.0044

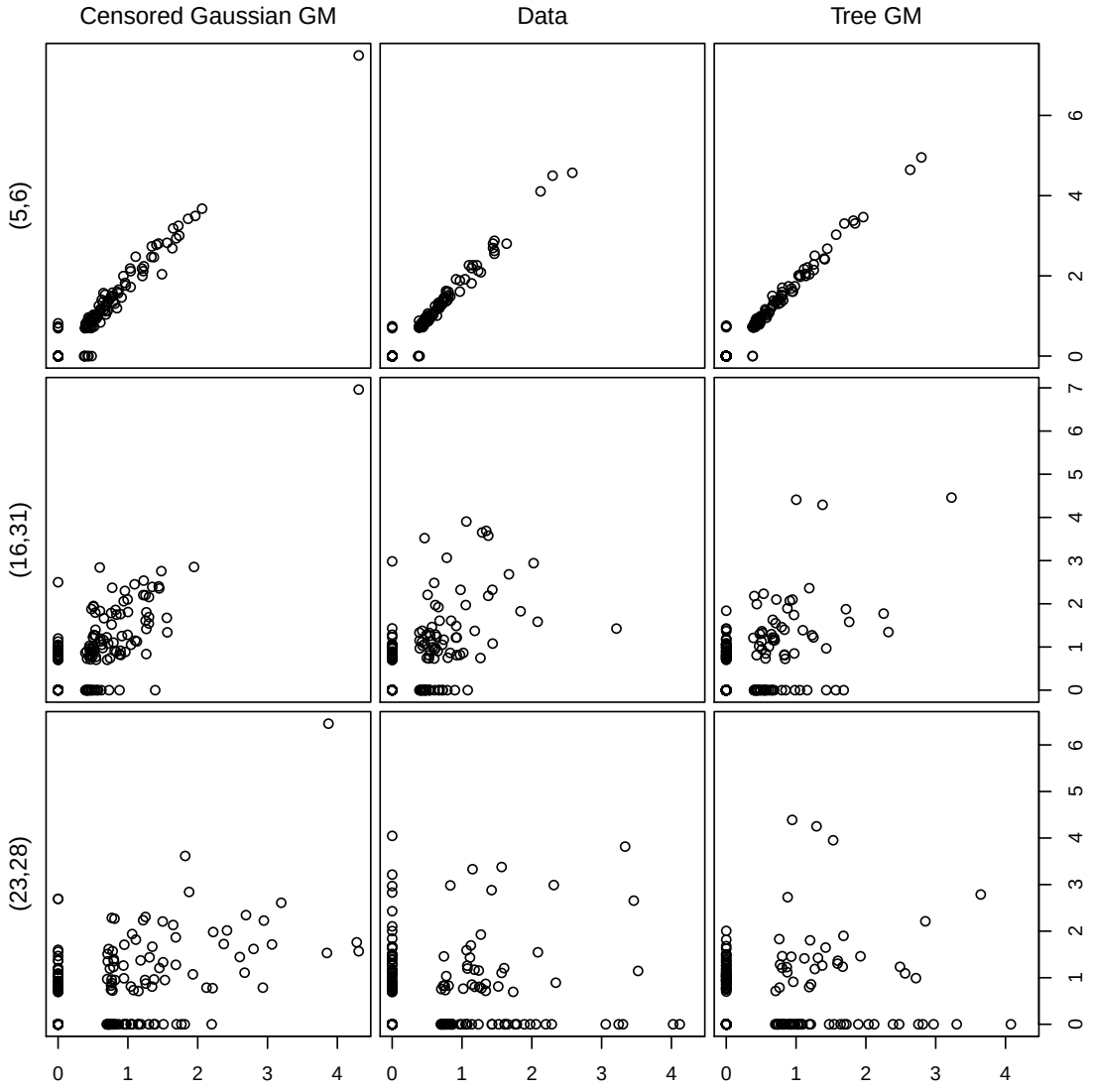

Figure 3.3: Scatterplots of the extreme flow data (middle column) and simulations from a censored Student copula graphical model (left column) and a tree graphical model (right column) on a log-scale for the pairs $(5, 6)$, $(2, 12)$, $(23, 28)$ sharing respectively the strongest, an average and the smallest correlation. Only $(5, 6)$ is connected in the tree.

The tree graphical model slightly outperforms the censored Student graphical model. Forcing the dependence structure to be a tree reduced drastically the number of parameters while modeling connected pair using a broader class of bivariate distributions than what the Student model permits. Although the Student model benefits from a more complex graph, the rough approximation that $\boldsymbol{X}^C$ factorizes according to a tree seems appropriate here. To assess the goodness of the fit, we simulate $n = 428$ realizations of $\boldsymbol{X}^C$ from the tree and censored Student graphical model with $\nu = 30$. Scatterplots of some pairs are compared in Figure 3.3. Both models appear to fit the extreme flows relatively well as few differences are noticeable.

As a further diagnostic, we test if observed pairs (X_i^C, X_j^C) take non-zero values with probability expected from the estimated graphical models. We perform binomial tests with level 0.05 for all pairs $i, j = 1, \ldots, d$ and report the percentage of time the test is rejected in the next table.

	Tree	$\mathrm{t}_{\nu=30}$	Gaussian	nbr of pairs
$\Pr(X_i > 0, X_j > 0)$ [1]	17%	7%	4%	465
$\Pr(X_i = 0, X_j > 0)$	13%	29%	13%	961

According to this comparison, the censored Gaussian copula graphical model is the most accurate to describe the occurrence of an extremal event. However, we have seen that its BIC remains larger than the tree and Student graphical model with $\nu = 30$ that seem more appropriate for modeling the strong dependence between large flows.

To sum up, we have illustrated how to infer the tail distribution of a random vector $\boldsymbol{X}$ using graphical models by performing inference on its censored version $\boldsymbol{X}^C$. We have seen that forcing the graph to have a simple structure enables us to express the probability density of $\boldsymbol{X}^C$ in terms of low-dimensional densities that can then be estimated separately. Alternatively, one can model $\boldsymbol{X}^C$ using a parametric

[1] In the case of the tree graphical model, this probability was estimated from 10 000 realizations

multivariate distribution and subsequently reduce the number of parameters by exploiting relations such as conditional independence.

These two approaches have been applied to extreme river flows $\boldsymbol{X}^C$ to which we fitted a tree and a censored Student copula graphical model. As the flows are strongly dependent, assuming block independence between the marginals of $\boldsymbol{X}^C$ is not realistic. In this context, conditional independence revealed itself a useful assumption to simplify $f_{\boldsymbol{X}^C}$. The fitted models satisfactorily replicated the data. In particular, the tree graphical model has a structure that is very simple and easily interpretable as it recovers almost exactly the actual river network, although no information on the location of the stations was included in the estimation.

We end by mentioning that the goal pursued in Chapter 5 is to characterize limiting tail densities that are homogeneous and factorize. The framework in this chapter was different because we assumed a non-asymptotic factorization of $f_{\boldsymbol{X}^C}$ and only passed to the limit to motivate homogeneous distributions for some of the marginals of $\boldsymbol{X}^C$.

4

One-Component Regular Variation

Contents

4.1 One-Component Regular Variation for Functions

Consider the commutative and associative operation

$$x \star y := T^{-1}\{T(x)T(y)\}, \quad x, y \in E, \tag{4.1}$$

where $T : E = [e_0, e_1) \to [1, \infty)$ is a diffeomorphism with $T(e_0) = 1$, for $e_0 \in \mathbb{R}$, $e_1 \in \mathbb{R} \cup \{\infty\}$. When $E = [0, \infty)$, possible operations include multiplication ($T \equiv \mathrm{id}$), addition ($T \equiv \exp$) and $x \star y = ||x, y||_p$ for $p > 0$, $T(x) = \exp(x^p)$. When $E = [0, e_1)$ and $e_1 < \infty$, $T(x) = (1 - e_1^{-1}x)^{-1}$ gives $x \star y = x + y - e_1 xy$.

We call a positive and measurable function u on E *regularly varying with decay* T and limit v if

$$\frac{u(t \star x)}{u(t \star e_0)} \to v(x) > 0, \tag{4.2}$$

where the arrow stands for pointwise convergence on the entire domain of definition as $t \uparrow e_1$. It follows that $u \circ T^{-1}$ is regularly varying, and thus $v(x) = T(x)^\alpha$ for some $\alpha \in \mathbb{R}$. The convergence moreover is uniform (see Resnick [1987]). We denote (4.2) by $u \in T\text{-RV}_\alpha$.

Let us extend this notion to the multivariate setting. Consider two measurable and non-negative functions u and v on $E \times \mathbb{R}^{d-1}$ such that $u(\cdot, \mathbf{1}) > 0$, $v(\cdot, \mathbf{1}) > 0$, and a non-negative function h on $\mathbb{R}^{d-1}$ satisfying $h(\mathbf{1}) = 1$. We call $\mathbf{1}$ the pivot and its choice is arbitrary.

Lemma 4.1.1 (Characterization)**.** *The following are equivalent:*

i. $u(t \star x, \boldsymbol{y})/u(t \star e_0, \mathbf{1}) \to v(x, \boldsymbol{y})$,

ii. $u(\cdot, \mathbf{1}) \in T\text{-RV}_\alpha$ *and* $u(t, \boldsymbol{y})/u(t, \mathbf{1}) \to h(\boldsymbol{y})$,

iii. $u(\cdot, \boldsymbol{y}) \in T\text{-RV}_\alpha$, $\forall \boldsymbol{y}$ *s.t.* $h(\boldsymbol{y}) > 0$, *and* $u(t, \boldsymbol{y})/u(t, \mathbf{1}) \to h(\boldsymbol{y})$,

iv. $u(t \star x, \boldsymbol{y})/u(t \star e_0, \mathbf{1}) \to T(x)^\alpha\, h(\boldsymbol{y})$ *uniformly in* x.

We say that u is *regularly varying in its first component*, written $u \in T\text{-RV}_\alpha^x(h)$, if (i.)–(iv.) hold. Condition $u(\cdot, \boldsymbol{y}) \in \text{RV}_\alpha$, $\forall \boldsymbol{y}$, corresponds to the uniform regular variation of Meerschaert [1993] if it is assumed further that the convergence is uniform in $\boldsymbol{y}$.

Proof. (i.) $\Rightarrow$ (ii.): to derive the first condition, set $\boldsymbol{y} = \mathbf{1}$; for the second, $x = e_0$. (ii.)$\Rightarrow$(iii.): let $\boldsymbol{y}$ s.t. $h(\boldsymbol{y}) > 0$. Since $u(t \star e_0, \boldsymbol{y})/u(t \star e_0, \mathbf{1}) \to h(\boldsymbol{y})$, $u(t \star e_0, \boldsymbol{y}) > 0$ for t sufficiently large. It follows that

$$\frac{u(t \star x, \boldsymbol{y})}{u(t \star e_0, \boldsymbol{y})} = \frac{u(t \star x, \boldsymbol{y})}{u(t \star x, \mathbf{1})} \frac{u(t \star x, \mathbf{1})}{u(t \star e_0, \mathbf{1})} \frac{u(t \star e_0, \mathbf{1})}{u(t \star e_0, \boldsymbol{y})} \to T(x)^\alpha.$$

(iii.) $\Rightarrow$ (iv.): as $h(\mathbf{1}) > 0$,

$$\frac{u(t \star x, \boldsymbol{y})}{u(t \star e_0, \mathbf{1})} = \frac{u(t \star x, \boldsymbol{y})}{u(t \star x, \mathbf{1})} \frac{u(t \star x, \mathbf{1})}{u(t \star e_0, \mathbf{1})} \to T(x)^\alpha h(\boldsymbol{y}),$$

and the convergence is uniform in x. (iv.) $\Rightarrow$ (i.): clear. $\square$

We now generalize the representation theorem for univariate regular variation (see Bingham et al. [1989] in the case $T = \mathrm{id}$, Jaroš and Kusano [2004] otherwise).

Proposition 4.1.2 (Multivariate Representation Theorem)**.** *Suppose that* $\alpha \in \mathbb{R}$ *and* $u(\cdot, \mathbf{1}) > 0$. *It holds* $u \in T\text{-RV}_\alpha^x(h)$ *if and only if*

$$u(x, \boldsymbol{y}) = c(x) \exp\left\{\int_{e_0}^x \alpha(z) \frac{T'(z)}{T(z)} dz\right\} q(x, \boldsymbol{y}), \tag{4.3}$$

for e_0 *sufficiently large and measurable functions s.t.* $c(t) \to c > 0$, $\alpha(t) \to \alpha$, $q(t, \boldsymbol{y}) \to h(\boldsymbol{y})$ *as* $t \uparrow e_1$.

As a consequence, $\forall \epsilon > 0$, $\exists c_1, c_2 > 0$ such that

$$c_1 T(x)^{\alpha - \epsilon} h(\boldsymbol{y}) < u(x, \boldsymbol{y}) < c_2 T(x)^{\alpha + \epsilon} h(\boldsymbol{y}), \tag{4.4}$$

for x sufficiently large and $\boldsymbol{y}$ satisfying $h(\boldsymbol{y}) \neq 0$.

Proof. For the direct implication, write $u(x, \boldsymbol{y}) = u(x, \boldsymbol{y})/u(x, \mathbf{1})\, u(x, \mathbf{1})$. Lemma 4.1.1 gives $q(t, \boldsymbol{y}) := u(t, \boldsymbol{y})/u(t, \mathbf{1}) \to h(\boldsymbol{y})$ and $u(x, \mathbf{1}) \in T\text{-RV}_\alpha$. The conclusion follows by applying the representation theorem on the regularly varying function $u\{T^{-1}(x), \mathbf{1}\}$. For the reverse, use the change of variable $z = t \star \tilde{z}$, $dz = T(t)T'(\tilde{z})/T'(t \star \tilde{z})$, to find

$$\int_{t \star e_0}^{t \star x} \alpha(z) T'(z)/T(z) dz = \int_{e_0}^x \alpha(t \star \tilde{z}) T'(\tilde{z})/T(\tilde{z}) d\tilde{z} \to \alpha \log T(x),$$

and thus $u(t \star x, \boldsymbol{y})/u(t \star e_0, \mathbf{1}) \to T(x)^\alpha h(\boldsymbol{y})$.

□

Another important result is Karamata's theorem, which relates the regular variation of a univariate function to that of its integral, see Bingham et al. [1989]. We generalize it for one-component regular variation. We only treat the case of a negative power index $-\alpha$ for $\alpha > 0$, but the result is easily reformulated when $\alpha \leq 0$. A required assumption is that $\bar{U}(x, \boldsymbol{y}) := \int_x^\infty u(z, \boldsymbol{y}) dz$ exists $\forall x, \boldsymbol{y}$, which is true for e_0 sufficiently large if $u(x, \boldsymbol{y})/T'(x) \in T\text{-RV}_{-\alpha-1}^x(h)$ as a consequence of (4.4).

Theorem 4.1.3 (Multivariate Karamata's Theorem)**.** *Suppose that* $\alpha > 0$, $u(\cdot, \mathbf{1}) > 0$ *and* $\bar{U}(x, \boldsymbol{y}) := \int_x^\infty u(z, \boldsymbol{y})dz$ *exists* $\forall x, \boldsymbol{y}$*. Then,*

$$\frac{u(x,\boldsymbol{y})}{T'(x)} \in T\text{-RV}^x_{-\alpha-1}(h) \quad \iff \quad \frac{T(t)}{T'(t)}\frac{u(t,\boldsymbol{y})}{\bar{U}(t,\mathbf{1})} \to \alpha\, h(\boldsymbol{y}). \tag{4.5}$$

In this case, $\bar{U} \in T\text{-RV}^x_{-\alpha}(h)$*, and its representation has coefficient* $c(x) \equiv \bar{U}(e_0, \mathbf{1})$*.*

Proof. To prove the direct implication, use the change of variable $z = t \star \tilde{z}$ to find

$$\begin{aligned}\frac{T(t\star e_0)}{T'(t\star e_0)}\frac{u(t\star e_0,\boldsymbol{y})}{\int_{t\star e_0}^\infty u(z,\mathbf{1})dz} &= \left\{\int_{e_0}^\infty \frac{u(t\star\tilde{z},\mathbf{1})}{u(t\star e_0,\mathbf{1})}\frac{T'(t\star e_0)}{T'(t\star\tilde{z})}\frac{T'(\tilde{z})T(t)}{T(t\star e_0)}d\tilde{z}\right\}^{-1}\frac{u(t\star e_0,\boldsymbol{y})}{u(t\star e_0,\mathbf{1})}\\ &\to \left\{\int_{e_0}^\infty T(\tilde{z})^{-\alpha-1}T'(\tilde{z})d\tilde{z}\right\}^{-1} h(\boldsymbol{y}) = \alpha h(\boldsymbol{y}),\end{aligned}$$

because $u(\cdot,\mathbf{1})/T'(\cdot) \in T\text{-RV}_{-\alpha-1}$ and $u(t\star e_0,\boldsymbol{y})/u(t\star e_0,\mathbf{1}) \to h(\boldsymbol{y})$ from Lemma 4.1.1. Limit and integral can be exchanged since the integrand is dominated by $cT(z)^{-\alpha-1+\epsilon}$ $T'(z)$, for some $c > 0$, $\epsilon \in (0, \alpha)$, and t sufficiently large thanks to (4.4).

For the reverse implication, let $\alpha(x) := \{u(x,\mathbf{1})/\bar{U}(x,\mathbf{1})\}/\ \{T'(x)/T(x)\}$, which satisfies $\alpha(t) \to \alpha$. Integrate $u(z,\mathbf{1})/\bar{U}(z,\mathbf{1}) = \alpha(z)T'(z)/T(z)$ from both sides between e_0 and x to obtain

$$\bar{U}(x,\mathbf{1}) = \bar{U}(e_0,\mathbf{1})\exp\left\{-\int_{e_0}^x \alpha(z)\frac{T'(z)}{T(z)}dz\right\}.$$

This is the representation of a T-regularly varying function with $c(x) = \bar{U}(e_0,\mathbf{1})$ thanks to Proposition 4.1.2. Hence,

$$\begin{aligned}&\frac{T'(t\star e_0)}{T'(t\star x)}\frac{u(t\star x,\boldsymbol{y})}{u(t\star e_0,\mathbf{1})} =\\ &\qquad \frac{T'(t\star e_0)}{T(t\star e_0)}\frac{\bar{U}(t\star e_0,\mathbf{1})}{u(t\star e_0,\mathbf{1})}\ \frac{T(t\star x)}{T'(t\star x)}\frac{u(t\star x,\boldsymbol{y})}{\bar{U}(t\star x,\mathbf{1})}\ \frac{T(t\star e_0)}{T(t\star x)}\ \frac{\bar{U}(t\star x,\mathbf{1})}{\bar{U}(t\star e_0,\mathbf{1})}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \to T(x)^{-\alpha-1}h(\boldsymbol{y}),\end{aligned}$$

which ends the proof of the equivalence in (4.5). Moreover, $\bar{U} \in T\text{-RV}_\alpha(h)$ because

$$\begin{aligned}\frac{\bar{U}(t\star x,\boldsymbol{y})}{\bar{U}(t\star e_0,\mathbf{1})} &= \bar{U}(t\star e_0,\mathbf{1})^{-1}\int_{t\star x}^\infty u(\tilde{x},\boldsymbol{y})d\tilde{x}\\ &= \frac{u(t\star e_0,\mathbf{1})}{\bar{U}(t\star e_0,\mathbf{1})}\frac{T(t\star e_0)}{T'(t\star e_0)}\int_x^\infty \frac{u(t\star z,\boldsymbol{y})}{u(t\star e_0,\mathbf{1})}\frac{T'(t\star e_0)}{T'(t\star z)}T'(z)dz\\ &\to T(x)^{-\alpha}h(\boldsymbol{y}).\end{aligned}$$

□

For instance, consider the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with probability density ϕ and survival function $\bar{\Phi}$. Since

$$\frac{\phi(t)}{2t\bar{\Phi}(t)} \to \frac{1}{2\sigma^2},$$

Theorem 4.1.3 gives $\phi(x)/T'(x) \in T\text{-RV}_{-(2\sigma^2)^{-1}-1}$ and $\bar{\Phi} \in T\text{-RV}_{-(2\sigma^2)^{-1}}$ for $T(x) = \exp(x^2)$. Let f be the probability density of the log-Cauchy distribution and $\bar{F}$ its survival function. As $t \log t \, f(t)/\bar{F}(t) \to 1$, it follows that $xf(x) \in \log\text{-RV}_{-2}$, and $\bar{F} \in \log\text{-RV}_{-1}$.

We now extend the standard multivariate regular variation in (1.34) to general decays. We denote the sign of $x \in \mathbb{R}$ by $\sigma_x \in \{-1, 0, 1\}$ and $\boldsymbol{\sigma_x} = (\sigma_{x_1}, \ldots, \sigma_{x_d})$ when $\boldsymbol{x} \in \mathbb{R}^d$. Operations between vectors are done componentwise, $x \star \boldsymbol{y} := (x\mathbf{1}) \star \boldsymbol{y} = (x \star y_1, \ldots, x \star y_d)$, and $\boldsymbol{e}_0 := e_0\mathbf{1}$. We say that $u : \mathbb{R}^d \to [0, \infty)$ satisfying $u(\lambda\mathbf{1}) > 0$, $\forall \lambda \in E$, is *regularly varying with decay* T if there exists v such that $v(\lambda\mathbf{1}) > 0$, $\forall \lambda \in E$, and

$$\frac{u\{(t \star |\boldsymbol{x}|)\boldsymbol{\sigma_x}\}}{u(t \star \boldsymbol{e}_0)} \to v(\boldsymbol{x}), \tag{4.6}$$

on $F := \{(-e_1, -e_0) \cup \{0\} \cup [e_0, e_1)\}^d \setminus \{\mathbf{0}\}$; by convention, $(t \star |x|)1_x = 0$ if $x = 0$.

Multivariate regular variation is easily expressed in terms of one-component regular variation by introducing the following change of variable. We say that $\phi : C \to (0, \infty) \times \Omega$, for $\Omega \subset \mathbb{R}^{d-1}$, defines a *radial system of coordinates* if it has the form

$$\phi : \boldsymbol{x} \mapsto \{r(\boldsymbol{x}), \boldsymbol{\theta}(\boldsymbol{x})\}, \quad \phi^{-1} : (r, \boldsymbol{\theta}) \mapsto \{r \star |\boldsymbol{\theta}^{-1}(\boldsymbol{\theta})|\}\boldsymbol{\sigma}_{\boldsymbol{\theta}^{-1}(\boldsymbol{\theta})}, \tag{4.7}$$

where $r(\cdot)$ and $\boldsymbol{\theta}(\cdot)$ satisfy $r\{(\lambda \star |\boldsymbol{x}|)\boldsymbol{\sigma_x}\} = \lambda \star r(\boldsymbol{x})$, $\boldsymbol{\theta}\{(\lambda \star |\boldsymbol{x}|)\boldsymbol{\sigma_x}\} = \boldsymbol{\theta}(\boldsymbol{x})$, $\forall \lambda \in E$. Examples include spherical coordinates on $\mathbb{R}^d$ and pseudo-polar coordinates defined by $r(\boldsymbol{x}) = ||\boldsymbol{x}||$, a norm, and $\boldsymbol{\theta}(\boldsymbol{x}) = \boldsymbol{x}_{1:d-1}/r(\boldsymbol{x})$ on $\mathbb{R}^d_+$. When $T \equiv \exp$, the latter translates into $r(\boldsymbol{x}) = \log(||e^{\boldsymbol{x}}||)$, $\boldsymbol{\theta}(\boldsymbol{x}) = \boldsymbol{x} - r(\boldsymbol{x})$.

It becomes clear that u is regularly varying if and only if $g(r, \boldsymbol{\theta}) := u\{\phi^{-1}(r, \boldsymbol{\theta})\}$ is regularly varying w.r.t. its first component since

$$\frac{g(t \star r, \boldsymbol{\theta})}{g(t \star \mathbf{1}_r, \mathbf{1}_{\boldsymbol{\theta}})} = \frac{u\{(t \star |\boldsymbol{x}|)\boldsymbol{\sigma_x}\}}{u(t \star \boldsymbol{e}_0)} \to v(\boldsymbol{x}), \tag{4.8}$$

for $\mathbf{1}_r = r(\boldsymbol{e}_0)$ and $\mathbf{1}_{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{e}_0)$, and in this case, Lemma 4.1.1 gives $v(\boldsymbol{x}) = T\{r(\boldsymbol{x})\}^{\alpha} h\{\boldsymbol{\theta}(\boldsymbol{x})\}$. In particular, v is homogeneous of order α, i.e., $v\{(\lambda \star |\boldsymbol{x}|)\boldsymbol{\sigma}_{\boldsymbol{x}}\} = T(\lambda)^{\alpha} v(\boldsymbol{x})$, $\forall \lambda \in E$, and we write (4.6) as $u \in T\text{-RV}_{\alpha}(v)$.

4.2 One-Component Regular Variation for Probability Distributions

So far, one-component regular variation has been treated for functions; we develop it further for distributions. A specific representation is found for regularly varying random vectors. Their limits moreover form the class of distributions that are homogeneous w.r.t. their first component. In a subsequent result, we show that one-component regular variation of the probability density implies, under an extra condition, one-component regular variation of the distribution.

For simplicity, we treat the case $T \equiv \text{id}$. Let $(X, \boldsymbol{Y})$ be a random variable with probability distribution μ on $[1, \infty) \times \mathbb{R}^{d-1}$. We call $\bar{F}(x, \boldsymbol{y}) = \mu\{[x, \infty) \times (-\boldsymbol{\infty}, \boldsymbol{y}]\}$ the *x-survival function*. Subscripts X or $\boldsymbol{Y}$ refer to the corresponding marginal distribution. Suppose that $\bar{F}_X > 0$. We are interested in the weak limit (see Billingsley [1995]) of $(t^{-1}X, \boldsymbol{Y}) \mid X \geq t$, equivalent to the weak limit of its distribution $\mu(tA, B)/\bar{F}_X(t)$, and of its x-survival function $\bar{F}(tx, \boldsymbol{y})/\bar{F}_X(t)$.

Let H and ν be two probability distributions on $\mathbb{R}^{d-1}$ and $[1, \infty) \times \mathbb{R}^{d-1}$ respectively such that $\nu_X \neq \delta_1$. We denote by P_{α} the Pareto distribution on $[1, \infty)$ with shape $\alpha > 0$.

Theorem 4.2.1 (Characterization for Probability Distributions)**.** *If $\alpha > 0$, $\bar{F}_X > 0$ and $\nu_X \neq \delta_1$, the following are equivalent:*

i. $(t^{-1}X, \boldsymbol{Y}) \mid X \geq t \overset{w}{\Rightarrow} \nu$,

ii. $\boldsymbol{Y} \mid X \geq t \overset{w}{\Rightarrow} H$ *and* $t^{-1}X \mid X \geq t \overset{w}{\Rightarrow} P_{\alpha}$,

iii. $\bar{F}(x, \boldsymbol{y}) = c(x) \exp\left\{-\int_{x_0}^{x} \alpha(z) z^{-1} dz\right\} Q(\boldsymbol{y} \mid x)$, *for measurable* $\alpha(t) \to \alpha$, $c(t) \to c$ *and a conditional cumulative distribution function* $Q(\boldsymbol{y} \mid t) \overset{w}{\Rightarrow} H$, $\forall x \geq x_0 \geq 1$, $\forall \boldsymbol{y}$,

iv. $(t^{-1}X, \boldsymbol{Y}) \mid X \geq t \overset{w}{\Rightarrow} P_\alpha \times H$.

We say that $(X, \boldsymbol{Y})$ is *regularly varying w.r.t. its first component*, written $(X, \boldsymbol{Y}) \in \mathrm{RV}^x_{-\alpha}(H)$, if (i.)–(iv.) are satisfied.

In the univariate case, the equivalence (i.)–(ii.) follows from Theorem 2.1.4 in Basrak [2000]. His proof, however, omits to show that the solution of the Cauchy's functional equation must be positive to have the form x^α. To see that $\nu_X \neq \delta_1$ is a necessary assumption, consider for instance $\bar{F}(x) = e^{-x}$. The equivalence of (i.) and (iv.) can be found in Lindskog et al. [2014] (Theorem 3.1 with $\{\lambda, (x, \boldsymbol{y})\} \mapsto (\lambda x, \boldsymbol{y})$) or in Heffernan and Resnick [2007] (Proposition 2 with $\alpha \equiv 1$, $\beta \equiv 0$). Our contribution is the representation in (iii.) and a short proof.

Proof. (i.) $\Rightarrow$ (ii.): on the one hand, $\boldsymbol{Y} \mid X \geq t \overset{w}{\Rightarrow} \nu_{\boldsymbol{Y}} =: H$. On the other, $t^{-1}X \mid X \geq t \overset{w}{\Rightarrow} \nu_{\boldsymbol{X}}$, or equivalently, $\bar{F}_X(tx)/\bar{F}_X(t) \overset{w}{\Rightarrow} \bar{G}(x)$, where $\bar{G}$ is the survival function of $\nu_{\boldsymbol{X}}$. Let $N \subset [1, \infty)$ be the countable set of discontinuities of $\bar{G}$. For any $x, y \in [1, \infty) \setminus N$ such that $xy \in [1, \infty) \setminus N$,

$$\bar{G}(xy) \leftarrow \frac{\bar{F}_X(txy)}{\bar{F}_X(t)} = \frac{\bar{F}_X(txy)}{\bar{F}_X(tx)} \frac{\bar{F}_X(tx)}{\bar{F}_X(t)} \rightarrow \bar{G}(x)\bar{G}(y).$$

Left-continuity of $\bar{G}$ guarantees that for any $x, y > 1$,

$$\bar{G}(x)\bar{G}(y) = \lim_{\epsilon \uparrow 0} \bar{G}(x+\epsilon)\bar{G}(y+\epsilon) = \lim_{\epsilon \uparrow 0} \bar{G}\{(x+\epsilon)(y+\epsilon)\} = \bar{G}(xy), \tag{4.9}$$

by choosing a sequence ϵ such that $x + \epsilon$, $y + \epsilon$ and their product are in $[1, \infty) \setminus N$. We now show that $\bar{G}$ is positive. Suppose, by contradiction, that it is not. Then $x_0 := \inf\{x \geq 1 : \bar{G}(x) = 0\} < \infty$. Since $\nu_X \neq \delta_0$, $x_0 > 1$. For $\epsilon > 0$ small enough, $(x_0 - \epsilon)^2 \geq x_0$, and thus, $0 = \bar{G}\{(x_0 - \epsilon)^2\} = \bar{G}(x_0 - \epsilon)\bar{G}(x_0 - \epsilon) > 0$, a contradiction. Since $\bar{G} > 0$, $\tilde{G}(x) := \log \bar{G}(e^x)$ is well-defined on $[0, \infty)$, measurable and satisfies the Cauchy's functional equation. Therefore, there exists $\alpha \in \mathbb{R}$ such that $\bar{G}(x) = x^{-\alpha}$ (see Bingham et al. [1989]). The only valid choice for a probability distribution is $\alpha > 0$, giving $\nu_X = P_\alpha$. (ii.) $\Rightarrow$ (iii.): since $\bar{F}_X \in \mathrm{RV}_{-\alpha}$ and $\bar{F}_{\boldsymbol{Y}|X \geq t} \overset{w}{\Rightarrow} H$, write $\bar{F}(x, \boldsymbol{y}) = \bar{F}_X(x)\bar{F}_{\boldsymbol{Y}|X \geq x}(\boldsymbol{y} \mid x)$ and apply the representation theorem on $\bar{F}_X$ to conclude. (iii.) $\Rightarrow$ (iv.): it follows that

$$\frac{\bar{F}(tx, \boldsymbol{y})}{\bar{F}_X(t)} = \frac{c(tx)}{c(t)} \exp\left\{-\int_t^{tx} \alpha(z) z^{-1} dz\right\} Q(\boldsymbol{y} \mid tx) \overset{w}{\Rightarrow} x^{-\alpha} H(\boldsymbol{y}),$$

because $\int_t^{tx} \alpha(z) z^{-1} dz = \int_1^x \alpha(t\tilde{z})\tilde{z}^{-1} dz \to \alpha \log x$. (iv.) $\Rightarrow$ (i.): clear. $\square$

Suppose further that $(X, \boldsymbol{Y})$ admits a probability density f w.r.t. the Lebesgue measure such that f_X and $f(\cdot, \mathbf{1})$ are positive. We want to know what the relation is between regular variation of f and of $(X, \boldsymbol{Y})$.

We start by answering the question in the univariate case. From Karamata's theorem (Theorem 4.5), $f_X \in \mathrm{RV}_{-\alpha-1}$ implies $\bar{F}_X \in \mathrm{RV}_{-\alpha}$ and its representation has coefficient $c(\cdot) \equiv 1$. Thus, $X \in \mathrm{RV}_{-\alpha}$ and its representation now determines exactly a probability distribution: $\bar{F}_X(x) = \exp\{-\int_1^x \alpha(z) z^{-1} dz\}$, $\forall x \geq 1$, for a non-negative and measurable $\alpha(t) \to \alpha > 0$. The following example shows that regular variation of X does not imply regular variation of f_X in general. Equalities and pointwise convergences between densities hold almost everywhere.

Example 4.2.2. Let $\bar{F}_X(x) = \exp\{-\int_1^x \alpha(y) y^{-1} dy\}$ on $[1, \infty)$ and $\alpha(x) = \sin(x) + 2$. Since $\bar{F}_X(x) = c(x) x^{-2}$ for $c(x) = \exp\{-\int_1^x y^{-1} \sin y \, dy\}$ satisfying $c(t) \to 1$, the representation theorem gives $\bar{F}_X \in \mathrm{RV}_{-2}$, and thus $X \in \mathrm{RV}_{-2}$. However, $t f_X(t)/\bar{F}_X(t) = \alpha(t) \not\to 2$, and from Karamata's theorem f_X is not regularly varying.

We now provide an answer in the multivariate case. If $f \in \mathrm{RV}^x(v)$ such that the limit v is non-null and the sequence is dominated by an integrable function, then by the dominated convergence theorem

$$f_{t^{-1}X, \boldsymbol{Y} | X \geq t}(x, \boldsymbol{y}) = c_t^{-1} \frac{f(tx, \boldsymbol{y})}{f(t, \mathbf{1})} \to c^{-1} v(x, \boldsymbol{y}), \tag{4.10}$$

where $c_t = \int_1^\infty \int_{\mathbb{R}^{d-1}} f(tx, \boldsymbol{y})/f(t, \mathbf{1}) dx d\boldsymbol{y} \to c = \int_1^\infty \int_{\mathbb{R}^{d-1}} v(x, \boldsymbol{y}) dx d\boldsymbol{y} < \infty$. Alternatively, we can assume v integrable and the sequence monotone instead of dominated. Let $\alpha > 0$ and H be a probability distribution on $\mathbb{R}^{d-1}$ with density h satisfying $h(\mathbf{1}) > 0$. According to Lemma 4.1.1, the limit in (4.10) has the form $\alpha x^{-\alpha-1} h(\boldsymbol{y})$, and thus $(X, \boldsymbol{Y}) \in \mathrm{RV}_{-\alpha}(H)$. This is an important result: we can guarantee regular variation of $(X, \boldsymbol{Y})$ simply by computing the limit of $f(tx, \boldsymbol{y})/f(t, \mathbf{1})$, a useful approach when $\bar{F}$ is intractable or the weak convergences in Theorem 4.2.1 are difficult to check. Whereas monotonicity or domination is sufficient to obtain (4.10), our next result reveals necessary and sufficient conditions.

Theorem 4.2.3 (Characterization for Probability Densities)**.** *Suppose that* $\alpha > 0$, $(X, \boldsymbol{Y})$ *admits a probability density* f *such that* f_X *and* $f(\cdot, \mathbf{1})$ *are positive, and* h *is a density satisfying* $h(\mathbf{1}) > 0$*. Then the following are equivalent:*

i. $f_{\boldsymbol{Y}|X}(\mathbf{1} \mid t) \to h(\mathbf{1})$ *and* $f \in \mathrm{RV}^x_{-\alpha-1}\{h(\cdot)/h(\mathbf{1})\}$,

ii. $f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid t) \to h(\boldsymbol{y})$ *and* $f_X \in \mathrm{RV}_{-\alpha-1}$,

iii. $f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid t) \to h(\boldsymbol{y})$ *and* $f_{X|\boldsymbol{Y}=\mathbf{1}} \in \mathrm{RV}_{-\alpha-1}$,

iv. $f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid t) \to h(\boldsymbol{y})$ *and* $f_{X|\boldsymbol{Y}=\boldsymbol{y}} \in \mathrm{RV}_{-\alpha-1}$, $\forall \boldsymbol{y}$ *s.t.* $h(\boldsymbol{y}) > 0$,

v. $f(x, \boldsymbol{y}) = \alpha(x)x^{-1} \exp\{-\int_1^x \alpha(z)z^{-1}dz\}\, q(\boldsymbol{y} \mid x)$*, for a positive and measurable* $\alpha(t) \to \alpha$ *and a conditional probability density* $q(\boldsymbol{y} \mid t) \to h(\boldsymbol{y})$, $\forall x \geq 1$, $\forall \boldsymbol{y}$,

vi. $f_{t^{-1}X,\boldsymbol{Y}|X\geq t}(x, \boldsymbol{y}) \to \alpha x^{-\alpha-1}h(\boldsymbol{y})$.

In this case, $(X, \boldsymbol{Y}) \in \mathrm{RV}^x_{-\alpha}(H)$.

From Scheffé's Lemma [Durrett, 2010], the convergence in (vi.) is also in L_1. In (i.), the first condition is necessary and can be replaced by $tf(t, \mathbf{1})/\bar{F}(t) \to \alpha h(1)$, and in (ii.) by $tf(t, \boldsymbol{y})/\bar{F}_X(t) \to \alpha h(\boldsymbol{y})/h(\mathbf{1})$.

Proof. (i.) $\Leftrightarrow$ (iv.): straightforward; use Lemma (4.1.1) for (iii.) $\Leftrightarrow$ (iv.) and (iv.) $\Leftrightarrow$ (i.). (ii.) $\Rightarrow$ (v.): write $f(x, \boldsymbol{y}) = f_X(x) f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid x)$. On the one hand, $f_X \in \mathrm{RV}_{-\alpha}$ and thus from Karamata's theorem (Theorem 4.1.3) $\bar{F}_X$ has a representation with $c(\cdot) \equiv 1$. Differentiate this representation to find one for f_X. On the other hand, $f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid t) \to h(\boldsymbol{y})$ by assumption. Altogether, this gives (v.), a well-defined probability density because ϵ and q are measurable, and Tonelli's theorem ensures that it integrates to 1. (v.) $\Rightarrow$ (vi.): integrate the representation to find $\bar{F}(x) = \exp\{-\int_1^x \alpha(z)z^{-1}dz\}$, and therefore

$$\frac{tf(tx, \boldsymbol{y})}{\bar{F}_X(t)} = \alpha(tx)x^{-1} \exp\left\{-\int_1^x \alpha(tz)z^{-1}dz\right\} q(\boldsymbol{y} \mid tx) \to \alpha x^{-\alpha-1}h(\boldsymbol{y}).$$

(vi.) $\Rightarrow$ (i.): first,

$$\frac{f(tx,\boldsymbol{y})}{f(t,\mathbf{1})} = \frac{tf(tx,\boldsymbol{y})}{\bar{F}_X(t)}\frac{\bar{F}_X(t)}{tf(t,\mathbf{1})} \to x^{-\alpha-1}\frac{h(\boldsymbol{y})}{h(\mathbf{1})}.$$

Second, apply Scheffé's Lemma on (vi.) to obtain $(X,\boldsymbol{Y}) \in \mathrm{RV}^x_{-\alpha}(H)$ and, in particular, $X \in \mathrm{RV}_{-\alpha}$. Karamata's theorem ensures that

$$f_{\boldsymbol{Y}|X}(\mathbf{1} \mid t) = \frac{tf(t,\mathbf{1})}{\bar{F}_X(t)}\,\frac{\bar{F}_X(t)}{tf_X(t)} \to h(\mathbf{1}).$$

□

The following example illustrates the use of Theorem 4.2.3, which holds also when the convergence is not uniform in $\boldsymbol{y}$.

Example 4.2.4. Consider

$$f(x,y) = \begin{cases} \frac{6}{5}x^{-2} & \text{if } y \in (0,x^{-1}), \\ \frac{6}{5}x^{-2}y & \text{else,} \end{cases} \text{ on } [1,\infty)\times(0,1].$$

The convergence $f(tx,y)/f(t,1) \to x^{-2}y$ gives $f \in \mathrm{RV}_{-2}\{h(\cdot)/h(1)\}$ for the cdf $H(y) = y^2$ with probability density $h(y) = 2y$ on $(0,1]$. Since $f(t,1)/f_X(t) \to 2 = h(1)$ (condition (i.), Theorem 4.2.3), $(X,\boldsymbol{Y}) \in \mathrm{RV}_{-\alpha}(H)$, i.e., $\bar{F}(tx,y)/\bar{F}_X(t) \to x^{-1}y^2$. The same conclusion is drawn by showing $f_X \in \mathrm{RV}_{-2}$ and $f_{Y|X}(y \mid t) \to 2y = h(y)$ (condition (ii.), Theorem 4.2.3).

The results of this section are easily extended to hold for the arbitrary decay T defined in (4.1). We write $(X,\boldsymbol{Y}) \in T\text{-}\mathrm{RV}^x_{-\alpha}(H)$ if $\{T(X),\boldsymbol{Y}\} \in \mathrm{RV}^x_{-\alpha}(H)$. From the continuous mapping theorem, this is equivalent to $\mu_{t\star} \overset{w}{\Rightarrow} \nu$, where

$$\mu_{\lambda\star}(A,B) := \frac{\mu(\lambda\star A, B)}{\mu_X(\lambda\star E)},$$

is a probability distribution on $E\times\mathbb{R}^d$ for every $\lambda \in E$. In this case, $\nu = P_\alpha \circ T \times H$, and in particular ν_X has survival function $T(x)^{-\alpha}$. We say that μ is homogeneous w.r.t. its first component if $\mu(\lambda\star A,B) = T(\lambda)^{-\alpha}\mu(A,B)$, $\forall\lambda\in E$. Theorem 4.2.1 characterizes such distributions.

Corollary 4.2.5 (One-Component Homogeneous Probability Distributions)**.** *The following are equivalent:*

i. μ *is homogeneous of order* $-\alpha$ *w.r.t. its first component,*

ii. $\bar{F}$ *is homogeneous of order* $-\alpha$ *w.r.t. its first component,*

iii. $\mu_{\lambda\star} = \mu$, $\forall \lambda \in E$,

iv. $X \perp\!\!\!\perp \boldsymbol{Y}$ *and* $X \sim P_\alpha \circ T$.

If $(X, \boldsymbol{Y})$ *admits a probability density* $f_{X,\boldsymbol{Y}}$*, an additional equivalent statement is:* $f_{X,\boldsymbol{Y}}$ *is homogeneous of order* $-\alpha - 1$ *w.r.t. its first component.*

Proof. (iii.) $\Rightarrow$ (iv.): use the equivalence between (i.) and (iv.) in Theorem 4.2.1. The other implications are clear. □

In the next section, we focus on one-component regular variation of $(X, \boldsymbol{Y})$ when X is the radius and $\boldsymbol{Y}$ the angle of a vector. Note the change in the notation.

4.3 Relation to Multivariate Regular Variation

Let $\boldsymbol{X}$ be a random vector and $(R, \boldsymbol{\Theta}) = \phi(\boldsymbol{X}) = \{r(\boldsymbol{X}), \boldsymbol{\theta}(\boldsymbol{X})\}$ its expression in polar coordinates with $T \equiv \mathrm{id}$ and suppose that $\bar{F}_R > 0$. Analogously to functions, we show that $(R, \boldsymbol{\Theta})$ is one-component regularly varying if and only if $\boldsymbol{X}$ is regularly varying in the sense

$$t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t \xrightarrow{d} \boldsymbol{Y}, \tag{4.11}$$

on $C_{||\cdot||_\infty} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_\infty \geq 1\}$ for some random vector $\boldsymbol{Y}$ such that $||\boldsymbol{Y}||_\infty \not\sim \delta_1$. This relation is well-known when ϕ is the change of variables into pseudo polar coordinates [Basrak et al., 2002].

Let $\alpha > 0$ and $\boldsymbol{Y} \sim \nu$, a probability distribution on $C_r = \{r(\boldsymbol{x}) \geq 1\}$, such that $r(\boldsymbol{Y}) \not\sim \delta_1$.

Proposition 4.3.1. *The following are equivalent:*

i. $t^{-1}\boldsymbol{X} \mid \boldsymbol{X} \in tC_r \overset{w}{\Rightarrow} \nu$ *on* C_r,

ii. $(R, \boldsymbol{\Theta}) \in \mathrm{RV}_{-\alpha}(H)$ *on* $[1, \infty) \times \Omega$.

In this case, ν is homogeneous of order $-\alpha$, i.e., $\nu(\lambda A) = \lambda^{-\alpha}\nu(A)$, $\forall\lambda \geq 1$, and $\nu\{\phi^{-1}(A_r \times A_{\boldsymbol{\theta}})\} = P_\alpha(A_r)H(A_{\boldsymbol{\theta}})$, for all Borel sets A, A_r, $A_{\boldsymbol{\theta}}$.

Proof. We only prove the direct implication; the reverse goes similarly. Since ϕ^{-1} is homogeneous of order 1 w.r.t. its first component, write $t^{-1}R, \boldsymbol{\Theta} \mid R \geq t = \phi(t^{-1}\boldsymbol{X}) \mid R \geq t$, which converges weakly to a probability distribution by the continuous mapping theorem. Hence, $(R, \boldsymbol{\Theta}) \in \mathrm{RV}_{-\alpha}(H)$. To show the homogeneity of ν, let $\boldsymbol{Y} = \phi^{-1}(R^\star, \boldsymbol{\Theta}^\star) \sim \nu$ and use Corollary 4.2.5:

$$\begin{aligned}\nu(\lambda A) &= \Pr\{\phi^{-1}(\lambda^{-1}R^\star, \boldsymbol{\Theta}^\star) \in A \mid R^\star \geq \lambda\}\Pr(R^\star \geq \lambda) \\ &= \lambda^{-\alpha}\Pr\{\phi^{-1}(R^\star, \boldsymbol{\Theta}^\star) \in A\} = \lambda^{-\alpha}\nu(A), \quad \forall\lambda \geq 1.\end{aligned}$$

□

We denote the multivariate regular variation in (i.) by $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\nu)$ or $\mathrm{RV}_{-\alpha}(\boldsymbol{Y})$. It is known and easy to show that regular variation w.r.t. another radial function $\tilde{r}(\cdot)$ is equivalent if there exists $c_1, c_2 > 0$ such that $c_1 r(\boldsymbol{x}) \leq \tilde{r}(\boldsymbol{x}) \leq c_2 r(\boldsymbol{x})$, $\forall\boldsymbol{x}$, and in this case, the limits are related as follows: $\nu_r(A) = \nu_{\tilde{r}}(c_2 A)/\nu_{\tilde{r}}(c_2 C_r)$.

Suppose that $\boldsymbol{X}$ admits a probability density $f_{\boldsymbol{X}}$ w.r.t. the Lebesgue measure such that $f_{\boldsymbol{X}}(\lambda\mathbf{1}) > 0$, $\forall\lambda \geq 1$ and let $f_{\boldsymbol{Y}}$ be a probability on C_r satisfying the same constraint. Similarly to what we did in the previous section, we now study the relation between regular variation of $\boldsymbol{X}$ and regular variation of $f_{\boldsymbol{X}}$ in the sense (4.6).

If $f_{\boldsymbol{X}} \in \mathrm{RV}(v)$ for some non-null function v and the sequence is dominated by an integrable function, then

$$f_{t^{-1}\boldsymbol{X}\mid\boldsymbol{X}\in tC_r}(\boldsymbol{x}) = \frac{f_{\boldsymbol{X}}(t\boldsymbol{x})}{f_{\boldsymbol{X}}(t\mathbf{1})}\left(\int_{C_r}\frac{f_{\boldsymbol{X}}(t\tilde{\boldsymbol{x}})}{f_{\boldsymbol{X}}(t\mathbf{1})}d\tilde{\boldsymbol{x}}\right)^{-1} \to \frac{v(\boldsymbol{x})}{\int_{C_r} v(\tilde{\boldsymbol{x}})d\tilde{\boldsymbol{x}}} =: f_{\boldsymbol{Y}}(\boldsymbol{x}), \qquad (4.12)$$

thus the limiting probability density $f_{\boldsymbol{Y}}$ is homogeneous and $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y})$ on C_r. As a comparison, Proposition 5.20 in Resnick [1987] has the same conclusion, however, its assumptions require the convergence to be uniform and the limiting density to be bounded on $\{\boldsymbol{x} \in \mathbb{R}^d : r(\boldsymbol{x}) = 1\}$.

The following result gives conditions that are equivalent to the convergence in (4.12) and it reveals the relation between regular variation of $f_{\boldsymbol{X}}$ and one-component

regular variation of $f_{R,\boldsymbol{\Theta}}(r,\boldsymbol{\theta}) = f_{\boldsymbol{X}}\{r\boldsymbol{\theta}^{-1}(\boldsymbol{\theta})\}J_{\phi^{-1}}(r,\boldsymbol{\theta})$, where $J_\phi = 1/J_{\phi^{-1}}$ is the Jacobian determinant of ϕ.

We assume that $r(\cdot)$ and $\boldsymbol{\theta}(\cdot)$ are differentiable so that their partial derivative w.r.t. x_i for all i are homogeneous of order 0 and -1 respectively, and consequently J_ϕ is homogeneous of order $1-d$. Consider a probability densities h on Ω with distribution H satisfying $h(\lambda\mathbf{1}_{\boldsymbol{\theta}}) > 0$, $\forall\lambda \geq 1$, for $\mathbf{1}_{\boldsymbol{\theta}} := \boldsymbol{\theta}(\mathbf{1})$. Suppose also that $f_R > 0$ and, say, $r(\mathbf{1}) = 1$.

Proposition 4.3.2. *The following are equivalent:*

i. $f_{\boldsymbol{X}} \in \mathrm{RV}_{-\alpha-d}\{f_{\boldsymbol{Y}}(\cdot)/f_{\boldsymbol{Y}}(\mathbf{1})\}$ *and* $t^{d-1}f_{\boldsymbol{X}|R}(t\mathbf{1} \mid t) \to \alpha^{-1}f_{\boldsymbol{Y}}(\mathbf{1})$,

ii. $f_{R,\boldsymbol{\Theta}} \in \mathrm{RV}^x_{-\alpha-1}\{h(\cdot)/h(\mathbf{1}_{\boldsymbol{\theta}})\}$ *and* $f_{\boldsymbol{\Theta}|R}(\mathbf{1}_{\boldsymbol{\theta}} \mid t) \to h(\mathbf{1}_{\boldsymbol{\theta}})$,

iii. $f_{t^{-1}\boldsymbol{X}|\boldsymbol{X}\in tC_r} \to f_{\boldsymbol{Y}}$.

In this case, $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \alpha r(\boldsymbol{y})^{-\alpha-1}h\{\boldsymbol{\theta}(\boldsymbol{y})\}J_\phi(\boldsymbol{y})$, $(R,\boldsymbol{\Theta}) \in \mathrm{RV}_{-\alpha}(H)$ *and* $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y})$.

Proof. (i.) $\Leftrightarrow$ (ii.): we start by showing that the first conditions in (i.) and (ii.) are equivalent. For the direct implication,

$$\frac{f_{R,\boldsymbol{\Theta}}(tr,\boldsymbol{\theta})}{f_{R,\boldsymbol{\Theta}}(t,\mathbf{1}_{\boldsymbol{\theta}})} = \frac{f_{\boldsymbol{X}}\{tr\boldsymbol{\theta}^{-1}(\boldsymbol{\theta})\}}{f_{\boldsymbol{X}}(t\mathbf{1})}\frac{J_{\phi^{-1}}(tr,\boldsymbol{\theta})}{J_{\phi^{-1}}(t,\mathbf{1}_{\boldsymbol{\theta}})} \to \frac{f_{\boldsymbol{Y}}\{r\boldsymbol{\theta}^{-1}(\boldsymbol{\theta})\}}{f_{\boldsymbol{Y}}(\mathbf{1})}\frac{J_{\phi^{-1}}(r,\boldsymbol{\theta})}{J_{\phi^{-1}}(1,\mathbf{1}_{\boldsymbol{\theta}})} =: v.$$

The limit $v(r,\boldsymbol{\theta})$ is integrable on $[1,\infty)\times\Omega$ since $f_{\boldsymbol{Y}}$ is a probability density. Moreover, it satisfies $v(\lambda,\mathbf{1}_{\boldsymbol{\theta}}) = f_{\boldsymbol{Y}}(\lambda\mathbf{1})/f_{\boldsymbol{Y}}(\mathbf{1})\,\lambda^{d-1} > 0$, $\forall\lambda \geq 1$. Together with Lemma 4.1.1, $v(r,\boldsymbol{\theta}) = r^{-\alpha-1}h(\boldsymbol{\theta})/h(\mathbf{1}_{\boldsymbol{\theta}})$ for $\alpha > 0$ and a probability density h. The reverse implication follows from

$$\frac{f_{\boldsymbol{X}}(t\boldsymbol{x})}{f_{\boldsymbol{X}}(t\mathbf{1})} = \frac{f_{R,\boldsymbol{\Theta}}\{tr(\boldsymbol{x}),\boldsymbol{\theta}(\boldsymbol{x})\}}{f_{R,\boldsymbol{\Theta}}(t,\mathbf{1}_{\boldsymbol{\theta}})}\frac{J_\phi(\boldsymbol{x})}{J_\phi(\mathbf{1})} \to r(\boldsymbol{x})^{-\alpha-1}\frac{h\{\boldsymbol{\theta}(\boldsymbol{x})\}}{h(\mathbf{1}_{\boldsymbol{\theta}})}\frac{J_\phi(\boldsymbol{x})}{J_\phi(\mathbf{1})},$$

and thus, $f_{\boldsymbol{Y}}(\boldsymbol{y}) := \alpha r(\boldsymbol{y})^{-\alpha-1}h\{\boldsymbol{\theta}(\boldsymbol{y})\}J_\phi(\boldsymbol{y})$ satisfies $f_{\boldsymbol{Y}}(\mathbf{1}) > 0$ and is a probability density on C_r. Besides that, $t^{d-1}f_{\boldsymbol{X}}(t\mathbf{1})/f_R(t1) \to \alpha^{-1}f_{\boldsymbol{Y}}(\mathbf{1})$ is clearly equivalent to $f_{\boldsymbol{\Theta}|R}(\mathbf{1}_{\boldsymbol{\theta}} \mid t) \to h(\mathbf{1}_{\boldsymbol{\theta}})$. (ii.) $\Leftrightarrow$ (iii.): apply Theorem 4.2.3. □

It is straightforward to extend Proposition 4.3.1 and 4.3.2 to hold for general decay T. We confine ourselves to stating a characterization of homogeneous probability distributions, that follows directly. We write $\boldsymbol{X} \in \text{T-RV}^x_{-\alpha}(H)$ if $\{T(X_1), \ldots, T(X_d)\} \in \text{RV}^x_{-\alpha}(\nu)$. By the continuous mapping theorem, this is equivalent to $\mu_{t\star} \overset{w}{\Rightarrow} \nu$, where

$$\mu_{\lambda\star}(A) := \mu(\lambda \star A)/\mu(\lambda \star C_r),$$

is a probability distribution on C_r for every $\lambda \in E$. In this case, ν is homogeneous of order $-\alpha$, i.e., $\mu(\lambda \star A) = T(\lambda)^{-\alpha}\mu(A)$, $\forall \lambda \in E$.

Corollary 4.3.3 (Homogeneous Probability Distributions)**.** *The following are equivalent:*

i. $\mu_{\boldsymbol{X}}$ *is homogeneous of order* $-\alpha$,

ii. $\mu_{R,\boldsymbol{\Theta}}$ *is homogeneous of order* $-\alpha$ *w.r.t. its first component,*

iii. $\mu_{\star\lambda} = \mu$, $\forall \lambda \in E$.

If $\boldsymbol{X}$ *admits a probability density* $f_{\boldsymbol{X}}$*, the following are also equivalent to the list above.*

(a) $f_{\boldsymbol{X}}$ *is homogeneous of order* $-\alpha - d$,

(b) $f_{R,\boldsymbol{\Theta}}$ *is homogeneous of order* $-\alpha - 1$ *w.r.t. its first component,*

(c) $f_{\boldsymbol{X}}(\boldsymbol{x}) = \alpha r(\boldsymbol{x})^{-\alpha} h\{\boldsymbol{\theta}(\boldsymbol{x})\} J_\phi(\boldsymbol{x})$.

Regular variation of $\boldsymbol{X}$ is better expressed in terms of its *survival function* than its cumulative distribution function. We define the former as

$$\bar{F}_{\boldsymbol{X}}(\boldsymbol{x}) := \Pr(\boldsymbol{\sigma}_x \boldsymbol{X} \geq |\boldsymbol{x}|).$$

Recall that multiplication is done componentwise and $\boldsymbol{\sigma}_{\boldsymbol{x}} = (\sigma_{x_1}, \ldots, \sigma_{x_d})$, where σ_{x_i} takes the value 1, 0 or -1 if x_i is positive, null or negative respectively. Any function $\bar{F} : \mathbb{R}^d \to [0, 1]$ non-increasing and continuous in the directions away from

$\mathbf{0}$ and satisfying $\lim_{||\boldsymbol{x}||\to\infty} \bar{F}(\boldsymbol{x}) = 0$ and $\bar{F}(\mathbf{0}) = 1$ determines a probability measure on $\mathbb{R}^d$. If $\boldsymbol{X}$ has values in $C_{||\cdot||_\infty}$, the last constraint becomes

$$\sum_{\boldsymbol{\sigma}\in\{-1,0,1\}^d\setminus\{\mathbf{0}\}} \bar{F}(\boldsymbol{\sigma})(-1)^{|\boldsymbol{\sigma}\neq 0|+1} = 1, \tag{4.13}$$

where $|\boldsymbol{x} \neq 0|$ is the number of non-null components of $\boldsymbol{x}$. The marginals of $\boldsymbol{X}$ have survival function $\bar{F}_{\boldsymbol{X}_A}(\boldsymbol{x}_A) = \bar{F}_{\boldsymbol{X}}(\boldsymbol{x}_A, \mathbf{0}_{A^c})$ for all $A \subseteq \{1,\dots,d\}$. In addition, $\boldsymbol{X}_t \overset{w}{\Rightarrow} \boldsymbol{X}$ if and only if $\bar{F}_{\boldsymbol{X}_t}(\boldsymbol{x}) \to \bar{F}_{\boldsymbol{X}}(\boldsymbol{x})$ for every continuous point $\boldsymbol{x}$ of $\bar{F}_{\boldsymbol{X}}$, written $\bar{F}_{\boldsymbol{X}_t} \overset{w}{\Rightarrow} \bar{F}_{\boldsymbol{X}}$. Hence, $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y})$ on $C_{||\cdot||_\infty}$ is equivalent to $\bar{F}(t\boldsymbol{x})/\Pr(||\boldsymbol{X}||_\infty \geq t) \overset{w}{\Rightarrow} \bar{G}(\boldsymbol{x})$, where $\bar{G}$ is the survival function of $\boldsymbol{Y}$.

Suppose that

$$\Pr(bX_i \geq t \mid ||\boldsymbol{X}||_\infty \geq t) \to c_{bi} > 0, \quad b \in \{-1,1\}, \forall i = 1,\dots,d. \tag{4.14}$$

For simplicity, we also assume that the sets $\{\boldsymbol{\sigma}\boldsymbol{x} \geq \mathbf{1}\}$ for $\boldsymbol{\sigma} \in \{-1,0,1\}^d$ are continuous sets of any limiting distribution. Let ν_{bi} be some probability distribution on $C_{bi} := \{\boldsymbol{x} \in \mathbb{R}^d : bx_i \geq 1\}$ with survival function $\bar{G}_{bi}$, $\forall b, i$.

Lemma 4.3.4. *It holds*

$$\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y}) \text{ on } C_{||\cdot||_\infty} \quad \Longleftrightarrow \quad \boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\nu_{bi}) \text{ on } C_{bi},\ \forall b, i, \tag{4.15}$$

and in this case $\boldsymbol{Y} \mid bY_i \geq 1 \sim \nu_{bi}$.

Proof. We start with the direct implication. For any $\boldsymbol{x} \in \mathbb{R}^d$ such that $bx_i \geq 1$,

$$\frac{\bar{F}(t\boldsymbol{x})}{\bar{F}_{X_i}(bt)} = \frac{\bar{F}(t\boldsymbol{x})}{\Pr(||\boldsymbol{X}||_\infty \geq t)} \frac{\Pr(||\boldsymbol{X}||_\infty \geq t)}{\bar{F}_{X_i}(bt)} \overset{w}{\Rightarrow} c_{bi}^{-1}\bar{G}(\boldsymbol{x}),$$

hence $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y} \mid bY_i \geq 1)$. For the reverse, if $bx_i \geq 1$, then

$$\frac{\bar{F}(t\boldsymbol{x})}{\Pr(||\boldsymbol{X}||_\infty \geq t)} = \frac{\bar{F}(t\boldsymbol{x})}{\bar{F}_{X_i}(bt)} \frac{\bar{F}_{X_i}(bt)}{\Pr(||\boldsymbol{X}||_\infty \geq t)} \overset{w}{\Rightarrow} \bar{G}_{bi}(\boldsymbol{x})c_{bi}.$$

The limit is a valid survival function on $C_{||\cdot||_\infty}$ because (4.13) is satisfied by $\bar{F}(t\boldsymbol{\sigma})/\Pr(||\boldsymbol{X}||_\infty \geq t)$, and thus by its pointwise limit $\bar{G}_{bi}(\boldsymbol{\sigma})c_{bi}$. □

A consequence of Lemma 4.3.4 is that a distribution μ satisfying (4.14) is homogeneous of order $-\alpha$ if and only if its survival function $\bar{F}$ is homogeneous of the same order. In this case, $\bar{F}$ corresponds up to a constant to the Möbius inverse of $V(\boldsymbol{x}) = \tilde{\nu}([\mathbf{0}, \boldsymbol{x}]^c)$ on $C_{||\cdot||_\infty} \cap \mathbb{R}^d_+$, which is also homogeneous of order $-\alpha$, where $\tilde{\nu}$ is the exponent measure defined in (1.12).

Another interest of Lemma 4.3.4 is to render the use of Proposition 4.3.2 possible when $f_R = f_{||\boldsymbol{X}||}$ is not known but f_{X_i} is for all i. Indeed, $f_{t^{-1}\boldsymbol{X}|||\boldsymbol{X}||\geq t} \to f_{\boldsymbol{Y}}$ can be checked on C_{bi}, $\forall b, i$, requiring only $f_{\boldsymbol{X}}$ and f_{X_i}.

We now ask whether the marginals of a regularly varying vector are also regularly varying and, in this case, if the marginals of the limit are the limits of the marginals. The following Lemma provides a positive answer. Without assuming (4.14), however, both statements are incorrect in general. Let $A \subseteq \{1, \dots, d\}$, $C^A_{||\cdot||_\infty} := \{\boldsymbol{x}_A \in \mathbb{R}^{|A|} : ||\boldsymbol{x}_A||_\infty \geq 1\}$.

Lemma 4.3.5. *If $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y})$ on $C_{||\cdot||_\infty}$, then*

$$\boldsymbol{X}_A \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y}_A \mid ||\boldsymbol{Y}_A||_\infty \geq 1) \quad \text{on } C^A_{||\cdot||_\infty}. \tag{4.16}$$

If $f \in \mathrm{RV}_{-\alpha}(v)$ for some $v \not\equiv 0$ and the sequence is either monotone and v is integrable or is dominated by an integrable function, then

$$f_{t^{-1}\boldsymbol{X}_A|||\boldsymbol{X}_A||_\infty \geq t} \to f_{\boldsymbol{Y}_A|||\boldsymbol{Y}_A||_\infty \geq 1} \quad \text{on } C^A_{||\cdot||_\infty},$$

and thus (4.16) holds.

Proof. Under (4.14), $\Pr(||\boldsymbol{X}_A||_\infty \geq t \mid ||\boldsymbol{X}||_\infty \geq t) \to c > 0$. For all ν-continuous Borel set B,

$$\frac{\Pr(\boldsymbol{X}_A \in tB)}{\Pr(||\boldsymbol{X}_A||_\infty \geq t)} = \frac{\Pr(\boldsymbol{X}_A \in tB)}{\Pr(||\boldsymbol{X}||_\infty \geq t)} \frac{\Pr(||\boldsymbol{X}||_\infty \geq t)}{\Pr\{\boldsymbol{X} \in t(C^A_{||\cdot||_\infty} \times \mathbb{R}^{|A^C|})\}},$$

and converges to $\Pr(\boldsymbol{Y}_A \in B \mid ||\boldsymbol{Y}_A||_\infty \geq 1)$, proving the first part of the lemma. The second part is a direct consequence of the monotone or dominated convergence theorem. □

Lemma 4.3.5 is specific to the Pareto decay ($T \equiv \mathrm{id}$). For instance, $\bar{F}(x, y) = 2/(e^{|x|} + e^{|y|})$ is homogeneous on $\mathbb{R}^2$ w.r.t. addition; its marginal $\bar{F}_X(x) = 2/(e^{|x|} + 1)$, although regularly varying, is not homogeneous.

5

Graphical Modeling of Extremes

In this chapter, we characterize homogeneous probability densities that factorize w.r.t. a graph. We present the particular case of a tree graphical model and define the Hüsler–Reiss graphical model. This allows us to construct valid high-dimensional limiting tail distributions from low-dimensional homogeneous functions. Eventually, we define asymptotic conditional independence and explain how to use this notion to infer tail distributions.

5.1 Factorization of the Limiting Tail Density

Extreme value theory aims at describing the distribution of an $\mathbb{R}^d$-random vector $\boldsymbol{X}$ when at least one of its marginals is large. Typically, it is assumed that $\boldsymbol{X}$ is regularly varying on $C_{||\cdot||_\infty} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_\infty \geq 1\}$, i.e.,

$$t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t \xrightarrow{d} \boldsymbol{Y},$$

and that $\boldsymbol{Y}$ admits a probability density $f_{\boldsymbol{Y}}$, known to be homogeneous. This excludes the case of asymptotically independent marginals [Resnick, 1987] which we do not treat here. In this section, we start by illustrating a procedure to obtain parametric homogeneous densities $f_{\boldsymbol{Y}}$, before explaining how to use graphical models to simplify $f_{\boldsymbol{Y}}$ in high dimensions.

Example 5.1.1. Let $\boldsymbol{X}$ be Student distributed with degree of freedom $\nu > 2$, mean $\mathbf{0}$ and covariance matrix $\frac{\nu}{\nu-2}\Sigma \subset \mathbb{R}^d \times \mathbb{R}^d$. We find

$$\frac{f_{\boldsymbol{X}}(t\boldsymbol{y})}{f_{\boldsymbol{X}}(t\mathbf{1})} \to \left(\frac{\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\right)^{-(\nu+d)/2}$$

and since the sequence is monotone and the limit integrates to $c < \infty$ on $C_{||\cdot||_\infty}$,

$$f_{t^{-1}\boldsymbol{X}\,|\,||\boldsymbol{X}||_\infty \geq t}(\boldsymbol{y}) \to f_{\boldsymbol{Y}}(\boldsymbol{y}) = c^{-1}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})^{-(\nu+d)/2},$$

as explained in (4.12), and thus $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y})$. Moreover, the censored versions and marginals of $f_{\boldsymbol{Y}}$ can be computed from the ones of the multivariate Student by passing to the limit as follows. The censored limiting density is the limit of the censored density, that is $f_{t^{-1}\boldsymbol{X}_A,|\boldsymbol{X}_A|<t\,|\,||\boldsymbol{X}||_\infty \geq t} \to f_{\boldsymbol{Y}_A,|\boldsymbol{Y}_A|<1}$, and Lemma 4.3.5 ensures that the marginals of the limit are the limits of the marginals, i.e., $f_{t^{-1}\boldsymbol{X}_A\,|\,||\boldsymbol{X}_A||_\infty \geq t} \to f_{\boldsymbol{Y}_A}$.

Up to a constant and a transformation of the univariate marginals, $f_{\boldsymbol{Y}}$ coincides with the density of the exponent measure of the extremal t distribution derived by Ribatet [2013]. We like to think of $f_{\boldsymbol{Y}}$ as a generalization of the Pareto distribution in the multivariate case.

We now focus on a complementary approach for modeling $f_{\boldsymbol{Y}}$ using graphical models. In short, graphical models offer a way to simplify a joint density by assuming conditional independence between some of the marginals. Independence and conditional independence between marginals is only meaningful on a product space [Dawid, 2001]. Since $\boldsymbol{Y}$ has values in $C_{||\cdot||_\infty}$, the range of Y_i depends on the other marginals. To remedy this problem, we work with $\boldsymbol{Y} \mid Y_k \geq 1$ which takes values in the product space $C_k = \{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\}$, for $k = 1, \ldots, d$.

Theorem 5.1.2. *Let $\mathcal{G} = (\{1, \dots, d\}, E)$ be a decomposable graph. The following are equivalent:*

i. $f_{\boldsymbol{Y}}$ is a homogeneous positive probability density on $C_{||\cdot||_\infty}$ and

$$Y_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Y}_{\{1,\dots,d\}\setminus\{i,j\}},\ |Y_k| \geq 1, \quad \forall (i,j) \notin E,\ \forall k = 1, \dots, d, \tag{5.1}$$

ii. $\mathcal{G}$ is connected and there exists a class $\{h_S\}_{S \in \mathcal{C} \cup \mathcal{D}}$ of homogeneous, positive and measurable functions on $\mathbb{R}^{|S|}$, integrable on $C^S_{||\cdot||_\infty}$, such that

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = c^{-1} \frac{\prod_{C \in \mathcal{C}} h_C(\boldsymbol{y}_C)}{\prod_{D \in \mathcal{D}} h_D(\boldsymbol{y}_D)}, \quad \boldsymbol{y} \in C_{||\cdot||_\infty}, \tag{5.2}$$

and for which the consistency constraint $h_D = \int_{\mathbb{R}^{|C \setminus D|}} h_C \, d\boldsymbol{y}_{C \setminus D}$ for all $D \subset C$, $D \in \mathcal{D}$, $C \in \mathcal{C}$ holds, where $\mathcal{C}$ is the set of maximal cliques of the graph, $\mathcal{D}$ is the multiset containing the separator sets and $c > 0$ is a normalizing constant.

In this case, h_S is proportional to the probability density of $\boldsymbol{Y}_S \mid ||\boldsymbol{Y}_S||_\infty \geq 1$ on $C^S_{||\cdot||_\infty}$ for all $S \in \mathcal{C} \cup \mathcal{D}$.

Proof. We start by proving that (i.) implies (ii.). By assumption, $f_{\boldsymbol{Y}}$ is homogeneous of order $-\alpha - d$ for some $\alpha > 0$. Consider its extension a.e. to $\mathbb{R}^d$ by homogeneity: for any $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, $h(\boldsymbol{y}) := ||\boldsymbol{y}||_\infty^{-\alpha-d} f_{\boldsymbol{Y}}(\boldsymbol{y}/||\boldsymbol{y}||_\infty)$. We clarify that h is integrable on $C_{||\cdot||_\infty}$, but not on $\mathbb{R}^d$. Now, define a candidate for the class $\{h_S\}_{S \in \mathcal{C} \cup \mathcal{D}}$ as

$$h_S(\boldsymbol{y}_S) := \int_{\mathbb{R}^{d-|S|}} h(\boldsymbol{y}_S, \boldsymbol{y}_{-S}) d\boldsymbol{y}_{-S}, \quad \boldsymbol{y}_S \in \mathbb{R}^{|S|},$$

using the notation $-S := \{i = 1, \dots, d : i \notin S\}$. Clearly, all h_S are positive, measurable, integrable on $C^S_{||\cdot||_\infty}$, satisfy the consistency constraints and it is easy to check that they are also homogeneous.

By assumption, $f_{\boldsymbol{Y}} > 0$ and $\boldsymbol{Y} \mid |Y_k| \geq 1$ satisfies the pairwise Markov property according to $\mathcal{G}$ for all $k = 1, \dots, d$ hence Hammersley–Clifford theorem yields

$$h \propto f_{\boldsymbol{Y} \,|\, |Y_k| \geq 1} = \frac{\prod_{C \in \mathcal{C}} f^k_C}{\prod_{D \in \mathcal{D}} f^k_D} \quad \text{on} \quad \{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\},$$

("$\propto$" denotes proportionality) where f^k_S is the density of the marginal $(\boldsymbol{Y} \mid |Y_k| \geq 1)_S$. Notice that f^k_S is not necessarily homogeneous because $f^k_S = \int f_{\boldsymbol{Y} \,|\, |Y_k| \geq 1} d\boldsymbol{y}_{-S}$

and the vector $\boldsymbol{y}_{-S}$ is integrated out under the constraint $|y_k| \geq 1$ when $k \notin S$. In the case $k \in S$, however, there is no such constraint and f_S^k is homogeneous and proportional to h_S.

Consider the sequence of cliques $C_1, \ldots, C_m$ and of separator sets $D_1, \ldots, D_m$ defined in (A.4). If $k \in D_1$, then

$$f_{\boldsymbol{Y}} \propto \frac{f_{C_1}^k f_{C_2 \cup \ldots \cup C_m}^k}{f_{D_1}^k} \propto \frac{h_{C_1} h_{C_2 \cup \ldots \cup C_m}}{h_{D_1}} \quad \text{on} \quad \{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\},$$

since D_1 is a subset of C_1 and of $C_1 \cup \ldots \cup C_m$ because the graph is necessarily connected. The relation above can be extended a.e. to $\mathbb{R}^d$ by homogeneity: for any $\boldsymbol{y} \in \mathbb{R}^d$ such that $y_k \neq 0$,

$$\begin{aligned} h &\propto y_k^{-\alpha-d} f_{\boldsymbol{Y}}(\boldsymbol{y}/y_k) \\ &\propto y_k^{-\alpha-d} \, \frac{h_{C_1}(\boldsymbol{y}_{C_1}/y_k) h_{C_2 \cup \ldots \cup C_m}(\boldsymbol{y}_{C_2 \cup \ldots \cup C_m}/y_k)}{h_{D_1}(\boldsymbol{y}_{D_1}/y_k)} = \frac{h_{C_1}(\boldsymbol{y}_{C_1}) h_{C_2 \cup \ldots \cup C_m}(\boldsymbol{y}_{C_2 \cup \ldots \cup C_m})}{h_{D_1}(\boldsymbol{y}_{D_1})}, \end{aligned}$$

for some $\alpha > 0$. It remains to decompose $h_{C_2 \cup \ldots \cup C_m}$ in the same manner. Up to a constant, $h_{C_2 \cup \ldots \cup C_m}$ is a homogeneous and positive probability density on $C_{||\cdot||_\infty}^{d-|C_1 \setminus D_1|}$ which factorizes on $\{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\}$ for all $k \in C_2 \cup \ldots \cup C_m$ because

$$h_{C_2 \cup \ldots} = \int h \, d\boldsymbol{y}_{C_1 \setminus D_1} \propto \int f_{\boldsymbol{Y} \,|\, |Y_k| \geq 1} d\boldsymbol{y}_{C_1 \setminus D_1} = \int \frac{\prod_{C \in \mathcal{C}} f_C^k}{\prod_{D \in \mathcal{D}} f_D^k} d\boldsymbol{y}_{C_1 \setminus D_1} = \frac{\prod_{C \neq C_1} f_C^k}{\prod_{D \neq D_1} f_D^k},$$

where the integrals above are taken over $\mathbb{R}^{|C_1 \setminus D_1|}$. Thus, $h_{C_2 \cup \ldots \cup C_m}$ can be decomposed by applying the same reasoning as for h, and so on for $h_{C_3 \cup \ldots \cup C_m}, \ldots, h_{C_m}$, giving (5.2).

For the reverse implication, it is straightforward to see that $f_{\boldsymbol{Y}}$ is a well-defined probability density on $C_{||\cdot||_\infty}$ by Tonelli's theorem and sequential integration on $C_1 \setminus D_1, \ldots, C_m \setminus D_m$. In addition, the density of $\boldsymbol{Y} \mid |Y_k| \geq 1$ clearly factorizes on $\{\boldsymbol{x} \in \mathbb{R}^d : |x_k| \geq 1\}$ for all k, thus Hammersley–Clifford theorem implies that $\boldsymbol{Y}$ satisfies the Markov property according to $\mathcal{G}$. □

Theorem 5.1.2 is an adaptation of Hammersley–Clifford theorem for homogeneous distributions. Further work could attempt to extend it for graphs that are not decomposable. If $f_{\boldsymbol{Y}}$ satisfies (5.2), we say that it *factorizes* w.r.t. the graph $\mathcal{G}$; the next example characterizes factorizations w.r.t. trees. We also mention that

there exists no positive homogeneous probability density $f_{\boldsymbol{Y}}$ factorizing w.r.t. a disconnected graph; in particular, the case $Y_1 \perp\!\!\!\perp Y_2$ is excluded.

Example 5.1.3 (Homogeneous Density Factorizing w.r.t. a Tree)**.** Let $\alpha > 0$ and $\mathcal{G} = (\{1, \dots, d\}, E)$ be a tree. The function $f_{\boldsymbol{Y}}$ is a homogeneous probability density on $C_{||\cdot||_\infty}$ of order $-\alpha - d$ and satisfies the Markov property according to the tree $\mathcal{G}$ if and only if $\mathcal{G}$ is connected and

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = c^{-1} \prod_{(i,j)\in E} \frac{h_{ij}(\boldsymbol{y}_{ij})}{y_i^{-\alpha-1} y_j^{-\alpha-1}} \prod_{i=1}^{d} y_i^{-\alpha-1}, \tag{5.3}$$

for a normalizing constant $c > 0$ and a class of homogeneous, positive and measurable densities $\{h_{ij}\}_{(i,j)\in E}$ satisfying $\int_{\mathbb{R}} h_{ij}(y, z)dz = \int_{\mathbb{R}} h_{ij}(z, y)dz = y^{-\alpha-1}$. Take for instance $h_{ij}(y, z) = \frac{1}{2}(|y| + |z|)^{-3}$. A homogeneity density of the form (5.3) defined through its angular component appears in Coles and Tawn [1991] in the case $E = \{(1, 2), (2, 3), \dots, (d - 1, d)\}$.

A flexible extreme value distribution parametrized by a covariance matrix is the Hüsler–Reiss distribution [Hüsler and Reiss, 1989], whose angular density has been derived in Engelke et al. [2015]. We show that conditional independence between two marginals corresponds to a specific constraint on the covariance matrix, analogously to Gaussian conditional independence.

Example 5.1.4 (Hüsler–Reiss and Conditional Independence)**.** The Hüsler–Reiss exponent measure has homogeneous probability density

$$f_{\boldsymbol{Y}|Y_k\geq 1}(\boldsymbol{y}) = y_k^{-2} \phi(\boldsymbol{y}_{-k} \mid y_k), \tag{5.4}$$

on $C_k = \{\boldsymbol{y} \geq \boldsymbol{0} : y_k \geq 1\}$, $\forall k = 1, \dots, d$, where ϕ denotes the probability density of the multivariate log-normal distribution on $\mathbb{R}^{d-1}$ with mean $(\log y_k - \frac{1}{2}\Gamma_{jk})_{j\neq k}$ and covariance matrix Σ_ϑ^{-1}, where

$$\Sigma_\vartheta = \frac{1}{2}\{\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}\}_{i,j\neq k},$$

and $\Gamma_{ij} = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{i,j}$ is the incremental variance defined by a correlation matrix $\Sigma \subset \mathbb{R}^{d\times d}$. The definition of Σ_ϑ ensures that $f_{\boldsymbol{Y}}$ is a well-defined probability density on $C_{||\cdot||_\infty}$.

The marginals $\boldsymbol{Y}_A \mid ||\boldsymbol{Y}_A||_\infty \geq 1$ and the censored densities can be computed using the properties of the log-normal distribution. It is known that two univariate marginals i and j of the log-normal distribution are conditionally independent given the rest exactly when the inverse covariance matrix has a 0 in the entry (i, j). This means here that

$$Y_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Y}_{\{1,\dots,d\}\setminus\{i,j\}},\, Y_k \geq 1 \quad \iff \quad \Sigma_{\vartheta,ij} = \Gamma_{ij} - \Gamma_{ik} - \Gamma_{jk} = 0, \tag{5.5}$$

for all $k = 1, \dots, d$.

We have seen that conditional independence between some marginals of $\boldsymbol{Y}$ enforces a factorization of $f_{\boldsymbol{Y}}$. We are now interested in finding assumptions on $\boldsymbol{X}$ under which its limiting tail density $f_{\boldsymbol{Y}}$ factorizes. This brings us to translate conditional independence into *asymptotic conditional independence* and generalize the Hammersley–Clifford theorem for densities that factorize in the limit. Although asymptotic independence has been widely discussed [Beirlant et al., 2004], there is to our knowledge little mention of asymptotic conditional independence.

We end by mentioning the success of the Gaussian graphical lasso [Friedman et al., 2008], introduced in Chapter 6, which has been able to tackle high-dimensional problems by imposing sparsity of the inverse covariance matrix. Future work could allow a similar approach for the Hüsler–Reiss distribution.

5.2 Asymptotic Graphical Models

Let $\boldsymbol{X}_t$ for $t \geq 1$ and $\boldsymbol{X}$ be $\mathbb{R}^d$-valued random vectors with almost everywhere (a.e.) continuous probability densities $f_{\boldsymbol{X}_t}$ and $f_{\boldsymbol{X}}$ w.r.t. a base measure μ_0, typically the Lebesgue, the counting measure, or a combination of the two. For disjoint subsets $A, B, C \subseteq \{1, \dots, d\}$, we call the marginals $\boldsymbol{X}_{t,A}$ and $\boldsymbol{X}_{t,B}$ *asymptotically conditionally independent* with respect to $\boldsymbol{X}_{C,t}$, written $\boldsymbol{X}_{t,A} \,\tilde{\perp\!\!\!\perp}\, \boldsymbol{X}_{t,B} \mid \boldsymbol{X}_{t,C}$, if

$$(f_{\boldsymbol{X}_{t,ABC}}\, f_{\boldsymbol{X}_{t,C}} - f_{\boldsymbol{X}_{t,AC}}\, f_{\boldsymbol{X}_{t,BC}}) d\mu_0 \overset{w}{\Rightarrow} 0. \tag{5.6}$$

Here, $f_{\boldsymbol{X}_t} d\mu_0 \overset{w}{\Rightarrow}$ stands for weak convergence of $\boldsymbol{X}_t$ defined as the convergence of $\int g f_{\boldsymbol{X}_t} d\mu_0$ for every g a.e. continuous and bounded. When $\boldsymbol{X}_t = \boldsymbol{X}_1$, $\forall t$, (5.6) coincides with conditional independence of random variables (see Lauritzen [1996]).

Let $\mathcal{G} = (V, E)$ be an undirected graph with set of nodes $V = \{1, \dots, d\}$ and set of edges E. We say that $\boldsymbol{X}_t$ satisfies the *asymptotic pairwise Markov property* according to $\mathcal{G}$ if $\{i, j\} \notin E \Leftrightarrow X_{t,i} \tilde{\perp\!\!\!\perp} X_{t,j} \mid \boldsymbol{X}_{t, V\setminus\{i,j\}}$. Moreover, we say that $f_{\boldsymbol{X}_t}$ *asymptotically factorizes* w.r.t. $\mathcal{G}$ if $\boldsymbol{X}_t \overset{w}{\Rightarrow} \boldsymbol{X}$ and $f_{\boldsymbol{X}}$ factorizes w.r.t. $\mathcal{G}$.

Proposition 5.2.1 (Asymptotic Hammersley-Clifford Theorem)**.** *Suppose that $\boldsymbol{X}_t \overset{w}{\Rightarrow} \boldsymbol{X}$ and that $f_{\boldsymbol{X}}$ is a.e. positive and bounded. Then the error $\epsilon_t := f_{\boldsymbol{X}_t} - f_{\boldsymbol{X}}$ satisfies*

$$(\epsilon_{t,V}\epsilon_{t,V\setminus\{i,j\}} - \epsilon_{t,V\setminus\{i\}}\,\epsilon_{t,V\setminus\{j\}}) d\mu_0 \overset{w}{\Rightarrow} 0, \quad \forall (i,j) \in E, \tag{5.7}$$

if and only if the following are equivalent:

i. $\boldsymbol{X}$ satisfies the pairwise Markov property according to $\mathcal{G}$,

ii. $f_{\boldsymbol{X}}$ factorizes w.r.t. to $\mathcal{G}$,

iii. $f_{\boldsymbol{X}_t}$ asymptotically factorizes w.r.t. $\mathcal{G}$ as $t \to \infty$,

iv. $\boldsymbol{X}_t$ satisfies the asymptotic pairwise Markov property according to $\mathcal{G}$ as $t \to \infty$.

Proof. (i.) $\Leftrightarrow$ (ii.): Hammersley–Clifford theorem. (ii.) $\Leftrightarrow$ (iii.): by definition. (iv.) $\Leftrightarrow$ (i.): it suffices to show relations of the form

$$f_{XY\boldsymbol{Z}} f_{\boldsymbol{Z}} = f_{X\boldsymbol{Z}} f_{Y\boldsymbol{Z}} \text{ a.e. } \iff (f_{X_tY_t\boldsymbol{Z}_t} f_{\boldsymbol{Z}_t} - f_{X_t\boldsymbol{Z}_t} f_{Y_t\boldsymbol{Z}_t}) d\mu_0 \overset{w}{\Rightarrow} 0. \tag{5.8}$$

Rewrite the right-hand side as

$$\begin{aligned} &(f_{XY\boldsymbol{Z}} f_{\boldsymbol{Z}} - f_{X\boldsymbol{Z}} f_{Y\boldsymbol{Z}}) d\mu_0 + (\epsilon_{X_t,Y_t,\boldsymbol{Z}_t}\epsilon_{\boldsymbol{Z}_t} - \epsilon_{X_t,\boldsymbol{Z}_t}\epsilon_{Y_t,\boldsymbol{Z}_t}) d\mu_0 \\ &+(\epsilon_{X_t,Y_t,\boldsymbol{Z}_t} f_{\boldsymbol{Z}} + \epsilon_{\boldsymbol{Z}_t} f_{X,Y,\boldsymbol{Z}} - \epsilon_{X_t,\boldsymbol{Z}_t} f_{Y,\boldsymbol{Z}} - \epsilon_{Y_t,\boldsymbol{Z}_t} f_{\boldsymbol{X},\boldsymbol{Z}}) d\mu_0 \overset{w}{\Rightarrow} 0. \end{aligned} \tag{5.9}$$

The error ϵ_t converges weakly to 0 because $(X_t, Y_t, \boldsymbol{Z}_t) \overset{w}{\Rightarrow} (X, Y, \boldsymbol{Z})$, and similarly does the last term in (5.9) because $f_{X,Y,\boldsymbol{Z}}$ is a.e. continuous and bounded. Moreover, the middle term vanishes by assumption (5.7), which proves (5.8). □

As mentioned in Lauritzen [1996], Example 3.11, the pairwise Markov property is in general not satisfied under weak convergence. Proposition 5.2.1 provides conditions under which it is.

If $f_{\boldsymbol{X}_t} \to f_{\boldsymbol{X}}$ and the sequence is dominated by an integrable function, then $f_{\boldsymbol{X}_{t,A}} \to f_{\boldsymbol{X}_A}$ for all $A \subseteq \{1, \dots, d\}$. Hence, $f_{\boldsymbol{X}}$ need not be bounded and, crucially, (5.7) holds directly. Similarly, if $f_{\boldsymbol{X}_t} \to f_{\boldsymbol{X}}$ in L_2, then $f_{\boldsymbol{X}_{t,A}} \to f_{\boldsymbol{X}_A}$ in L_2 by Jensen's inequality, and we apply Hölder's inequality to show that (5.7) is satisfied.

5.3 Asymptotic Graphical Modeling of Extremes

We focus here on modeling the limiting tail density of a random vector $\boldsymbol{X}$ with probability density $f_{\boldsymbol{X}}$ using graphical models. Suppose that

$$f_{t^{-1}\boldsymbol{X} \,|\, ||\boldsymbol{X}||_\infty \geq t} \to f_{\boldsymbol{Y}} > 0,$$

such that $f_{\boldsymbol{Y}}$ is a probability density on $C_{||\cdot||_\infty}$. Recall that this condition is stronger than $\boldsymbol{X} \in \mathrm{RV}_{-\alpha}(\boldsymbol{Y})$, i.e., $t^{-1}\boldsymbol{X} \mid ||\boldsymbol{X}||_\infty \geq t \overset{w}{\Rightarrow} \boldsymbol{Y}$. As a consequence, the sequence of the density of

$$\boldsymbol{X}_t := t^{-1}\boldsymbol{X} \mid |X_k| \geq 1,$$

converges to the density of $\boldsymbol{Y} \mid |Y_k| \geq 1$, for all $k = 1, \dots, d$ (see Lemma 4.3.4). Let $\mathcal{G} = (\{1, \dots, d\}, E)$ be a graph with set of cliques $\mathcal{C}$. From Proposition 5.2.1 and Theorem 5.1.2, $\boldsymbol{X}_t$ satisfies the asymptotic pairwise Markov property according to $\mathcal{G}$, i.e.,

$$X_i \,\tilde{\perp\!\!\!\perp}\, X_j \mid \boldsymbol{X}_{-ij}, |X_k| \geq 1, \quad (i,j) \notin E, \quad k = 1, \dots, d,$$

if and only if

$$f_{\boldsymbol{Y}} = \frac{\prod_{C \in \mathcal{C}} h_C}{\prod_{D \in \mathcal{D}} h_D} \quad \text{on } C_{||\cdot||_\infty}, \tag{5.10}$$

where $\{h_S\}_{S \in \mathcal{C} \cup \mathcal{D}}$ is a class of homogeneous densities for which

$$h_D(\boldsymbol{y}_D) = \int_{\mathbb{R}^{|C \setminus D|}} h_C(\boldsymbol{y}_D, \boldsymbol{y}_{C \setminus D}) d\boldsymbol{y}_{C \setminus D}. \tag{5.11}$$

By Lemma (4.3.5),

$$f_{t^{-1}\boldsymbol{X}_S \,|\, ||\boldsymbol{X}_S||_\infty \geq t} \to f_{\boldsymbol{Y}_S \,|\, ||\boldsymbol{Y}_S||_\infty \geq 1} = c^{-1} h_S,$$

on $C^S_{||\cdot||_\infty}$ for some constant $c > 0$, $\forall S \in \mathcal{C} \cup \mathcal{D}$. In other words, h_S corresponds to the limiting tail density of $\boldsymbol{X}_S$, providing a way of inferring the low-dimensional factors in (5.10) from observations of $\boldsymbol{X}_S$. This leads us to the following possible strategy to model the limiting tail distribution of an i.i.d. sample of $\boldsymbol{X}$.

- Transform the positive and negative part of each marginal that exceeds a large threshold to unit Pareto to obtain $\boldsymbol{X}^C := (X_1 1_{|X_1| \geq 1}, \ldots, X_d 1_{|X_d| \geq 1})$.
- Select a decomposable graph $\mathcal{G}$ by testing asymptotic conditional independence between the marginals of $\boldsymbol{X}$.
- For all $S \in \mathcal{C} \cup \mathcal{D}$, select a model for the homogeneous density $h_S \propto f_{\boldsymbol{Y}_S \,|\, ||\boldsymbol{Y}_S|| \geq 1}$ with unit Pareto marginals; estimate its censored version using a sample from $\boldsymbol{X}^C_S$.
- This determines a valid limiting density $f_{\boldsymbol{Y}}$ of the form (5.10).

In low dimensions, there is a rich class of models for extreme value distributions [Gudendorf and Segers, 2010]. As we have seen in (1.12), if the exponent measure admits a density, then it is homogeneous and thus provides a model for h_S.

In a parametric approach, a refinement is to impose the parameters of $h_S(\cdot; \boldsymbol{\theta}_C)$ to be consistent with the ones of $h_D(\cdot; \boldsymbol{\theta}_D)$ whenever $D \subset C$ as in (5.11) and to estimate them from sufficient statistics. Example 5.1.1 illustrated a procedure to find consistent parametric families.

Recall that the procedure described in this chapter excludes the case of asymptotically independent variables as opposed to Chapter 3, but it has the benefit of being based on a proper limiting tail density.

Asymptotic graphical modeling for extremes necessitates testing asymptotic conditional independence, which can be difficult in a non-parametric setting. As a first approximation, we can easily test conditional independence of the binary

vector $\boldsymbol{B} = (1_{|X_1| \geq t}, \ldots, 1_{|X_d| \geq t})$ for some large t [Nagarajan et al., 2013]. Further efforts are needed to build a test for asymptotic conditional independence in the parametric continuous case — for instance, by relying upon (5.5).

6
Modeling Website Visits

We propose a simple model for the number of hits to Internet websites and show it to describe well a data set consisting of the visits of about 20 000 users to the 99 most visited websites in the Unites States in a specific month. We assume that the random vector of hits $\boldsymbol{X} = (X_1, \ldots, X_d)$ is distributed according to a censored multivariate normal (or Student) distribution with marginals transformed to be discrete Pareto IV and, following the ideas of graphical models, we enforce sparsity on the inverse covariance matrix to reduce dimensionality and visualize the dependence structure as a graph. The model easily includes covariates and is useful to comprehend the behavior of Internet users as a function of their age and gender.

We study a data set from Nielsen Holdings N.V., an information and measurement company, containing the number of visits of 19 436 users to the 99 most visited websites in the United States during one month. The table below displays hits of two users to the top 10 websites.

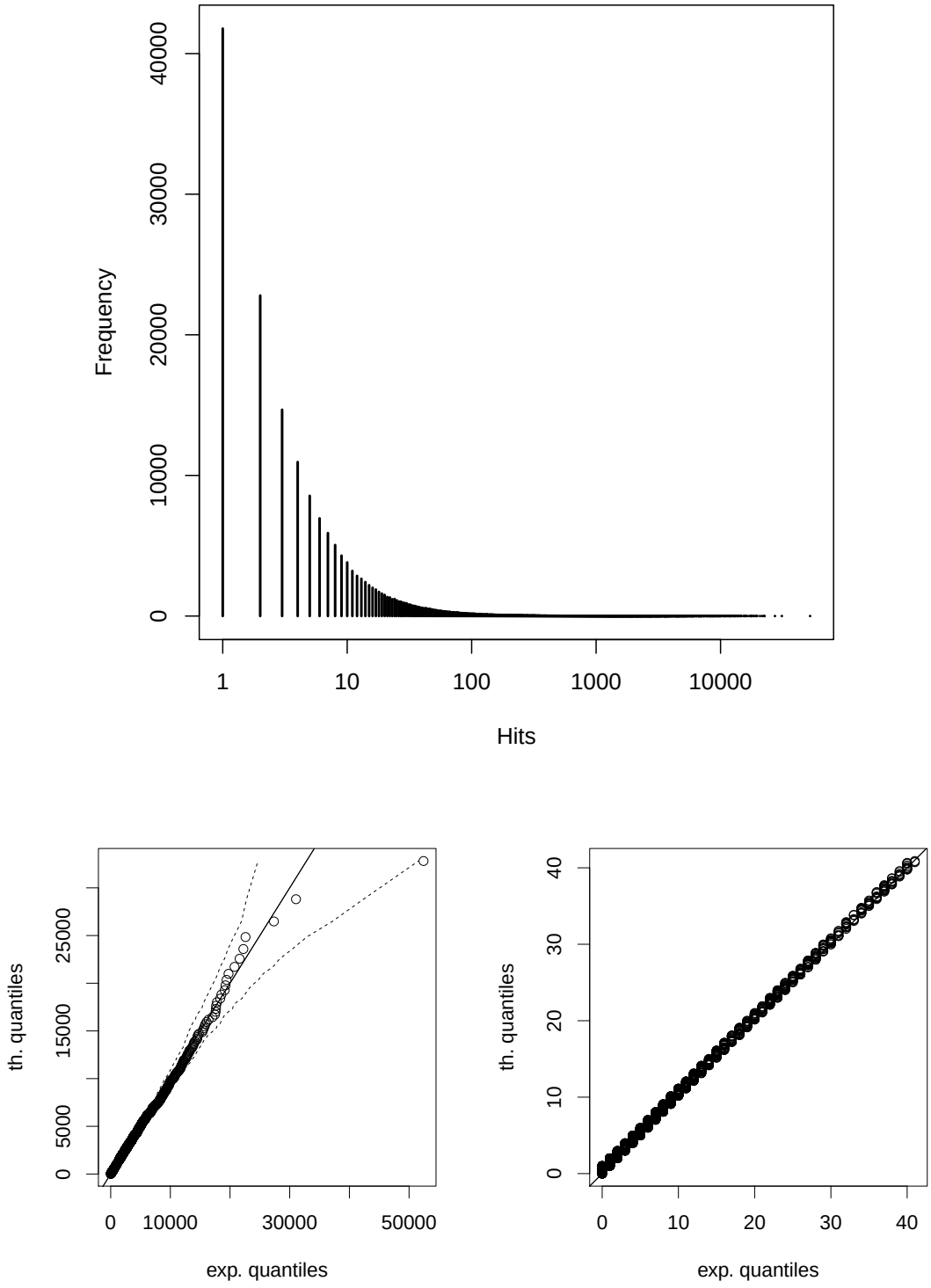

Figure 6.1: Frequency table of about 207 000 positive hits to any of the 99 most visited websites (top), QQ-plot for a discrete PIV fitted to these data (left) and the same plot for quantiles smaller than 0.8 (right). The dashed lines denote 95% pointwise confidence intervals.

Website	user 1	user 2
Google	2	1
Yahoo!	155	668
Facebook	0	0
YouTube	0	2
MSN/WindowsLive/Bing	0	0
AOL Media Network	0	0
Amazon	0	0
Ask Search Network	1	0
Wikipedia	0	0
eBay	0	0

We are treating the number of visits as i.i.d. realizations $\boldsymbol{x} = \{\boldsymbol{x}^{(k)}\}_{k=1}^{n}$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ for $d = 99$ and $n = 19\,436$. Modeling these data seems challenging because the marginals of $\boldsymbol{X}$ have very heavy-tailed distributions exhibiting peaks at zero and are strongly dependent. Nevertheless, we will present a simple multivariate distribution that fits the data relatively well. Before starting, we leave aside about 10% of the observations and refer to them as test data.

6.1 The Discrete Pareto IV Distribution

First, we seek to describe the distribution of the number of hits to any of the 99 websites, written X_{tot}. A frequency table of $X_{\text{tot}} \mid X_{\text{tot}} \geq 1$ is displayed on top of Figure 6.1 on a log-scale and reveals a surprisingly regular decay. The probability mass on the value 0, however, forms an irregular peak: 88.1% of hits are 0 while only 2.4% and 1.3% are 1 and 2 respectively. A flexible 4 parameter discrete family distribution is the discrete Pareto IV (D-PIV) defined by its probability mass function $p(k) = F(k) - F(k-1)$ on $\mathbb{N} = \{1, 2, \ldots\}$, where

$$F(x) = 1 - \left\{1 + \xi \frac{(x+\mu)^\beta - \mu^\beta}{\sigma}\right\}^{-1/\xi} 1_{\{0 \leq x < e_1\}},$$

for $\xi \in \mathbb{R}$, $\mu \geq 0$, $\beta > 0$, $\sigma > 0$, and, if $\xi \geq 0$, $e_1 = \infty$. When $\xi > 0$, F is the cumulative distribution function (cdf) of a Pareto IV distribution [Arnold, 2015]. When $\xi < 0$, F has a finite endpoint

$$e_1 = \left(\mu^\beta + \frac{\sigma}{|\xi|}\right)^{1/\beta} - \mu,$$

and when $\xi = 0$, it is extended by continuity to $F(x) = 1 - \exp\left\{(x+\mu)^\beta/\sigma - \mu^\beta/\sigma\right\}$. In the case $\mu = 0$, $\beta = 1$, the D-PIV coincides with a discrete generalized Pareto distribution which is closely related to the Zipf–Mandelbrot distribution. Zipf-based distributions are commonly used to describe discrete real data such as word frequencies [Booth, 1967], city sizes [Gabaix, 1999], size of companies Axtell [2001] and website hits [Clauset et al., 2009].

We fit this family distribution to $X_{\text{tot}} \mid X_{\text{tot}} \geq 1$ by maximizing numerically the log-likelihood using the function `optim` of `R` with starting parameters $(\xi, \sigma, \beta, \mu) = (0.1, 1, 0.1, 1)$ [R Core Team, 2015]. Even though there are more than 200 000 observations, the maximization is fast because the likelihood only needs to be evaluated once at each observed integer value. The D-PIV fits the data well as shown in the bottom of Figure 6.1. This is confirmed by a discrete Kolmogorov–Smirnov goodness-of-fit test in package `dgof` [Arnold and Emerson, 2011] applied to the test data set which gives a p-value of 0.14. Maximum likelihood estimates and approximate 95% confidence intervals based on asymptotic normality are displayed in the table below. We mention that the normality assumption leads to an approximation error.

ξ	σ	β	μ
$-0.17_{\pm 0.03}$	$0.20_{\pm 0.06}$	$0.07_{\pm 0.02}$	$1.13_{\pm 0.04}$

Since $\xi < 0$, the fitted distribution has a finite endpoint e_1, The latter is roughly 1.7 times the largest observation in the data set which is 52 419 hits to Facebook.

For each website $i = 1, \ldots, d$, we now fit a D-PIV and two embedded models corresponding to the cases $\mu = 0$ and $(\mu = 0,\ \beta = 1)$ to $X_i \mid X_i \geq 1$ and select among them according to the Bayesian information criteria (BIC). The following table shows the percentage of discrete Kolmogorov–Smirnov tests whose p-values are smaller than 0.05 and the average sample size.

	Data Set	Rejected	Av. Sample size
D-PIV	training	5.05%	2090
D-PIV	test	6.06%	234

The rejection rate is close to 5%, indicating a good fit which supports the adequacy of the D-PIV family distribution to model $X_i \mid X_i \geq 1$. The saturated model and the embedded models in the case $\mu = 0$ and $\mu = 0$, $\beta = 1$ were selected 25, 59 and 15 times respectively.

6.2 Censored Student Copula Graphical Models

We are now interested in modeling the dependence structure of $\boldsymbol{X} = (X_1, \ldots, X_d)$. The problem cannot easily be reduced to lower dimension by assuming independence between blocks of marginals because the dependence between them is rather strong. To illustrate this, we define a binary vector $\boldsymbol{B}$ by $B_i = 1_{\{X_i \geq 1\}}$, where 1_A is the indicator function, and perform χ^2 tests of independence with level 0.05 for all pairs; the test rejects independence 98% of the time. Graphical models provide a way to simplify a joint density by making assumptions of conditional independence between marginals as we will discuss later. We use `ci.test` of package `bnlearn` to test conditional independence $B_i \perp\!\!\!\perp B_j \mid \boldsymbol{B}_{-i,-j}$ [Scutari, 2010]. Pairwise conditional independence is rejected only 33% of the time, suggesting that there might be many conditional independence relations to exploit to reduce dimensionality.

Is there a simple multivariate distribution that can be chosen for $\boldsymbol{X}$? There are extensions of the Pareto IV in the multivariate case but they typically have limited dependence structure or are intractable (Arnold [2015], Chapter 6). As explained by copula theory, the marginals of any multivariate continuous distribution can be transformed to obtain another distribution with marginals of arbitrary continuous distribution while preserving the dependence structure [Sklar, 2010, Nelsen, 2006]. In the discrete case, copulas suffer from limitations such as non-identifiability [Genest and Nešlehová, 2007].

An additional difficulty for modeling $\boldsymbol{X}$, as we have seen, is that its distribution is fundamentally different on 0 than on positive integers, requiring a discrete

multivariate distribution that accounts for this mixture. The multivariate zero inflated Poisson distribution is such an instance [Li et al., 1999, Liu and Tian, 2015].

The multivariate Gaussian and Student distributions are some of the few known multivariate distributions with explicit multivariate marginal and conditional densities. Their dependence structure is well understood and has some flexibility without being overdetermined as it only involves pairwise interactions. We now explain how they can be used for modeling $\boldsymbol{X}$. Let $\boldsymbol{t} \in \mathbb{R}^d$ be a vector of thresholds and suppose that $\boldsymbol{Z}$ follows a centered Student distribution with degree of freedom $\nu \in (2, \infty]$ and covariance matrix $\frac{\nu}{\nu-2}\Sigma$ such that Σ has diagonal $\mathbf{1} \in \mathbb{R}^d$. By convention, $\boldsymbol{Z} \sim \mathcal{N}(0, \Sigma)$ in the case $\nu = \infty$. We obtain a vector $\boldsymbol{X}$ with values in $(\{0\} \times \mathbb{N}_+)^d$ satisfying $X_i \mid X_i \geq 1 \sim F_{\text{D-PIV}}$ for all i by applying to $\boldsymbol{Z}$ the following procedure: censor Z_i when it falls below t_i, set censored values to 0, transform non-censored Z_i appropriately and round them. More precisely, define $\boldsymbol{X}$ by

$$X_i = \begin{cases} \lfloor Y_i \rfloor + 1 & \text{if } Z_i \geq t_i, \\ 0 & \text{else,} \end{cases} \qquad Y_i := F^{-1}_{\text{PIV}_i}\{\Phi^{(\nu)}_{Z|Z\geq t_i}(Z_i \mid Z_i \geq t_i)\},$$

for all $i = 1, \ldots, d$, where $\Phi^{(\nu)}_{Z|Z\geq t_i}$ is the cdf of a truncated Student and F_{PIV_i} is such that $F_{\text{PIV}_i}(k) - F_{\text{PIV}_i}(k-1)$ is the probability mass function of the D-PIV fitted to $X_i \mid X_i \geq i$.

This formulation provides a simple yet non-trivial probability distribution with values in $\mathbb{N}^d$ accounting for the irregularity of the distribution between null and positive values. It is frequently used to model multivariate rainfall data where zeros occur when no rain is measured as in Bell [1987], Allcroft and Glasbey [2003]. It also appears in multivariate extreme value analysis with zeros corresponding to non-extremal events, such as a modeling of extremal oceanographic data in Bortot et al. [2000].

In theory, we could express the probability mass function of $\boldsymbol{X}$ as a sum of terms involving the joint cdf of $\boldsymbol{Z}$, but the expression is intractable when d is large. We are thus treating positive values of $\boldsymbol{X}$ as continuous, making the working assumption that $X_i \approx Y_i + \frac{1}{2}$ when $Z_i \geq t_i$. A similar approximation for the multivariate normal copula in the case of discrete data is studied in Nikoloulopoulos [2016]. This is a

reasonable assumption here because the range of positive values is relatively large. Another benefit is that it yields a tractable expression for conditional distributions: for any disjoint sets $A, B, C, D \subseteq \{1, \dots, d\}$, $\boldsymbol{k} \in \mathbb{N}_+^d$,

$$\Pr\{\boldsymbol{X}_{AB} = (\boldsymbol{k}_A, \boldsymbol{0}) \mid \boldsymbol{X}_{CD} = (\boldsymbol{k}_C, \boldsymbol{0})\} = \frac{\Phi^{(\nu)}\left(\boldsymbol{t}_{BD} \mid \boldsymbol{Z}_{AC} = \boldsymbol{z}_{AC}\right)}{\Phi^{(\nu)}\left(\boldsymbol{t}_B \mid \boldsymbol{Z}_{AC} = \boldsymbol{z}_{AC}\right)} \phi^{(\nu)}(\boldsymbol{z}_A \mid \boldsymbol{z}_C), \tag{6.1}$$

where $z_i = \Phi^{-1}_{Z_i | Z_i \geq t_i}\{F_{\text{PIV}}(\boldsymbol{k}_i - \frac{1}{2})\}$. The quantities $\Phi^{(\nu)}(\cdot \mid \boldsymbol{z}_F)$ and $\phi^{(\nu)}(\cdot \mid \boldsymbol{z}_F)$ denotes respectively the joint cdf and density function of $\boldsymbol{Z}_E \mid \boldsymbol{Z}_F = \boldsymbol{z}_F$ for some disjoint sets $F, E \subset \{1, \dots, d\}$ and coincides with the joint cdf $\Phi^{(\nu)}(\cdot)$ and density $\phi^{(\nu)}(\cdot)$ of a non-centered Student distribution on $\mathbb{R}^{|E|}$ [Ding, 2016]. The joint cdf $\Phi^{(\nu)}(\cdot)$ can be evaluated using the package `mvtnorm` [Genz et al., 2008]. Notice that one can sample from $\boldsymbol{Z}_E \mid \boldsymbol{Z}_F$ using (6.1), for instance by applying inverse transform sampling recursively.

We refer to the model above as the censored Student copula model and we now present a possible way of estimating it. Recall that the number of visits $\boldsymbol{x} = \{\boldsymbol{x}^{(k)}\}_{k=1}^n$ are assumed to be sampled from $\boldsymbol{X}$. The parameters t_i can directly be estimated as $\hat{t}_i = \Phi^{-1}_{(\nu)}(n_i/n)$ where n_i is the number of observations such that $x_i^{(k)} \geq 1$. It remains to estimate the matrix Σ to fully determine the distribution. If a sample $\boldsymbol{z}$ from $\boldsymbol{Z}$ was observed, its maximum likelihood estimate in the Gaussian case would be

$$\hat{S} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{z}^{(i)} \boldsymbol{z}^{(i)T}.$$

However, only a censored transformation of $\boldsymbol{z}$ is observed and different methods are possible in this case [Lee and Scott, 2012, Schemper et al., 2013]. We found satisfactory performance and efficiency of estimating $\rho = \Sigma_{ij}$ for all pairs (i, j) separately by maximizing the pairwise likelihood $\ell_{ij}(\rho) = \sum_{k=1}^n \log \Pr(X_i = x_i^{(k)}, X_j = x_j^{(k)})$, where

$$\begin{aligned}
\Pr(X_i = 0, X_j = 0) &= \Pr(Z_i \leq t_i, Z_j \leq t_j), \\
\Pr(X_i = x_i, X_j = x_j) &= c_i(x_i) c_j(x_j) \Pr(Z_i = z_i, Z_j = z_j), \\
\Pr(X_i = x_i, X_j = 0) &= c_i(x_i) \Pr(Z_i = z_i, Z_j \leq t_j),
\end{aligned}$$

$$z_i = M_i^{-1}\{F_{X_i|X_i\geq 1}(x_i)\}, \quad M_i(z) = \Pr(Z_i \leq z \mid Z_i \geq t_i)$$
$$m_i(z) = M_i'(z), \quad c_i(x_i) = \frac{f_{X_i|X_i\geq 1}(x_i)}{m_i(z_i)},$$

where we abuse notation and write $\Pr(Z_i = z) := \frac{\partial}{\partial z}\Pr(Z_i \leq z)$. All together, this gives a consistent estimator $\hat{\Sigma}^C$ of Σ defined by $\hat{\Sigma}^C_{ij} = \arg\max_{\rho\in[-1,1]} \ell_{ij}(\rho)$.

We compute Σ^C in the case $\nu = \infty$ for $n \approx 20\,000$ users and display the largest positive and negative correlations Σ^C_{ij} in the table below with a 95% confidence interval based on asymptotic normality of the maximum likelihood estimator.

i	j	ρ_{ij}
TurboTax	Intuit	$0.91_{\pm 0.01}$
YouTube	YouTube-NoCookie	$0.86_{\pm 0.00}$
YouTube	Vevo	$0.78_{\pm 0.01}$
Expedia	TripAdvisor	$0.75_{\pm 0.02}$
YouTube-NoCookie	Vevo	$0.74_{\pm 0.01}$
Lowe-s	The Home Depot	$0.71_{\pm 0.02}$
eBay	PayPal	$0.68_{\pm 0.02}$
Facebook	Zynga	$0.68_{\pm 0.02}$
Blogger	WordPress	$0.67_{\pm 0.02}$
U.S. Internal Revenue Service	TurboTax	$0.65_{\pm 0.03}$
American Express	Vevo	$-0.11_{\pm 0.06}$
Comcast Digital Entertainment	Road Runner	$-0.10_{\pm 0.05}$
Vevo	Discover	$-0.08_{\pm 0.06}$

The marked dependence between the websites above is sensible: Turbo Tax and Intuit are tax softwares, YouTube and Vevo video websites, Expedia and TripAdvisor travel websites, Lowe-s and The Home Depot home improvement stores, eBay is an e-commerce company who owned PayPal, an online payment operator. Facebook, a social networking service, had a partnership with Zynga, an online video game provider; Blogger and WordPress are blog publishing services.

Although there are many other positively correlated pairs of websites, only three pairs exhibit a significant negative correlation according to our model. Visitors of Vevo are less interested in American Express and Discover, two financial service companies, as are visitors of Road Runner, a running shoe store, in Comcast, a mass media company.

Two characteristics are available on users: their age, which takes tabulated values and is censored above 65, and their gender, which is binary. These covariates could be included in the analysis by regressing the parameters of the model with respect to them. We prefer here to treat them on the same terms as the other variables in a fully multivariate analysis. Let Z_a and Z_b be Student distributed. For age, we suppose that X_a is censored from above when $Z_a \geq t_a$ for some t_a and $X_a = F_a^{-1}(\Phi^{(\nu)}_{Z_a|Z_a \leq t_a}(Z_a \mid Z_a \leq t_a)\}$, where F_a^{-1} is the empirical quantile function of $\boldsymbol{x}_a$. Note that since $\boldsymbol{x}_a$ is discrete but treated as continuous, it is a better approximation to transform it to uniform using $\{F_a(x) + F_a(x-1)\}/2$ instead of its empirical cdf F_a. For gender, we assume that $X_b = 1_{\{Z_b \geq t_b\}}$ for some $t_b \in \mathbb{R}$. Following the procedure explained previously, we obtain an estimate for the threshold vector $\boldsymbol{t} \in \mathbb{R}^{d+2}$ and the matrix $\Sigma \in \mathbb{R}^{d+2} \times \mathbb{R}^{d+2}$. The underlying random vector is now $\boldsymbol{Z} = (Z_1, \ldots, Z_{99}, Z_a, Z_b)$ which is Student distributed with mean $\mathbf{0}$, covariance matrix $\frac{\nu}{\nu-2}\Sigma$ and degree of freedom $\nu > 2$.

The next table shows with which websites age and gender share the strongest correlation in the case $\nu = \infty$, with the convention $X_a = 1$ for female and $X_a = 0$ for male.

i	j	Σ^C_{ij}
age	Legacy	$0.34_{\pm 0.03}$
age	American Express	$0.29_{\pm 0.04}$
age	WhitePages	$0.28_{\pm 0.03}$
age	Shopzilla	$0.27_{\pm 0.03}$
age	SuperPages	$0.27_{\pm 0.03}$
age	Vevo	$-0.30_{\pm 0.02}$
age	YouTube-NoCookie	$-0.23_{\pm 0.02}$
age	Disney DOL	$-0.22_{\pm 0.03}$
age	YouTube	$-0.21_{\pm 0.02}$
age	Tumblr	$-0.20_{\pm 0.03}$

i	j	Σ^C_{ij}
gender	Pinterest	$0.31_{\pm 0.04}$
gender	JCPenney	$0.30_{\pm 0.04}$
gender	Macy's	$0.29_{\pm 0.04}$
gender	Allrecipes	$0.28_{\pm 0.03}$
gender	Everyday Health	$0.27_{\pm 0.03}$
gender	ESPN	$-0.25_{\pm 0.03}$
gender	Turner-SI	$-0.22_{\pm 0.03}$
gender	Big Lead Sports	$-0.21_{\pm 0.04}$
gender	NFL	$-0.21_{\pm 0.03}$
gender	CNET	$-0.17_{\pm 0.04}$

The most visited websites by older persons were Legacy, an online memorial provider, American Express, a financial service company, White Pages, a contact information provider, Shopzilla, a shopping website and Experian, a global information services group. Younger persons mostly hit to Vevo, YouTube, Disney, a mass media and entertainment company, and Tumblr, a social networking website.

Women were more frequently on Pinterest, a photo sharing website, JCPenney and Macy's, two department store chains, Allrecipes, a social networking service focused on food, and Everyday Health, a company producing content on wellness. Men, on the other hand, preferred sport related networks and CNET, a technology news website.

From a model selection point of view, it is natural to try to represent the matrix Σ with fewer parameters. As we have seen, there are very few independent pairs of marginals so Σ is unlikely to be sparse. Following the ideas of graphical models, one can try to exploit sparsity of the inverse matrix Σ^{-1} which translates, in the Gaussian case, into conditional independence relations between some of the marginals: $\Sigma^{-1}_{ij} = 0$ if and only if $Z_i \perp\!\!\!\perp Z_j \mid \boldsymbol{Z}_{\{1,\dots,d\}\setminus\{i,j\}}$ (see Lauritzen [1996]). The constraint is a little different for the Student distribution: $\Sigma^{-1}_{ij} = 0$ means that Z_i and Z_j are conditionally uncorrelated given the rest of the vector [Finegold and Drton, 2011]. Besides reducing the number of parameters, these assumptions allow us to visualize the dependence structure as a graph $\mathcal{G} = (V, E)$ defined as follows: the set of nodes V is $\{1, \dots, d\}$, i.e., each node corresponds to a marginal of $\boldsymbol{Z}$, and the set of edges E is determined by

$$(i, j) \notin E \implies \Sigma^{-1}_{ij} = 0. \tag{6.2}$$

An efficient procedure to estimate a sparse matrix Σ^{-1} given a sample $\boldsymbol{z}$ i.i.d. drawn from $\mathcal{N}(0, \Sigma)$ is proposed in Friedman et al. [2008]. It solves the convex optimization problem called Gaussian graphical lasso which consists in maximizing the log-likelihood of a multivariate Gaussian with an additional penalization to enforce sparsity. More precisely, consider the following estimator for Σ^{-1}:

$$\hat{\Theta} = \underset{\Theta \succ 0}{\arg\max} \quad \log\det\Theta - \mathrm{tr}(\hat{S}_n\Theta) - \lambda_n||\Theta||_1, \tag{6.3}$$

where $\hat{S}_n = n^{-1}\boldsymbol{z}^T\boldsymbol{z}$ is the empirical covariance matrix, $\lambda_n > 0$ is a regularization parameter, "tr" denotes the trace of a matrix, "$\succ$" means positive definite and $||\Theta||_1 := \sum_{ij} |\Theta_{ij}|$. After slightly reformulating (6.3), Ravikumar et al. [2011] shows that for any consistent estimator $\hat{S}_n$ of Σ (such as $\hat{\Sigma}^C$) and under an additional

assumption (called incoherence or irrepresentability condition), the solution $\hat{\Theta}$ is a consistent estimator of Σ^{-1} and it correctly detects null entries as $n \to \infty$ for fixed d. When $\nu < \infty$, solving (6.3) still makes sense as $\hat{\Theta}$ is the closest matrix to Σ^{-1} in terms of a Bregman divergence. Note that if one is interested in estimating the graph $\mathcal{G}$ only, Meinshausen and Bühlmann [2006] presents a procedure to estimate it consistently in the Gaussian case. There is a rich literature on Gaussian graphical lasso [Banerjee et al., 2008, Yuan, 2010, Cai et al., 2016], copula Gaussian graphical models [Dobra and Lenkoski, 2011, Xue and Zou, 2012, Liu et al., 2012], Student graphical lasso [Finegold and Drton, 2011] and Gaussian lasso in the censored case [Johnson, 2009], including several variants such as the adaptive lasso which penalizes coefficients differently Zou [2006] and a decomposition of Σ^{-1} into the sum of a sparse and a low-rank matrix [Chandrasekaran et al., 2012].

Coming back to our data analysis, we compute $\hat{S}_n = \hat{\Sigma}^C$ for various degrees of freedom ν and solve (6.3) for several regularization parameters λ using the package `huge` [Zhao et al., 2012] to obtain matrices $\hat{\Sigma}$ whose inverse are sparse. We then choose ν and λ by minimizing two scores: the average negative log-likelihood

$$\ell(\hat{\Sigma}) = -\frac{1}{n}\sum_{k=1}^{n}\left\{\log \Phi^{(\nu)}(\boldsymbol{t}_{A_k^c} \mid \boldsymbol{z}_{A_k}^{(k)}) + \log \phi^{(\nu)}(\boldsymbol{z}_{A_k}^{(k)}) + \sum_{i \in A_k} \log c_i(x_i^{(k)})\right\}, \qquad (6.4)$$

where $A_k = \{i : x_i^{(k)} \neq 0\}$, and the sum of average pairwise negative log-likelihood

$$\ell_{\text{pairwise}}(\hat{\Sigma}) = \sum_{i<j} \ell_{ij}(\hat{\Sigma}_{ij}).$$

We compute ℓ and ℓ_{pairwise} from the test data set using 400 and about 20 000 observations respectively and display them in the table below. We also report the percentage of null entries in the upper triangular part of $\hat{\Sigma}^{-1}$ referred to as sparsity.

		$\hat{\Sigma}^C$	$\hat{\Sigma}_{\lambda=0.0075}$	$\hat{\Sigma}_{\lambda=0.02}$	$\hat{\Sigma}_{\lambda=0.1}$
$\nu = \infty$	sparsity	0%	25%	50%	73%
	ℓ_{pairwise}	3079.33	**3079.31**	3079.36	3081.78
	ℓ	23.14	**23.07**	23.09	23.43
$\nu = 60$	sparsity	0%	24%	50%	75%
	ℓ_{pairwise}	3079.79	3079.78	3079.83	3082.15
	ℓ	23.20	23.12	23.11	23.43
$\nu = 30$	sparsity	0%			
	ℓ_{pairwise}	3080.19			
	ℓ	23.24			

Both scores agree on selecting the censored Gaussian copula graphical model with $\lambda = 0.0075$, corresponding to a sparsity of 25%. Figure 6.2 shows the corresponding graph, plotting only edges (i, j) such that $\hat{\Sigma}^{-1}_{ij}$ are among the 5% largest non-null entries in absolute values. An edge between i and j should be interpreted as X_i carrying relevant information to predict X_j, and X_i and X_j being positively dependent. We identify several meaningful groups such as financial companies (Chase, Discover, Citibank, Capital One), news (Fox News, CNN, ABCNEWS, USATODAY, NYTimes, Tribune Newspaper) and sports (Turner-SI, NFL Internet Network, Big Lead Sports, ESPN). The graph also illustrates which websites a user tends to visit depending on its age and gender and we recognize some of the relations discussed earlier.

We now present some diagnostics of the selected model. We start by testing if all pairs from the real data set have the probability expected from the model of occurring in certain regions by performing binomial tests with level 0.05. The results are reported in the next table.

Region	Rejected
$\Pr(X_i > 0, X_j > 0)$	10.4%
$\Pr(X_i = 0, X_j > 0)$	7.7%

The percentage of rejection is above 5%, which is the rate of a correct model.

[1]The following websites do not have any neighbor in the graph in Figure 6.2: Apple, Glam Media, Weather Channel Network, Bank of America, Adobe, Wells Fargo, NBC Universal Sites, Napster, ShopAtHome, Hewlett Packard, Experian.

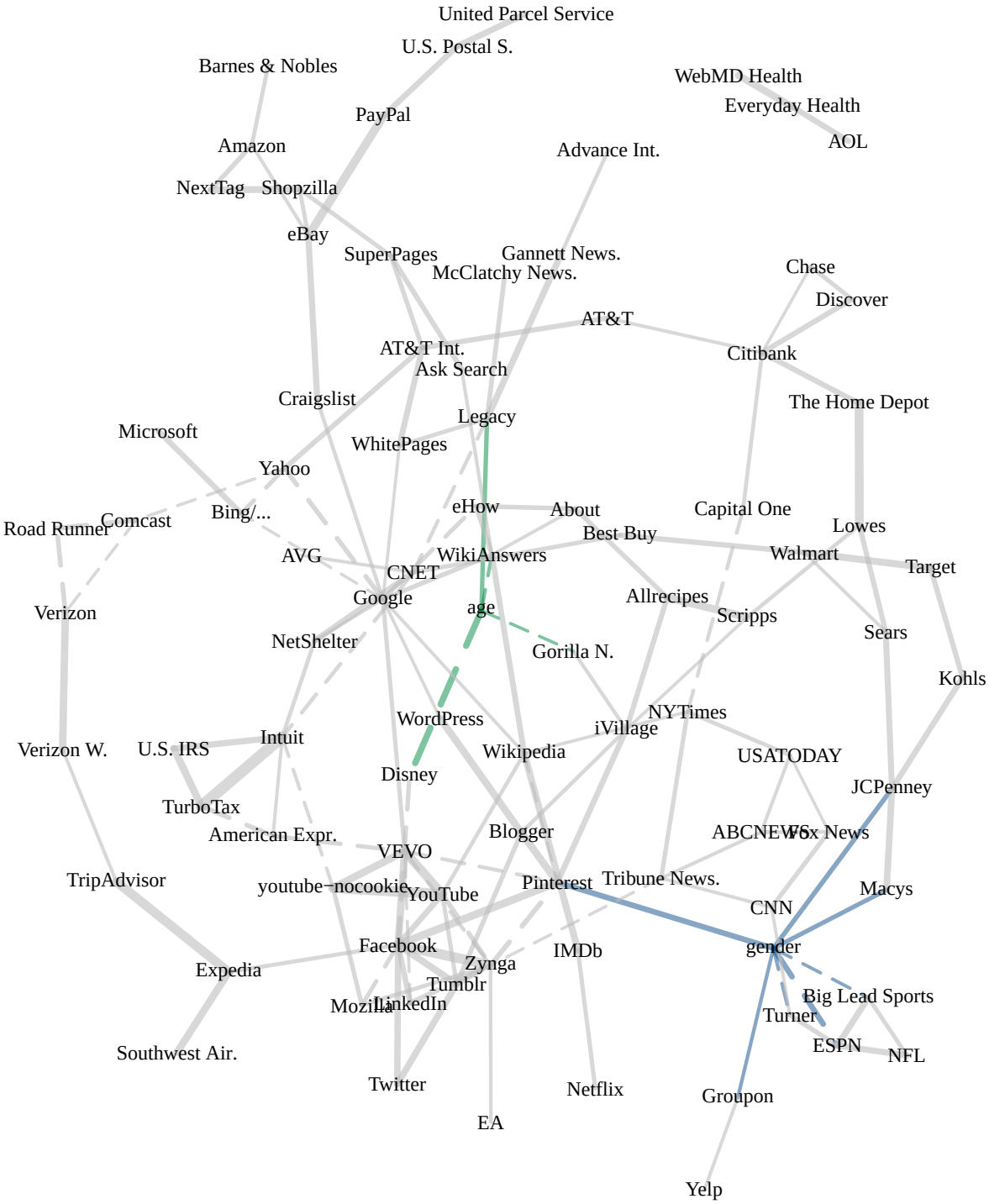

Figure 6.2: Graph illustrating the dependence structure of $\boldsymbol{X} = (X_1, \ldots, X_{99}, X_a, X_b)$, the vector counting hits to the 99 most visited websites in the US in a given month; X_a is the age (in green) and X_b the gender of a user (in blue; 1 for female, 0 for male). We fitted a censored Gaussian copula graphical model to $\boldsymbol{X}$ with sparse inverse covariance matrix Σ^{-1}. The width of an edge between i and j is proportional to the interaction coefficient Σ^{-1}_{ij} on a log-scale and shows how predictive X_i is for X_j. We plotted only edges having the 5% largest non-null coefficients in absolute values. Dashed edges correspond to positive coefficients, i.e., negative partial correlations.[1]

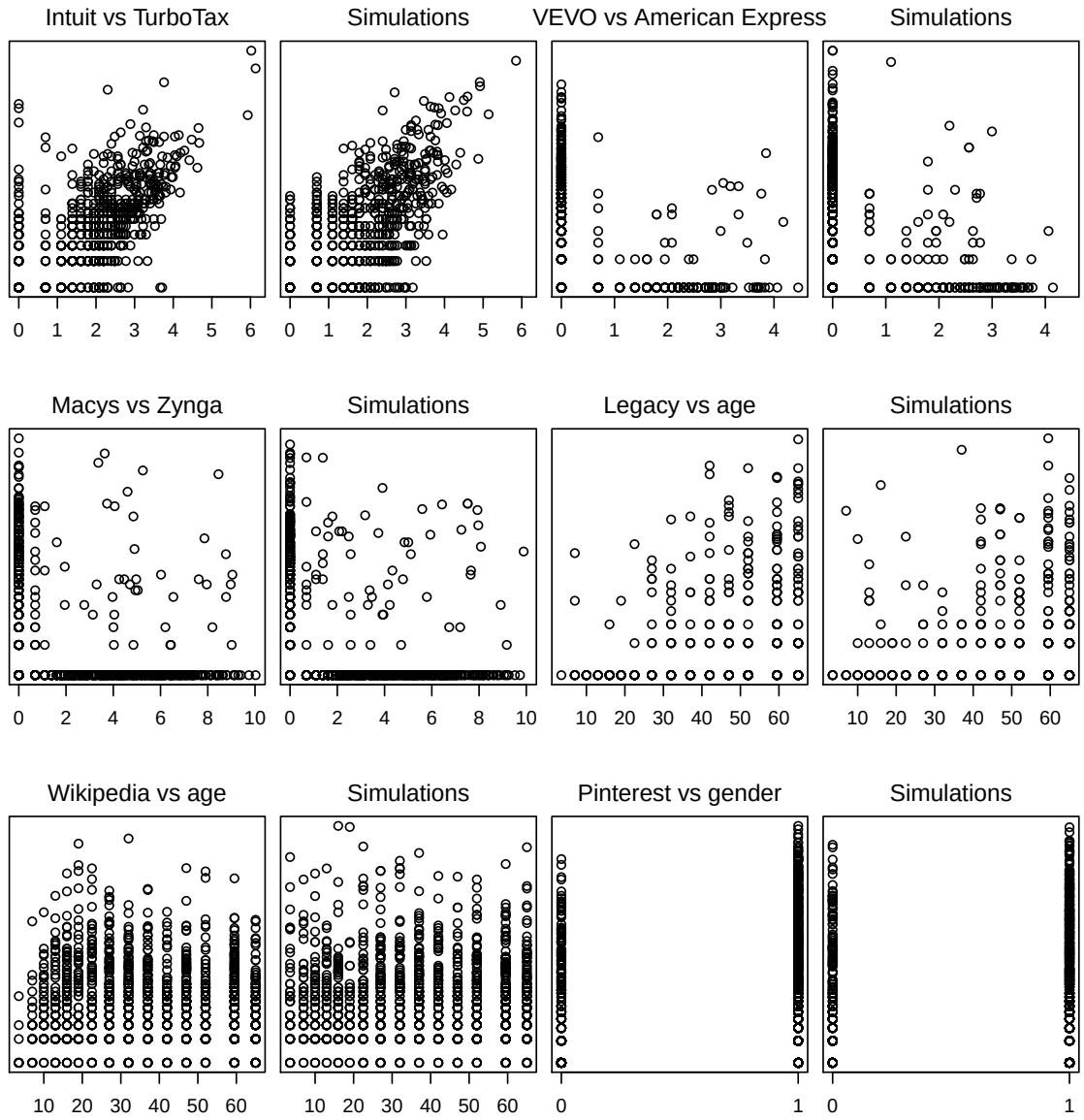

Figure 6.3: Several scatterplots of the data (first and third columns) and simulations from a censored Gaussian copula graphical model (second and fourth columns). The quantities plotted are the age X_a, the gender X_b (1 for female, 0 for male) and the transformation $\log(X_i + 1)$ of the number of hits X_i to website i.

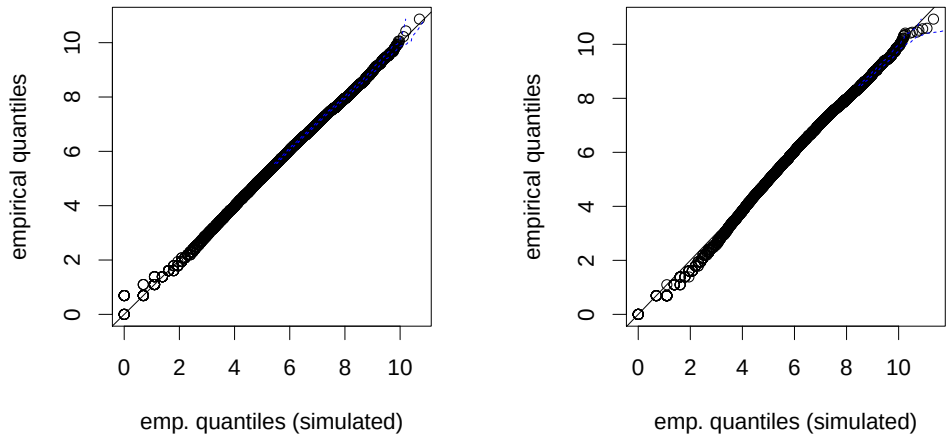

Figure 6.4: Diagnostic QQ-plots for the number of hits to the 5 (left) and 99 (right) most visited websites. Empirical quantiles of the data are plotted against simulated data from the censored Gaussian copula graphical model on a log scale. The dashed lines denote 95% confidence intervals for the upper tail based on 100 simulations from the model.

As a further diagnostic, we draw about 20 000 realizations from the censored Gaussian copula graphical model and compare them to the data. Figure 6.3 displays a few scatterplots chosen as follows: most positively, most negatively and least correlated pairs of hits, the most positively and less correlated pairs involving age and the most correlated pair involving gender. Replicates from the model appear to be relatively similar to the data. Figure 6.4 shows QQ-plots for the number of hits to the 5 (on the left) and 99 (on the right) most visited websites. Confidence intervals for the upper part of the distribution are built by simulating 100 times from the model and indicate that the tail is well fitted. The quantiles in the plot on the right-hand side form a slightly concave line instead of a straight one, revealing some inaccuracies.

We end by assessing the performance of the graphical model by comparing it to several alternative techniques in the following experiment: we remove observations from a marginal X_i in the test data set and try to predict if $X_i > 0$ by estimating $\Pr(X_i > 0 \mid \boldsymbol{X}_{-i})$. For $i = 1, \ldots, 99$, this amounts to predicting if a user visits website i given its age, gender and the number of hits to the other websites

during the month. For gender, we try to guess if the user is a female; for age, if the user is older than 35. To gain efficiency, we make the working assumption that $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{\{1,\dots,d\}\setminus\{i,j\}}$ if $\Sigma^{-1}_{ij} = 0$, a relation which holds for $\boldsymbol{Z}$ but only approximately for $\boldsymbol{X}$. We choose the score function

$$1_{\{X_i>0\}} \log \hat{p} + (1 - 1_{\{X_i>0\}}) \log(1 - \hat{p}),$$

where $\hat{p}$ is the predicted probability that $X_i > 0$, and compare the predictions of the model to various methods. First, we naively approximate $\Pr(X_i > 0 \mid \boldsymbol{X}_{-i})$ by $\Pr(X_i > 0)$. Second, we fit an Ising model — a probabilistic graphical model for binary data — to $\boldsymbol{B} = (1_{X_1>0}, \dots, 1_{X_{d+2}>0})$ using the `R` package `IsingFit` and compute $\Pr(X_i > 0 \mid \boldsymbol{B}_{-i})$. Third, we train the decision tree algorithms found in packages `rpart`, `tree`, `ct` and `randomForest` (RF) to predict $X_i > 0 \mid \boldsymbol{X}_{-i}$ [Van Borkulo and Epskamp, 2014, Therneau et al., 2015, Hothorn et al., 2006, Liaw and Wiener, 2002]. The next table shows the average score and the percentage of correct guessing computed for about 600 predictions from the test data set.

	Naive	$\hat{\Sigma}^C$	$\hat{\Sigma}_{\lambda=0.0075}$	$\hat{\Sigma}_{\lambda=0.02}$	$\hat{\Sigma}_{\lambda=0.1}$
Average score	0.334	**0.254**	0.255	**0.254**	0.257
Correctly guessed	86.8%	89.6%	89.6%	89.3%	89.9%

	Ising	`rpart`	`tree`	`ct`	RF
Average score	0.267	0.324	0.293	0.287	∞[2]
Correctly guessed	89.3%	88.5%	88.5%	88.1%	89.8%

The censored Gaussian copula is a simple probabilistic model that provides an approximation of the full joint distribution of the data and is thus valuable to comprehend its dependence structure as a whole. The experiment above suggests that decision trees that were specifically trained for certain predictions do not outperform the model, at least when they are used without further tuning. The Ising model performs relatively well and seems a suitable alternative if only the fact that a website is visited is important in the analysis, but not the number of hits. On average, the censored Gaussian gives the best performance, showing its ability to capture relevant information in this data set.

[2] `RandomForest` returns 7 infinite values; its average score computed from finite values only is 0.221.

7 Discussion

As a research topic for my Master's thesis at EPFL in 2013, Anthony Davison suggested to apply the ideas of graphical models to multivariate extreme value distributions. Modeling extremes that propagate through a graph seemed a very intuitive and appealing idea to me and I began with a reading of Resnick [1987], Tawn [1990], Ballani and Schlather [2011] and Cooley et al. [2010] to understand which models were being used in high dimensions. When I arrived at Columbia University as a visiting student, I focused on the extremal Dirichlet model and tried unsuccessfully to impose a factorization of its density — this was doomed to fail since extreme value distributions can only factorize trivially as shown by Papastathopoulos and Strokorb [2016]. I also investigated vine copulas which is a type of graphical models that can theoretically represent any continuous density, but I did not pursue this direction, becoming skeptical about assumptions made on conditional distributions for practical use. Richard Davis saw the need for a definition of asymptotic conditional independence (see Section 5.2) and the main result of my Master's thesis was a factorization of the limiting tail density with respect to a graph, as stated in Theorem 5.1.2, part (ii.). However, this statement was only motivated by some approximations and it was not clear that the low-dimensional densities were the limiting tail densities of the corresponding marginals, although I relied heuristically on this fact for estimation.

In parallel, I was following a lecture by Thomas Mikosch on regular variation which later incited me to rely on this notion for justifying the statement above. Once arrived in Oxford, I started to read Bingham et al. [1989], and Lauritzen [1996] with the help of my supervisor Robin Evans. I was fascinated by some results on regularly varying functions such as Karamata's theorem and the representation theorem, which characterize the behavior of certain functions evaluated far from the origin. I started to clarify the relation between regular variation for functions and distributions as presented in Chapter 4 before focusing on the multivariate case. I was quite dissatisfied with the notion of multivariate regular variation in Resnick [1987] because the two previous theorems were not generalized and, in trying to extend them, I fell on the notion of one-component regular variation which seemed fundamental and for which Karamata's theorem and the representation theorem find a natural generalization in the multivariate case as shown in Chapter 4.

My progress in formulating a valid limiting tail density that factorizes had been quite unsatisfactory, when Robin put his finger on the issue: the limiting density is defined on the truncated cone $C_{||\cdot||_\infty} = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_\infty \geq 1\}$, whereas conditional independence is only meaningful on a product space. This rapidly enabled us to formulate a factorization of the censored limiting tail density and to enforce it by assuming asymptotic conditional independence between some pairs of marginals, see Proposition 5.2.1.

During a poster session at the department, I presented the modeling of the extreme river flow data in Asadi et al. [2015] using graphical models, which later became Chapter 3. I found it to be a rather promising illustration because the selected graph recovers almost exactly the river network and the model describes tail dependence relatively well in this case.

As I wanted to gain experience in data analysis, I pursued an idea based on the generalization of regular variation for various decays in Chapter 4: instead of studying the limit of $t^{-1}X \mid X \geq t$, one can consider more general forms of rescaling, e.g. $X - t \mid X \geq t$. I thus reviewed a series of decays such as $\exp(-\alpha x^p)$ and more complex constructions, and fitted them to several data sets including

the extreme flow data, the word frequency data in Chapter 2 and the website hits data in Chapter 6. The Pareto IV distribution appeared to fit the website data quite well (Section 6.1). Nevertheless, my initial goal was not reached because the method based on the generalized Pareto distribution (GPD) was the most robust for estimating extreme quantiles.

I thus turned my attention to the assumption underlying the GPD method: the fact that a distribution belongs to the maximum domain of attraction. As the website hits and word frequency data consist of discrete values, I questioned the use of the GPD, a continuous distribution, for modeling discrete data and started to search for general rescaling operations from $\mathbb{N}$ to $\mathbb{N}$. My attempts were unfruitful and I opted for the formulation in Chapter 2 that considers a discrete random variable X as the discretization of a random variable whose distribution is in the MDA (a condition that I call D-MDA). Under this condition, I could motivate — without an exact justification — the approximation of the tail distribution of X by a discrete generalized Pareto distribution (D-GPD).

Once in New York for a second visit in the fall of 2015, this loose justification was immediately challenged by Richard and Gennady Samorodnitsy and I pursued the experiments in Section 2.3 on extreme quantile estimation to bring some evidence. I realized that the D-GPD method is essentially the same as the GPD method, except when the data have many tied observations, in which case the D-GPD works better. I then compared the D-GPD to several discrete distributions: it outperformed all of them except the Zipf–Mandelbrot distribution, which delivered similar estimates as the D-GPD.

When I came back to Oxford, I read Shimura [2012] and relied on his ideas to ground the use of the D-GPD for modeling tails of distributions in D-MDA (Section 2.2). I then noticed that an alternative assumption was that the probability mass function (instead of the survival function) was in D-MDA, which motivates the use of a generalized Zipf distribution (GZD). I was pleased to fall back on a variant of the Zipf–Mandelbrot distribution because I had not been able to outperform it in the experimental part and it is commonly used for modeling various data

sets, without, to my knowledge, having been proposed in extreme value theory. I mention that I only compared the D-GPD, GZD and GPD methods empirically and future work could fill this gap by bringing theoretical contributions, relying for instance on second order conditions [Alves et al., 2007].

Regarding the website hits, I had started to model their dependence structure by considering bivariate pairs. I spent some time searching for a simple discrete bivariate distribution but eventually turned my attention to bivariate copulas, treating non-zero values as continuous and zero values as censored observations. As the Gaussian and Student copulas delivered the best fit to pairs of website hits, it was straightforward to extend the model to larger dimensions before applying a Gaussian lasso to shrink the inverse covariance matrix, following the ideas of Friedman et al. [2008] and Ravikumar et al. [2011].

While I was writing up the introduction of this thesis, I noticed an unexpected issue with the fact that a censored limiting tail density factorizes: the requirement that the density is homogeneous imposes additional involved constraints on the low-dimensional functions (see Example A.3.2). This difficulty does not appear when working with the uncensored limiting tail densities: they can easily be constructed from low-dimensional functions (see Example A.3.2).

In the need for concrete examples, I examined the density of the Hüsler–Reiss exponent measure more closely. The latter can be expressed in terms of a multivariate log-normal density on $C_k^+ = \{\boldsymbol{x} \in \mathbb{R}_+^d : |\boldsymbol{x}_k| \geq 1\}$ for every $k = 1, \ldots, d$, and thus, similarly to the Gaussian graphical model, can factorize non-trivially on each of these sets. More generally, the limiting tail density on $C_{||\cdot||_\infty}$ can equivalently be expressed on C_k for all k, and conditional independence is meaningful on C_k since they are product spaces. It was thus sensible to prove Theorem 5.1.2, revealing the relation between asymptotic conditional independence relations of the form $X_i \tilde{\perp\!\!\!\perp} X_j \mid \boldsymbol{X}_{-i,j}$, $|X_k| \geq 1$ for all k, and a factorization of the limiting tail density as a product of low-dimensional homogeneous densities h_S. As believed in my Master's thesis, h_S truly corresponds to the limiting tail density of $\boldsymbol{X}_S$ under these assumptions.

I end this discussion by suggesting future research in the context of graphical modeling of extremes. Two distinct approaches were presented in Section 1.3: the first restricts its attention to graphs with a simple structure such as trees. It has the benefit of estimating the low-dimensional densities using a heterogeneous class of models, but it does not provide direct access to conditional distributions. An extension of the analysis on extreme rivers in Chapter 3 could select a graph with cliques of size less than three and combine models capturing three-way interactions to determine the joint distribution. Would such a graphical model outperform the tree and censored Student graphical models that are only based on pairwise interactions? Besides that, possible improvements could arise if the graph is selected using higher order depence summarizes such as the extremal coefficients in Schlather and Tawn [2003].

The second approach consists in modeling the joint distribution using a parametric family that factorizes when its parameters take certain values. Such a framework is very convenient as it allows graph selection and parameter estimation to be performed at once. An instance is the Hüsler–Reiss model which has thus the potential to tackle high-dimensional problems and could be relied upon to provide statistical tests for asymptotic conditional independence.

This thesis focused on combining the two fields of multivariate extreme value theory and graphical models from a theoretical point of view. The relevance of the formulation presented here remains to be assessed for pratical applications. Rigorous comparison of existing methods should be untertaken, such as comparing estimates of $\Pr(X_1 \geq y_1 \mid \boldsymbol{X}_{-1} = \boldsymbol{x}_{-1})$, for some large y_1, in a similar manner to what has been done in the experimental part of Chapter 2 in the univariate case.

A Appendix

A.1 Probability Theory

This section, mainly based on Klenke [2014], presents some definitions and results in probability theory.

A.1.1 Measure Theory

In what follows, we introduce fundamental notions such as a measure, a density with respect to a measure and a random variable. Let Ω be some non-empty set and 2^Ω its power set.

Definition A.1.1 (σ-Algebra)**.** A *σ-algebra* on Ω is a collection $\Sigma \subset 2^\Omega$ such that

- $\Omega \in \Sigma$,
- if $A \in \Sigma$, then $\Omega \setminus B \in \Sigma$ (closed under complementation),
- if $A_1, A_2, \ldots \in \Sigma$, then $\cup_{i \in \mathbb{N}} A_i \in \Sigma$ (closed under countable unions).

A σ-algebra is also closed under countable intersections.

For example, $\{\Omega, \emptyset\}$ is the smallest σ-algebra on Ω and 2^Ω is the largest.

Definition A.1.2 (Dynkin's System)**.** A *Dynkin's system* is a collection $\mathcal{D} \subset 2^\Omega$ satisfying

- $\Omega \in \mathcal{D}$,
- if $A, B \in \mathcal{D}$ and $B \subset A$, then $A \setminus B \in \mathcal{D}$,
- if $A_1, A_2, \ldots \in \mathcal{D}$, then $\cup_{i \in \mathbb{N}} A_i \in \mathcal{D}$.

Definition A.1.3. A σ-algebra $\sigma(\mathcal{A})$ is *generated* by $\mathcal{A}$ when it is the smallest σ-algebra such that $\mathcal{A} \subset \sigma(\mathcal{A})$. Similarly, $\delta(\mathcal{A})$ denotes the smallest Dynkin's system containing $\mathcal{A}$.

Proposition A.1.4 (Dynkin's Proposition)**.** *Let $\mathcal{A} \subset 2^\Omega$ be a collection of sets closed under finite intersection. Then,*

$$\sigma(\mathcal{A}) = \delta(\mathcal{A}).$$

Definition A.1.5 (Topology)**.** A collection $\tau \subset \Omega$ is a *topology* on Ω if

- $\emptyset$, $\Omega \in \tau$,
- τ is closed under any unions (non necessarily countable),
- τ is closed under finite intersections.

The pair (Ω, τ) is called a topological space, the sets in τ are the *open sets* and their complementary with respect to Ω are the *closed sets*. A topological space is *Hausdorff* if, for any $x, y \in \Omega$, $x \neq y$, there exist $U, V \in \tau$ such that $x \in U$, $y \in V$ and $U \cap V = \emptyset$.

Definition A.1.6. Let (Ω, τ) be a topological space. Then the *Borel σ-algebra* is

$$\mathcal{B}(\Omega) := \sigma(\tau).$$

The sets in $\mathcal{B}(\Omega)$ are called the Borels.

For example, $\sigma(\{A \subset \mathbb{R} : A \text{ is an open set}\}) = \mathcal{B}(\mathbb{R}^n)$ and $\sigma(\{(\infty, x] : x \in \mathbb{Q}^n\}) = \mathcal{B}(\mathbb{R}^n)$.

Let $\Sigma \subset 2^\Omega$ be a σ-algebra.

Definition A.1.7 (Measure)**.** A function $\mu : \Sigma \to [0, \infty]$ is a *measure* if it is

- non-negative: $\mu(A) \geq 0$ for any $A \in \Sigma$,
- countable additive: for any countable collection of pairwise disjoint sets $\{A_i\}_{i=1}^{\infty} \subset \Sigma$,
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i),$$
- $\mu(\emptyset) = 0$.

The sets in Σ are called the *measurable sets* and the pair (Ω, Σ) a *measurable space*. The measure μ is called a *finite measure* if $\mu(\Omega) < \infty$ and a *probability measure* if $\mu(\Omega) = 1$. The triple (Ω, Σ, μ) is referred to as a measure space, or a probability space when μ is a probability measure.

We denote the space of finite and probability measures on (Ω, Σ) by $\mathcal{M}_f(\Omega)$ and $\mathcal{M}_1(\Omega)$ respectively. Examples of measures are:

- the *counting measure* defined as follows: $\mu(A)$ is the number of elements in A,
- the *Lebesgue measure* λ on $\mathbb{R}$ which is invariant by translation and satisfies $\lambda([0, 1]) = 1$,
- the *Dirac measure* δ_x defined as follows: $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise.

Let μ be a measure on (Ω, Σ).

Proposition A.1.8 (Properties of a Measure)**.** *It holds that*

- *μ is monotonic: if A, B are measurable and $A \subseteq B$, then $\mu(A) \leq \mu(B)$,*
- *μ is continuous from below: for any measurable sets $A_1, A_2, \ldots$ such that $A_i \subseteq A_{i+1}$, $\forall i$, then $\mu(\cup_{i \in \mathbb{N}} A_i) = \lim_{i \to \infty} \mu(A_i)$,*
- *μ is continuous from above: for any measurable sets $A_1, A_2, \ldots$ such that $A_{i+1} \subseteq A_i$, $\forall i$, and $\mu(A_i) < \infty$ for some i, then $\mu(\cap_{i \in \mathbb{N}} A_i) = \lim_{i \to \infty} \mu(A_i)$.*

Condition $\mu(A_i) < \infty$ for some i is necessary: consider for example the measure space $(\mathbb{N}, 2^{\mathbb{N}}, \mu)$ with $\mu(A) = \sum_{w \in A} 1$. We find $A_n = \{n, n+1, \ldots\} \downarrow \emptyset$ but $\mu(A_n) = \infty$ and $\mu(\emptyset) = 0$.

Definition A.1.9 (Strictly Positive Measure)**.** Let $\{\Omega, \mathcal{B}(\Omega)\}$ be a measurable space where the Borel σ-algebra was generated from the open sets of a Hausdorff topological space (Ω, τ). A measure μ on this space is said to be *strictly positive* if

$$\mu(U) > 0 \text{ for any } U \in \tau,\, U \neq \emptyset.$$

Definition A.1.10 (Cumulative Distribution Function)**.** A non-decreasing, right-continuous function $F : \mathbb{R} \to [0, 1]$ satisfying

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1,$$

is a *cumulative distribution function*, or cdf. A *survival function* is defined as $\bar{F} := 1 - F$.

Since a cdf is a non-decreasing function, it has at most countably many discontinuities. If $\{\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu\}$ is a probability space, then $F_\mu : x \mapsto \mu(-\infty, x]$ is called the cdf of μ.

Lemma A.1.11 (Measure on Intersection-Closed Generator)**.** *Let $(\mathbb{R}, \Sigma, \mu)$ be a probability space. If $\mathcal{E} \subset \Sigma$ is closed under intersection and $\sigma(\mathcal{E}) = \Sigma$, then μ is uniquely determined by the values $\mu(E)$ for $E \in \mathcal{E}$.*

For instance, $\mathcal{E} = \{(-\infty, x] : x \in \mathbb{R}\}$ is closed under intersection and $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$, hence, any probability measure μ on $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ is determined by its cdf $F_\mu(x) = \mu(-\infty, x]$ for $x \in \mathbb{R}$.

Proposition A.1.12 (Bijection between Measures and Cumulative Distribution Functions)**.** *The mapping $\mu \mapsto F_\mu$ is a bijection from the set of probability measures on $\mathcal{M}_1\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ to the set of cumulative distribution functions.*

Definition A.1.13 (Random Variable)**.** Let $(\Omega, \Sigma, \Pr)$ be a probability space and $(E, \mathcal{E})$ a measurable space. *A random variable* is a function $X : \Omega \to E$ which is *measurable*, i.e., $\forall A \in \mathcal{E}$, the preimage $X^{-1}(A) \in \mathcal{E}$.

We use the notation $X^{-1}(A) = \{X \in A\} = \{w \in \Omega : X(w) \in A\}$. Unless specified otherwise, $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a random vector with values in $(E, \mathcal{E}) = \{\mathbb{R}^d, \mathcal{B}(R^d)\}$ and X is a random variable with values in $(E, \mathcal{E}) = \{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$. In this case, the probability measure $\mu_X := \Pr \circ X^{-1}$ is called the *distribution of* X, $F_X(x) = \Pr(X \leq x)$ is the *cumulative distribution function (cdf)* of X and we say that X follows the distribution μ_X, written $X \sim \mu_X$.

We say that $X_1, X_2, \ldots$ are *identically distributed* if $\mu_{X_i} = \mu_{X_j}$ for all $i, j \in \mathbb{N}$, also written $X_i \overset{d}{=} X_j$.

Let μ be a measure on $(E, \mathcal{E})$. The *integral* w.r.t. to a measure μ is first defined for simple functions of the form $f(x) = \sum_{i=1}^{k} \alpha_i 1_{A_i}(x)$ for $\alpha_i \in (0, \infty)$, $\forall i$, where $1_A(x)$ is 1 if $x \in A$ and 0 otherwise:

$$\int f d\mu := \sum_{i=1}^{k} \alpha_i \mu(A_i).$$

Second, it is defined for any measurable function $f : E \to [0, \infty)$:

$$\int f d\mu := \sup \left\{ \int g d\mu : g \text{ is a simple function s.t. } g \leq f \right\}.$$

Finally, the integral for any *integrable function*, i.e., functions of the set

$$\mathcal{L}^1(\mu) = \{f : E \to \mathbb{R} \cup \{\pm\infty\} \text{ s.t. } f \text{ is measurable and } \int |f| d\mu < \infty\},$$

is

$$\int f d\mu := \int f^+ d\mu - \int f^- d\mu,$$

where $f^+(x) = \max\{f(x), 0\}$ and $f^-(x) = \max\{-f(x), 0\}$ are the positive and negative parts respectively. We write $\int_A f d\mu := \int (f 1_A) d\mu$ for $A \in \mathcal{E}$. In addition, we say that $f \leq g$ almost everywhere, abbreviated "a.e." or "μ-a.e.", if $\mu(\{f > g\}) = 0$.

The integral satisfies the following properties: for any $f, g \in \mathcal{L}^1(\mu)$,

- (Monotonicity) if $f \leq g$ a.e., then $\int f d\mu \leq \int g d\mu$,
  in particular, if $f = g$ a.e., then $\int f d\mu = \int g d\mu$,
- (Triangular inequality) $|\int f d\mu| \leq \int |f| d\mu$,

- (Linearity) if $a, b \in \mathbb{R}$, then $af + bg \in \mathcal{L}^1(\mu)$, and

$$\int (af + bg)d\mu = a \int f d\mu + b \int g d\mu.$$

For any $1 \leq p < \infty$, we define

$$\mathcal{L}^p(\mu) := \{f : E \to \mathbb{R} \cup \{\pm\infty\} \text{ s.t. } f \text{ is measurable and } \left(\int |f|^p d\mu\right)^{1/p} < \infty\}.$$

Definition A.1.14 (Density w.r.t. a Measure)**.** Let μ be a measure on the measurable space $(E, \mathcal{E})$ and $f : E \to [0, \infty)$ be measurable. We say that the measure ν defined by

$$\nu(A) := \int_A f d\mu, \quad A \in \mathcal{E}, \tag{A.1}$$

admits a *density* f w.r.t. μ. If a random variable $X \sim \nu$, then f is called the density of X.

The expression in (A.1) is a well-defined measure.

- If μ is the counting measure on $\mathbb{N}$, we call f a *probability mass*. For example, the Poisson distribution has probability mass function $f(k) = \lambda^k e^{-\lambda}/k!$ for $k \in \mathbb{N}$.

- If μ is the Lebesgue measure on $E = \mathbb{R}$, we call f a *probability density*. For instance, the normal distribution $\nu = \mathcal{N}$ has probability density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

- If a random variable X has a cdf F which is differentiable, then μ_X has admits a probability density $f = F'$ w.r.t. Lebesgue.

Proposition A.1.15 (Probability Densities and Cdfs)**.** *A distribution μ has probability density f if and only if its cdf F is absolutely continuous. In this case, F is almost everywhere differentiable and its derivative coincides with f a.e.*

Definition A.1.16 (Mean and Variance)**.** The *mean* of a random vector $X \in \mathcal{L}^1(\mu)$ is

$$E(X) := \int X d\mu.$$

If $X \sim \mu_X$, then $E(X) := \int x d\mu_X$. If X admits a probability density $f_X \in \mathcal{L}^1$, then $E(X) = \int x f_X(x) dx$.

If $X \in \mathcal{L}^2(\mu)$, its *variance* is $\mathrm{Var}(X) := E(X^2) - E(X)^2$. If $X, Y \in \mathcal{L}^2(\mu)$, the *covariance* of X and Y is

$$\mathrm{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}].$$

A.1.2 Convergences

We state here several results such as monotone convergence, dominated convergence, Scheffé's lemma, and define weak convergence between random variables.

Proposition A.1.17 (Monotone Convergence)**.** *Let $\{f_i\}_{i\in\mathbb{N}}$ be a sequence of measurable functions in $\mathcal{L}^1(\mu)$ such that $0 \le f_i(x) \le f_{i+1}(x)$, $\forall x, \forall i$ and $f_n \to f$ a.e. for some $f \in \mathcal{L}^1(\mu)$. Then,*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

Proposition A.1.18 (Dominated Convergence)**.** *Let f be measurable and $\{f_i\}_{i\in\mathbb{N}}$ be a sequence in $\mathcal{L}^1(\mu)$ such that $f_n \to f$ a.e. If there exists $0 \le g \in \mathcal{L}^1(\mu)$ such that $|f_n| \le g$ a.e. $\forall n$, then $f \in \mathcal{L}^1(\mu)$ and*

$$\int f_n d\mu \to \int f d\mu.$$

Let (E, τ) be a topological space with Borel σ-algebra $\mathcal{E} = \mathcal{B}(E) := \sigma(\tau)$ and a complete metric d.

Definition A.1.19 (Radon Measure)**.** A measure μ is called *Radon measure* if

- (σ-finite) there exists $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{E}$ such that $E = \bigcup_{i\in\mathbb{N}} \mu(A_i)$ and $\mu(A_i) < \infty$ $\forall i$,
- $\forall x \in E$, there exists an open set U containing x such that $\mu(U) < \infty$,

- $\mu(A) = \sup\{\mu(K) : K \subseteq A \text{ and } K \text{ is compact }\}, \forall A \in \mathcal{E}$.

The set of Radon measures on $(E, \mathcal{E})$ is denoted by $\mathcal{M}(E)$. $\mathcal{M}_1(E)$ is the set of Radon measures such that $\mu(E) = 1$.

Definition A.1.20 (Vague Convergence)**.** If $\{\mu_i\}_{i\in\mathbb{N}}$ and μ are some Radon measures on $(E, \mathcal{E})$, the sequence μ_t is said to *converge vaguely* to μ if

$$\int f d\mu_t \to \int f d\mu,$$

for any $f : E \to \mathbb{R}$ continuous and with compact support. We denote it by $\mu_t \overset{v}{\Rightarrow} \mu$.

Definition A.1.21 (Weak Convergence)**.** If $\{\mu_i\}_{i\in\mathbb{N}}$ and μ are some finite measures on $(E, \mathcal{E})$, the sequence μ_t is said to *converge weakly* to μ if

$$\int f d\mu_t \to \int f d\mu,$$

for any $f : E \to \mathbb{R}$ continuous and bounded. We denote it by $\mu_t \overset{w}{\Rightarrow} \mu$.

Proposition A.1.22 (Portemanteau Theorem)**.** *Let $\{\mu_i\}_{i\in\mathbb{N}}$ be probability measures on $(E, \mathcal{E})$. The following are equivalent.*

- $\mu_n \overset{w}{\Rightarrow} \mu$,
- $\mu_n(A) \to \mu(A)$ *for all* $A \in \mathcal{B}(E)$ *s.t.* $\mu(\partial A) = 0$.

Definition A.1.23 (Weak Convergence of Cdfs)**.** Let $\{F_n\}_{n=1}^{\infty}$ and F be some cdfs on $\mathbb{R}$. Then, F_n converges *strongly* against F, written $F_n \to F$, if

$$\lim_{n\to\infty} F_n(x) = F(x), \quad x \in \mathbb{R}.$$

It *converges in distribution* or *weakly* against F, written $F_n \xrightarrow{d} F$, if

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ for any continuity point } x \text{ of } F.$$

Clearly, strong convergence implies convergence in distribution.

Proposition A.1.24 (Weak Convergence Equivalence)**.** *Let $\{F_n\}_{n=1}^{\infty}$ and F be some cdfs corresponding to the probability measures $\{\mu_n\}_{i\in\mathbb{N}}$ and μ on $\mathbb{R}$. It holds*

$$F_{\mu_n} \overset{d}{\Longrightarrow} F_\mu \quad \textit{if and only if} \quad \mu_n \overset{w}{\Longrightarrow} \mu.$$

Let $\{X_i\}_{i\in\mathbb{N}}$ and X be random variables with values in E. We say that X_n *converges in distribution* to X, written $X_n \xrightarrow{d} X$, if $\mu_{X_n} \overset{w}{\Rightarrow} \mu_X$, i.e., the distributions of X_n converge weakly. We sometimes denote this by $X_n \xrightarrow{d} \mu_X$. Convergence in distribution is also called convergence in law or weak convergence and denoted by "$\overset{\mathcal{L}}{\Rightarrow}$", or "$\overset{w}{\Rightarrow}$".

Proposition A.1.25 (Continuous Mapping Theorem)**.** *Let (E,d) and (E',d') be two metric spaces, $f : E \to E'$ be a measurable function and U_f be the set of discontinuities of f. Suppose that $\{\mu_i\}_{i\in\mathbb{N}}$ and μ are some probability measures and that $\mu(U_f) = 0$. If $\mu_n \overset{w}{\Rightarrow} \mu$, then $\mu_n \circ f^{-1} \overset{w}{\Rightarrow} \mu \circ f^{-1}$.*

In particular, if $\{X_i\}_{i\in\mathbb{N}}$ and X are random variables with values in E, $\Pr(X \in U_f) = 0$ and $X_n \xrightarrow{d} X$, then $f(X_n) \xrightarrow{d} f(X)$.

Example A.1.26 (Strong Limit of Measures is Not Necessarily a Measure)**.** Consider the cdf defined by $F_n(x) := F(x+n)$ for any $n \in \mathbb{N}$. Then, $F_n(x) \to 1$, which is not a cdf since it is required that $\lim_{x\to-\infty} F_n(x) = 0$. This means that a pointwise limit of a sequence of measure is not necessarily a measure. A different way of seeing this is to consider the sequence $\mu_n = \delta_{-n}$, where

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else .} \end{cases}$$

Then, $\delta_{-n} \to \delta_{-\infty}$ pointwise, where

$$\delta_{-\infty} = \begin{cases} 1 & \text{if } \inf A = -\infty, \\ 0 & \text{else .} \end{cases}$$

However, $\delta_{-\infty}$ is not a measure because it is not σ-additive: namely $\{[-k,-k+1)\}_{k\in\mathbb{N}}$ is a sequence of disjoint sets but

$$\delta_{-\infty}\left\{\bigcup_{k\in\mathbb{N}}[-k,-k+1)\right\} = 1 \neq 0 = \sum_{k\in\mathbb{N}} \delta_{-\infty}\left\{[-k,-k+1)\right\}.$$

Example A.1.27 (Weak Limit of Measures is Not Necessarily a Measure)**.** Consider the cdf defined by $F_n(x) := F(x-n)$ for any $n \in \mathbb{N}$. It holds $F_n(x) \to 0$ pointwise.

By contradiction, suppose that there exists a cdf F such that $F_n \xrightarrow{d} F$. Let $x \to \infty$ be a sequence of continuity points of F. A contradiction is found as

$$1 = \lim_{x\to\infty} F(x) = \lim_{x\to\infty} \lim_{n\to\infty} F_n(x) = 0.$$

Example A.1.28 (Weak Convergence vs Strong Convergence)**.** Let $\delta_{n^{-1}}$ be the Dirac measure at n^{-1} and F_n its cdf. We find $F_n \to F$ for all $x \neq 0$, where F is the corresponding to δ_0, and thus $F_n \xrightarrow{d} F_{\delta_0}$ because 0 is a discontinuity point. As a comparison, $\delta_{n^{-1}}$ converges strongly to

$$\delta_{0_+} = \begin{cases} 1 & \text{if } 0 \in \bar{A}, \\ 0 & \text{else,} \end{cases}$$

which is not a measure.

Proposition A.1.29 (Scheffé's Lemma)**.** *If f_n converges pointwise to a probability density f, then f_n converges weakly to f and, equivalently, the corresponding cdfs F_n converge weakly to F.*

Definition A.1.30 (Convergence in Probability)**.** A sequence $\{X_i\}_{i\in\mathbb{N}}$ *converges in probability* to X if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| \geq \epsilon) = 0.$$

Convergence in probability implies weak convergence. The opposite, however, is not true: let $X \sim \mathcal{N}(0, 1)$ and $X_i = -X$ for $i \in \mathbb{N}$. Then, $X_n \xrightarrow{d} X$ but $\Pr(|X_n - X| \geq \epsilon) = \Pr(|X| \geq \epsilon/2) \neq 0$, so X_n does not converge in probability to X.

A.1.3 Independence and Conditional Independence

We define independence and conditional independence for random variables and briefly introduce graphical models [Lauritzen, 1996]. Let $(\Omega, \Sigma, \Pr)$ be a probability space.

Definition A.1.31 (Mutual Independence)**.** A collection of sets $\{A_i\}_{i=1}^d \subset \Sigma$ is *mutually independent* if

$$\Pr\left(\bigcap_{i=1}^d A_i\right) = \prod_{i=1}^d \Pr(A_i).$$

Let $\boldsymbol{X}$ be a random vector with values in $E \subseteq \mathbb{R}^d$ with probability distribution $\mu_{\boldsymbol{X}}$ and joint cdf $F_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr(\boldsymbol{X} \leq \boldsymbol{x})$.

Proposition A.1.32 (Mutual Independence)**.** *The following are equivalent.*

- $F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^d F_{X_i}(x_i)$, $\forall \boldsymbol{x}$,
- $\{X_1 \leq x_1\}, \ldots, \{X_d \leq x_d\}$ *are mutually independent events for all* $\boldsymbol{x} \in \mathbb{R}^d$,
- $\{X_1 \in A_1\}, \ldots, \{X_d \in A_d\}$ *are mutually independent events for all* $\{A_i\}_{i=1}^d \subseteq \mathcal{B}(\mathbb{R}^d)$.

In this case, we write $X_i \perp\!\!\!\perp X_j$, $\forall i, j$.

Let μ_0 be some base measure, e.g., the Lebesgue measure on $\mathbb{R}^d$ or the counting measure on $\mathbb{N}^d$, and suppose further that $\mu_{\boldsymbol{X}}$ admits a continuous density $f_{\boldsymbol{X}}$ w.r.t. μ_0. (All functions on a discrete space are considered to be continuous.) In this case, the following can be added to the list of equivalent conditions in Proposition (A.1.32): $f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^d f_{X_i}(x_i)$, $\forall \boldsymbol{x}$.

Definition A.1.33 (Conditional Independence)**.** Let A, B and C be disjoint subsets of $\{1, \ldots, d\}$. Two marginals $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are called *conditionally independent* with respect to $\boldsymbol{X}_C$, written $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_C$, if a.e.,

$$f_{\boldsymbol{X}_C} f_{\boldsymbol{X}_{A,B,C}} = f_{\boldsymbol{X}_{A,C}} f_{\boldsymbol{X}_{B,C}},$$

or equivalently, $f_{\boldsymbol{X}_{A,B|C}} = f_{\boldsymbol{X}_{A|C}} f_{\boldsymbol{X}_{B|C}}$, a.e., for all $\boldsymbol{x}$ such that $f_{\boldsymbol{X}_C}(\boldsymbol{x}_C) > 0$.

Example A.1.34 (Multivariate Gaussian)**.** Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, i.e., $\boldsymbol{X}$ follows the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ defined by its probability density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}.$$

It holds $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$. Every marginal is also normally distributed: $\boldsymbol{X}_A \sim \mathcal{N}(\boldsymbol{\mu}_A, \Sigma_{AA})$, $\forall A \subset \{1, \ldots, d\}$, using the notation $\Sigma_{AB} := (\Sigma_{ij})_{i \in A, j \in B}$. For any disjoint set A, B, the conditional distributions satisfy

$$\boldsymbol{X}_A \mid \boldsymbol{X}_B = \boldsymbol{x}_B \sim \mathcal{N}\left\{\boldsymbol{\mu}_A + \Sigma_{AB}\Sigma_{BB}^{-1}(\boldsymbol{x}_B - \boldsymbol{\mu}_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right\}.$$

It holds $X_i \perp\!\!\!\perp X_j$ if and only if $\Sigma_{ij} = 0$. Moreover,

$$X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{\{1,\ldots,d\}\setminus\{i,j\}} \text{ if and only if } \Sigma_{ij}^{-1} = 0.$$

Conditional independence does not imply independence because there are covariance matrices satisfying $\Sigma_{ij}^{-1} = 0$ and $\Sigma_{ij} \neq 0$. Independence does not imply conditional independence: if $\Sigma_{ij} = 0$, then X and Y are independent but $X \mid \max(X, Y)$ is not independent of $Y \mid \max(X, Y)$.

Example A.1.35 (Multivariate Student Distribution)**.** A random vector $\boldsymbol{X}$ follows the multivariate Student distribution or t distribution, written $\boldsymbol{X} \sim t_d(\boldsymbol{\mu}, \Sigma, \nu)$, if it admits a probability density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\Gamma\left[(\nu + d)/2\right]}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}|\Sigma|^{1/2}}\left[1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]^{-(\nu+d)/2},$$

for $\boldsymbol{x} \in \mathbb{R}^d$. If $\nu > 1$, $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and if $\nu > 2$, $\mathrm{Cov}(X_i, X_j) = \frac{\nu}{\nu-2}\Sigma_{ij}$. Any marginal also follows a Student distribution: $\boldsymbol{X}_A \sim t_d(\boldsymbol{\mu}_A, \Sigma_{AA}, \nu)$, for $A \subseteq \{1, \ldots, d\}$. Moreover, the conditional distributions are again Student distributed [Ding, 2016]:

$$\begin{aligned}
\boldsymbol{X}_B \mid \boldsymbol{X}_A = \boldsymbol{x}_A &\sim t_{|B|}(\boldsymbol{\mu}_{B|A}, \Sigma_{B|A}, \nu + |A|),\\
\Sigma_{B|A} &= \frac{\nu + \delta_A}{\nu + |A|}\tilde{\Sigma}_{B|A},\\
\delta_A &= (\boldsymbol{x}_A - \boldsymbol{\mu}_A)^T\Sigma_{AA}^{-1}(\boldsymbol{x}_A - \boldsymbol{\mu}_A),\\
\tilde{\Sigma}_{B|A} &= \Sigma_B - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB},\\
\boldsymbol{\mu}_{B|A} &= \boldsymbol{\mu}_A + \Sigma_{BA}\Sigma_{AA}^{-1}(\boldsymbol{x}_A - \boldsymbol{\mu}_A).
\end{aligned}$$

If $\Sigma_{ij}^{-1} = 0$, then X_i and X_j are uncorrelated given $\boldsymbol{X}_{\{1,\ldots,d\}\setminus\{i,j\}}$ [Finegold and Drton, 2011].

Definition A.1.36 (Pairwise Markov Property According to a Graph)**.** A random vector $\boldsymbol{X}$ satisfies the *pairwise Markov property* according to an undirected graph $\mathcal{G} = (V, E)$ with set of nodes $V = \{1, \dots, d\}$ if

$$(i, j) \notin E \quad \Longrightarrow \quad X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{V \setminus \{i,j\}}.$$

For example, if $\boldsymbol{X}$ satisfies the pairwise Markov property according to the graph in Figure A.1, it holds $X_1 \perp\!\!\!\perp X_i \mid X_2$ for $i = 3, 4, 5, 6$, and $X_6 \perp\!\!\!\perp X_i \mid \boldsymbol{X}_{4,5}$ for $i = 1, 2, 3$.

Definition A.1.37 (Clique)**.** Let $\mathcal{G} = (V, E)$ be a graph. A *clique* of $\mathcal{G}$ is a subgraph $\mathcal{G}' = (V', E')$ which is complete, i.e., $(i, j) \in E'$, $\forall i, j \in V'$ such that $i \neq j$. A *maximal clique* of $\mathcal{G}$ is a clique that cannot be extended by including more vertices of V.

For example, the maximal cliques of the graph in Figure A.1 are $\{1, 2\}$, $\{2, 3, 4, 5\}$ and $\{4, 5, 6\}$. The set $\{2, 3, 4\}$ is a clique but not a maximal clique.

Proposition A.1.38 (Hammersley–Clifford Theorem)**.** *Suppose that $f_{\boldsymbol{X}}$ is positive. Then, $\boldsymbol{X}$ satisfies the pairwise Markov property according to a graph $\mathcal{G}$ if and only if $f_{\boldsymbol{X}}$* factorizes *according to $\mathcal{G}$, i.e., there exist $k > 0$, and functions ϕ_C such that*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = k^{-1} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{x}_C), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{A.2}$$

where $\mathcal{C}$ denotes the set of maximal cliques of the graph $\mathcal{G}$.

The functions ϕ_C are not unique.

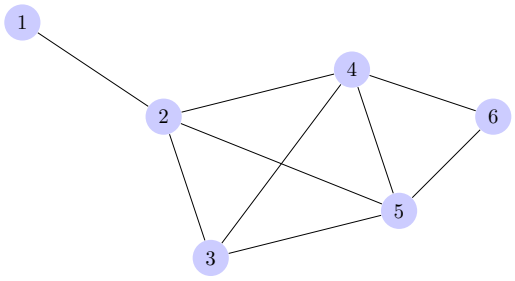

Figure A.1: A decomposable graph.

Under certain constraints on the graph, the functions ϕ_C can be expressed in terms of the marginals of $\boldsymbol{X}$. This requires the notion of decomposability.

Definition A.1.39 (Decomposition)**.** A triple (A, B, C) of disjoints subsets of $V = \{1, \dots, d\}$ is said to form a *decomposition* of $\mathcal{G}$ into the components $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$ if

- $V = A \cup B \cup C$,
- C separates A from B (i.e., every path from A to B intersects C),
- C is a complete subset.

The decomposition is called *proper* if A and B are both non empty. A graph $\mathcal{G}$ is *decomposable* if it is complete or if there exists a proper decomposition (A, B, C) into decomposable subgraphs $\mathcal{G}_{A\cup C}$ and $\mathcal{G}_{B\cup C}$.

For instance, $(\{2, 3, 4, 5\}, \{4, 5, 6\}, \{4, 5\})$ is a proper decomposition of the graph in Figure A.1.

If $\mathcal{G}$ is decomposable, then (A.2) can be written

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\prod_{C\in\mathcal{C}} f_{\boldsymbol{X}_C}(\boldsymbol{x}_C)}{\prod_{D\in\mathcal{D}} f_{\boldsymbol{X}_D}(\boldsymbol{x}_D)}, \tag{A.3}$$

where $\mathcal{C}$ is the set of the maximal cliques of $\mathcal{G}$ and $\mathcal{D} = \{D_1, \dots, D_m\}$ for $m = |\mathcal{C}|$ is a multiset containing intersections between some of the maximal cliques called *separator sets*, which we now define.

As the graph is decomposable, it can be shown that there exists at least a maximal clique C_1 that intersects only one other maximal clique C_j. Let $D_1 := C_1 \cap C_j$. Now consider the subgraph $\mathcal{G}' = \mathcal{G} \setminus (C_1 \setminus D_1)$. Again, there exists at least one maximal clique C_2 in $\mathcal{G}'$ such that $D_2 := C_2 \cap C_j \neq \emptyset$ for only one maximal clique C_j of $\mathcal{G}'$. Repeating the procedure gives an ordering of the maximal cliques and separator sets

$$C_1, \dots, C_m, \quad D_1, \dots, D_{m-1}, D_m = \emptyset. \tag{A.4}$$

Notice that the same separator set can appear several times in the sequence. The factorization in (A.3) can equivalently be written as

$$f_{\boldsymbol{X}} = \prod_{i=1}^{m} f_{C_i\setminus D_i \,|\, D_i},$$

which can be used to sample from $\boldsymbol{X}$.

For example, a possible sequence of maximal cliques and intersecting sets for the graph in Figure A.1 is

$$\mathcal{C} = (\{1,2\},\{2,3,4,5\},\{4,5,6\}), \quad \mathcal{D} = (\{2\},\{4,5\},\emptyset).$$

If $f_{\boldsymbol{X}}$ factorizes w.r.t. this graph, one can write

$$f_{\boldsymbol{X}} = \frac{f_{\boldsymbol{X}_{12}} f_{\boldsymbol{X}_{2345}} f_{\boldsymbol{X}_{456}}}{f_{\boldsymbol{X}_2} f_{\boldsymbol{X}_{45}}} = f_{\boldsymbol{X}_{1|2}} f_{\boldsymbol{X}_{23|45}} f_{\boldsymbol{X}_{456}}.$$

Definition A.1.40 (Kruskal's Algorithm)**.** Let $\boldsymbol{w} = \{w_{ij}\}_{(i,j)\in V\times V}$ be a collection of weights for $V = \{1,\dots,d\}$. *Kruskal's algorithm* computes a minimal spanning tree, i.e., a tree $\mathcal{G} = (V,E)$ such that the sum of the weights w_{ij} over its edges $(i,j) \in E$ is minimized. More precisely, it solves the problem

$$\underset{\text{tree } \mathcal{G}=(V,E)}{\arg\min} \quad \sum_{(i,j)\in E} w_{ij}. \tag{A.5}$$

The algorithm starts with $E = \emptyset$ and updates E and $\boldsymbol{w}$ as follows:

- select the smallest weight w_{ij} in $\boldsymbol{w}$;
- check if $\mathcal{G} = (V, E \cup \{(i,j)\})$ is a tree; if it is, update $E \leftarrow E \cup \{(i,j)\}$,
- remove w_{ij} from $\boldsymbol{w}$;
- end if $|E| = d-1$ and return $\mathcal{G} = (V,E)$ which is a tree satisfying (A.5);

[Kruskal, 1956].

A.2 Statistical Theory

This section gathers some basic notions in statistical theory such as maximum likelihood estimation, confidence intervals, goodness of fit tests and independence tests [Davison, 2003, Panaretos, 2016].

A.2.1 Maximum Likelihood Estimation

Let $\boldsymbol{x} = \{x^{(k)}\}_{k=1}^{n}$ be i.i.d. realizations from a random variable X with parametric probability density f_{θ_0} for $\theta_0 \in \Theta \subseteq \mathbb{R}^p$.

Definition A.2.1 (Likelihood)**.** The *likelihood* is a function $\mathcal{L} : \Theta \to \mathbb{R}$ defined as

$$\mathcal{L}(\theta; \boldsymbol{x}) = \prod_{k=1}^{n} f_\theta(x_k).$$

The notation $\mathcal{L}(\cdot; \boldsymbol{x})$ is used to signal its dependence with the observations. The *log-likelihood* is $\ell = \log \mathcal{L}$.

Definition A.2.2 (Maximum Likelihood Estimator)**.** A *maximum likelihood estimator* $\hat{\theta}$ is

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} \ \ell(\theta; \boldsymbol{x}),$$

if a maximum exists.

A maximum likelihood estimator aims at estimating the true parameter θ_0 from the observations $\boldsymbol{x}$, and may not exist or be unique. It fulfills attractive properties when the parametric density f_θ satisfies the conditions stated below.

Definition A.2.3 (Regularity Conditions I)**.** The regularity conditions I for f_θ are:

- Identification: $\theta \neq \theta_0$ if and only if $f_\theta \neq f_{\theta_0}$.
- Compactness: Θ is compact.
- Continuity: $g(\theta) := f_\theta(X)$ is a.e. continuous on Θ.
- Dominance: there exists a function $B(x)$ which is integrable w.r.t. f_{θ_0} and satisfies $|\log f_\theta(x)| < B(x)$, $\forall \theta \in \Theta$.

For conditions not requiring the parameter space to be compact, see e.g. Anastasiou and Reinert [2017].

Proposition A.2.4 (Consistency of the Maximum Likelihood Estimator)**.** *If f_θ satisfies the regularity conditions I, then the maximum likelihood estimator $\hat{\theta}$ is consistent, i.e.,*

$$\hat{\theta}_{mle} \xrightarrow{p} \theta_0,$$

as $n \to \infty$.

We use the following notation involving the differential operator ∇:

$$\nabla_\theta f_\theta = \left(\frac{\partial}{\partial \theta_1} f_\theta, \ldots, \frac{\partial}{\partial \theta_p} f_\theta \right), \quad \nabla_{\theta\theta} f_\theta = \left(\frac{\partial^2}{\partial \theta_i \theta_j} f_\theta \right)_{1 \le i,j \le p}.$$

Definition A.2.5 (Regularity Conditions II)**.** The regularity conditions II for f_θ are:

- θ_0 lies in the interior of Θ,
- $g(\theta) = f_\theta(x)$ is positive and twice continuously differentiable in a neighborhood N of θ_0,
- $\int \sup_{\theta \in N} ||\nabla_\theta f_\theta(x)|| dx < \infty$ and $\int \sup_{\theta \in N} ||\nabla_{\theta\theta} f_\theta(x)|| dx < \infty$,
- $I = E\left[\nabla_\theta \log f_{\theta_0}(x) \, \{\nabla_\theta \log f_{\theta_0}(x)\}^T \right]$ exists and is nonsingular,
- $E\left\{ \sup_{\theta \in N} ||\nabla_{\theta\theta} \log f_\theta(x)|| \right\} < \infty$.

Proposition A.2.6 (Asymptotic Normality of the Maximum Likelihood Estimator)**.** *If f_θ satisfies the regularity conditions I and II, then the maximum likelihood satisfies*

$$\sqrt{n}(\hat{\theta}_{mle} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}),$$

where I is the Fisher information matrix defined above. In addition,

$$I = E\left\{ -\nabla_{\theta\theta} \log f_{\theta_0}(x) \right\}.$$

Confidence Intervals based on Asymptotic Normality

Consider a quantity q_0 depending on the parameter θ_0 through $q_0 = g(\theta_0)$, where g is a continuously differentiable function. If f_θ satisfies the regularity conditions I and II, then $\hat{q}_{\text{mle}} := g(\hat{\theta}_{\text{mle}})$ is a consistent estimator of q_0 and it holds

$$\sqrt{n}(\hat{q}_{\text{mle}} - q_0) \xrightarrow{d} \mathcal{N}\left\{0, (\nabla_\theta\, q_0)^T\, I^{-1}\, \nabla_\theta\, q_0\right\}. \tag{A.6}$$

When n is large, this can be used to build a confidence intervals for q_0 as we now explain. A consistent estimator of I is the Hessian matrix of the log-likelihood evaluated at $\hat{\theta}$, i.e.,

$$\hat{I} = n^{-1} \sum_{k=1}^{n} \nabla_\theta \log f_{\hat{\theta}}(x)\ \{\nabla_\theta \log f_{\hat{\theta}}(x)\}^T,$$

or alternatively,

$$\hat{I} = -n^{-1} \sum_{k=1}^{n} \nabla_{\theta\theta} \log f_{\hat{\theta}}(x).$$

These two estimators are called outer product of gradients and Hessian estimators respectively, and are typically computed numerically.

At present, we can rely on (A.6) to obtain the approximation

$$\hat{q}_{\text{mle}} \dot{\sim} \mathcal{N}(q_0, \sigma_q^2),$$

for large n, where $\sigma_q^2 = n^{-1}(\nabla_\theta\, q_0)^T\, I\, \nabla_\theta\, q_0$. We can then compute a two-sided confidence interval with approximative level α by replacing I and q_0 by their respective estimates $\hat{I}$ and $\hat{q}_{\text{mle}}$:

$$\hat{q}_{\text{mle}} \pm \hat{\sigma}_q \Phi^{-1}(1 - \alpha/2).$$

This means that the probability that $\hat{q} = \hat{q}(\boldsymbol{x})$ lies in the interval above is approximatively $1 - \alpha$ when n is large.

Example A.2.7 (Confidence intervals for Extreme Quantiles)**.** This example briefly illustrates how to obtain confidence intervals for extreme quantiles and probability of regions far from the origin. We mention that other techniques can deliver more

accurate results such as profile likelihood confidence intervals [Davison and Smith, 1990].

Let q_α be the α-quantile of X and the *return level* of level p defined by $y_p := q_{1-1/p}$. Suppose that $X - u \mid X \geq u$ follows a generalized Pareto distribution (GPD) (or a D-GPD and $u \in \mathbb{N}$) and let us write $p_u = \Pr(X \geq u)$ and $z = (1-\alpha)/p_u$. Then,

$$y_p = y_p(\sigma, \xi) = u + \frac{\sigma}{\xi}\left(z^{-\xi} - 1\right).$$

Moreover,

$$\nabla_{(\sigma,\xi)}\, y_p = \left[\frac{z^{-\xi}-1}{\xi},\ \frac{\sigma}{\xi^2}\left\{1 - z^{-\xi}(1+\xi\log z)\right\}\right].$$

From the asymptotic normality of the maximum likelihood estimator, one finds the approximation

$$\hat{y}_p \,\dot{\sim}\, \mathcal{N}\left\{y_p, \nabla\hat{y}_p^T\, \hat{I}^{-1}\, \nabla\hat{y}_p\right\},$$

where $\hat{I}$ is the estimator defined earlier.

A similar asymptotic relation is derived for $\Pr(X \geq k) = p_u(1+\xi x/\sigma)^{-1/\xi}$ for some $k \geq u$ using

$$\nabla_{(\sigma,\xi)} \Pr(X \geq k) = p_u(1+\xi x/\sigma)^{-1/\xi-1}\left\{\frac{x}{\sigma^2}, -\frac{x}{\sigma\xi} + \frac{(1+\xi x/\sigma)\log(1+\xi x/\sigma)}{\xi^2}\right\}.$$

A.2.2 Goodness of Fit Tests

Definition A.2.8 (Kullback–Leibler Divergence)**.** The *Kullback–Leibler divergence* between two probability densities f and g w.r.t. to a measure μ is

$$d_{\mathrm{KL}}(f||g) = \int_{-\infty}^{\infty} f \log\left(\frac{f}{g}\right) d\mu.$$

The Kullback–Leibler divergence satisfies $d_{\mathrm{KL}}(f||g) \geq 0$; moreover, $d_{\mathrm{KL}}(f||g) = 0$ if and only $f = g$. It is not a metric as it does fulfill the triangle inequality and is not symmetric in general.

Pearson's χ^2-test is a test for goodness of fit. Consider a random variable X with values in a finite set $\mathcal{X}$ and probability mass function p_θ. Suppose that $\boldsymbol{x} = \{x^{(k)}\}_{k=1}^n$ are i.i.d. realizations of X. Then,

$$n\sum_{i\in\mathcal{X}} \frac{\{\hat{p}(i) - p_\theta(i)\}^2}{p_\theta(i)} \underset{n\to\infty}{\dot{\sim}} \chi^2_{|\mathcal{X}|-|\theta|}, \tag{A.7}$$

where $\hat{p}(i) = n_i/n$ and n_i counts the number of observations taking value i. Here, "$\underset{n\to\infty}{\dot{\sim}}$" means that the law of the sequence converges weakly to a χ^2 distribution with degree of freedom $|\mathcal{X}| - |\theta|$.

The G^2-*test* is closely related: the left-hand side in (A.7) is simply replaced by

$$2n \sum_{i\in\mathcal{X}} \hat{p}(i) \log\left\{\frac{\hat{p}(i)}{p_\theta(i)}\right\},$$

which equals $2n\, d_{\mathrm{KL}}(\hat{p}||p_\theta)$.

In practice, the Pearson's χ^2 and G^2-tests are very similar. They are known to be inaccurate when $p_\theta(i)$ is small for some values $i \in \mathcal{X}$.

Let $\boldsymbol{x} = \{x^{(k)}\}_{k=1}^n$ be a sample i.i.d. drawn from a distribution with cumulative distribution function F.

Definition A.2.9 (Empirical Distribution Function)**.** The *empirical distribution function* is

$$\hat{F}(x; \boldsymbol{x}) = n^{-1} \sum_{k=1}^{n} 1_{x^{(k)}\leq x}.$$

Proposition A.2.10 (Kolmogorov Theorem)**.**

$$\sqrt{n} \sup_x |\hat{F}(x; \boldsymbol{x}) - F(x)| \xrightarrow{d} \sup_t |B\{F(t)\}|, \tag{A.8}$$

where $B(t)$ is the Brownian bridge (see e.g. Klenke [2014]).

The *Kolmogorov–Smirnov test* is a goodness of fit test which relies on the fact that (A.8) holds under the null hypothesis that F generated the data. In the continuous case, the test is distribution free, i.e., the distribution under the null does not depend on F, since $\sup_t |B\{F(t)\}| = \sup_{t\in[0,1]} |B(t)|$, which makes it attractive. Two alternative tests known to be more powerful and sensitive in the tails are the Anderson–Darling test and the Cramer–von Mises test. Recall that the power of a test is the probability of rejecting the null hypothesis when the alternative hypothesis is true. The Kolmogorov–Smirnov test also exists for discrete distributions, without being distribution free anymore in this case [Arnold and Emerson, 2011].

A.2.3 Conditional Independence Testing

Let $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \{x^{(k)}, y^{(k)}, \boldsymbol{z}^{(k)}\}_{k=1}^{n}$ be i.i.d. realizations from a random vector $(X, Y, \boldsymbol{Z})$.

The Discrete Case

Let $p_{X,Y,\boldsymbol{Z}}$ be the joint probability mass function of $(X, Y, \boldsymbol{Z})$ which are discrete random variables taking values in finite sets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ respectively.

We start by presenting the G^2-*test* which can be used for testing independence between X and Y. It is based on the mutual information

$$I(X,Y) := d_{\mathrm{KL}}(p_{X,Y}||p_X p_Y) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{X,Y} \log\left(\frac{p_{X,Y}}{p_X p_Y}\right),$$

where d_{KL} is the Kullback–Leibler divergence defined in (A.2.8). Notice that $I(X,Y) \geq 0$ and that $I(X,Y) = 0$ if and only if $X \perp\!\!\!\perp Y$. Under $X \perp\!\!\!\perp Y$, the plug-in estimator of $I(X,Y)$ is asymptotically χ^2 distributed, more precisely

$$2n\hat{I}(X,Y) = 2n\sum_{x,y} \hat{p}_{X,Y}(x,y) \log\frac{\hat{p}_{X,Y}(x,y)}{\hat{p}_X(x)\hat{p}_Y(y)} \underset{n\to\infty}{\overset{\cdot}{\sim}} \chi^2_{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)},$$

where $\hat{p}_{XY}(x,y) = n_{xy}/n$, n_{xy} is the number of time (x,y) is observed, and similarly for $\hat{p}_X$ and $\hat{p}_Y$.

We now turn our attention to testing conditional independence. Consider the conditional mutual information

$$I(X,Y \mid \boldsymbol{Z}) := \mathbb{E}_{\boldsymbol{Z}}\{I(X,Y) \mid \boldsymbol{Z}\} = \sum_{x,y,\boldsymbol{z}} p_{XY\boldsymbol{Z}} \log\frac{p_{XY\boldsymbol{Z}}\, p_{\boldsymbol{Z}}}{p_{X\boldsymbol{Z}}\, p_{Y\boldsymbol{Z}}}.$$

which satisfies $I(X,Y \mid Z) \geq 0$ and $I(X,Y \mid \boldsymbol{Z}) = 0$ exactly when $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$, and thus measures how strongly X and Y are dependent conditionally on $\boldsymbol{Z}$. Under the hypothesis $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$,

$$2n\,\hat{I}(X,Y \mid \boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} \hat{p}_{X,Y,\boldsymbol{Z}} \log\frac{\hat{p}_{X,Y,\boldsymbol{Z}}\,\hat{p}_{\boldsymbol{Z}}}{\hat{p}_{X,\boldsymbol{Z}}\, p_{Y,\boldsymbol{Z}}} \underset{n\to\infty}{\overset{\cdot}{\sim}} \chi^2_{(|\mathcal{X}_X|-1)(|\mathcal{X}_Y|-1)|\mathcal{X}_{\boldsymbol{Z}}|},$$

[Kullback, 1997]. This provides a useful tool for testing conditional independence when d is not particularly large. When d is large, we can get some insight into the dependence structure by computing $\hat{I}(X_i, X_i \mid \boldsymbol{X}_A)$ for lower dimensional sets A.

The Continuous Nonparametric Case

The conditional mutual information $I(X, Y \mid \boldsymbol{Z})$ can be used as a measure of conditional dependence in the continuous case as well by replacing $\hat{p}_{XY\boldsymbol{Z}}$ by kernel density estimators. See e.g. Székely et al. [2007] for testing $X \perp\!\!\!\perp Y$ and Huang [2010], Zhang et al. [2012] for testing $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$.

The Continuous Parametric Case

Suppose that $\boldsymbol{X}$ has probability density $f_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \Theta$. Let $\Omega_0 \subseteq \Omega$ be the set such that $\boldsymbol{\theta} \in \Omega_0$ if and only if $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-\{i,j\}}$. Testing conditional independence corresponds to testing if the true parameter lies in Ω_0. Two common examples are:

- the multivariate Gaussian distribution: if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}} \Leftrightarrow \Sigma^{-1}_{i,j} = 0$,
- the probabilistic Ising model: the probability mass function of $\boldsymbol{X}$ is

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_{\boldsymbol{\theta}}^{-1} \exp\left\{ \sum_{i=1}^{d} \theta_i x_i + \sum_{i,j=1}^{d} \theta_{ij} x_i x_j \right\},$$

  and it holds $\theta_{ij} = 0 \Leftrightarrow X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-\{i,j\}}$.

A.3 Supplementary Material

Example A.3.1 (Useful Quantities for Censored Bivariate Distributions)**.** We provide here the quantities needed to sample from a random vector (Y_1^C, Y_2^C) distributed according to the bivariate distributions on $F_+^2 = (\{0\} \cup [1,\infty))^2$ in Example 3.2.1 and 3.2.2.

We start by the first example. Let $g(x) = \int_0^1 h(x,y)dy$, $G(x) = \int_1^x g(\bar{x})d\bar{x}$, $c = c_{uu} + 2c_{ud}$, for $c_{uu} = \int_1^\infty \int_1^\infty h(x,y)dxdy$, $c_{ud} = \int_1^\infty \int_0^1 h(x,y)dxdy$, and $r(x,y) = \int_1^x h(\bar{x},y)d\bar{x}$. Then, for $x, y \geq 1$,

$$\begin{aligned}
\Pr(Y_1^C = 0) &= p_{00} + (1-p_{00})c^{-1}c_{ud},\\
\Pr(Y_1^C = 0 \mid Y_2^C = y) &= \frac{g(y)}{(c_{uu}+c_{ud})y^{-2}},\\
\Pr(Y_1^C \leq x \mid Y_1^C > 0, Y_2^C = y) &= \frac{r(x,y)}{(c_{uu}+c_{ud})y^{-2} - g(y)},\\
\Pr(Y_1^C = 0 \mid Y_2^C = 0) &= \frac{1}{1+(p_{00}^{-1}-1)c^{-1}c_{ud}},\\
\Pr(Y_1^C \leq x \mid Y_1^C > 0, Y_2^C = 0) &= c_{ud}^{-1}G(x).
\end{aligned}$$

In particular, when h is the exponent measure of the extremal t distribution,

$$\begin{aligned}
h(x,y) &= (x^2 - 2\rho xy + y^2)^{-3/2},\\
g(x) &= x^{-2}\frac{\rho\left(\sqrt{1-2\rho x + x^2} - x\right) + 1}{(1-\rho^2)\sqrt{-2\rho x + x^2 + 1}},\\
G(x) &= \frac{\sqrt{2-2\rho}+\rho}{1-\rho^2} - \frac{\rho + \sqrt{1-2\rho x + x^2}}{1-\rho^2}x^{-1},\\
c_{uu} &= \frac{-2+\sqrt{2-2\rho}}{-1+\rho^2}, \quad c_{ud} = \frac{1}{1+\sqrt{2-2\rho}-\rho},\\
r(x,y) &= \frac{\frac{x}{\sqrt{x^2-2\rho xy+y^2}} - \frac{1}{\sqrt{1-2\rho y+y^2}} - \rho y\left\{\frac{1}{\sqrt{x^2-2\rho xy+y^2}} - \frac{1}{\sqrt{1-2\rho y+y^2}}\right\}}{(1-\rho^2)y^2}
\end{aligned}$$

When h is the exponent measure of the Hüsler–Reiss distribution,

$$
\begin{aligned}
h(x,y) &= x^{-3/2}y^{-3/2}\,\exp\left\{-\frac{\log^2(y/x)}{2\rho}\right\},\\
g(x) &= e^{\rho/8}\sqrt{2\pi\rho}\,x^{-2}\,\Phi\left(\frac{\rho-2\log x}{2\sqrt{p}}\right),\\
G(x) &= e^{\rho/8}\sqrt{2\pi\rho}\,x^{-1}\left\{2x\Phi\left(\frac{\sqrt{\rho}}{2}\right)-\Phi\left(\frac{\rho-2\log x}{2\sqrt{\rho}}\right)-x\Phi\left(\frac{\rho+2\log x}{2\sqrt{\rho}}\right)\right\},\\
c_{uu} &= 2e^{\rho/8}\sqrt{2\pi\rho}\left\{1-\Phi\left(\frac{\sqrt{p}}{2}\right)\right\},\quad c_{ud}=e^{\rho/8}\sqrt{2\pi\rho}\left\{2\Phi\left(\frac{\sqrt{p}}{2}\right)-1\right\},\\
r(x,y) &= e^{\rho/8}\sqrt{2\pi\rho}\,y^{-2}\left\{\Phi\left(\frac{\rho+2\log x-2\log y}{2\sqrt{\rho}}\right)-\Phi\left(\frac{\rho-2\log y}{2\sqrt{\rho}}\right)\right\}.
\end{aligned}
$$

The same quantities for the censored bivariate copula in Example 3.2.2 are

$$
\begin{aligned}
\Pr(Y_1^C=0) &= p_{00}+(1-p_{00})\frac{s-C(s,t)}{1-C(s,t)},\\
\Pr(Y_1^C=0\mid Y_2^C=y) &= \frac{\partial}{\partial v}C(s,\tilde{v}),\\
\Pr(Y_1^C\le x\mid Y_1^C>0,Y_2^C=y) &= \frac{\frac{\partial}{\partial y}C(\tilde{u},\tilde{v})-\frac{\partial}{\partial y}C(s,\tilde{v})}{1-\frac{\partial}{\partial y}C(s,\tilde{v})},\\
\Pr(Y_1^C=0\mid Y_2^C=0) &= \frac{1}{1+(p_{00}^{-1}-1)\frac{t-C(s,t)}{1-C(s,t)}},\\
\Pr(Y_1^C\le x\mid Y_1^C>0,Y_2^C=0) &= \frac{C(\tilde{u},t)-C(s,t)}{t-C(s,t)}.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\Pr(Y_1^C\ge 1,Y_2^C\ge 1) &= (1-p_{00})\frac{1-s-t+C(s,t)}{1-C(s,t)},\\
\Pr(Y_1^C\ge 1,Y_2^C=0) &= (1-p_{00})\frac{t-C(s,t)}{1-C(s,t)}.
\end{aligned}
$$

After having implemented the functions above in `R`, we checked the following equation:

$$
\begin{aligned}
\Pr(Y_1^C\le x\mid Y_1^C>0)\Pr(Y_1^C>0) &= \int_1^\infty \Pr(Y_1^C\le x\mid Y_1^C>0,Y_2^C=y)\\
&\quad\Big\{1-\Pr(Y_1^C=0\mid Y_2^C=y)\Big\}\Pr(Y_2^C=y\mid Y_2^C>0)\Big\{1-\Pr(Y_2^C=0)\Big\}\,dy\\
&\quad+\Pr(Y_1^C\le x\mid Y_1^C>0,Y_2^C=0)\Big\{1-\Pr(Y_1^C=0\mid Y_2^C=0)\Big\}\Pr(Y_2^C=0).
\end{aligned}
$$

Example A.3.2 (Factorization of a Homogeneous Limiting Tail Density)**.** We illustrate here a limiting tail density that factorizes non-trivially. Let

$$h(x,y,z) = \frac{(x+y)^{-3}(y+z)^{-3}}{y^{-2}},$$

which is integrable on $C^+_{||\cdot||_\infty} = \{\boldsymbol{x} \in \mathbb{R}^3_+ : ||\boldsymbol{x}||_\infty \geq 1\}$. Define a random vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ with values in $C^+_{||\cdot||_\infty}$ and probability density $h(\boldsymbol{x})/\int_{C^+_{||\cdot||_\infty}} h(\boldsymbol{x})d\boldsymbol{x}$. Then the following conditional independence relation holds:

$$Y_1 \perp\!\!\!\perp Y_3 \mid Y_2,\, Y_k \geq 1, \quad k = 1,2,3,$$

Moreover, $t^{-1}\boldsymbol{Y} \mid ||\boldsymbol{Y}||_\infty \geq t \stackrel{d}{=} \boldsymbol{Y}$ because

$$\frac{\partial}{\partial \boldsymbol{x}} \Pr(\boldsymbol{Y} \leq t\boldsymbol{x} \mid ||\boldsymbol{Y}||_\infty \geq t) = \frac{t^3 h(tx,ty,tz)}{\int_{tC_{||\cdot||_\infty}} h(x,y,z)dxdydz} = \frac{h(x,y,z)}{\int_{C_{||\cdot||_\infty}} h(\bar{x},\bar{y},\bar{z})d\bar{x}d\bar{y}d\bar{z}},$$

by a change of variable $(x,y,z) = (t\bar{x}, t\bar{y}, t\bar{z})$ and from the homogeneity of h. Notice that $(Y_1, Y_2) \mid \max(Y_1, Y_2) \geq 1$ has, up to a constant, homogeneous probability density $(x+y)^{-3}$, and thus equals $t^{-1}(Y_1, Y_2) \mid \max(Y_1, Y_2) \geq t$ in distribution for any $t > 0$. Similarly, the probability density of $(Y_1, Y_3) \mid \max(Y_1, Y_3) \geq 1$ is proportional to

$$\frac{-3x^2 + 3z^2 + (x^2 + 4xz + z^2)\log\frac{x}{z}}{(x-z)^5},$$

which is homogeneous as expected.

Example A.3.3 (Prohibitive Constraints on Censored Homogeneous Limiting Densities that Factorize)**.** Suppose that $f_{\boldsymbol{Y}^C}$ is a density on $F^3_+ = (\{0\} \cup [1,\infty))^3$ satisfying

$$f_{\boldsymbol{Y}^C} = \frac{f_{\boldsymbol{Y}^C_{12}} f_{\boldsymbol{Y}^C_{23}}}{f_{\boldsymbol{Y}^C_2}},$$

and that there exists a homogeneous probability density $h(x,y,z)$ of order -4 on $C^+_{||\cdot||_\infty}$ for which

$$f_{\boldsymbol{Y}^C}(\boldsymbol{y}_A, \boldsymbol{0}_{A^c}) = (1-p_0)\int_{|\boldsymbol{y}_{A^c}| \leq 1} h(\boldsymbol{y}_A, \boldsymbol{y}_{A^c})d\boldsymbol{y}_{A^c}, \quad \boldsymbol{y}_A \geq \boldsymbol{1},\ A \subseteq \{1,2,3\} \setminus \{\emptyset\},$$
$$f_{\boldsymbol{Y}^C}(\boldsymbol{0}) = p_0,$$

for $p_0 \in (0,1)$. Let $h_{12}(x,y) := \int_0^\infty h(x,y,z)dz$ and $h_{23}(y,z) := \int_0^\infty h(x,y,z)dx$. We compute

$$\begin{aligned} f_{\boldsymbol{Y}_{12}^C}(x,y) &= (1-p_0)h_{12}(x,y), \quad x,y \geq 1, \\ f_{\boldsymbol{Y}_{12}^C}(x,0) &= (1-p_0)\int_0^1 h_{12}(x,y)dy, \quad x \geq 1, \\ f_{\boldsymbol{Y}_2^C}(y) &= (1-p_0)y^{-2}, \quad y \geq 1, \end{aligned}$$

and find

$$h(x,y,z) = \frac{f_{\boldsymbol{Y}^C}(x,y,z)}{1-p_0} = \frac{h_{12}(x,y)h_{23}(y,z)}{y^{-2}}, \quad x,y,z \geq 1, \tag{A.9}$$

$$\int_0^1 h(x,y,z)dy = \frac{f_{\boldsymbol{Y}^C}(x,0,z)}{1-p_0} \propto \int_0^1 h_{12}(x,y)dy \int_0^1 h_{23}(y,z)dy, \quad x,z \geq 1, \tag{A.10}$$

for $x,y,z \geq 1$, where "$\propto$" means proportional to. Since h is homogeneous of order -4 and h_{12} and h_{23} are homogeneous of order -3, for any $0 < \epsilon < 1$,

$$\begin{aligned} \int_0^1 h(x,y,z)dy &= \int_0^\epsilon h(x,y,z)dy + \int_\epsilon^1 \epsilon^{-4}h(x/\epsilon, y/\epsilon, z/\epsilon)dy \\ &= \int_0^\epsilon h(x,y,z)dy + \int_\epsilon^1 \epsilon^{-4}\frac{h_{12}(x/\epsilon, y/\epsilon)h_{23}(y/\epsilon, z/\epsilon)}{(y/\epsilon)^{-2}} \\ &= \int_0^\epsilon h(x,y,z)dy + \int_\epsilon^1 \frac{h_{12}(x,y)h_{23}(y,z)}{y^{-2}}, \end{aligned}$$

using (A.9). Letting $\epsilon \to 0$ and relying on (A.10), we find

$$\int_0^1 \frac{h_{12}(x,y)h_{23}(y,z)}{y^{-2}}dy \propto \int_0^1 h_{12}(x,y)dy \int_0^1 h_{23}(y,z)dy,$$

This is an involved constraint that is not satisfied by any homogeneous function h_{12} and h_{23} in general. In the case $h(\cdot,\cdot) := h_{12} = h_{23}$ and $x = z$, the constraint reads

$$\int_0^1 y^2 h(x,y)^2 dy \propto \left\{\int_0^1 h(x,y)dy\right\}^2. \tag{A.11}$$

For instance, when $h(x,y) = (x+y)^{-3}$ the ratio between the left and right-hand side in (A.11) is

$$\frac{2x\{1+5x(1+2x)\}}{15(1+x)(1+2x)^2} \not\propto 1.$$

Bibliography

British National Corpus, Version 3 BNC XML edition, 2007.

D. J. Allcroft and C. A. Glasbey. A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 52(4):487–498, 2003.

M. I. F. Alves, M. I. Gomes, L. de Haan, and C. Neves. A note on second order conditions in extreme value theory: linking general and heavy tail conditions. *REVSTAT Statistical Journal*, 5(3):285–304, 2007.

A. Anastasiou and G. Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218, 2017.

C. W. Anderson. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, 7:99–113, 1970.

C. W. Anderson. Local limit theorems for the maxima of discrete random variables. *Mathematical Proceedings of the Cambridge Philosophical Society*, 88(1):161–165, 1980.

B. C. Arnold. *Pareto distribution*. Wiley Online Library, 2015.

T. B. Arnold and J. W. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39, 2011.

P. Asadi, A. C. Davison, and S. Engelke. Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050, 2015.

R. L. Axtell. Zipf distribution of US firm sizes. *Science*, 293:1818–1820, 2001.

F. Ballani and M. Schlather. A construction principle for multivariate extreme value distributions. *Biometrika*, 98(3):633–645, 2011.

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

B. Basrak. *The Sample Autocorrelation Function of Non-Linear Time Series*. PhD thesis, Rijksuniversteit Groningen, 2000.

B. Basrak, R. A. Davis, and T. Mikosch. A characterization of multivariate regular variation. *The Annals of Applied Probability*, 12(3):908–920, 2002.

C. Beaumel and V. Bellamy. Les statistiques sur les naissances en 2014. *Insee Résultats*, 171, 2015.

J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, D. De Waal, and C. Ferro. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2004.

T. L. Bell. A space-time stochastic model of rainfall for satellite remote-sensing studies. *Journal of Geophysical Research: Atmospheres*, 92(D8):9631–9643, 1987.

J. M. Berger and B. Mandelbrot. A new model for error clustering in telephone circuits. *IBM Journal of Research and Development*, 7(3):224–236, 1963.

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1995.

N. H. Bingham and A. J. Ostaszewski. Topological regular variation: III. Regular variation. *Topology and its Applications*, 157(13):2024–2037, 2010.

N. H. Bingham, C. M. Goldie, and J. L. Teugles. *Regular Variation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1989.

M.-O. Boldi and A. C. Davison. A mixture model for multivariate extremes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69(2):217–229, 2007.

A. D. Booth. A "law" of occurrences for words of low frequency. *Information and Control*, 10(4):386–393, 1967.

P. Bortot, S. G. Coles, and J. Tawn. The multivariate Gaussian tail model: an application to oceanographic data. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 49(1):31–49, 2000.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

A. Buddana and T. J. Kozubowski. Discrete Pareto distributions. *Economic Quality Control*, 29(2):143–156, 2014.

T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):2005–2013, 2012.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

S. G. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer, 2001.

S. G. Coles and J. A. Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B. Methodological*, 53(2):377–392, 1991.

D. Cooley, R. A. Davis, and P. Naveau. The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9): 2103–2117, 2010.

A. C. Davison. *Statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2003.

A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B. Methodological*, 52(3):393–442, 1990.

A. P. Dawid. Separoids: a mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):335–372, 2001.

L. de Haan and S. I. Resnick. Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 40(4):317–337, 1977.

S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistical Review*, 73(1):111–129, 2005.

P. Ding. On the conditional distribution of the multivariate t distribution. *The American Statistician*, 70(3):293–295, 2016.

A. Dobra and A. Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.

R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*. Applications of Mathematics. Springer, 1997.

S. Engelke, A. Malinowski, Z. Kabluchko, and M. Schlather. Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society. Series B. Methodological*, 77(1):239–265, 2015.

M. Falk and R. Michel. Testing for tail independence in extreme value models. *Annals of the Institute of Statistical Mathematics*, 58(2):261–290, 2006.

M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *The Annals of Applied Statistics*, 5(2A):1057–1080, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

X. Gabaix. Zipf's Law and the Growth of Cities. *The American Economic Review*, 89(2): 129–132, 1999.

C. Genest and J. Nešlehová. A primer on copulas for count data. *Astin Bulletin*, 37(02): 475–515, 2007.

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. mvtnorm: multivariate normal and t distributions, 2008.

N. Gissibl and C. Klüppelberg. Max-linear models on directed acyclic graphs. *arXiv preprint arXiv:1512.07522*, 2015.

G. Gudendorf and J. Segers. Extreme-value copulas. In *Copula theory and its applications*, pages 127–145. Springer, 2010.

B. E. Hamilton, J. A. Martin, M. J. Osterman, S. C. Curtin, and M. TJ. Births: Final data for 2014. *National Vital Statistics Reports*, 64(1), 2015.

J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.

J. E. Heffernan and S. I. Resnick. Limit laws for random vectors with an extreme component. *Ann. Appl. Probab.*, 17(2):537–571, 2007. ISSN 1050-5164. doi: 10.1214/105051606000000835. URL `http://dx.doi.org/10.1214/105051606000000835`.

S. A. Hitz and J. R. Evans. One-component regular variation and graphical modeling of extremes. *Journal of Applied Probability*, 53(3):733–746, 2016.

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

T.-M. Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

J. Hüsler and R.-D. Reiss. Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286, 1989.

J. Jaroš and T. Kusano. Self-adjoint differential equations and generalized Karamata functions. *Bulletin. Classe des Sciences Mathématiques et Naturelles. Sciences Mathématiques*, 29:25–60, 2004.

B. A. Johnson. On lasso for censored data. *Electronic Journal of Statistics*, 3:485–506, 2009.

J. Karamata. Sur un mode de croissance régulière. Théorèmes fondamentaux. *Bulletin de la Société Mathématique de France*, 61:55–62, 1933.

A. Klenke. *Probability theory*. Universitext. Springer, second edition, 2014.

T. J. Kozubowski, A. K. Panorska, and M. L. Forister. A discrete truncated Pareto distribution. *Statistical Methodology*, 26:135–150, 2015.

H. Krishna and P. Singh Pundir. Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, 6(2):177–188, 2009.

J. B. Kruskal, Jr. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956.

S. Kullback. *Information theory and statistics*. Dover Publications, 1997.

J. Lafferty, H. Liu, and L. Wasserman. Sparse nonparametric graphical models. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 27(4): 519–537, 2012.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

A. W. Ledford and J. A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187, 1996.

G. Lee and C. Scott. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9): 2816–2829, 2012.

C.-S. Li, J.-C. Lu, J. Park, K. Kim, P. A. Brinkley, and J. P. Peterson. Multivariate zero-inflated Poisson models and their applications. *Technometrics*, 41(1):29–38, 1999.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3): 18–22, 2002.

F. Lindskog, S. I. Resnick, and J. Roy. Regularly varying measures on metric spaces: hidden regular variation and hidden jumps. *Probab. Surv.*, 11:270–314, 2014. ISSN 1549-5787. doi: 10.1214/14-PS231. URL `http://dx.doi.org/10.1214/14-PS231`.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4): 2293–2326, 2012.

Y. Liu and G.-L. Tian. Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics & Data Analysis*, 83:200–222, 2015.

B. Mandelbrot. Contribution à la théorie mathématique des jeux de communication. *Publications de l'Institut de statistique de l'Université de Paris*, 2(nos. 1-2):124, 1953.

K. Maulik, S. Resnick, and H. Rootzén. Asymptotic independence and a network traffic model. *Journal of Applied Probability*, 39(04):671–699, 2002.

M. M. Meerschaert. Regular variation and generalized domains of attraction in $\mathbf{R}^k$. *Statistics & Probability Letters*, 18(3):233–239, 1993.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian networks in R with applications in systems biology*. Use R! Springer, 2013.

R. B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, second edition, 2006.

B. New, C. Pallier, M. Brysbaert, and L. Ferrand. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524, 2004.

A. K. Nikoloulopoulos. Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic Environmental Research and Risk Assessment*, 30(2):493–505, 2016.

V. M. Panaretos. *Statistics for mathematicians*. Compact Textbook in Mathematics. Birkhäuser/Springer, 2016.

I. Papastathopoulos and K. Strokorb. Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15, 2016.

J. Pickands, III. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131, 1975.

F. Prieto, E. Gómez-Déniz, and J. M. Sarabia. Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention*, 71: 38–49, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2015.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

S. I. Resnick. *Extreme values, regular variation, and point processes*. Applied Probability. A Series of the Applied Probability Trust. Springer, 1987.

S. I. Resnick and D. Zeber. Transition kernels and the conditional extreme value model. *Extremes*, 17(2):263–287, 2014. ISSN 1386-1999. doi: 10.1007/s10687-014-0182-0. URL `http://dx.doi.org/10.1007/s10687-014-0182-0`.

M. Ribatet. Spatial extremes: max-stable processes at work. *Journal de la SFdS. Journal de la Socíeté Française de Statistique*, 154(2):156–177, 2013.

M. Schemper, A. Kaider, S. Wakounig, and G. Heinze. Estimating the correlation of bivariate failure times under censoring. *Statistics in Medicine*, 32(27):4781–4790, 2013.

M. Schlather and J. A. Tawn. A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156, 2003.

M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

J. Segers. Regularly varying Markov trees. Presentation at the conference on conditional independence structures and extremes, Munich, 2016.

E. Seneta. An interpretation of some aspects of Karamata's theory of regular variation. *Publications de l'Institut de Mathématique*, 15(29):111–119, 1973.

T. Shimura. Discretization of distributions in the maximum domain of attraction. *Extremes. Statistical Theory and Applications in Science, Engineering and Economics*, 15(3):299–317, 2012.

A. Sklar. Fonctions de répartition a n dimensions et leurs marges. *Annales de l'I.S.U.P.*, 54(1-2):3–6, 2010.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

J. A. Tawn. Modelling multivariate extreme value distributions. *Biometrika*, 77(2): 245–253, 1990.

T. Therneau, B. Atkinson, and B. Ripley. R package rpart: Recursive partitioning for classification, regression and survival trees, 2015.

E. Thibaud and T. Opitz. Efficient inference and simulation for elliptical Pareto processes. *Biometrika*, 102(4):855–870, 2015.

C. Van Borkulo and S. Epskamp. R package IsingFit: Fitting Ising models using the eLasso method, 2014.

J. Wadsworth, J. A. Tawn, A. Davison, and D. M. Elton. Modelling across extremal dependence classes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 79(1):149–175, 2017.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

J. Yan. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The `huge` package for high-dimensional undirected graph estimation in `R`. *Journal of Machine Learning Research*, 13:1059–1062, 2012.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.