Statistical inference for tail dependence structures

by

Michaël Lalancette

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Statistical Sciences
University of Toronto

Statistical inference for tail dependence structures

Michaël Lalancette
Doctor of Philosophy
Graduate Department of Statistical Sciences
University of Toronto
2022

Abstract

In multivariate extreme value theory, the study of tail dependence seeks to understand the dependence structure of multivariate data among those observations that are considered extreme, typically by having at least one of their components take a large value. This thesis offers solutions to two inferential problems concerning tail dependence.

Existing work often assumes that observed variables are tail dependent, i.e., that observing multiple extreme values simultaneously is roughly as likely as observing at least one extreme. Real data, however, suggests that this is a restrictive assumption even in the bivariate case; the probability of simultaneous extremes can often be significantly smaller than the probability of a single extreme, while being non-negligible. In the first part of this thesis, a novel method is introduced to construct parametric models for bivariate tails that can be agnostic to the presence or absence of tail dependence. A class of M-estimators is constructed for the models and is theoretically justified. The model construction, inference methodology and asymptotic theory are then extended to the case where the tails of a spatial process are of interest.

The conditional tail dependence structure of a moderate- to high-dimensional random vector can be encoded in the edges of an extremal graph, where each vertex represents an observed variable. Learning general extremal graphs in a fully data-driven way is an important open problem. In the second part of this thesis, a family of algorithms is introduced to solve this task by borrowing tools from Gaussian graphical model selection. For two such algorithms which are based on L^1 regularization,

consistency of the estimated graph is established in a general setting. No assumptions are made on the structure of the underlying graph, other than connectedness, and the number of variables is allowed to be exponentially larger than the effective sample size. Along the way, a general concentration result is proved for the empirical extremal variogram, which has widespread applicability in multivariate extreme value theory.

À mes grands-parents.

Acknowledgments

I first owe my supervisor, Stanislav Volgushev, a great dept of gratitude for his unrelenting, but always kind, support. I have learned more by working with him these past four years than I thought I ever could, even if I probably only scratched the surface of his knowledge. I also thank Sebastian Engelke for taking me under his wing and being an amazing mentor and collaborator since the beginning of my degree. I look forward to being only a train ride away from him.

I am very grateful of the financial assistance that both Stanislav and Sebastian have offered me that supported my research travels. Funding from the Fonds de recherche du Québec – Nature et Technologies and from the Ontario Graduate Scholarships is also gratefully acknowledged.

The Department of Statistical Sciences proved to be a very stimulating research environment, where I immediately felt welcome and valued. I am thankful for my time here. Among the many faculty members with whom I had enjoyable conversations, I especially thank Sebastian Jaimungal and Keith Knight for agreeing to serve on my supervisory committee. I will keep fond memories of the badminton matches with Stanislav, Radu V. Craiu, Dehan Kong and Linbo Wang.

For making great friends and colleagues, I thank all my peers in the department from the past five years. From day one, I have felt elevated from being surrounded by such a strong group of students. I should particularly mention Cédric Beaulac, Yuxiang Gao, Mufan Li, Jeffrey Negrea, Arvind Shrivats, Yanbo Tang and Robert Zimmerman. Acknowledgements also go to my friends at the University of Geneva for making my stay there most enjoyable.

Des remerciements spéciaux sont dûs à mes parents pour leur amour et leur support inconditionnel. J'espère qu'ils me pardonneront de ne pas avoir appelé aussi souvent que j'aurais dû, surtout dans les derniers temps durant lesquels j'écrivais cette thèse. Une considération toute spéciale va à mon grand-père, qui m'accompagne en pensée.

Finalement, tout ce travail n'aurait été possible sans Emy. Sa bonne humeur, son amour, sa patience incommensurable m'ont gardé la tête hors de l'eau pendant ces cinq années. Je n'aurais besoin de rien de moins qu'une seconde thèse pour exprimer toute la gratitude que je lui dois.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"*Juillet s'en vient*
Notre cœur dégèle"

Neige

The main chapters of this thesis are formed of two research papers that were written during the past four years. While Chapters 2 and 3 are not directly related to each other, they both fall under the following narrative: they offer new methodologies for the estimation of certain aspects of the dependence structure between the extremes of two or more random variables. They both strongly rely on the framework of extreme value theory and as such, the current chapter is mostly devoted to that topic. It should be seen as a gentle introduction to the subject, going back almost a century and slowly building up to the content of Chapters 2 and 3. Those two chapters are left mostly unchanged from the respective source materials (the published version of the former, the current draft manuscript for the latter). Each is self-contained with its own introduction, a breakdown of the chapter as well as a review of the relevant literature. For that reason, the present chapter is by no means meant to be exhaustive. It does not elaborate on the numerous applications of extreme value theory, nor does it lose itself in precise mathematical statements. Rather, we attempt to contextualize the contribution of those later chapters into the existing body of literature, only introducing the necessary material and referencing an exclusive selection of the most relevant work.

In particular, the content of Chapter 2 comes up in Section 1.2.3. After touching the subject of graphical modeling, this introduction builds up to the content of Chapter 3 in Section 1.3.2.

This chapter is finally closed with an attribution section, where the role of each co-author in writing the two papers that became the present thesis is precisely described

and acknowledged.

1.1 Extreme value theory

In multiple areas of application of statistics, rare events can have catastrophic consequences. Since they are typically not represented in most of the available data, traditional statistical methods fail to produce valid inferences on the small probabilities of those occurrences. In environmental science, understanding the distribution of unusual meteorological phenomena is crucial in building the necessary infrastructure to protect populations, such as dikes against flooding. In insurance, the risk associated to extreme simultaneous claims must be accurately assessed in order to correctly manage actuarial reserves.

Extreme value theory offers a framework to statistically infer the probability of small events that consist in one or a collection of random variables taking on a large value. On a high level, it does so by studying the far, typically unobserved tail regions of probability distributions. In contrast to more traditional statistical methods which use the bulk of data to infer central features of the data-generating distribution, extreme value methods use extreme observations and extrapolation tools to infer the properties of the latter outside the range of available data.

Mathematically speaking, suppose that we are interested in a random variable X, and especially in the tail regions of its distribution. This could be motivated by a number of questions, such as:

1. If we were to observe independent realizations $X_1, \dots, X_n$ of X, what should we expect about the behavior of their maximum $\max_i X_i$?
2. For a large, predefined threshold x, what is the probability $\mathbb{P}(X > x)$?
3. For a small number $q > 0$, what is a $(1-q)$th quantile of X, that is a number x such that $\mathbb{P}(X > x) = q$?

1.1.1 Sample maxima

The Fisher–Tippett–Gnedenko theorem, perhaps the most famous result in extreme value theory, offers a direct and surprisingly clean answer to the first question above. This result, which bears many names including the first extreme value theorem or the extremal types theorem, goes as far back as Fisher and Tippett (1928), and a partial version was even discovered by Fréchet (1927). It states the following. Let $X_1, \dots X_n$ be independent realizations of X and suppose that there exist sequences of numbers

$a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\frac{\max_{1\le i\le n} X_i - b_n}{a_n} \rightsquigarrow Z, \quad n \to \infty \tag{1.1}$$

for a non-degenerate random variable Z, where $\rightsquigarrow$ denotes convergence in distribution (or weak convergence). Then the distribution function G of Z has to be of the form $G(z) = G_\gamma((z-\mu)/\sigma)$ for some $\mu \in \mathbb{R}$, $\sigma > 0$ and $\gamma \in \mathbb{R}$, where

$$G_\gamma(z) := \exp\{-(1+\gamma z)^{-1/\gamma}\}, \quad 1+\gamma z > 0.$$

Hereafter, for $\gamma = 0$, $(1+\gamma z)^{-1/\gamma}$ is understood as the limit e^{-z}. This defines the family of generalized extreme value (GEV) distributions $\{G_\gamma\}$, which is parameterized by the shape parameter γ. For positive shape, this is also known as the family of Fréchet distributions, while negative shape gives rise to the family of negative Weibull distributions; G_0 is also known as the Gumbel distribution. Informally, the shape parameter characterizes the heaviness of the upper tail of the random variable X and is therefore usually called the tail index, or extreme value index, of X. If it is positive, then X is unbounded above and is said to be heavy-tailed. If it is negative, then X is bounded above. If $\gamma = 0$, then X can either be bounded or unbounded (but light-tailed), depending on the sequences a_n and b_n that make (1.1) hold.

The theorem in its original form guarantees that GEV distributions are the only non-degenerate laws that can arise as the limit of normalized maxima of independent and identically distributed (iid) samples. What it does not tell is for which distributions do normalized maxima converge weakly to a GEV distribution, although sufficient conditions were obtained by von Mises (1936).

Let F be the distribution function of X. Note that (1.1) holds if and only if

$$F^n(a_n x + b_n) \longrightarrow G_\gamma(x), \quad n \to \infty \tag{1.2}$$

for some $\gamma \in \mathbb{R}$; by the choice of a_n and b_n, the location and scale parameters μ and σ of the limiting distribution can be assumed to be 0 and 1, respectively. The set of distributions F that satisfy the above property is called the max-domain of attraction of G_γ. In later work (Gnedenko, 1943), an exact characterization of the max-domain of attraction of each GEV distribution was obtained. For a complete and modern statement, we refer to Theorem 1.2.1 of de Haan and Ferreira (2006) or to Chapter 1 of Resnick (1987). For $\gamma \neq 0$, the domain of attraction condition ties in elegantly to the theory of regular variation. An interesting fact is that with the right choice of constants a_n, b_n, G_γ satisfies (1.2) with equality, i.e., $G_\gamma^n(a_n x + b_n) = G_\gamma(x)$. This property is known as max-stability, in parallel with the standard notion of stability,

and in fact the family of GEV distributions are the only max-stable distributions, up to location and scale.

Since information on the tail heaviness of X is contained in the value of the tail index, an interesting problem is to estimate γ. The relation in (1.1) suggests the following strategy. Split the n observations into k blocks of approximately equal size, which is roughly n/k. Then the maxima of each block are independent, as long as the observations are, and they can be considered to be approximate observations from the distribution of $a_{n/k}Z + b_{n/k}$, i.e., a location-scale transformation of the GEV distribution G_γ. The tail index can then be estimated by maximum likelihood or by methods based on moments and probability-weighted moments (for details, we refer to Beirlant et al., 2004, Section 5.1). This general approach is usually termed block maxima (BM) inference.

The larger n/k is, the closer the distribution of the block maxima is to the limiting GEV, but the smaller k, the number of such approximate observations, is. This leads to a bias-variance trade-off that makes picking the number k, the effective sample size, generally difficult. Virtually every extreme value analysis requires some version of this choice.

1.1.2 Threshold exceedances

To answer the last two questions posed at the beginning of the present section, we now consider the problem of tail estimation. It would in theory be sufficient to know the value of the survival function $\bar{F} := 1 - F$ of X at high levels. Based on the iid sample $X_1, \ldots, X_n$, a natural estimate of $1 - F(x)$ is

$$1 - \widehat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i > x\},$$

the proportion of observations that exceed x. If x is beyond the range of observed data, that is no observations exceed it, then the above estimate is zero, and we are left with the conclusion that the probability that X exceeds x is approximately null (or simply that it is very small). Is it possible to somehow make this inference more precise for large x?

The answer turns out to be affirmative, and the tool that allows it is the concept of conditional tail. For any $u < x$, we may write

$$\bar{F}(x) = \mathbb{P}(X > x) = \mathbb{P}(X > u)\mathbb{P}(X - u > x - u \,|\, X > u) =: \bar{F}(u)\bar{F}_u(x - u), \quad (1.3)$$

where $\bar{F}_u$ represents the conditional survival function of $X - u$ given that $X > u$. The

Pickands–Balkema–de Haan theorem (Balkema and de Haan, 1974; Pickands, 1975), often called the second extreme value theorem, offers a very surprising insight on conditional survival functions. It states that for a large class of survival functions $\bar{F}$ (or equivalently, of distributions F), there exist a positive function σ and a number $\gamma \in \mathbb{R}$ such that

$$\bar{F}_u(\sigma(u)y) \longrightarrow (1+\gamma y)^{-1/\gamma}, \quad 1+\gamma y > 0, \tag{1.4}$$

as u tends to the upper endpoint of the distribution of X; either ∞ if X is unbounded, or the upper limit of its support if it is bounded. It also states that this is the only possible form for a non-degenerate limit. For a fixed, large u, this suggests the approximation

$$\bar{F}_u(x-u) \approx \Big(1+\gamma\frac{x-u}{\sigma}\Big)_+^{-1/\gamma},$$

where a_+ denotes the positive part $\max\{a, 0\}$. Here, σ is dependent on the choice of u while γ is a property of $\bar{F}$, and more specifically of the tail of the random variable X, only. That is to say, given $X > u$, $X - u$ is approximately distributed according to a generalized Pareto (GP) distribution with scale parameter $\sigma = \sigma(u)$ and shape parameter γ. The implications are quite strong: (almost) no matter how the distribution of X is shaped, its conditional tail can be approximated by a two-parameter family of distributions. Assuming that u is chosen so that a proportion of the observations $X_1, \ldots, X_n$ exceed it, then the parameters σ and γ can be estimated via, for instance, maximum likelihood on the exceedances $X_i - u \,|\, X_i > u$.

Coming back to the tail probability in (1.3), an estimate can now be computed as

$$\widehat{\bar{F}}(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > u\}\Big(1+\widehat{\gamma}\frac{x-u}{\widehat{\sigma}}\Big)_+^{-1/\widehat{\gamma}},$$

where $\widehat{\sigma}$ and $\widehat{\gamma}$ are the maximum likelihood estimates. Note that once the latter two parameters have been estimated, any value x can in principle be plugged in $\widehat{\bar{F}}$. A non-trivial estimate of the whole tail of X is thus obtained by extrapolating from the most extreme observations in the sample.

This general approach is usually called peaks over threshold (POT) inference. It requires the choice of the threshold u which determines the observations to be used for fitting. Typically, u will be chosen as a sample quantile of order $1 - k/n$, for a number k that is large but such that k/n is small. Similar to the block maxima approach, one then needs to select the effective sample size k according to a bias-variance trade-off: smaller k means less data (so higher variance), but it also means that the selected observations are more extreme (so closer to the limiting GP model, leading to a lower bias).

The punchline of this section is that the conditions under which the limiting relation (1.4) holds are exactly the same as the necessary and sufficient conditions for (1.1). That is, the set of distributions for which the threshold exceedances are asymptotically GP distributed consists exactly of the max-domains of attraction of the GEV distributions G_γ. Moreover, the shape parameter γ that appears in (1.4) is none other than the tail index of the random variable X. So loosely speaking, whether we are studying sample maxima or extreme observations, we are learning about the heaviness of the upper tail of the distribution of interest through its tail index γ. Note here that there are a number of different inferential procedures not mentioned above for the tail index, such as the popular Hill (Hill, 1975) and Pickands (Pickands, 1975) estimators, or for other properties of the tail (such as extreme quantiles). For a comprehensive overview, we refer to Chapters 4 and 5 of Beirlant et al. (2004) and Chapter 4 of de Haan and Ferreira (2006).

1.2 Tail dependence

The most severe casualties are commonly associated to multiple phenomena that cannot be summarized into a single quantity. A plausible situation is that two variables are observed and the interest is in the probability that at least one exceeds a high threshold. For example, the discharge is measured at two locations along a river, and an extreme at one location could be sufficient for a flooding event. Other severe risks are due to joint extreme events. Simultaneous extremes in two or more climate variables in a given region, such as temperature, humidity or wind speed, can have devastating consequences.

The univariate extreme value methods, while successful at estimating tail properties of single random variables, do not take into account the possible dependence between high values taken by different variables. The study of tail dependence aims to fill that gap by proposing and estimating various models and representations for the dependence between extremes. It can be motivated by questions such as:

1. If we were to repeatedly observe (potentially non-independent) random variables $X_1, \ldots, X_d$, how correlated would the d maxima be? Should we expect them to occur simultaneously?

2. How should we expect "multivariate extreme observations" to behave?

In a lot of work, including the one presented in this thesis, those questions are formalized in terms of the copula of the multivariate observations. The object of interest is then purely described by the dependence structure between the observed

variables; the conclusions would be unchanged if the marginal distributions of some of the variables were altered, for instance by changing the measuring units.

Analogously to the univariate case, two general approaches have arisen to investigate the tail dependence of Euclidean data, and in particular to answer the questions above. They are based on multivariate sample maxima and on multivariate threshold exceedances, respectively.

1.2.1 Multivariate maxima

Let $\boldsymbol{X} := (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent copies, with $\boldsymbol{X}_t := (X_{t1}, \ldots, X_{td})$. If

$$\left(\frac{\max_{1 \leq t \leq n} X_{t1} - b_{n1}}{a_{n1}}, \ldots, \frac{\max_{1 \leq t \leq n} X_{td} - b_{nd}}{a_{nd}} \right) \rightsquigarrow \boldsymbol{Z} \tag{1.5}$$

for a non-degenerate random vector $\boldsymbol{Z}$, then $\boldsymbol{X}$ (or its distribution) is said to be in the max-domain of attraction of $\boldsymbol{Z}$ (or, equivalently, of the distribution of $\boldsymbol{Z}$). The family of *multivariate extreme value* (MEV) distributions which arise as possible limits in (1.5) has been completely characterized; see for instance (de Haan and Resnick, 1977) or Section 5.4 of Resnick (1987). They correspond exactly, as in the univariate setting, to the class of multivariate max-stable distributions. The marginal distributions of $\boldsymbol{Z}$ themselves have to be location-scale transformations of GEV distributions. Denoting those marginals by $G_1, \ldots, G_d$, the joint distribution function G of $\boldsymbol{Z}$ can be expressed as

$$G(\boldsymbol{z}) := \exp\{-L(-\log G_1(z_1), \ldots, -\log G_d(z_d))\},$$

where $L : [0, \infty)^d \to [0, \infty)$ is called the *stable tail dependence function* (of $\boldsymbol{Z}$). It is convex, component-wise non-decreasing, homogeneous (for $a > 0$, $L(a\boldsymbol{z}) = aL(\boldsymbol{z})$) and satisfies the bounds $\max_i z_i \leq L(\boldsymbol{z}) \leq \sum_i z_i$. Conversely, any such function corresponds to a MEV distribution. The stable tail dependence function is equivalent to the copula of $\boldsymbol{Z}$, in that it characterizes the latter while containing no information on the marginals $G_1, \ldots, G_d$. Other than their copulas themselves, many equivalent objects have been introduced to characterize the dependence structure of MEV distributions. Examples include the function Ω (Sibuya, 1960), the Pickands dependence function (Pickands, 1981), and various spectral measures (e.g., de Haan and Resnick, 1977); we refer to Chapter 6 of de Haan and Ferreira (2006) for an exposition.

Inference for multivariate extreme values is typically produced by first estimating the marginal tails, e.g., by fitting individual GEV distributions, and subsequently

inferring the tail dependence structure by margin-free, rank-based methods. This thesis develops methodologies for the latter task.

1.2.2 Tail dependence through the functions L and R

If $\boldsymbol{X}$ has marginal distribution functions $F_1, \ldots, F_d$ and is in the max-domain of attraction of a MEV distribution $\boldsymbol{Z}$ with stable tail dependence function L, then as $q \downarrow 0$,

$$q^{-1}\mathbb{P}\big(F_1(X_1) > 1 - qx_1 \text{ or } \ldots \text{ or } F_d(X_d) > 1 - qx_d\big) \longrightarrow L(\boldsymbol{x}), \quad \boldsymbol{x} \in [0,\infty)^d. \tag{1.6}$$

In fact, this limiting relation is equivalent to the copula of $\boldsymbol{X}$ being attracted by the extreme value copula of $\boldsymbol{Z}$; under the assumption that each marginal F_i is in the max-domain of attraction of the distribution G_i of Z_i, (1.6) is equivalent to (1.5).

A consequence of (1.6) is that as $q \downarrow 0$,

$$q^{-1}\mathbb{P}\big(F_1(X_1) > 1 - qx_1, \ldots, F_d(X_d) > 1 - qx_d\big) \longrightarrow R(\boldsymbol{x}), \quad \boldsymbol{x} \in [0,\infty)^d, \tag{1.7}$$

for a certain function R which, however, does not quite fully characterize the tail dependence structure; R can always be deduced from L by an inclusion-exclusion formula, but the converse fails in certain cases. Note that L and R are equivalent in dimension $d = 2$, being linked by the relation $L(x, y) = x + y - R(x, y)$.

The relation (1.6) suggests that the tail dependence structure can be inferred by studying observations where at least one variable is large and estimating L. To an extent, features of the tail dependence structure can also be learned from studying the rarer data points where all components are simultaneously large and estimating R. Hereon, "large" should be understood relatively to each variable's marginal distribution, as in (1.6) and (1.7).

A natural non-parametric estimator of L can be derived from the tail empirical copula process of iid data. Its asymptotic behavior is well understood, both in the bivariate (Huang, 1992; Drees and Huang, 1998) and general multivariate (Einmahl et al., 2012; Fougères et al., 2015) settings. Finite sample guarantees were even demonstrated in arbitrary dimension by Goix et al. (2015). Einmahl et al. (2008, 2012) show how this estimator can be leveraged to fit parametric tail models based on the stable tail dependence function via M-estimation.

1.2.3 Bivariate tail dependence and asymptotic independence

Suppose that X and Y are independent random variables. It is trivial to show that the random vector (X, Y) satisfies (1.6) and (1.7) with $L(x, y) = x + y$ and

$R(x, y) = 0$. Now, suppose instead that (X, Y) is normally distributed with correlation $\rho \in (-1, 1)$. One can show that, again, (1.6) and (1.7) are satisfied with the same functions $L(x, y) = x + y$ and $R(x, y) = 0$ (see for instance Ledford and Tawn, 1997, Appendix A). As a matter of fact, a multitude of bivariate distributions can be seen to have this particular pair of L and R functions. In this case, X and Y are said to be *asymptotically independent*. Roughly speaking, X and Y are asymptotically independent if the probability that they simultaneously take an extreme value is of smaller order than the probability of one single extreme.

Modeling the joint tail of bivariate data through the stable tail dependence function (or equivalently the function R) then amounts to pooling all the asymptotically independent distributions together. This approach does not separate the joint tails of distributions which, by all measures, should be distinguished. For example, see Figures 2.10 to 2.12 (except the left panel in the latter) or Ledford and Tawn (1997), Figure 1, panels (b)-(d), for sample clouds from different asymptotically independent distributions.

Mathematically, for such distributions, the only information contained in the functions L and R is that, by (1.7),

$$P(F_1(X) > 1 - qx, F_2(Y) > 1 - qy) \tag{1.8}$$

is of order $o(q)$ as $q \downarrow 0$. There is generally a certain amount of "second order tail dependence" that is not captured by this definition. In the case of the bivariate normal distribution with correlation $\rho \in (-1, 1)$, for instance, (1.8) is asymptotically proportional to $q^{2/(1+\rho)}(-\log q)^{-\rho/(1+\rho)}$. For other distributions, this is often simplified to $q^{1/\eta}$, for some parameter $\eta \in (0, 1]$, multiplied by a logarithmic term in q (or more generally, any slowly varying function of q). The *coefficient of tail dependence* η, due to Ledford and Tawn (1996), is equal to $1/2$ in case of perfect independence. Larger (smaller) values are generally interpreted as positive (negative) tail association, with asymptotically dependent distributions (that have $R(x, y) \neq 0$) necessarily having $\eta = 1$.

The approach put forward by Ledford and Tawn (1996, 1997) is to model (1.8) for a fixed, small q. For a number of exactly specified, smooth models, they develop inference procedures based on censored likelihoods. Under considerably more general settings, other papers focus on the estimation of the coefficient η which characterizes the strength of association (Peng, 1999; Draisma et al., 2004).

Draisma et al. (2004) assume a tail expansion reminiscing of (1.7), but where the factor q^{-1} is replaced by whichever scale makes the limit non-trivial. This essentially reduces to (1.7) under asymptotic dependence, but it becomes a representation of

the second order tail dependence otherwise. Under that assumption, they develop a non-parametric estimator of far bivariate tail probabilities which is provably consistent under both asymptotic dependence and independence.

In Chapter 2, which has been published in the form of Lalancette et al. (2021), we adopt the general setting of Draisma et al. (2004), particularly their tail expansion. While the limiting function c that appears therein is merely a tool that they use, we treat it as the inferential object and show that it can be used to construct meaningful parametric models for bivariate tails that are agnostic to the presence of asymptotic dependence. In particular, we show that popular models from the literature (the family of inverted max-stable distributions (Wadsworth and Tawn, 2012, Example 2) and a random scale construction (Wadsworth et al., 2017; Engelke et al., 2019b)) can be rephrased as models for c. We revisit a certain tail empirical copula process introduced in Draisma et al. (2004) which can be seen as a non-parametric estimator of c. Adapting the ideas of Einmahl et al. (2012), we derive a family of M-estimators for parametric models that are based on this function. We further show how the methodology from Einmahl et al. (2016) can be adapted, using our bivariate M-estimators, to fit certain types of tail processes over a spatial domain (Wadsworth and Tawn, 2012).

1.2.4 Multivariate threshold exceedances

Let us step back to the general multivariate setting. Yet another representation of the tail dependence structure of the random vector $\boldsymbol{X} \in \mathbb{R}^d$ is through multivariate threshold exceedances. A generally accepted definition of an extreme observation of $\boldsymbol{X}$ is one where at least one of the variables $X_1, \ldots, X_d$ exceeds a high quantile of its marginal distribution. Equation (1.6) implies the existence of a random vector $\boldsymbol{Y}$ such that as $q \downarrow 0$,

$$\left\{ \frac{q}{1 - F(\boldsymbol{X})} \;\middle|\; \max_i F_i(X_i) > 1 - q \right\} \rightsquigarrow \boldsymbol{Y}, \tag{1.9}$$

where $F(\boldsymbol{X})$ denotes the standardized random vector $(F_1(X_1), \ldots, F_d(X_d))$. In other words, when the marginals of $\boldsymbol{X}$ are transformed to unit Pareto distributions, the scaled observations where at least one component is extreme are approximately distributed as $\boldsymbol{Y}$. The family of distributions that can arise as the above limit, termed multivariate Pareto, have been fully characterized (Rootzén and Tajvidi, 2006; Rootzén et al., 2018a). The law of $\boldsymbol{Y}$ in (1.9) contains the same information as the stable tail dependence function L appearing in (1.6); its distribution function can be uniquely expressed in terms of L.

Multivariate peaks-over-threshold inference can therefore be defined as using the extreme observations to estimate the multivariate Pareto distribution of $\boldsymbol{Y}$, offering a new way to infer the tail dependence structure of $\boldsymbol{X}$. Likelihood-based methods are investigated in Rootzén et al. (2018b). In the second half of this thesis, we tackle this estimation problem with the use of graphical models.

1.3 Graphical models

An undirected graph is a couple $G := (V, E)$, where the *vertex set* V (also called *nodes*) is an arbitrary, usually finite set which will be taken here to be $\{1, \dots, d\}$ for an integer $d \geq 3$. The *edge set* $E \subseteq V \times V$ is a subset of unordered pairs of elements of V, representing edges that link the vertices in V; see Figure 3.1 for visual representation of simple undirected graphs. For the remainder of this section, $G = (V, E)$ will be considered a fixed, undirected graph.

A random vector $\boldsymbol{W} \in \mathbb{R}^d$ is said to be a (undirected) *graphical model* on G (or with respect to G) if it satisfies the pairwise Markov property

$$W_i \perp W_j \,|\, \boldsymbol{W}_{\setminus\{i,j\}} \iff (i,j) \notin E,$$

i.e., pairs of variables that are not connected in the graph are independent conditionally on the values taken by all the other variables. Informally, a graphical model on G can be seen as a distribution where all the dependence flows through the connections in G.

The usefulness of graphical models for statistics stems from the famous Hammersley–Clifford theorem, first developed by Hammersley and Clifford (1971). Suppose that $\boldsymbol{W}$ has a positive density (with respect to some product measure on $\mathbb{R}^d$). Then the theorem states that $\boldsymbol{W}$ is a graphical model on G if and only if its density can be factorized into a product of functions of certain groups of variables determined by the structure of G. These groups, called the cliques of G, correspond to the fully connected groups of nodes. See Theorem 3.9 in Lauritzen (1996) or Section 1.7 of Maathuis et al. (2019).

By the Hammersley–Clifford theorem, it is enjoyable to know the graphical structure underlying a high-dimensional distribution. This is especially so if the graph is sparse, in the sense that it has significantly less than the total possible number of edges, $\binom{d}{2} \approx d^2/2$. Example of computational tasks that are simplified by a known, parsimonious graphical structure are Monte Carlo (and MCMC) sampling (see for instance Section 7.3 of Lauritzen (1996) or Section 5.3.5 of Maathuis et al. (2019)) and likelihood inference (see for instance Chapters 4 to 6 of Lauritzen (1996) or Chapters 9 and 10 of Maathuis et al. (2019)).

To enjoy the computational gains provided by a sparse graphical model, however, the graph needs to be known. Since it is rarely given *a priori*, this need has spawned an area of research known as graphical model selection, whose goal is to infer the underlying graphical structure of given data. This is very difficult to do in all generality, especially in high dimensions, as it requires detecting conditional independence. Yet, in the Gaussian case, the problem admits a beautiful simplification.

1.3.1 Gaussian graphical models

Suppose that $\boldsymbol{W} \sim N(\mu, \Sigma)$ and let $\Theta := \Sigma^{-1}$ be its precision matrix. It is well known that $W_i \perp W_j \,|\, \boldsymbol{W}_{\setminus\{i,j\}}$ if and only if $\Theta_{ij} = 0$. The implication is that the graphical structure of a multivariate Gaussian distribution can be read off of the sparsity pattern of its precision matrix. This simple fact reduces the problem to estimating simple moments, namely pairwise covariances, of the data rather than testing for conditional independence.

A variety of methods has therefore been developed to obtain a sparse (meaning that it contains exact zeros) estimator of the precision matrix of multivariate Gaussian data. Alternatively, some methods directly estimate the support of the precision matrix (the set of its entries which are non-zero). Either can thereupon be turned into an estimate of the underlying graph. The most popular such algorithms are certainly neighborhood selection (Meinshausen and Bühlmann, 2006), based on the lasso for linear regression, and the graphical lasso (Yuan and Lin, 2007; Friedman et al., 2008).

1.3.2 Extremal graphical models

A majority of the distributions that arise and are to be inferred in multivariate extreme value theory belong to one of two families. The MEV distributions emerge in (1.5) as weak limits of multivariate sample maxima, and the multivariate Pareto distributions appear in (1.9) as weak limits of multivariate threshold exceedances. To efficiently fit high-dimensional tail dependence models, one may wonder whether it is possible for those distributions to have a non-trivial conditional independence structure.

As it turns out, a major hurdle for graphical modeling of extremes is that for MEV distributions with a positive continuous density, conditional independence is only possible under independence (Papastathopoulos and Strokorb, 2016). For a certain class of MEV models that exhibit certain singularities, directed graphical models are investigated in Gissibl and Klüppelberg (2018), Klüppelberg and Lauritzen (2019) and Améndola et al. (2022).

Instead, Engelke and Hitz (2020) consider multivariate Pareto distributions. From

its definition as a limit in (1.9), it is easy to see that such a random vector $\boldsymbol{Y}$ is supported on the set $\mathcal{L} := [0,\infty)^d \setminus [0,1]^d$. Since this is not a product space, the variables $Y_1,\ldots,Y_d$ will typically not satisfy standard conditional independence relations. To overcome this issue, Engelke and Hitz (2020) propose a new notion of conditional independence, denoted by $\perp_e$, and define an *extremal graphical model* on a connected graph G as a multivariate Pareto distribution $\boldsymbol{Y}$ satisfying

$$Y_i \perp_e Y_j \,|\, \boldsymbol{Y}_{\setminus\{i,j\}} \iff (i,j) \notin E.$$

They further justify their definition by showing that it leads to a Hammersley–Clifford-type factorization theorem, thus allowing for efficient model building and fitting.

Their work opens up a new question: is it possible to use data from a distribution $\boldsymbol{X}$ satisfying (1.9) to estimate the graph structure of the multivariate Pareto limit $\boldsymbol{Y}$? In their paper, they describe an algorithm for stepwise selection of block graph models. Engelke and Volgushev (2020) show that when the extremal graph structure is a tree, it can be learned consistently by constructing a minimum spanning tree. As edge weights, they use either of two different summary statistics: the *empirical variogram* defined therein and the empirical version of the *extremal correlation coefficient* (Coles et al., 1999, denoted as χ).

In Chapter 3, we adopt a certain parametric model for multivariate Pareto distributions, the Hüsler–Reiss model. We show that by using the empirical variogram of Engelke and Volgushev (2020), it is possible to consistently learn the graphical structure of multivariate Pareto distributions under this parametric assumption. Our methodology makes it possible to consistently learn any connected graph even when the dimension grows almost exponentially fast in the effective sample size. It is based on estimating a collection of sparse precision matrices, borrowing tools from Gaussian graphical model selection, and averaging through a majority voting procedure. Along the way, we obtain a concentration result for the empirical variogram which is applied, for instance, in proving high-dimensional consistency of the tree learning algorithm of Engelke and Volgushev (2020). This result and its proof are already available in the form of a preprint (Engelke et al., 2021).

1.4 Attribution of the work in Chapters 2 and 3

Chapter 2 is based upon the published version of Lalancette et al. (2021). For this work, credit goes to Sebastian Engelke and Stanislav Volgushev for the original idea and the initial direction of the project. The final form of the methodologies presented is mostly mine, and I am responsible for producing all the numerical results and (the

final versions of) all the theoretical results and their proofs. Credit is due to SE and SV for overall guidance and supervision during the whole process. Crucially, the expertise of SV was invaluable in helping me derive the weak convergence results found in Section 2.7 and parts of Section 2.8. The proofs of Sections 2.9 and 2.10 are entirely mine. The writing all originated from a draft of mine, but was then collaboratively reworked by SE, SV and myself, with special credit going to SE for parts of Section 2.1.

Chapter 3 is based on a draft manuscript which is to be submitted for publication shortly (Engelke et al., 2022), and part of which is derived from a preprint (Engelke et al., 2021). Once again, credit goes to SE for suggesting the idea to develop extremal graph learning methods based on L^1 penalization, which originated while he was working on the paper Engelke and Hitz (2020). It evolved into the final methodology that is presented here. The theoretical backbone of this paper, Theorem 3.3 as well as it's proof which spans Sections 3.11 and 3.12, is my own work. Preliminary versions of Lemma 3.1 and Corollary 3.1, and the general idea to simplify my prior proof with the use of those two results, was the fruit of a collaborative effort between SV and myself. The consistency results for the neighborhood selection and graphical lasso algorithms and their proofs (Sections 3.9 and 3.10), are also due to collaborative work between SV and myself. Credit goes to SE for producing the figures appearing in Section 3.5, while other numerical results (Sections 3.6 and 3.8) are mine. Similarly to Chapter 2, the writing was collaborative work which originated from a draft written by myself.

Chapter 2

Rank-based estimation under asymptotic dependence and independence, with applications to spatial extremes

2.1 Introduction

Assessing the frequency of extreme events is crucial in many different fields such as environmental sciences, finance and insurance. The most severe risks are often associated to a combination of extreme values of several different variables or the joint occurrence of an extreme phenomenon across different spatial locations. Statistical methods for accurate modeling of such multivariate or spatial dependencies between rare events is provided by extreme value theory. Applications include the analysis of extreme flooding (Keef et al., 2009; Asadi et al., 2015; Engelke and Hitz, 2020), risk diversification between stock returns (Poon et al., 2004; Zhou, 2010) and climate extremes (Westra and Sisson, 2011; Zscheischler and Seneviratne, 2017).

Extremal dependence between largest observations of two random variables X and Y with distribution functions F_1 and F_2, respectively, can take many different forms. A classical assumption to measure and model this dependence is multivariate regular variation (cf., Resnick, 1987), which is equivalent to the existence of the stable tail dependence function

$$L(x,y) := \lim_{t \downarrow 0} \frac{1}{t} \mathbb{P}\left(F_1(X) \geq 1 - tx \text{ or } F_2(Y) \geq 1 - ty\right), \quad x, y \in [0, \infty); \tag{2.1}$$

see Huang (1992) and de Haan and Ferreira (2006). This condition allows a first

broad classification regarding extremal dependence of bivariate random vectors into two different regimes. If $L(x, y) = x + y$, X and Y are said to be asymptotically independent; in this case the joint exceedance probability is negligible compared to the marginal exceedance probabilities. Otherwise, a stronger form of extremal dependence, called asymptotic dependence, holds and the joint exceedance probability is of the same order as the probability of one of the components exceeding a high threshold.

Most of the existing probabilistic and statistical theory deals with asymptotic dependence. A variety of methods exists, including non-parametric estimation (Huang, 1992; Einmahl and Segers, 2009; Guillotte et al., 2011), bootstrap procedures (Peng and Qi, 2008; Bücher and Dette, 2013), parametric approaches including likelihood estimation (Ledford and Tawn, 1996; de Haan et al., 2008; Padoan et al., 2010; Dombry et al., 2017) and M-estimation (Einmahl et al., 2008; Engelke et al., 2015). See also Einmahl et al. (2012, 2016) for inference in the d-dimensional and spatial setting. There is a rich literature on multivariate tail models (see for instance Gumbel, 1960; Tawn, 1988; Hüsler and Reiss, 1989, among many others) and generalizations to spatial domains (Brown and Resnick, 1977; Smith, 1990; Schlather, 2002).

Recent studies have shown that in many applications such as spatial precipitation (Le et al., 2018) and significant wave height (Wadsworth and Tawn, 2012), dependence tends to become weaker for the largest observations and asymptotic independence is therefore the more appropriate regime. In this case, the stable tail dependence function in (2.1) does not contain information on the degree of asymptotic independence and is therefore not suitable for statistical modeling. A remedy to this problem was proposed by Ledford and Tawn (1996, 1997) who introduced a more flexible condition on the joint exceedance probabilities. Their setting implies the existence of

$$c(x, y) := \lim_{t \downarrow 0} \frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right), \quad x, y \in [0, \infty), \tag{2.2}$$

where q is a suitable measurable function that makes the limit nontrivial. Necessarily, q is regularly varying at zero with index $1/\eta \in [1, \infty)$. The residual tail dependence coefficient η describes the strength of residual dependence in the tail and $\eta < 1$ implies asymptotic independence. One speaks about positive and negative association between extremes if $\eta > 1/2$ and $\eta < 1/2$, respectively. Early works focus on estimating the degree of asymptotic independence η and various estimators have been proposed and studied (Ledford and Tawn, 1997; Peng, 1999; Draisma et al., 2004). A more complete description of the residual extremal dependence structure is given by the function c in (2.2); in fact, the value of η can be deduced from c (see Section 2.2 below). Several parametric families exist for multivariate (e.g., Weller and Cooley,

2014) and spatial applications (e.g., Wadsworth and Tawn, 2012). Other statistical approaches for modeling asymptotic independence are also related to this function, including hidden regular variation (Resnick, 2002; Heffernan and Resnick, 2007) and the conditional extreme value model (Heffernan and Tawn, 2004). Note that (2.2) includes the asymptotic dependence case if $\lim_{t \downarrow 0} q(t)/t > 0$, and the function $c(x, y) \propto x + y - L(x, y)$ then contains the same information as L.

Since it is typically not known a priori whether asymptotic dependence or independence is present in a data set, recent parametric models are able to capture both regimes as different sub-sets of the parameter space (e.g., Ramos and Ledford, 2009; Wadsworth et al., 2017; Huser et al., 2017; Engelke et al., 2019b; Huser and Wadsworth, 2019). Most of the literature in this domain is concerned with constructing parametric models, and estimation is usually based on censored likelihood and discussed informally while no theoretical treatment of the corresponding estimators is provided. Moreover, it is typically assumed that extreme observations used to construct estimators already follow a limiting model, and the bias which results from this type of approximation is ignored.

This chapter is motivated by a lack of generic estimation methods that work under both asymptotic dependence and independence and have a thorough theoretical justification. We first revisit a non-parametric, rank-based estimator of the function c in (2.2) which appeared in (Draisma et al., 2004) and provide a new fundamental result on its asymptotic properties, which completely removes any smoothness assumptions on c. This result is the crucial building block for the second major contribution of this chapter: a new M-estimation framework that is applicable across dependence classes.

M-estimators for the stable tail dependence function L have been proposed by Einmahl et al. (2008, 2012, 2016). Under asymptotic dependence, c can be recovered from L and thus any method for estimating L also yields an estimator for c. However, this is no longer true under asymptotic independence. Estimators of L can therefore not be used to fit statistical models with asymptotic independence or models bridging both dependence classes. We define a new class of M-estimators based on c for parametric extreme value models that can be applied regardless of the dependence class. A major challenge under asymptotic independence is due to the fact that the scaling function q is unknown. Additionally, c loses some of the regularity (such as concavity) that it enjoys under asymptotic dependence. Nevertheless, we are able to prove asymptotic normality of our estimators under weak regularity conditions, which are shown to be satisfied for popular models such as the class of inverted max-stable distributions (see Wadsworth and Tawn, 2012).

The challenges described above become even greater for spatial data. Even at

the level of pairwise distributions, real data can exhibit asymptotic dependence at locations that are close but asymptotic independence at locations that are far apart. This necessitates models that can incorporate both, asymptotic dependence and independence at the same time. Estimation in such models calls for methods that can deal with both regimes simultaneously, and we show that our findings in the bivariate case can be leveraged to construct estimators in this setting.

In Section 2.2, we provide the necessary background on asymptotic dependence and independence for bivariate distributions, discuss an extension to the spatial setting, and provide several examples. The estimation methodology is introduced in Section 2.3, while theoretical results are collected in Section 2.4. The methodology is illustrated via simulation studies in Section 2.5, while an application to extreme rainfall data is given in Section 2.6. Section 2.7 contains the proofs of all the main results found in Section 2.4, with a number of technical lemmas deferred to Section 2.8. Sections 2.9 and 2.10 present proofs of several claims from different examples. A brief discussion of computational complexity in spatial estimation is given in Section 2.11 and additional simulation results appear in Section 2.12.

2.2 Multivariate extreme value theory

2.2.1 Bivariate models

Let (X, Y) be a bivariate random vector with joint distribution function F and marginal distribution functions F_1 and F_2, respectively. There is a variety of approaches to describe the joint tail behavior of (X, Y).

The assumption of multivariate regular variation (cf., Resnick, 1987) is classical in extreme value theory and the stable tail dependence function in (2.1) has been extensively studied. Its margins are normalized, $L(x, 0) = L(0, x) = x$, and it satisfies $x \vee y \leq L(x, y) \leq x + y$ for all $x, y \in [0, \infty)$. Moreover, it is a convex and homogeneous function of order one, the latter meaning that $L(tx, ty) = tL(x, y)$ for all $t > 0$. The importance of stable tail dependence functions stems from their connection to max-stable distributions. A bivariate random vector (Z_1, Z_2) has max-stable dependence with standard uniform margins iff its distribution function is given by

$$\mathbb{P}\left(Z_1 \leq x, Z_2 \leq y\right) = \exp\{-L(-\log x, -\log y)\}, \quad x, y \in [0, 1], \tag{2.3}$$

where L is the stable tail dependence function pertaining to (Z_1, Z_2). Note that any max-stable distribution associated with L satisfies (2.1) with that same L, this follows after a simple Taylor expansion. Two examples of max-stable distributions

(equivalently, stable tail dependence functions) that will repeatedly appear in the chapter are as follows.

(i) The bivariate Hüsler–Reiss distribution (Hüsler and Reiss, 1989; Engelke et al., 2015) is defined by

$$L(x,y) = x\Phi\Big(\lambda + \frac{\log x - \log y}{2\lambda}\Big) + y\Phi\Big(\lambda + \frac{\log y - \log x}{2\lambda}\Big),$$

where Φ is the standard normal distribution function and $\lambda \in [0,\infty]$ parametrizes between perfect independence ($\lambda = \infty$) and dependence ($\lambda = 0$).

(ii) The asymmetric logistic distribution (Tawn, 1988), is given by

$$L(x,y) = (1-\nu)x + (1-\phi)y + (\nu^r x^r + \phi^r y^r)^{1/r}, \quad \nu,\phi \in [0,1], r \geq 1.$$

Note that $\nu = \phi = 1$ yields the classical logistic model (Gumbel, 1960).

While multivariate regular variation and max-stability have been very popular due to their nice theoretical properties, they are not informative under asymptotic independence which limits their use in many applications.

Assumption (2.2) allows for more flexible tail models since the limiting function c is non-trivial even under asymptotic independence and contains information on the structure of residual extremal dependence in the vector (X,Y). For the sake of identifiability, we scale q such that $c(1,1) = 1$. We will refer to c and q as the survival tail function and the scaling function associated to (X,Y). It turns out that q has to be regularly varying of order $1/\eta \in [1,\infty)$ and that c is a homogenous funcion of order $1/\eta$, that is,

$$c(tx,ty) = t^{1/\eta}c(x,y), \quad t > 0;$$

see for example Draisma et al. (2004) or Lemma 2.2. Note that the extremal dependence coefficient (see Coles et al., 1999) can be defined as $\chi := \lim_{t\downarrow 0} q(t)/t$. Asymptotic independence is then equivalent to $\chi = 0$, or $q(t) = o(t)$, whereas asymptotic dependence corresponds to $\chi > 0$.

Furthermore, the homogeneity property of c implies a spectral representation. More precisely, there exists a finite measure H on $[0,1]$ such that

$$c(x,y) = \int_{[0,1]} \Big(\frac{x}{1-w} \wedge \frac{y}{w}\Big)^{1/\eta} H(dw), \quad x,y \in [0,\infty);$$

see Theorem 1 in Ramos and Ledford (2009) or Lemma 2.6.

We provide several examples that illustrate the concepts discussed above without going too deeply into technical details. A more thorough discussion of the corresponding

theory is given throughout Section 2.4.

Example 2.1 (*Domain of attraction of max-stable distributions*)**.** Suppose that (X, Y) satisfies (2.1) for a stable tail dependence function L which is not everywhere equal to $x + y$. Then (2.2) holds with $q(t) = \chi t$ and $c(x, y) = (x + y - L(x, y))/\chi$, where the extremal dependence coefficient χ is positive. We further note that (2.1) holds whenever (X, Y) is in the max domain of attraction of a max-stable random vector Z satisfying (2.3), see de Haan and Ferreira (2006) for a definition and additional details.

Example 2.2 (*Inverted max-stable distributions*)**.** The family of inverted max-stable distributions (e.g., Wadsworth and Tawn, 2012, Definition 2) is parametrized by all stable tail dependence functions. More precisely, let G be the distribution function of a bivariate distribution with max-stable dependence, uniform margins and stable tail dependence function L. A random vector (X, Y) with uniform marginal distributions is said to have an inverted max-stable distribution with stable tail dependence L if $(1-X, 1-Y) \sim G$. Assuming that L satisfies a quadratic expansion (see Example 2.8), the law of (X, Y) satisfies (2.2) with

$$q(t) = t^{L(1,1)}, \quad c(x, y) = x^{\dot{L}_1(1,1)} y^{\dot{L}_2(1,1)},$$

where $\dot{L}_j$ denotes the j-th directional partial derivative of L from the right, $j = 1, 2$. Any such stable tail dependence function satisfies $L(1, 1) = \dot{L}_1(1, 1) + \dot{L}_2(1, 1) \in (1, 2]$, and therefore this is an asymptotically independent model with $\eta = 1/L(1, 1)$.

Any existing parametric class of stable tail dependence functions can be used to define a parametric class of inverted max-stable distributions. In particular we consider the two families discussed earlier

(i) Provided that $\lambda > 0$, the inverted Hüsler–Reiss distribution has

$$q(t) = t^{2\theta}, \quad c(x, y) = (xy)^{\theta}, \tag{2.4}$$

where $\theta := \Phi(\lambda) \in (1/2, 1]$.

(ii) The inverted asymmetric logistic distribution has

$$q(t) = t^{\theta_1+\theta_2}, \quad c(x, y) = x^{\theta_1} y^{\theta_2}, \tag{2.5}$$

where $\theta_1 := 1 - \nu + \nu^r(\nu^r + \phi^r)^{1/r-1}$ and $\theta_2 := 1 - \phi + \phi^r(\nu^r + \phi^r)^{1/r-1}$. Note that by suitable choices of the parameters r, ν, ϕ any value of $(\theta_1, \theta_2) \in (0, 1]^2$ such that $\theta_1 + \theta_2 \in (1, 2]$ can be obtained.

Example 2.3 (*A random scale construction*)**.** Bivariate random scale constructions are a popular way of creating distributions with rich extremal dependence structures; see Engelke et al. (2019b) and references therein for an overview. They are random vectors of the form $(X, Y) = R(W_1, W_2)$ where the radial variable R is assumed independent of the angular variables W_j, $j \in \{1, 2\}$. This motivates the following model with parameters $\alpha_R, \alpha_W > 0$:

$$(X, Y) = R(W_1, W_2), \quad R \sim \text{Pareto}(\alpha_R), W_j \sim \text{Pareto}(\alpha_W) \tag{2.6}$$

where W_1, W_2 are independent and a Pareto(α) distribution has distribution function $1 - x^{-\alpha}$ for $x \geq 1$. By Example 2.9 below, (X, Y) satisfies (2.2) with functions q and c depending only on the value of the ratio $\lambda := \alpha_R/\alpha_W$. In particular, we obtain asymptotic dependence if $\lambda < 1$ and asymptotic independence otherwise. Detailed expressions for q and c are provided in Example 2.9.

2.2.2 Spatial models

Spatial extreme value theory is an extension of multivariate extremes to continuous index sets. It is particularly useful for modeling extremes of natural phenomena over spatial domains and examples include heavy rainfall, high wind speeds and heatwaves (e.g., Davison and Gholamrezaee, 2012; Le et al., 2018).

Let $\mathcal{T}$ be a spatial domain (typically a subset of $\mathbb{R}^2$) and $Y = \{Y(u) : u \in \mathcal{T}\}$ be a stochastic process whose extremal behavior we are interested in. We impose the condition in (2.2) on all bivariate margins of Y so that for each pair $s = (u, u')$ of locations, and all $x, y \in [0, \infty)$ the limit

$$c^{(s)}(x, y) := \lim_{t \downarrow 0} \frac{1}{q^{(s)}(t)} \mathbb{P}\left(F^{(u)}(Y(u)) \geq 1 - tx, F^{(u')}(Y(u')) \geq 1 - ty\right) \tag{2.7}$$

exists and is non-trivial; here $F^{(u)}$ is the distribution function of $Y(u)$. Similarly to the bivariate case, $q^{(s)}$ must be regularly varying with index $1/\eta^{(s)} \in [1, \infty)$ and $c^{(s)}$ is homogeneous of order $1/\eta^{(s)}$.

In applications, spatial extreme value theory can be linked to multivariate extreme value theory through the fact that spatial processes are usually measured at a finite set of locations. However, generic multivariate models do not take into account the additional structure arising from spatial features of the domain. Statistical models for processes, in contrast, make use of geographical information to construct structured, low-dimensional parametric models (see, e.g., Engelke and Ivanovs, 2021).

Similarly to max-stable distributions in (2.3), max-stable processes play an impor-

tant role in modeling spatial extremes. The stochastic process $Z = \{Z(u) : u \in \mathcal{T}\}$ is called max-stable if all its finite dimensional distributions are max-stable, which implies in particular that for each pair $s = (u, u')$, the bivariate margin $(Z(u), Z(u'))$ satisfies (2.3) with stable tail dependence function $L^{(s)}$. Hence (2.7) follows for any max-stable process Z for which $(Z(u), Z(u'))$ are not independent for all $u, u' \in \mathcal{T}$.

Brown–Resnick processes (Brown and Resnick, 1977) provide an important subclass of max–stable processes. A Brown–Resnick process $\mathcal{B} = \{\mathcal{B}(u) : u \in \mathcal{T}\}$ is parametrized by a variogram function $\gamma : \mathcal{T}^2 \to \mathbb{R}_+$, and any pair $(\mathcal{B}(u), \mathcal{B}(u'))$ is a bivariate Hüsler–Reiss distribution with parameter $\lambda = \sqrt{\gamma(u, u')}/2$ (Hüsler and Reiss, 1989). Parametric models can be constructed by imposing a parametric form for γ. An example when $\mathcal{T} \subset \mathbb{R}^d$ is the fractal family of variograms given by $\gamma(s) = (\|s_1 - s_2\|/\beta)^\alpha$, where $s = (s_1, s_2)$, $\|\cdot\|$ is the Euclidean norm and $\alpha \in (0, 2]$, $\beta > 0$ are the model parameters (Kabluchko et al., 2009). We next discuss two classes of processes for which (2.7) holds.

Example 2.4 (*Domain of attraction of max-stable processes*)**.** A process $Y = \{Y(u) : u \in \mathcal{T}\}$ is in the max-domain of attraction of the max-stable process Z if there exist sequences of continuous functions $a_n, b_n : \mathcal{T} \to \mathbb{R}$ such that

$$\{\max_{i=1,\dots,n} Y_i(\cdot) - a_n(\cdot)\}/b_n(\cdot) \rightsquigarrow Z(\cdot), \quad n \to \infty \tag{2.8}$$

for i.i.d. copies $Y_1, Y_2, \dots$ of the process Y where weak convergence takes place in the space of continuous functions on $\mathcal{T}$ equipped with the supremum norm; see de Haan et al. (2001) and Chapter 9 of de Haan and Ferreira (2006) for the special case $\mathcal{T} = [0, 1]$.

(2.8) implies that any pair $(Y(u), Y(u'))$ with $u \neq u' \in \mathcal{T}$ is in the max-domain of attraction of the pair $(Z(u), Z(u'))$. If every such pair is not independent, (2.7) holds for all $s = (u, u')$ by the discussion in Example 2.1.

While max-stable processes allow for flexible spatial dependence structures, they can only be used as models for asymptotic dependence. This often violates the characteristics observed in real data, especially for locations $u, u' \in \mathcal{T}$ that are far apart. To model data in such cases, asymptotically independent spatial models have been constructed that satisfy (2.7) and where the residual tail dependence coefficients $\eta^{(s)}$ vary with the distance between the pair s of locations. Most of the models are identifiable from the bivariate margins so that statistical methods for $c^{(s)}$ will provide estimators for spatial tail dependence parameters; see Section 2.3.3 for the methodology. A broad class of asymptotically independent stochastic processes are the inverted max-stable processes (Wadsworth and Tawn, 2012).

Example 2.5 (*Inverted max-stable processes*)**.** Let $Z = \{Z(u) : u \in \mathcal{T}\}$ be a process with max-stable dependence, uniform margins and bivariate tail dependence functions $L^{(s)}$. The process $Y = \{1 - Z(u) : u \in \mathcal{T}\}$ is called inverted max-stable. For a pair $s \in \mathcal{T}^2$, assuming that $L^{(s)}$ satisfies the smoothness condition mentioned in Example 2.2, Y satisfies (2.7) with

$$q^{(s)}(t) = t^{L^{(s)}(1,1)}, \quad c^{(s)}(x,y) = x^{\dot{L}_1^{(s)}(1,1)} y^{\dot{L}_2^{(s)}(1,1)},$$

so that $\eta^{(s)} = 1/L^{(s)}(1,1)$ is a (usually non-constant) function on $\mathcal{T}^2$. In particular, if a Brown–Resnick process is parametrized by a variogram function $\gamma : \mathcal{T}^2 \to \mathbb{R}_+$ then the corresponding inverted Brown–Resnick process has $1/\eta^{(s)} = 2\Phi(\sqrt{\gamma(s)}/2)$.

2.3 Estimation

In this section we present the proposed estimators. First, we recall the non-parametric estimator of a survival tail function from Draisma et al. (2004) in Section 2.3.1. Using this as building block, we construct M-estimators for bivariate survival tail functions (Section 2.3.2) and leverage those estimators to introduce methodology for spatial tail estimation (Section 2.3.3).

2.3.1 Non-parametric estimators of survival tail functions

Recall that (X, Y) is a random vector with joint distribution function F that satisfies (2.2), and assume that its marginal distribution functions F_1 and F_2 are continuous. Denoting by Q the joint distribution function of $(1 - F_1(X), 1 - F_2(Y))$, we can rephrase (2.2) as

$$\frac{Q(tx, ty)}{q(t)} = c(x,y) + O(q_1(t)), \quad x, y \in [0, \infty), \tag{2.9}$$

for some function $q_1(t) \to 0$ as $t \to 0$. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent samples from F. Since F_1, F_2 are unknown, the observations $(1 - F_1(X_i), 1 - F_2(Y_i))$ are not available and can not be used to construct a feasible estimator of Q. A typical solution to this problem is to replace F_j by its empirical counterpart $\widehat{F}_j$, which leads to the estimator

$$\widehat{Q}_n(x,y) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{ n\widehat{F}_1(X_i) \geq n + 1 - \lfloor nx \rfloor, n\widehat{F}_2(Y_i) \geq n + 1 - \lfloor ny \rfloor \right\}; \tag{2.10}$$

see Huang (1992); Drees and Huang (1998); Einmahl et al. (2008, 2012) among others for related approaches in the estimation of stable tail dependence functions.

Given $\widehat{Q}_n$ and the expansion in (2.9), an intuitive plug-in estimator of the function c is given by

$$\widehat{c}_n(x,y) = \frac{\widehat{Q}_n\left(kx/n, ky/n\right)}{q(k/n)}, \tag{2.11}$$

where we set $t = k/n$ in (2.9) for an intermediate sequence $k = k_n$ such that $k \to \infty$, $k/n \to 0$. Note, however, that this estimator is infeasible under asymptotic independence since the function q is in general unknown. A simple remedy is to recall that we considered the normalization $c(1,1) = 1$ and construct a ratio type estimator

$$\widetilde{c}_n(x,y) := \frac{\widehat{c}_n(x,y)}{\widehat{c}_n(1,1)} = \frac{\widehat{Q}_n\left(kx/n, ky/n\right)}{\widehat{Q}_n\left(k/n, k/n\right)} \tag{2.12}$$

to cancel out the unknown scaling factor $q(k/n)$. This leads to a fully non-parametric estimator of c, which is interesting in its own right. Some comments on the asymptotic properties of this estimator will be provided in Section 2.4.1.

Remark. In practice, and especially in a spatial context, it is sometimes appropriate to select directly the effective number of observations used for estimating c (Wadsworth and Tawn, 2012). This can be achieved by selecting $k = \widehat{k}$ such that $nQ_n(\widehat{k}/n, \widehat{k}/n) = m$ for a given value m. This leads to a data-dependent parameter $\widehat{k}$ which will also be covered by our theory.

2.3.2 M-estimation in (bivariate) parametric model classes

While the non-parametric estimators from the previous section possess attractive theoretical properties, they have certain practical drawbacks. For any finite sample size n they are neither continuous nor homogeneous, hence they are not admissible survival tail functions. Additionally, it is difficult to use purely non-parametric estimators in spatial settings. A solution to this problem, which also yields easily interpretable estimators, is to fit parametric models.

In what follows, assume that c belongs to a family $\{c_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^p$ and the true parameter $\theta_0 \in \Theta$ is unknown. Our aim is to leverage the non-parametric estimators from Section 2.3.1 to construct an estimator for θ_0. For stable tail dependence functions which are only informative under asymptotic dependence such a program was carried out in Einmahl et al. (2008, 2012). A direct application

of the corresponding ideas in our setting would be to estimate θ through

$$\breve{\theta} := \arg\min_{\theta\in\Theta} \left\| \int_{[0,T]^2} g(x,y)c_\theta(x,y)dxdy - \int_{[0,T]^2} g(x,y)\widetilde{c}_n(x,y)dxdy \right\|,$$

for an integrable vector-valued weight function $g : \mathbb{R}^2 \to \mathbb{R}^q$, where $\|\cdot\|$ denotes the Euclidean norm. However, as we will discuss in Remark 2.4.1, the use of $\widetilde{c}_n$ would place unnecessarily strong assumptions on the function c in the case of asymptotic dependence. Hence we propose to consider the following alternative approach. Define

$$\Psi_n^*(\theta,\zeta) := \zeta \int_{[0,T]^2} g(x,y)c_\theta(x,y)dxdy - \int_{[0,T]^2} g(x,y)\widehat{Q}_n(kx/n, ky/n)dxdy \quad (2.13)$$

and let

$$(\widehat{\theta}_n, \widehat{\zeta}_n) := \arg\min_{\theta\in\Theta,\zeta>0} \|\Psi_n^*(\theta,\zeta)\|. \quad (2.14)$$

To understand the rationale of this approach, note that $\widehat{c}_n(x,y)$ is proportional to $\widehat{Q}_n(kx/n, ky/n)$ but the proportionality constant involves q and is thus unknown. We thus essentially propose to add this unknown normalization factor as an additional scale parameter ζ. More precisely, write μ_{L} for the Lebesgue measure on $[0,T]^2$, let

$$\Psi_n(\theta,\sigma) = \sigma \int g c_\theta d\mu_{\mathrm{L}} - \int g\widehat{c}_n d\mu_{\mathrm{L}},$$

and note that Ψ_n^* and Ψ_n are linked through $\Psi_n^*(\theta,\zeta) = q(k/n)\Psi_n(\theta,\zeta/q(k/n))$. Thus $(\widehat{\theta}_n, \widehat{\zeta}_n)$ minimizes $\|\Psi_n^*\|$ if and only if $(\widehat{\theta}_n, \widehat{\zeta}_n/q(k/n))$ minimizes $\|\Psi_n\|$. Furthermore, under suitable assumptions on g and Θ we have $\sigma\int gc_\theta d\mu_{\mathrm{L}} = \int gc_{\theta_0}d\mu_{\mathrm{L}}$ if and only if $\theta = \theta_0$ and $\sigma = 1$. Hence, if $\widehat{c}_n$ is close to c_{θ_0}, it is expected that $\widehat{\theta}_n$ will be close to θ_0 and that $\widehat{\zeta}_n/q(k/n)$ will be approximately 1.

Note that (2.13) only involves an integral of $\widehat{Q}_n$ while $\widetilde{c}_n$ also involves point-wise evaluation of this function. Since integration acts as smoothing, it can be expected that studying Ψ_n^* will require less regularity conditions than working with $\breve{\theta}$; see Section 2.4.1 for additional details.

2.3.3 Parametric estimation for spatial tail models

In this section, we show how the bivariate estimation procedures discussed earlier can be leveraged to obtain two different estimators for parametric spatial models, which can include both asymptotic dependence and independence. Assume that we observe n independent copies $Y_1, \ldots, Y_n$ of a spatial process Y at a finite set of locations $u_1, \ldots, u_d \in \mathcal{T}$. Denote the corresponding observations by $X_1, \ldots, X_n$ where $X_i = (X_i^{(1)}, \ldots, X_i^{(d)}) := (Y_i(u_1), \ldots, Y_i(u_d))$ are independent copies of the random

vector $X = (X^{(1)}, \ldots, X^{(d)}) := (Y(u_1), \ldots, Y(u_d)) \in \mathbb{R}^d$; see Einmahl et al. (2016) for a similar framework.

Let $\mathcal{P}$ denote the set of all subsets of $\{1, \ldots, d\}$ of size 2 interpreted as ordered pairs, so that elements of $\mathcal{P}$ will take the form $s = (s_1, s_2)$ with $s_1 < s_2$. In what follows, we will need to repeatedly make use of vectors $x \in \mathbb{R}^{|\mathcal{P}|}$ that are indexed by all pairs $s \in \mathcal{P}$. For such vectors we will assume that the pairs in $\mathcal{P}$ are ordered in a pre-specified order and will write $x^{(s)}$ for the entry of the vector x that corresponds to pair s.

Assume that for each pair s the random vector $(X^{(s_1)}, X^{(s_2)})$ satisfies (2.9) with scale function $q^{(s)}$ and survival tail function $c^{(s)}$. Following the ideas laid out in Section 2.3.1, define $\widehat{Q}_n^{(s)}$ as in (2.10) but based on the bivariate observations $(X_i^{(s_1)}, X_i^{(s_2)}), i = 1, \ldots, n$. We now discuss two parametric estimators for the functions $c^{(s)}$.

Assume that we start with a parametric model $\{c_\theta : \theta \in \widetilde{\Theta}\}$, $\widetilde{\Theta} \subseteq \mathbb{R}^{\tilde{p}}$, for bivariate survival tail functions and that each $c^{(s)}$ falls in this class. This implies that $\widetilde{\Theta}$ can be linked to a spatial parameter space $\Theta \subseteq \mathbb{R}^p$ through the relations $c^{(s)} = c_{h^{(s)}(\vartheta_0)}$, where $h^{(s)} : \Theta \to \widetilde{\Theta}$ for each pair s. To make this idea more concrete, consider the following example, which we will revisit in Sections 2.5.2 and 2.6.

Example 2.6. If the process Y is an inverted Brown–Resnick process on $\mathbb{R}^2$ (see Example 2.5) then X has an inverted Hüsler–Reiss distribution and the bivariate survival tail functions are of the form $c^{(s)}(x, y) = (xy)^{\theta^{(s)}}$, for some $\theta^{(s)} \in (1/2, 1)$. This determines the parametric class $\widetilde{\Theta}$. A more specific model of Brown–Resnick processes corresponds to the sub-family of fractal variograms (Kabluchko et al., 2009; Engelke et al., 2015), where

$$\theta^{(s)} = h^{(s)}((\alpha, \beta)) = \Phi\left(\frac{(\|u_{s_1} - u_{s_2}\|/\beta)^{\alpha/2}}{2}\right), \quad s \in \mathcal{P}, \tag{2.15}$$

where $u_j \in \mathbb{R}^2$ is the coordinate of the location j; see Section 2.6 for more motivation of this particular parametrization. The global parameter ϑ thus takes the form $\vartheta = (\alpha, \beta)$ and $\Theta = (0, 2] \times (0, \infty)$.

Given the setting above, we can thus compute parametric estimators $\widehat{\theta}_n^{(s)}, s \in \mathcal{P}$, by the methods for bivariate estimation discussed in Section 2.3.2, i.e., $(\widehat{\theta}_n^{(s)}, \widehat{\zeta}_n^{(s)})$ is the minimizer of $\left\|\Psi_n^{*(s)}(\theta, \zeta)\right\|$, where $\Psi_n^{*(s)}$ is defined as Ψ_n^* in (2.13) with $\widehat{Q}_n^{(s)}$ and an intermediate sequence $k^{(s)}$ replacing $\widehat{Q}_n$ and k. We obtain an estimator of the spatial parameter by least squares minimization,

$$\widehat{\vartheta}_n := \arg\min_{\vartheta \in \Theta} \sum_{s \in \mathcal{P}} \left\| h^{(s)}(\vartheta) - \widehat{\theta}_n^{(s)} \right\|^2. \tag{2.16}$$

As an alternative, one may use the relations $h^{(s)}$ between the spatial and bivariate parameters and minimize all the objective functions $\Psi_n^{*(s)}$ simultaneously, leading to the estimator

$$(\widetilde{\vartheta}_n, \widetilde{\zeta}_n) := \operatorname*{arg\,min}_{\vartheta \in \Theta, \zeta \in \mathbb{R}_+^{|\mathcal{P}|}} \sum_{s \in \mathcal{P}} \left\| \Psi_n^{*(s)}(h^{(s)}(\vartheta), \zeta^{(s)}) \right\|^2, \tag{2.17}$$

A theoretical analysis of the estimators $\widehat{\vartheta}_n$ and $(\widetilde{\vartheta}_n, \widetilde{\zeta}_n)$ is provided in Theorem 2.5. We further remark that the computational complexity of the proposed estimators is much lower than that of methods based on full likelihood and it compares favorably to pairwise likelihood. Additional details regarding the latter point can be found in Section 2.11.

Remark. At first glance the minimization problem in (2.17) seems to be computationally intractable since it contains $|\mathcal{P}| + \dim(\Theta)$ parameters and since $|\mathcal{P}|$ can be very large even for moderate dimension d. However, a closer inspection reveals that for given ϑ, partially minimizing the objective function in (2.17) over $\zeta \in \mathbb{R}_+^{|\mathcal{P}|}$ has the exact solution

$$\widehat{\zeta}_n^{(s)}(\vartheta) = \frac{\sum_{j=1}^q \int g_j(x,y) \widehat{Q}_n^{(s)}(k^{(s)}x/n, k^{(s)}y/n) dxdy}{\sum_{j=1}^q \int g_j(x,y) c_{h^{(s)}(\vartheta)}(x,y) dxdy},$$

whenever the right-hand side is positive for all s. This is satisfied if for instance g is positive everywhere and each $\widehat{Q}_n^{(s)}$ is based on at least one data point. Thus only numerical optimization over ϑ, which is usually low-dimensional, is required.

2.4 Theoretical results

We now present our main results on the asymptotic distributions of the various estimators introduced in Section 2.3. First, functional central limit theorems are stated for $\widehat{c}_n$, followed by our main result on the bivariate M-estimator. Finally, asymptotic normality of the processes $\widehat{c}_n^{(s)}$ and of the two parametric estimators in the spatial setting is established. The proofs of all main results are deferred to Section 2.7.

2.4.1 The bivariate setting

All results in this section will be derived under the following fundamental assumption.

Condition 2.1. (i) (2.9) holds uniformly on $\mathcal{S}^+ = \{(x,y) \in [0,\infty)^2 : x^2 + y^2 = 1\}$ with a function $q_1(t) = O(1/\log(1/t))$ and the function q is such that $\chi := \lim_{t \downarrow 0} q(t)/t \in [0,1]$ exists.

(ii) As $n \to \infty$, $m = m_n := nq(k/n) \to \infty$ and $\sqrt{m} q_1(k/n) \to 0$.

We note that in the proofs, (2.9) is required to hold locally uniformly on $[0,\infty)^2$, but by Lemma 2.2 uniformity on $\mathcal{S}^+$ implies uniformity over compact subsets of $[0,\infty)^2$. Condition 2.1(ii) is a standard assumption that makes certain bias terms negligible. It is not a model assumption; under Condition 2.1(i), there always exists a sequence k such that Condition 2.1(ii) is satisfied and thus all of the following discussion will focus on Condition 2.1(i). Notably and in contrast to most of the existing literature involving non-parametric estimation, Condition 2.1 does not assume any differentiability of the function c. In fact, our proofs show that all the regularity required on c can be derived from (2.9). Considering Section 2.3.1, it is possible to use a data-dependent value $\widehat{k}$. In following results, when this is done, we will assume that there is an unknown sequence k that satisfies Condition 2.1(ii), that m is defined as therein, and that $\widehat{k}$ is chosen so that $n\widehat{Q}_n(\widehat{k}/n, \widehat{k}/n) = m$.

We next discuss this condition in the examples introduced in Section 2.2.1. Proofs for the claims in the examples below can be found in Sections 2.9 and 2.10.

Example 2.7 (*Example 2.1, continued*)**.** Most of the literature on asymptotic analysis of estimators of the stable tail dependence function L or related quantities under domain of attraction conditions makes some version of the following assumption

$$\frac{1}{t}\mathbb{P}\left(F_1(X) \geq 1 - tx \text{ and } F_2(Y) \geq 1 - ty\right) - R(x,y) = O(\widetilde{q}_1(t)) \quad x, y \in [0,\infty); \tag{2.18}$$

for a non-vanishing function R on $[0,\infty)^2$ where $\widetilde{q}_1(t) = o(1)$, see for instance condition (C2) in Einmahl et al. (2008) or the discussion in Bücher et al. (2019). A simple computation involving the inclusion-exclsion formula further shows that this is equivalent to assuming that convergence in (2.1) takes place with rate $O(\widetilde{q}_1(t))$ and that $L(x,y) = x + y - R(x,y)$. Clearly (2.18) implies Condition 2.1(i) with $q(t) = tR(1,1)$, $c(x,y) = R(x,y)/R(1,1)$ and $q_1(t) = \widetilde{q}_1(t)$.

Example 2.8 (*Example 2.2, continued*)**.** Let (X,Y) be a bivariate inverted max-stable distribution and assume that there exists a constant $C < \infty$ such that for all $u, v > 0$,

$$\left|L(1+u, 1+v) - L(1,1) - \dot{L}_1(1,1)u - \dot{L}_2(1,1)v\right| \leq C\left(u^2 + v^2\right),$$

where $\dot{L}_j$ represent the directional partial derivatives of L from the right. In particular, it suffices for L to be twice differentiable. Then the random vector (X,Y) satisfies Condition 2.1(i) with $q(t) = t^{L(1,1)}$, $c(x,y) = x^{\dot{L}_1(1,1)}y^{\dot{L}_2(1,1)}$ and $q_1(t) = 1/\log(1/t)$. Moreover, $\dot{L}_j(1,1) \in (0,1]$ and $\dot{L}_1(1,1) + \dot{L}_2(1,1) = L(1,1) \in (1,2]$.

Example 2.9 (*Example 2.3, continued*)**.** Let (X,Y) be a random scale construction as defined in (2.6) and set $\lambda = \alpha_R/\alpha_W$. Then (X,Y) satisfies Condition 2.1(i) with

functions q, c and q_1 determined by λ as in Table 2.1 below.

Range of λ	$q(t)$	$c(x,y)$	$q_1(t)$
$(0,1)$	$K_\lambda t$	$\frac{2-\lambda}{2(1-\lambda)}\mu - \frac{\lambda}{2(1-\lambda)}\mu^{1/\lambda}\mathcal{M}^{1-1/\lambda}$	$t^{1/\lambda-1}$
1	$\frac{K_\lambda t}{\log(1/t)+\log\log(1/t)}$	$\mu\Big(1+\frac{1}{2}\log\Big(\frac{\mathcal{M}}{\mu}\Big)\Big)$	$1/\log(1/t)$
$(1,2)$	$K_\lambda t^\lambda$	$\frac{\lambda}{2(\lambda-1)}\mu\mathcal{M}^{\lambda-1} - \frac{2-\lambda}{2(\lambda-1)}\mu^\lambda$	$t^{(\lambda-1)\wedge(2-\lambda)}$
2	$K_\lambda t^2\log(1/t)$	$\mu\mathcal{M}$	$1/\log(1/t)$
$(2,\infty)$	$K_\lambda t^2$	$\mu\mathcal{M}$	$t^{\lambda-2}$

Table 2.1: Tail expansion of the random scale model in (2.6), here we set $\mu := x \wedge y, \mathcal{M} := x \vee y$, and K_λ is a positive constant given in (2.53).

Asymptotic theory for non-parametric estimators

In this section we consider the estimator $\widehat{c}_n$ from (2.11). Since the process convergence results differ under asymptotic dependence and independence, we discuss these settings separately. Our first result deals with asymptotic independence.

Theorem 2.1 (Asymptotic normality of $\widehat{c}_n$ under asymptotic independence)**.** *Assume Condition 2.1. Then under asymptotic independence, i.e., when $\chi = 0$,*

$$W_n := \sqrt{m}(\widehat{c}_n - c) \rightsquigarrow W,$$

in $\ell^\infty([0,T]^2)$, for any $T < \infty$. Here, W is a centered Gaussian process with covariance structure given by $\mathbb{E}\left[W(x,y)W(x',y')\right] = c(x \wedge x', y \wedge y')$. The same remains true if k is replaced by $\widehat{k}$ as described after Condition 2.1.

Note that process convergence of the estimator $\widetilde{c}_n$ from (2.12) can be obtained from the above result through a straightforward application of the functional delta method. This will not be needed in the theory for M-estimators in the next section and details are omitted for the sake of brevity.

Asymptotic properties of $\widehat{c}_n$ were considered in Draisma et al. (2004). However, the proof of the corresponding result (Lemma 6.1) in the latter reference makes the additional assumption that the partial derivatives of c exist and are continuous on $[0,T]^2$ (cf. Peng, 1999, Theorem 2.2). In contrast, we are able to show that no condition on existence or continuity of partial derivatives is required. This is a considerable strengthening of the result which further allows to handle many interesting examples that were not covered by the results of Draisma et al. (2004). Indeed, both the popular class of inverted max-stable distributions in Example 2.2 and the random scale construction in Example 2.3 lead to functions c that fail to have continuous or even bounded partial derivatives. Before moving on to discussing results under asymptotic dependence, we briefly comment on some of the main ideas of the proof.

Remark (Main ideas of the proof of Theorem 2.1). The proof relies on the decomposition

$$\widehat{c}_n(x,y) - c(x,y) = \left\{ \frac{Q_n(\frac{ku_n(x)}{n}, \frac{kv_n(y)}{n})}{q(k/n)} - c(u_n(x), v_n(y)) \right\} + \big(c(u_n(x), v_n(y)) - c(x,y)\big),$$

where

$$u_n(x) := \frac{n}{k} U_{n,\lfloor kx \rfloor} \quad \text{and} \quad v_n(y) := \frac{n}{k} V_{n,\lfloor ky \rfloor},$$

and $U_{n,k}$ and $V_{n,k}$ denote the kth order statistics of $1 - F_1(X_1), \ldots, 1 - F_1(X_n)$ and $1 - F_2(Y_1), \ldots, 1 - F_2(Y_n)$, respectively with $U_{n,0} = V_{n,0} = 0$. The core difficulty is to show that the difference $c(u_n(x), u_n(y)) - c(x,y)$ is negligible. Under the assumption of the existence and continuity of partial derivatives of c on $[0,T]^2$ made in Draisma et al. (2004) this is a direct consequence of the fact that under asymptotic independence $\sqrt{m}(u_n(x) - x) = o_{\mathbb{P}}(1)$. Dropping this assumption considerably complicates the theoretical analysis. The proof strategy is to derive bounds on increments of $c(x,y)$ for x, y close to 0 where the partial derivatives of c can become unbounded (see Lemmas 2.7 and 2.8) and to combine those bounds with subtle results on weighted weak convergence of $u_n(x) - x$ as a process in x; see Lemma 2.3 where we essentially leverage the findings of Csörgő and Horváth (1987).

We next turn to the case of asymptotic dependence. Results on convergence of $\widehat{c}_n$ in the space ℓ^∞ are well known under this regime; they are equivalent to similar results about estimated stable tail dependence functions (cf. Huang, 1992). However, they require the existence and continuity of partial derivatives of L or, equivalently, c. As shown in Einmahl et al. (2008, 2012), the latter condition is restrictive and in fact not necessary to derive asymptotic normality of M-estimators.

The treatment of M-estimators in Einmahl et al. (2008, 2012) involves a direct analysis of certain integrals without using process convergence in $\ell^\infty([0,T]^d)$. While this approach could be transferred to our setting, we will instead follow a strategy put forward in Bücher et al. (2014) and prove weak convergence of $\widehat{c}_n$ with respect to the hypi-metric introduced therein. This approach will turn out to generalize much more easily when we deal with spatial estimation problems. Convergence with respect to this metric holds without any assumptions on the existence of partial derivatives and is sufficiently strong to guarantee convergence of integrals which is needed for the analysis of M-estimators.

Let $\dot{c}_1$ denote the partial derivative of c with respect to x from the left and $\dot{c}_2$ denote its partial derivative with respect to y from the right. Under asymptotic dependence, $c(x,y) \propto x + y - L(x,y)$ is concave since L is convex (de Haan and Ferreira, 2006, Proposition 6.1.21), hence those directional partial derivatives exist everywhere on $(0,\infty)^2$, by Theorem 23.1 of Rockafellar (1970). The definition can be extended to

$[0,\infty)^2$ be setting $\dot{c}_1(0,y)$ to be the derivative from the right instead of from the left.

To describe the limiting distribution, recall that $\chi = \lim_{t\to 0} q(t)/t \in [0,1]$ is positive only in the case of asymptotic dependence. For $(x,y),(x',y') \in [0,\infty)^2$, define

$$\Lambda((x,y),(x',y')) = \begin{bmatrix} c(x\wedge x', y\wedge y') & \chi c(x\wedge x', y) & \chi c(x, y\wedge y') \\ \chi c(x\wedge x', y') & \chi(x\wedge x') & \chi^2 c(x, y') \\ \chi c(x', y\wedge y') & \chi^2 c(x', y) & \chi(y\wedge y') \end{bmatrix}, \tag{2.19}$$

and let $(W, W^{(1)}, W^{(2)})$ be an $\mathbb{R}^3$-valued, zero mean Gaussian process on $[0,\infty)^2$ with covariance function Λ. Note that W is the limiting process in Theorem 2.1, that $W^{(1)}(x,y)$ is constant in y and that $W^{(2)}(x,y)$ is constant in x.

Theorem 2.2 (Asymptotic normality of $\widehat{c}_n$ under asymptotic dependence)**.** *Assume Condition 2.1. Then under asymptotic dependence, i.e., when $\chi > 0$,*

$$W_n \rightsquigarrow B := W - \dot{c}_1 W^{(1)} - \dot{c}_2 W^{(2)}$$

in $(L^\infty([0,T]^2), d_{hypi})$, for any $T < \infty$. Here, W_n is defined as in Theorem 2.1. The same remains true if k is replaced by $\widehat{k}$ as described after Condition 2.1.

Note that weak convergence in the above theorem takes place in $(L^\infty([0,T]^2), d_{\text{hypi}})$ where $L^\infty([0,T]^2)$ corresponds to equivalence classes of functions in $\ell^\infty([0,T]^2)$ with respect to the hypi-(semi-)metric d_{hypi}, see Bücher et al. (2014) for additional details.

The proof of Theorem 2.2 follows by adapting the arguments given in Bücher et al. (2014) for the function L and builds on the fact that under asymptotic dependence the function c is differentiable almost everywhere. Note however that, in contrast to similar results in Bücher et al. (2014), our limiting process is stated without appealing to lower semi-continuous extensions. This type of statement is inspired by the representation of certain integrals in Einmahl et al. (2012) and is possible in the bivariate setting due to concavity of c under asymptotic dependence. Additional comments on the representation of the limiting process are given in Section 2.4.1 below.

Remark. In order to obtain asymptotic results for our M-estimator, weak convergence of $\int g W_n d\mu_{\text{L}}$ to $\int g B d\mu_{\text{L}}$ is sufficient. Under asymptotic dependence, this is seen to follow from Theorem 2.2 (see the proof of Theorem 2.3). However, this process convergence result is not necessary. An approach that is used in Einmahl et al. (2012) is to write an expression for the random vector $\int g W_n d\mu_{\text{L}}$ and directly work out its weak limit. With this strategy, $\dot{c}_j$ may be defined as left or right derivatives without problem as $\int \dot{c}_j W^{(j)} d\mu_{\text{L}}$ will be unchanged. In contrast, proving weak hypi-convergence of W_n to B makes our results more general and more easily generalized to the spatial framework. The cost of doing so is that the directional derivatives $\dot{c}_j$ must be chosen

in a specific way; see Lemma 2.9.

Remark. Recall that under asymptotic independence, process convergence of $\widetilde{c}_n$ could be obtained from Theorem 2.1 by a simple application of the delta method. This is no longer the case in the general setting of Theorem 2.2 because weak convergence with respect to the hypi-metric does not imply convergence of $W_n(1,1)$, unless the limiting process B has sample paths which are a.s. continuous in $(1,1)$. The latter happens only if the partial derivatives of c exist and are continuous in $(1,1)$. Under this additional assumption convergence of $\widetilde{c}_n$ with respect to the hypi-metric can be obtained.

Asymptotic theory for bivariate M-estimators

Equipped with the process convergence tools from the previous section, we proceed to analyze the M-estimator introduced in Section 2.3.2. Consistency is established by standard arguments, and for the sake of brevity we do not state the corresponding results here. In the present section, we focus on the asymptotic distribution. Define the objective function $\Psi : \Theta \times \mathbb{R}_+ \to \Psi(\Theta \times \mathbb{R}_+) \subseteq \mathbb{R}^q$ by

$$\Psi(\theta, \sigma) := \sigma \int g c_\theta \, d\mu_{\mathrm{L}} - \int g c \, d\mu_{\mathrm{L}}. \tag{2.20}$$

Clearly, $\Psi(\theta_0, 1) = 0$. In addition, assume that $(\theta_0, 1)$ is a unique, well separated zero of Ψ and let $J_\Psi(\theta, \sigma)$ denote the Jacobian matrix of Ψ for points $(\theta, \sigma) \in \Theta \times \mathbb{R}_+$ where it exists.

Define $\Gamma((x,y),(x',y'))$ as $c(x \wedge x', y \wedge y')$ under asymptotic independence and as

$$(1, -\dot{c}_1(x,y), -\dot{c}_2(x,y))\Lambda((x,y),(x',y'))(1, -\dot{c}_1(x',y'), -\dot{c}_2(x',y'))^\top$$

otherwise, where Λ is defined in (2.19). Recall from the previous section that these directional derivatives always exist when $\chi > 0$ since in this case c is concave. In fact, $\Gamma((x,y),(x',y'))$ is the covariance between $W(x,y)$ and $W(x',y')$ (under asymptotic independence) or between $B(x,y)$ and $B(x',y')$ (under asymptotic dependence). Hence in those two regimes,

$$A := \int_{[0,T]^4} g(x,y) g(x',y')^\top \Gamma((x,y),(x',y')) dx dy dx' dy' \in \mathbb{R}^{q \times q}$$

is the covariance matrix of the random vector $\int g W d\mu_{\mathrm{L}}$ or $\int g B d\mu_{\mathrm{L}}$, respectively. We are now ready to state the main result of this section: asymptotic normality of $(\widehat{\theta}_n, \widehat{\zeta}_n)$, which holds under both asymptotic dependence and independence.

Theorem 2.3 (Asymptotic normality of $\widehat{\theta}_n$)**.** *Assume that Ψ has a unique, well separated zero at $(\theta_0, 1)$ and is differentiable at that point with Jacobian $J := J_\Psi(\theta_0, 1)$ of full rank $p+1$, $p = \dim(\Theta)$. Further assume Condition 2.1. Then the estimators $(\widehat{\theta}_n, \widehat{\zeta}_n)$ defined in (2.14) satisfy*

$$\sqrt{m}\Big(\Big(\widehat{\theta}_n, \frac{n\widehat{\zeta}_n}{m}\Big) - (\theta_0, 1)\Big) \rightsquigarrow N(0, \Sigma)$$

where $\Sigma := (J^\top J)^{-1} J^\top A J (J^\top J)^{-1}$. The same remains true if k is replaced by $\widehat{k}$ as described after Condition 2.1.

While for simplicity the estimator is defined as an exact minimizer, the same result can be obtained for an approximate minimizer. Precisely, it is obvious from the proof of Theorem 2.3 that as long as $\Psi_n^*(\widehat{\theta}_n, \widehat{\zeta}_n) = \inf_{\theta,\zeta} \Psi_n^*(\theta, \zeta) + o_{\mathbb{P}}(\sqrt{m}/n)$, the conclusion still holds. Finally, recall that the coefficient of tail dependence η can be recovered from the function c since the latter is homogeneous of order $1/\eta$, and this relation always holds. Therefore, inside the assumed parametric model, η can be represented as a function $\eta(\theta)$. The asymptotic distribution of the resulting estimator can be obtained by a direct application of the delta method and details are omitted for the sake of brevity.

2.4.2 The spatial setting

In this section we assume the framework of Section 2.3.3 and establish asymptotic properties of the estimators therein. For each pair $s \in \mathcal{P}$, let $k^{(s)}$ be an intermediate sequence and define

$$\widehat{c}_n^{(s)}(x, y) := \frac{\widehat{Q}_n^{(s)}\left(k^{(s)}x/n, k^{(s)}y/n\right)}{q^{(s)}(k^{(s)}/n)}.$$

From Section 2.4.1, the asymptotic distribution of $\widehat{c}_n^{(s)}$ is known under suitable conditions. However, as the spatial estimators $\widehat{\vartheta}_n$ and $\widetilde{\vartheta}_n$ are based on all pairs, a joint convergence statement about all processes $\widehat{c}_n^{(s)}$ is necessary. This will require an additional assumption which we present and discuss next.

Let $F^{(1)}, \ldots, F^{(d)}$ denote the marginal distribution functions of the random vector X, which itself consists of the spatial process Y evaluated at d different locations. In order to obtain the asymptotic covariance between different processes $\widehat{c}_n^{(s)}$, we need to ensure that certain multivariate tail probabilities converge. Partition the set $\mathcal{P}$ into $\mathcal{P}_I$ and $\mathcal{P}_D$, consisting of the asymptotically independent and asymptotically dependent pairs, respectively. In the formulation of the following assumption, $s = (s_1, s_2)$ and $s^i = (s_1^i, s_2^i)$ are used to denote pairs. For brevity, $x^i = (x_1^i, x_2^i)$ is also used to denote

a point in $[0,\infty)^2$.

Condition 2.2. For every $s \in \mathcal{P}$, $(X^{(s_1)}, X^{(s_2)})$ satisfies Condition 2.1(i) with functions $q^{(s)}, q_1^{(s)}, c^{(s)}$ and $\chi^{(s)} := \lim_{t\downarrow 0} q^{(s)}(t)/t$ exists. Intermediate sequences $k^{(s)}$ are chosen so that $m^{(s)} := nq^{(s)}(k^{(s)}/n) \to \infty$ and $\sqrt{m^{(s)}}q_1^{(s)}(k^{(s)}/n) \to 0$. For pairs $s^1, s^2 \in \mathcal{P}$, points $x^1, x^2 \in [0,\infty)^2$ and sets J of two-dimensional vectors with entries in $\{1,2\}$, let

$$\Gamma_n\left(s^1, s^2, x^1, x^2; J\right) = \frac{n}{\sqrt{m^{(s^1)}m^{(s^2)}}}\mathbb{P}\Big(F^{(s^i_j)}(X^{(s^i_j)}) \geq 1 - \frac{k^{(s^i)}x^i_j}{n}, \quad (i,j) \in J\Big).$$

We assume that the sequences $k^{(s)}$ are chosen such that the limits

$$\begin{aligned}
\Gamma^{(s^1,s^2)}(x^1,x^2) &:= \lim_{n\to\infty} \Gamma_n\left(s^1, s^2, x^1, x^2; \{(1,1),(1,2),(2,1),(2,2)\}\right), \ s^1, s^2 \in \mathcal{P},\\
\Gamma^{(s^1,s^2,j)}(x^1,x^2) &:= \chi^{(s^2)} \lim_{n\to\infty} \Gamma_n\left(s^1, s^2, x^1, x^2; \{(1,1),(1,2),(2,j)\}\right), \ s^1 \in \mathcal{P}, s^2 \in \mathcal{P}_D,\\
\Gamma^{(s^1,j^1,s^2,j^2)}(x^1,x^2) &:= \chi^{(s^1)}\chi^{(s^2)} \lim_{n\to\infty} \Gamma_n\left(s^1, s^2, x^1, x^2; \{(1,j^1),(2,j^2)\}\right), \ s^1, s^2 \in \mathcal{P}_D,
\end{aligned}$$

exist for all $j, j^i \in \{1,2\}$, and that the convergence is locally uniform over $x^1, x^2 \in [0,\infty)^2$.

We next discuss the above condition in three special cases of particular interest. The first two are processes in the domain of attraction of max-stable processes and inverted max-stable processes. The third one is a mixture process appearing in Wadsworth and Tawn (2012), which can have asymptotically dependent and independent pairs simultaneously.

Example 2.10 (*Example 2.4, continued*)**.** If Y is in the max-domain of attraction of a max-stable process, then X is in the max-domain of attraction of a max-stable distribution G on $\mathbb{R}^d$ with stable tail dependence function

$$L(x_1,\ldots,x_d) := \lim_{t\downarrow 0} \frac{1}{t}\mathbb{P}\left(F^{(1)}(X^{(1)}) \geq 1 - tx_1 \text{ or } \ldots \text{ or } F^{(d)}(X^{(d)}) \geq 1 - tx_d\right), \quad x_j \geq 0;$$

see (2.1). If moreover the convergence is locally uniform over $(x_1,\ldots,x_d) \in [0,\infty)^d$ and if every pair is asymptotically dependent, then Condition 2.2 holds. Note that this is automatically satisfied if Y itself is max-stable. The sequences $k^{(s)}$ can be chosen all equal to k, say, and for every pair s, $m^{(s)}/k \to \chi^{(s)} > 0$. The sequences $m^{(s)}$ can also be chosen all asymptotically equivalent to m, say, by choosing $k^{(s)} = m/\chi^{(s)}$. The limiting covariance terms can all be deduced from L by straightforward calculations.

Example 2.11 (*Example 2.5, continued*)**.** If Y is an inverted max-stable process, then X has an inverted max-stable distribution, and we assume that the associated stable tail dependence function L is component-wise strictly increasing. The latter is

trivially satisfied if X has a positive density. Then if all the pairwise functions $L^{(s)}$ satisfy the quadratic expansion introduced in Example 2.8, Condition 2.2 is satisfied and the sequences $k^{(s)}$ can be chosen so that the $m^{(s)}$ are all equal, that is, for every pair $s \in \mathcal{P}$, $m^{(s)} = m$ for some intermediate sequence m. Here, $\mathcal{P}_D$ is empty so the only required covariance terms are (see Section 2.9)

$$\Gamma^{(s^1,s^2)}(x^1, x^2) = \begin{cases} c^{(s)}(x_1^1 \wedge x_1^2, x_2^1 \wedge x_2^2), & s^1 = s^2 = s, \\ 0, & s^1 \neq s^2 \end{cases}.$$

For instance, any inverted Brown–Resnick process (or rather the implied inverted d-dimensional Hüsler–Reiss distribution corresponding to the d observed locations) satisfies Condition 2.2 as long as the aforementioned d-variate distribution has a density. The latter can easily be checked (e.g., Engelke and Hitz, 2020, Corollary 2).

Example 2.12 (*Wadsworth and Tawn (2012), Section 4*)**.** Let Z be a max-stable process and Z' be an inverted max-stable process, both with unit Fréchet margins. Suppose that Z' satisfies the monotonicity condition stated in Example 2.11, and additionally that none of its pairwise distributions $(Z'(u_1), Z'(u_2))$ is perfectly independent. Let $a \in (0, 1)$ and define the process Y by

$$Y(u) := \max\{aZ(u), (1-a)Z'(u)\}.$$

Then Y also has unit Fréchet margins. If Z becomes independent at a certain spatial distance, the process Y transitions between asymptotic dependence and independence at that distance. An instance of such a max-stable process Z is found in the second example after Theorem 1 of Schlather (2002), assuming that the Radius R of the random disks is bounded (see also Davison et al., 2012a, eq. (23) and the discussion that precedes).

The process Y can be shown to satisfy Condition 2.2 if the sequences $k^{(s)}$ are chosen so that the $m^{(s)}$ are all equal. The terms $\Gamma^{(s^1,s^2)}$, $\Gamma^{(s^1,s^2,j)}$ and $\Gamma^{(s^1,j^1,s^2,j^2)}$ are mostly determined by the process Z, as in Example 2.10; see Section 2.9 for details.

Joint distribution of non-parametric estimators

The joint limiting behavior of the processes $\widehat{c}_n^{(s)}$ relies on $\big((W^{(s)})_{s\in\mathcal{P}}, (W^{(s,j)})_{s\in\mathcal{P}_D, j\in\{1,2\}}\big)$, a collection of centered Gaussian processes on $[0,\infty)^2$. The covariance between $W^{(s)}(x,y)$ and $W^{(s')}(x',y')$ is given by $\Gamma^{(s,s')}((x,y),(x',y'))$, the covariance between $W^{(s)}(x,y)$ and $W^{(s',j)}(x',y')$ takes the form $\Gamma^{(s,s',j)}((x,y),(x',y'))$, and the covariance between $W^{(s,j)}(x,y)$ and $W^{(s',j')}(x',y')$ is equal to $\Gamma^{(s,j,s',j')}((x,y),(x',y'))$. For $s \in \mathcal{P}_I$,

let $B^{(s)} = W^{(s)}$ and for $s \in \mathcal{P}_D$, let

$$B^{(s)} = W^{(s)} - \dot{c}_1^{(s)} W^{(s,1)} - \dot{c}_2^{(s)} W^{(s,2)},$$

where $\dot{c}_j^{(s)}$ are defined similarly to $\dot{c}_j$ in Section 2.4.1.

Theorem 2.4 (Asymptotic normality of $\widehat{c}_n^{(s)}$)**.** *Assume Condition 2.2. Then*

$$\big(W_n^{(s)}\big)_{s\in\mathcal{P}} := \big(\sqrt{m^{(s)}}(\widehat{c}_n^{(s)} - c^{(s)})\big)_{s\in\mathcal{P}} \rightsquigarrow \big(B^{(s)}\big)_{s\in\mathcal{P}}$$

in the product space $(L^\infty([0,T]^2), d_{hypi})^{|\mathcal{P}|}$*, for any* $T < \infty$*. The same remains true if each* $k^{(s)}$ *is replaced by the data-dependent sequence* $\widehat{k}^{(s)}$ *as described after Condition 2.1.*

The preceding result can be applied in all generality as long as the four-dimensional tails of the spatial process of interest are sufficiently smooth. The admissible settings include, but are far from limited to, Examples 2.10 to 2.12.

According to Bücher et al. (2014), convergence in the hypi-metric is equivalent to uniform convergence when the limit is a continuous function. The process $B^{(s)}$ clearly has almost surely continuous sample paths under asymptotic independence, as well as under asymptotic dependence if the partial derivatives of c exist everywhere and are continuous. It follows that in those cases $W_n^{(s)}$ converges in $(\ell^\infty([0,T]^2), \|\cdot\|_\infty)$. In fact, one may replace the product space in the result above by $\otimes_{s\in\mathcal{P}}\mathbb{D}^{(s)}$, where $\mathbb{D}^{(s)}$ represents either $\ell^\infty([0,T]^2)$ equipped with the supremum distance (if $s \in \mathcal{P}_I$ or c has continuous partial derivatives) or $L^\infty([0,T]^2)$ equipped with the hypi-metric (otherwise). In particular, for processes where every pair is asymptotically independent such as inverted max-stable processes, the hypi-metric can be replaced by the supremum distance everywhere.

Asymptotics for parametric estimators

We now show how Theorem 2.4 leads to asymptotic results for the parametric estimators $\widehat{\vartheta}_n$ and $\widetilde{\vartheta}_n$ introduced in (2.16) and (2.17). Recall the setting of Section 2.3.3, and in particular the functions $h^{(s)} : \Theta \to \widetilde{\Theta}$ and the relation $c^{(s)} = c_{h^{(s)}(\vartheta_0)}$. Similarly to the bivariate setting, define

$$\Psi^{(s)} : \widetilde{\Theta} \times \mathbb{R}_+ \to \mathbb{R}^q, \quad \Psi^{(s)}(\theta, \sigma) = \sigma \int g c_\theta d\mu_{\mathrm{L}} - \int g c^{(s)} d\mu_{\mathrm{L}}.$$

In the bivariate setting, we required Ψ to be differentiable and have a unique well-separated zero. In the spatial setting we need a comparable assumption.

Condition 2.3. For every pair $s \in \mathcal{P}$, the functions $\Psi^{(s)}$ and $h^{(s)}$ are continuously differentiable at the points $(h^{(s)}(\vartheta_0), 1)$ and ϑ_0, respectively, with Jacobian matrices $J_{\Psi^{(s)}}(h^{(s)}(\vartheta_0), 1)$ and $J_{h^{(s)}}(\vartheta_0)$ of full ranks $\widetilde{p}+1$ and p. Additionally (i) or (ii) holds.

(i) The functions $\Psi^{(s)}$ and $\vartheta \mapsto (h^{(s)}(\vartheta) - h^{(s)}(\vartheta_0))_{s\in\mathcal{P}}$ have a unique, well separated zero at the points $(h^{(s)}(\vartheta_0), 1)$ and ϑ_0, respectively.

(ii) The function $(\vartheta, \sigma) \mapsto (\Psi^{(s)}(h^{(s)}(\vartheta), \sigma^{(s)}))_{s\in\mathcal{P}}$ as a function on $\Theta \times \mathbb{R}_+^{|\mathcal{P}|}$ has a unique, well separated zero at the point $(\vartheta_0, 1, \ldots, 1)$.

Assuming both parts of Condition 2.3, we now introduce the notation that is needed to define the limiting covariance matrices of the two estimators. In the following, elements of a vector $x \in \mathbb{R}^{q|\mathcal{P}|}$ are ordered by pair $s \in \mathcal{P}$ first, and then by dimension $j \in \{1, \ldots, q\}$. The same convention is used when ordering the rows or columns of a matrix.

Letting $B^{(s)}$ denote the limiting Gaussian processes appearing in Theorem 2.4, consider the matrix $A \in \mathbb{R}^{q|\mathcal{P}|\times q|\mathcal{P}|}$ with blocks of the form

$$A^{(s,s')} := \int_{[0,T]^4} g(x,y)g(x',y')^\top \mathbb{C}\mathrm{ov}\left(B^{(s)}(x,y); B^{(s')}(x',y')\right) dxdydx'dy'.$$

Let $\mathcal{D} \in \mathbb{R}^{\widetilde{p}|\mathcal{P}|\times q|\mathcal{P}|}$ be a block-diagonal matrix with blocks given by

$$\mathcal{D}^{(s)} := \left[\left(J_{\Psi^{(s)}}(h^{(s)}(\vartheta_0), 1)^\top J_{\Psi^{(s)}}(h^{(s)}(\vartheta_0), 1)\right)^{-1} J_{\Psi^{(s)}}(h^{(s)}(\vartheta_0), 1)^\top\right]_{1:\widetilde{p},1:q} \in \mathbb{R}^{\widetilde{p}\times q}, \tag{2.21}$$

where $s \in \mathcal{P}$ and $[M]_{1:\widetilde{p},1:q}$ indicates the sub-matrix consisting of rows 1 to $\widetilde{p}$ and columns 1 to q of the matrix M. Define $J_1 \in \mathbb{R}^{\widetilde{p}|\mathcal{P}|\times p}$ by stacking the matrices $J_{h^{(s)}}(\vartheta_0)$, $s \in \mathcal{P}$, on top of each other. Denote by $(e^{(s)})^\top$ the unit vector in $\mathbb{R}^{|\mathcal{P}|}$ with a one in the position corresponding to the pair s and let $J_2 \in \mathbb{R}^{q|\mathcal{P}|\times(p+|\mathcal{P}|)}$ be obtained by stacking the matrices

$$J_{\Psi^{(s)}}(h^{(s)}(\vartheta_0), 1) \begin{bmatrix} J_{h^{(s)}}(\vartheta_0) & 0 \\ 0 & e^{(s)} \end{bmatrix} \in \mathbb{R}^{q\times(p+|\mathcal{P}|)}, \quad s \in \mathcal{P},$$

on top of each other. Finally, define

$$\Sigma_1 = (J_1^\top J_1)^{-1} J_1^\top \mathcal{D} A \mathcal{D}^\top J_1 (J_1^\top J_1)^{-1}, \qquad \Sigma_2 = (J_2^\top J_2)^{-1} J_2^\top A J_2 (J_2^\top J_2)^{-1}.$$

Theorem 2.5 (Asymptotic normality of the estimators of ϑ)**.** *Assume Condition 2.2 and suppose that the sequences $m^{(s)}$ are all asymptotically equivalent to m, say. Then*

under Condition 2.3(i), the estimator defined in (2.16) satisfies

$$\sqrt{m}\left(\widehat{\vartheta}_n - \vartheta_0\right) \rightsquigarrow N(0, \Sigma_1)$$

and under Condition 2.3(ii), the estimators defined in (2.17) satisfy

$$\sqrt{m}\Big(\Big(\widetilde{\vartheta}_n, \frac{n\widetilde{\zeta}_n}{m}\Big) - (\vartheta_0, 1, \ldots, 1)\Big) \rightsquigarrow N(0, \Sigma_2),$$

where Σ_1 and Σ_2 are as above. The same remains true if each $k^{(s)}$ is replaced by the data-dependent sequence $\widehat{k}^{(s)}$, based on the same sequence m, as described after Condition 2.1.

The assumption of asymptotic equivalence of all $m^{(s)}$ can be substantially relaxed. Otherwise, a simple way to satisfy it is to select one m and use data-driven sequences $\widehat{k}^{(s)}$.

2.5 Simulations

2.5.1 Bivariate distributions

In this section we illustrate the performance of the proposed methodology for bivariate data. We simulate samples from the bivariate vector $(X + X', Y + Y')$, where (X, Y) is the signal and (X', Y') is and independent noise vector. We consider three different models for the bivariate distributions (X, Y).

(M1) The inverted Hüsler–Reiss model from Example 2.2(i) with unit Fréchet margins, whose corresponding class of functions c takes the form $c_\theta(x, y) = (xy)^\theta$ where $\theta \in (1/2, 1]$.

(M2) The inverted asymmetric logistic model from Example 2.2(ii) with fixed $r = 2$ and unit Fréchet margins. We fit the full parametric model $\{c_\theta(x, y) = x^{\theta_1} y^{\theta_2} : \theta \in \Theta\}$, where $\Theta := \{(\theta_1, \theta_2) \in (0, 1]^2 : \theta_1 + \theta_2 > 1\}$, even though due to our choice of r the only attainable parameters are approximately the square $[0.7, 1]^2$; see Figure 2.4.

(M3) The random scale construction from Example 2.3 where we fix $\alpha_W = 1$ and vary α_R. The collection of possible functions $c = c_\lambda$, $\lambda \in (0, 2)$ is given in Table 2.1.

Figures 2.10 to 2.12 show realizations of models M1–M3 corresponding to different parameter values and rescaled to unit exponential margins for illustration.

As a noise vector we simulate samples of (X', Y'), where X' and Y' are independent with Pareto distribution function $1 - 1/x^4$, $x \geq 1$. Note that this tail is lighter than that of the marginal distributions in all three models; it can be shown that this additive noise does not affect the functions q and c of (X, Y).

All of the results that follow are based on 1000 simulation repetitions and samples of size $n = 5000$. In all the simulations, we use the same weight function (represented by g in (2.13)), which we now describe. Consider the following rectangles: $I_1 := [0,1]^2$, $I_2 := [0,2]^2$, $I_3 := [1/2, 3/2]^2$, $I_4 := [0,1] \times [0,3]$ and $I_5 := [0,3] \times [0,1]$. The function $g : \mathbb{R}^2 \to \mathbb{R}^5$ is given by

$$g(x,y) := \left(\mathbb{1}\left\{(x,y) \in I_1\right\} / a_{1,\theta_{\mathrm{REF}}}, \ldots, \mathbb{1}\left\{(x,y) \in I_5\right\} / a_{5,\theta_{\mathrm{REF}}}\right)^\top \tag{2.22}$$

where $a_{j,\theta_{\mathrm{REF}}} := \int_{I_j} c_{\theta_{\mathrm{REF}}} d\mu_{\mathrm{L}}$ and θ_{REF} is simply a reference point in the parameter space that ensures that all components of g have comparable magnitude. In the three models above, the reference points are 0.6, $(0.6, 0.6)$ and 1, respectively. The rectangles are chosen in order to capture various aspects of the function c: I_3 contains information about the unknown scale ζ (recall that we scale c so that $c(1,1) = 1$). The rectangles I_1, I_2 are geared towards determining homogeneity properties of c since $I_2 = 2I_1$ and are especially useful for estimating η. The rectangles I_4, I_5 are informative about asymmetry of the function c with respect to its arguments. Different choices of the weight function would be possible, and the best choice will be different for each model under consideration and even for each specific parameter value within a given model class. Nevertheless, the aforementioned choice seems close to optimal for all the models considered here. In Section 2.12, a sensitivity analysis is carried out where we repeat the simulation study with different weight functions that are constructed by considering only some of the rectangles $I_1, \ldots, I_5$ instead of all five. See also Einmahl et al. (2008, 2012) for a related discussion in the estimation of stable tail dependence functions.

The inverted Hüsler–Reiss model (M1)

Figure 2.1 shows the effect of k on the estimation performance of $\widehat{\theta}_n$ from (2.14) in terms of absolute bias and root MSE for the three parameter values $\theta = 0.6$, 0.75, and 0.9. We observe that for larger values of θ (or smaller values of η, corresponding to more independence in the extremes) larger values of k lead to the best RMSE. This is in line with our theory as, for fixed k, smaller η corresponds to smaller values of m and hence larger asymptotic variance.

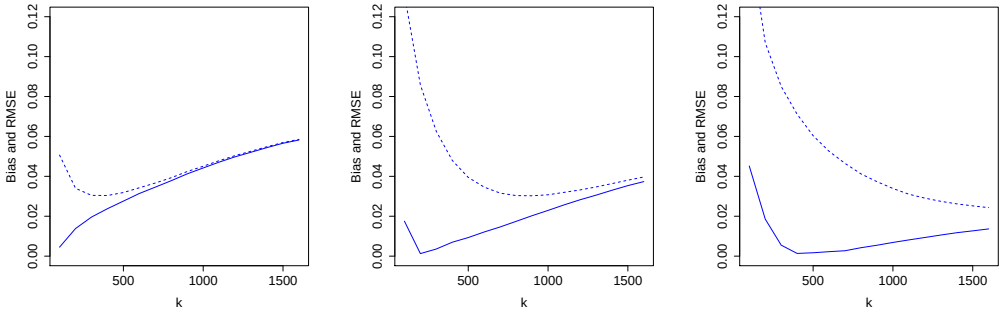

Figure 2.1: Absolute bias (solid lines) and RMSE (dashed lines) of the M-estimator of θ as a function of k, based on 1 000 samples of size 5 000 from model M1 with parameter values 0.6, 0.75 and 0.9, from left to right.

An analysis of $\widehat{\theta}_n$ for a finer range of parameter values is provided in Figure 2.2. Motivated by the findings in Figure 2.1 we fix $k = 800$; this choice leads to reasonable performance across all parameter values. Overall the results are satisfactory, with a more pronounced negative bias for smaller values of θ and more variance for increasing θ.

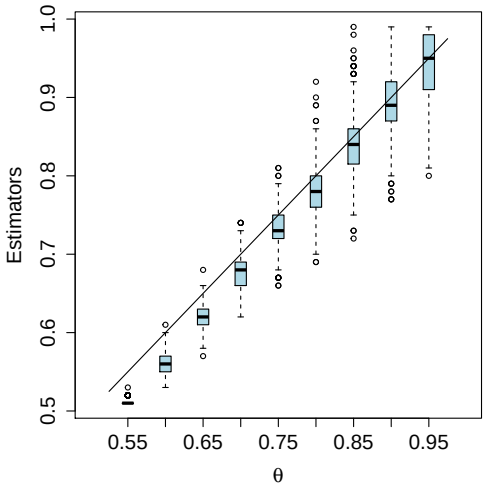

Figure 2.2: Box plots of the M-Estimators of θ based on 1 000 samples of size 5 000 for each parameter value.

The inverted asymmetric logistic model (M2)

Figure 2.3 shows the impact of k on estimated parameter values for three different choices of θ. Since here the parameter is two-dimensional, we consider (and estimate) the Euclidean bias and RMSE of the estimator $\widehat{\theta}_n$, defined as $\|\mathbb{E}[\widehat{\theta}_n - \theta]\|$ and $(\mathbb{E}\|\widehat{\theta}_n - \theta\|^2)^{1/2}$, respectively.

Similarly to the pattern observed in Figure 2.1 we see that smaller values of η necessitate larger values of k in order to achieve a good balance between bias and variance.

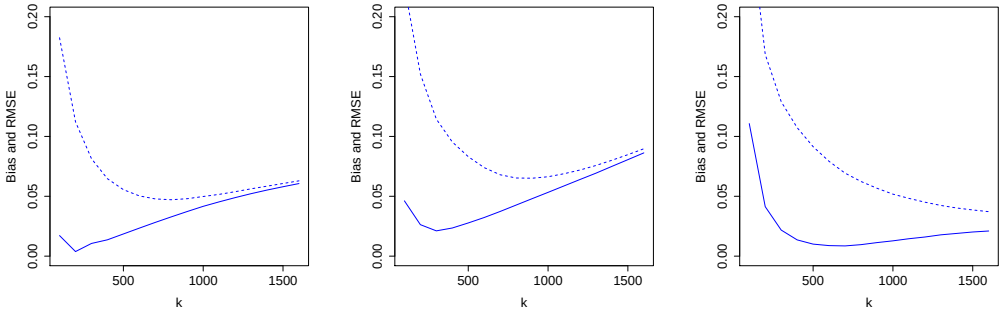

Figure 2.3: Absolute bias (solid lines) and RMSE (dashed lines) of the M-estimator of θ as a function of k, based on 1 000 samples of size 5 000 from model M2 with parameter θ equal to $(0.72, 0.72)$, $(0.75, 0.91)$ and $(0.91, 0.91)$, from left to right. In the original parametrization, the corresponding values of (ν, ϕ) are $(0.94, 0.94)$, $(0.44, 0.94)$ and $(0.31, 0.31)$, respectively.

Figure 2.4 shows the performance of the proposed M-estimator for a range of different parameters (θ_1, θ_2) with Euclidean bias in the left panel and RMSE in the right panel; the value $k = 800$ is fixed throughout. Since the relation $(\nu, \phi) \mapsto (\theta_1, \theta_2)$ is not easily invertible, we selected a grid of values of $(\nu, \phi) \in [0, 1]^2$, calculated all the corresponding points θ and kept the values for which $\theta_j \leq 0.95$, $j = 1, 2$.

We observe that the estimators perform better for parameter values close to the diagonal, with larger bias and variance for more asymmetric parameter values. The overall estimation accuracy is reasonably good, with worst case RMSE values around 0.07.

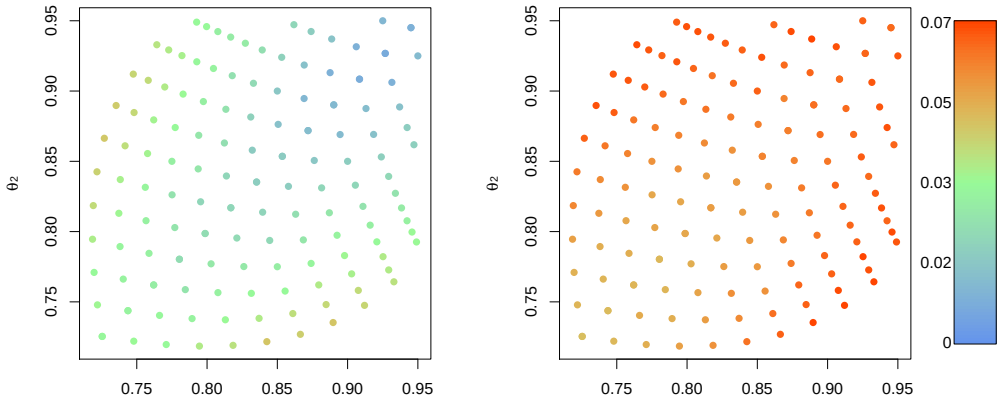

Figure 2.4: Absolute bias (left) and RMSE (right) of the M-estimator of $\theta = (\theta_1, \theta_2)$ as a function of θ, based on 1 000 samples of size 5 000 from model M2.

The Pareto random scale model (M3)

Figure 2.5 shows the effect of k on the performance of our M-estimator $\widehat{\lambda}_n$ in terms of absolute bias and root MSE for the three parameter values $\lambda = 0.4$, 1, and 1.6. We notice that the estimator is considerably more biased at $\lambda = 1$ than at other parameter

values. This is expected as, according to Table 2.1, the bias function q_1 vanishes only at a logarithmic rate when $\lambda = 1$, compared to a polynomial rate elsewhere. Moreover, like in the other models, we observe that for more independent data (characterized by larger λ), larger values of k are required to drive down the variance of the estimator.

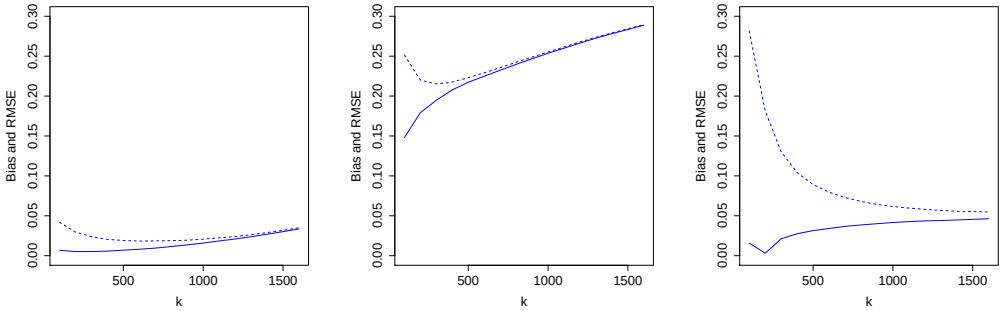

Figure 2.5: Absolute bias (solid lines) and RMSE (dashed lines) of the M-estimator of λ as a function of k, based on 1 000 samples of size 5 000 from model M3 with parameter values 0.4, 1 and 1.6, from left to right.

An analysis of $\widehat{\lambda}_n$ for a finer range of parameter values is provided in Figure 2.6. Motivated by Figure 2.5 we fix $k = 400$, which approximately minimizes the maximal RMSE. Overall the estimator is very precise for small values of λ, but incurs a bias around $\lambda = 0.8$ where it struggles to distinguish between values slightly smaller and slightly larger than 1. This phenomenon is not completely unexpected; a close look at Table 2.1 reveals that c_λ has almost (but not quite) a symmetry around the point $\lambda = 1$, e.g. $c_{0.8}$ is very similar in shape to $c_{1.2}$. This point also corresponds to the transition between asymptotic dependence and independence, which makes estimation challenging.

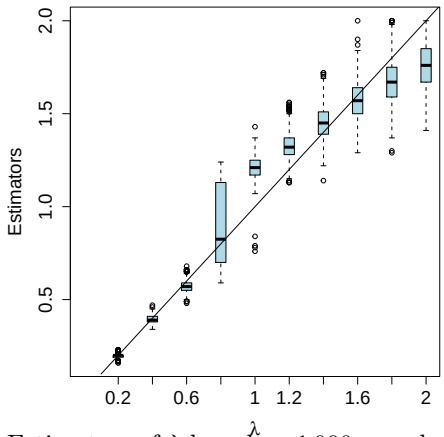

Figure 2.6: Box plots of the M-Estimators of λ based on 1 000 samples of size 5 000 for each parameter value.

2.5.2 Spatial models

In this section we illustrate the performance of the proposed methodology for spatial data. The candidate class for c_θ results from inverted Brown–Resnick processes with fractal variograms (see Example 2.6) and takes the form

$$c_\vartheta^{(s)}(x,y) = (xy)^{\theta^{(s)}}, \quad \theta^{(s)} = \theta(\Delta^{(s)};\vartheta) := \Phi\Big(\frac{1}{2}(\Delta^{(s)}/\beta)^{\alpha/2}\Big), \quad s \in \mathcal{P}, \tag{2.23}$$

where $\vartheta = (\alpha,\beta) \in (0,2] \times \mathbb{R}_+$ and $\Delta^{(s)}$ is the Euclidean distance between the two locations in pair s (measured in units of latitude). Motivated by the data application in the following section, the true parameter values are set as $\vartheta_0 = (1,3)$ and the values for $\Delta^{(s)}$ are obtained from 40 randomly sampled pairs of locations in that data set; see Figure 2.14 for a histogram of the distances in this sample.

To evaluate the performance of our estimators we simulate 1000 independent data sets, each of size 5000, of an inverted Brown–Resnick process with unit Fréchet margins and fractal variogram from (2.15) with $\alpha = 1, \beta = 3$. Following the bivariate simulations, to each of the 40 components of the data we add an independent random variable with Pareto distribution function $1 - 1/x^4$, $x \geq 1$. Using the same weight function g as in the bivariate simulations (see (2.22)), we compute the two estimators introduced in (2.16) and (2.17). Since the performance of both estimators turns out to be very similar, we only report results for the least squares estimator from (2.16) here and defer all simulations for the estimator (2.17) to Section 2.12.

Following the discussion in Section 2.3.1, we fix a value m and select each $k^{(s)}$ such that $\widehat{Q}_n^{(s)}(k^{(s)}/n, k^{(s)}/n) = m$. The first two panels of Figure 2.7 show the absolute bias and RMSE of the estimators $\widehat{\alpha}$ and $\widehat{\beta}$, respectively, as functions of $m \in \{75, 100, \ldots, 500\}$. We observe that the RMSE for both estimators is relatively large across all values of m. Interestingly, this does not result in a bad performance in estimating the function $\theta(\cdot;\vartheta)$. Indeed, the last panel of Figure 2.7 shows averaged (over simulation runs) values for $\sup_{0\leq\Delta\leq 3} |\theta(\Delta;\widehat{\vartheta}) - \theta(\Delta;\vartheta)|$ and indicates a good overall performance; note that the observed values of Δ are all smaller than 3 (see Figure 2.14). This can be explained by the fact that different values of (α,β) can lead to somewhat similar curves in the range of interest. This is further illustrated in the left panel of Figure 2.8 where a random sample of 50 estimated functions $\theta(\Delta;\widehat{\vartheta})$ is displayed.

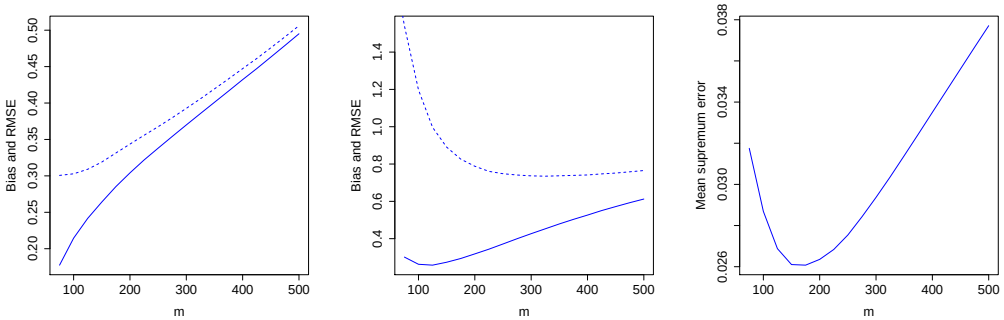

Figure 2.7: Left and middle columns: Bias (solid line) and RMSE (dotted line) of the estimators of the two spatial parameters α (left) and β (middle) as a function of m. Right: Mean of the supremum error $\sup_{0\le\Delta\le 3} |\theta(\Delta;\widehat{\alpha},\widehat{\beta}) - \theta(\Delta;\alpha,\beta)|$ as a function of m.

We conclude this section by fixing $m = 150$ and comparing the performance of estimators for $\theta^{(s)}$ based on a bivariate sample at a given distance and the spatial estimator discussed above. Boxplots corresponding to five pairs of stations with distances $\Delta^{(s)} \approx 0.5, 1, \ldots, 2.5$ are shown in the left panel of Figure 2.8. As expected from the theory, using the spatial estimator is advantageous as it allows to combine information from different distances and leads to a reduced variance.

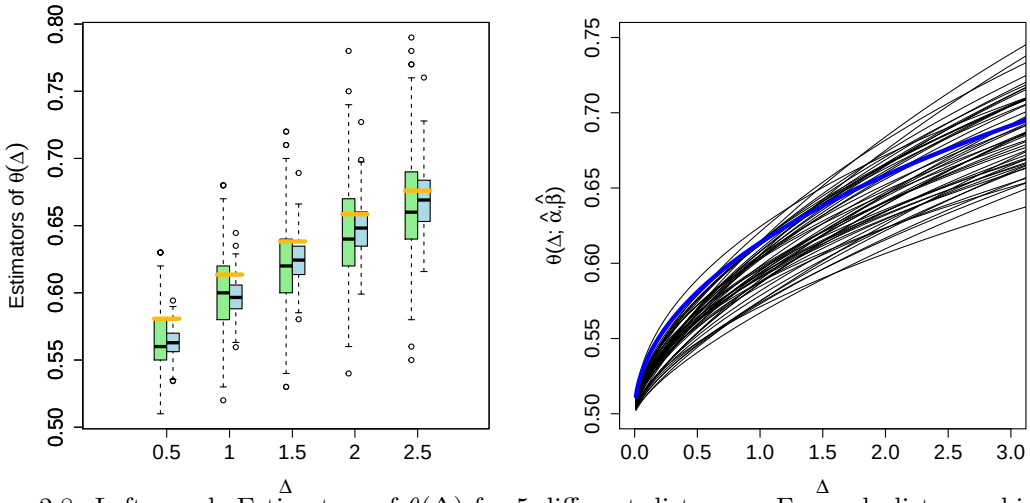

Figure 2.8: Left panel: Estimators of $\theta(\Delta)$ for 5 different distances. For each distance, bivariate M-estimator $\widehat{\theta}_n^{(s)}$ (green) and spatial estimator $\theta(\Delta^{(s)};\widehat{\alpha},\widehat{\beta})$ (blue) based on the $d = 40$ locations. Right panel: 50 sampled curves $\theta(\cdot;\widehat{\alpha},\widehat{\beta})$. Blue represents the true curve $\theta(\cdot;\alpha,\beta)$.

2.6 Application to rainfall data

In a data set introduced in Le et al. (2018), rainfall was measured daily from 1960 to 2009 at a set of 92 different locations in the state of Victoria, southeastern Australia, for a total of $n = 18\,263$ measurements. The conclusions in that paper are that an asymptotically independent model is suitable. A subset of 40 locations, for a total of

780 pairs, was randomly sampled; see the right panel of Figure 2.9. To the data at those selected locations we fit the same tail model as in Section 2.5.2, given in (2.23). The weight function g that we use is the same as before and as in Section 2.5.2, we make use of Section 2.3.1 by fixing a value m and choosing each $k^{(s)}$ accordingly.

We set $m = 400$. The left panel of Figure 2.9 shows the 780 pairwise estimators $\widehat{\theta}_n^{(s)}$ plotted against the distances $\Delta^{(s)}$. Despite some estimates at the boundary of the parameter space, the results do not provide much evidence for asymptotic dependence, whereas all estimates are away from the boundary for distances of at least 0.3 units of latitude, strongly suggesting asymptotic independence at these distances. Our two estimators (2.16) and (2.17) of (α, β) yield estimates $(\widehat{\alpha}, \widehat{\beta})$ of (1.55, 2.24) and (1.56, 2.24), respectively. They are extremely similar, as hinted by the simulation study from Section 2.5.2. The curve $\theta(\cdot; \widehat{\alpha}, \widehat{\beta})$ corresponding to the least squares estimator is also shown in the left panel of Figure 2.9. The middle panel of Figure 2.9 displays similar curves for the least squares estimator when m varies from 200 to 1 000. It shows that the estimated curve is robust with respect to the choice of m.

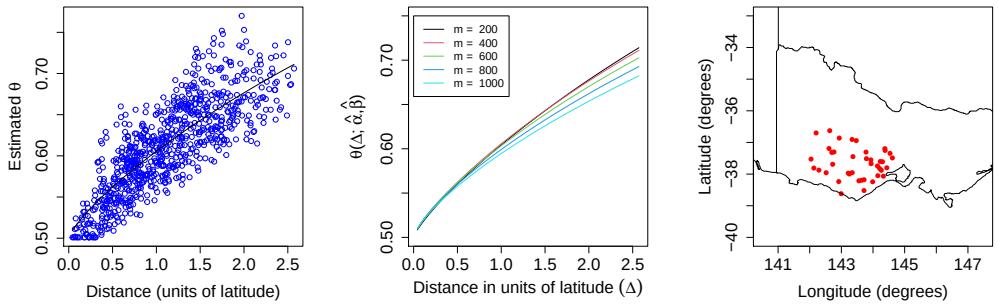

Figure 2.9: Left: Estimated parameters $\widehat{\theta}_n^{(s)}$ against the distances $\Delta^{(s)}$. The black line represents the estimated curve $\theta(\cdot; 1.55, 2.24)$. Middle: Estimated curve $\theta(\cdot; \widehat{\alpha}, \widehat{\beta})$ for the least squares estimator with different values of m. Right: The 40 sampled locations in the state of Victoria, southeastern Australia.

2.7 Proofs of main results

In this section are collected the proofs of Theorems 2.1 to 2.5. A number of more technical results, which are instrumental in the following, are collected in Section 2.8.

2.7.1 Bivariate estimation

For the proofs concerning the bivariate estimators, we assume the framework of Sections 2.3.1 and 2.3.2, we define the transformed random variables $U = 1 - F_1(X)$, $V = 1 - F_2(Y)$ and note that Q is the distribution function of the random vector (U, V). Define the transformed observations $U_i = 1 - F_1(X_i)$, $V_i = 1 - F_2(Y_i)$ and

denote by $U_{n,1},\dots,U_{n,n}$ and $V_{n,1},\dots,V_{n,n}$ the ordered versions thereof. Additionally define $U_{n,0}=V_{n,0}=0$. For an intermediate sequence k, define the random functions u_n and v_n by

$$u_n(x)=\frac{n}{k}U_{n,\lfloor kx\rfloor}\quad\text{and}\quad v_n(y)=\frac{n}{k}V_{n,\lfloor ky\rfloor},$$

for $(x,y)\in[0,T]^2$. Recalling that $m=nq(k/n)$, it allows us to write

$$\widehat{c}_n(x,y)=\frac{n}{m}Q_n\left(\frac{k}{n}u_n(x),\frac{k}{n}v_n(y)\right)$$

where

$$Q_n(x,y):=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{U_i\le x,V_i\le y\right\}$$

denotes the empirical distribution function of $(U_1,V_1),\dots,(U_n,V_n)$. We begin by discussing technical results that will be used in the proof of both Theorem 2.1 and Theorem 2.2. Consider the decomposition

$$\begin{aligned}W_n(x,y)=&\sqrt{m}\Big(\frac{n}{m}Q_n\Big(\frac{k}{n}u_n(x),\frac{k}{n}v_n(y)\Big)-\frac{n}{m}Q\Big(\frac{k}{n}u_n(x),\frac{k}{n}v_n(y)\Big)\Big)\\&+\sqrt{m}\Big(\frac{n}{m}Q\Big(\frac{k}{n}u_n(x),\frac{k}{n}v_n(y)\Big)-c(u_n(x),v_n(y))\Big)\\&+\sqrt{m}\big(c(u_n(x),v_n(y))-c(x,y)\big).\end{aligned}$$

For the second term in the above decomposition, note that

$$\sqrt{m}\Big(\frac{n}{m}Q\Big(\frac{k}{n}x,\frac{k}{n}y\Big)-c(x,y)\Big)=O\left(\sqrt{m}q_1\left(\frac{k}{n}\right)\right)=o(1)$$

uniformly over all $(x,y)\in[0,2T]^2$; here the last equation follows from Condition 2.1(ii). By Corollary 2.1 we have $\mathbb{P}(u_n(T)\vee v_n(T)\le 2T)\to 1$, and thus

$$\sup_{x,y\in[0,T]}\sqrt{m}\Big|\frac{n}{m}Q\Big(\frac{k}{n}u_n(x),\frac{k}{n}v_n(y)\Big)-c(u_n(x),v_n(y))\Big|=o_{\mathbb{P}}(1).$$

Next define for all $x,y\in[0,2T]$

$$H_n(x,y):=\sqrt{m}\Big(\frac{n}{m}Q_n\Big(\frac{k}{n}x,\frac{k}{n}y\Big)-\frac{n}{m}Q\Big(\frac{k}{n}x,\frac{k}{n}y\Big)\Big).\tag{2.24}$$

By Lemma 2.4 this process converges, in $\ell^\infty([0,2T]^2)$, to the process W from Theorem 2.1 and by Corollary 2.1 u_n and v_n converge uniformly in probability to the identity function $I:[0,2T]\to[0,2T]$. Therefore, the triple (H_n,u_n,v_n) converges

jointly in distribution to (W, I, I). This implies

$$\sup_{x,y\in[0,T]} \left| H_n(u_n(x), v_n(y)) - H_n(x, y) \right| = o_{\mathbb{P}}(1). \tag{2.25}$$

Indeed, consider the map

$$f : \begin{cases} \ell^\infty([0, 2T]^2) \times \mathcal{V}[0, T] \times \mathcal{V}[0, T] \to \mathbb{R} \\ (a, b_1, b_2) \mapsto \sup_{x,y\in[0,T]} |a(b_1(x), b_2(y)) - a(x, y)| \end{cases}$$

where $\mathcal{V}[0, T] := \{g \in \ell^\infty([0, T]) : g([0, T]) \subset [0, 2T]\}$ and assume that the product space is equipped with the norm $\|a\|_\infty + \|b_1\|_\infty + \|b_2\|_\infty$. Observe that f is continuous at points (a, b_1, b_2) where a is a continuous function and that the sample paths of W are almost surely continuous. Thus, by the continuous mapping theorem, with probability converging to 1,

$$\sup_{x,y\in[0,T]} \left| H_n(u_n(x), v_n(y)) - H_n(x, y) \right| = f(H_n, u_n, v_n) \rightsquigarrow f(W, I, I) = 0.$$

Since the limit is constant a.s. (2.25) follows. Combining the equations above, we find

$$W_n(x, y) = H_n(x, y) + \sqrt{m}\big(c(u_n(x), v_n(y)) - c(x, y)\big) + o_{\mathbb{P}}(1), \tag{2.26}$$

where the term $o_{\mathbb{P}}(1)$ is uniform on $[0, T]^2$, and we recall that $H_n \rightsquigarrow W$ in $\ell^\infty([0, 2T]^2)$.

Proof of Theorem 2.1

Define

$$S_n(x, y) := \sqrt{m}\big(c(u_n(x), v_n(y)) + c(x, y)\big).$$

In light of (2.26) it suffices to prove that $S_n \xrightarrow{P} 0$ uniformly on $[0, T]^2$. From here on it is more convenient to study component-wise increments. That is, we write

$$\begin{aligned} S_n(x, y) &= \sqrt{m}(c(u_n(x), y) - c(x, y)) + \sqrt{m}(c(u_n(x), v_n(y)) - c(u_n(x), y)) \\ &=: S_n^{(a)}(x, y) + S_n^{(b)}(x, y) \end{aligned}$$

and we will show that both $S_n^{(a)}$ and $S_n^{(b)}$ converge to 0 in probability, starting with $S_n^{(a)}$.

By assumption, since with probability converging to 1 we have $u_n(x) \in [0, 2T]$ for every $x \le T$, we can write

$$S_n^{(a)}(x, y) = \sqrt{m}\,(c(u_n(x), y) + c(x, y)) \tag{2.27}$$

$$= \sqrt{m}\left\{\frac{n}{m}Q\left(\frac{k}{n}u_n(x), \frac{k}{n}y\right) - \frac{n}{m}Q\left(\frac{k}{n}x, \frac{k}{n}y\right) + O_{\mathbb{P}}\left(q_1\left(\frac{k}{n}\right)\right)\right\}$$
$$= \frac{n}{\sqrt{m}}\left(Q\left(\frac{k}{n}u_n(x), \frac{k}{n}y\right) - Q\left(\frac{k}{n}x, \frac{k}{n}y\right)\right) + o_{\mathbb{P}}(1) \tag{2.28}$$

uniformly on $[0,T]^2$, since the sequence m was chosen so that $\sqrt{m}q_1(k/n) \to 0$. We will use both (2.27) and (2.28) as representations of $S_n^{(a)}$ throughout the proof.

Let $\beta_n = (m/k)/(\log(k/m))$. From there, partition $[0,T]^2$ in $\Theta_n^{(1)} = [0, 1/k) \times [0,T]$, $\Theta_n^{(2)} = [1/k, \beta_n) \times [0,T]$ and $\Theta_n^{(3)} = [\beta_n, T] \times [0,T]$ (if $\beta_n < 1/k$, $\Theta_n^{(2)}$ is empty). These sets represent the "small", "intermediate" and "large" values of x, respectively. We will prove that the suprema of $S_n^{(a)}$ on $\Theta_n^{(1)}$, $\Theta_n^{(2)}$ and $\Theta_n^{(3)}$ all converge to 0 in probability. (2.28) yields

$$\begin{aligned}
\sup_{(x,y)\in\Theta_n^{(1)}} |S_n^{(a)}(x,y)| &= \frac{n}{\sqrt{m}} \sup_{0\le x<1/k} \left|Q\left(\frac{k}{n}u_n(x), \frac{k}{n}y\right) - Q\left(\frac{k}{n}x, \frac{k}{n}y\right)\right| + o_{\mathbb{P}}(1) \\
&= \frac{n}{\sqrt{m}} \sup_{0\le x<1/k} Q\left(\frac{k}{n}x, \frac{k}{n}y\right) + o_{\mathbb{P}}(1) \\
&\le \frac{n}{\sqrt{m}}\frac{1}{n} + o_{\mathbb{P}}(1) \\
&= \frac{1}{\sqrt{m}} + o_{\mathbb{P}}(1),
\end{aligned}$$

where we have once again used the facts that $u_n(x) = 0$ whenever $x < 1/k$ and that $Q(0,\cdot) = Q(\cdot,0) = 0$, in addition to the fact that $Q(u,v) \le u$. This proves that $\sup_{\Theta_n^{(1)}} |S_n^{(a)}| \to 0$ in probability.

Using (2.28) again, the supremum of $S_n^{(a)}$ on $\Theta_n^{(2)}$ can be expressed as

$$\begin{aligned}
\sup_{1/k\le x<\beta_n} \left|S_n^{(a)}(x,y)\right| &= \sup_{1/k\le x<\beta_n} \frac{n}{\sqrt{m}} \left|Q\left(\frac{k}{n}u_n(x), \frac{k}{n}y\right) - Q\left(\frac{k}{n}x, \frac{k}{n}y\right)\right| + o_{\mathbb{P}}(1) \\
&\le \sup_{1/k\le x<\beta_n} \frac{n}{\sqrt{m}} \left|\frac{k}{n}u_n(x) - \frac{k}{n}x\right| + o_{\mathbb{P}}(1) \\
&= \sup_{1/k\le x<\beta_n} \frac{k}{\sqrt{m}}|u_n(x) - x| + o_{\mathbb{P}}(1) \\
&= O_{\mathbb{P}}\left(\sup_{1/k\le x<\beta_n} \sqrt{\frac{k}{m}}\varphi(x)\right) + o_{\mathbb{P}}(1),
\end{aligned}$$

where we have used Lipschitz continuity of Q and Lemma 2.3. The last bound holds for any function φ that satisfies the conditions in Lemma 2.3, but from now on we use $\varphi(x) := \sqrt{x\log\log(1/x)}$ on $(0,B]$ and $\varphi(x) := \sqrt{x}$ on $(B,T]$, where $B > 0$ is chosen small enough so that φ is well defined and non-decreasing. By monotonicity,

the supremum is attained at $x = \beta_n$. We then have

$$\sup_{1/k \le x < \beta_n} \left| S_n^{(a)}(x,y) \right| = O_{\mathbb{P}}\left(\sqrt{\frac{k}{m}\beta_n \log\log(1/\beta_n)} \right) + o_{\mathbb{P}}(1)$$

because since $\beta_n \to 0$, eventually $\beta_n \le B$, so eventually $\varphi(\beta_n) = \sqrt{\beta_n \log\log(1/\beta_n)}$. The last display converges in probability to 0 since

$$\frac{k}{m}\beta_n \log\log(1/\beta_n) = \frac{\log\log\left(\frac{k}{m}\log(k/m)\right)}{\log(k/m)} \longrightarrow 0$$

as $k/m \to \infty$, which proves that $\sup_{\Theta_n^{(2)}} |S_n^{(a)}| \to 0$ in probability.

Finally, when considering large values of x, Lemma 2.3 and a combination of Lemmas 2.7 and 2.8 imply that

$$\begin{aligned}
\sup_{\beta_n \le x \le T} \left| S_n^{(a)}(x,y) \right| &= \sup_{\beta_n \le x \le T} \sqrt{m} |c(u_n(x),y) - c(x,y)| \\
&\lesssim \sqrt{m} \sup_{\beta_n \le x \le T} |u_n(x) - x| r(x \vee u_n(x)) \\
&= O_{\mathbb{P}}\left(\sqrt{\frac{m}{k}} \sup_{\beta_n \le x \le T} \varphi(x) r(x \vee u_n(x)) \right),
\end{aligned}$$

where $r(x) = (x \log(1/x))^{-1}$. By monotonicity of φ, the inside of the $O_{\mathbb{P}}$ term can be upper bounded by

$$\sqrt{\frac{m}{k}} \sup_{\beta_n \le x \le T} \varphi(x \vee u_n(x)) r(x \vee u_n(x))$$

and since with probability converging to 1, for every $x \le T$, $u_n(x) \le 2T$, this can in turn be upper bounded (with probability converging to 1) by

$$\sqrt{\frac{m}{k}} \sup_{\beta_n \le x \le 2T} \varphi(x) r(x).$$

It can easily be checked (e.g. by differentiation) that the function $\varphi \times r$ is decreasing. Thus, the above supremum is attained at β_n. Finally, elementary computations yield

$$\sqrt{\frac{m}{k}} \varphi(\beta_n) r(\beta_n) \lesssim \sqrt{\frac{\log\log((k/m)^2)}{\log(k/m)}} \longrightarrow 0.$$

Overall, we have shown that $S_n^{(a)} \xrightarrow{P} 0$ uniformly over $[0,T]^2$. Note that all the bounds we derived are uniform over all values of $y \in [0,T]$, although it was removed from the notation for parsimony. In order to deal with $S_n^{(b)}$, we recall once again that with probability converging to 1, we have $u_n(x) \le 2T$ for every $x \le T$. Therefore,

with probability converging to 1,

$$\begin{aligned}\sup_{(x,y)\in[0,T]^2} \left|S_n^{(b)}(x,y)\right| &= \sup_{(x,y)\in[0,T]^2} \sqrt{m}|c(u_n(x),v_n(y)) - c(u_n(x),y)| \\ &\leq \sup_{x\in[0,2T],y\in[0,T]} \sqrt{m}|c(x,v_n(y)) - c(x,y)|.\end{aligned}$$

This can be shown to converge in probability to 0 using the exact same proof as for $S_n^{(a)}$. We finally conclude that $S_n \xrightarrow{P} 0$ in $\ell^\infty([0,T]^2)$, and the proof for deterministic $k = k_n$ is complete. It remains to show that the result continues to hold if we replace the deterministic sequence $k = k_n$ by data-dependent $\widehat{k}$ as outlined in Section 2.3.1. This is established in Section 2.7.1. □

Proof of Theorem 2.2

In view of (2.26), we require the joint asymptotic behavior of H_n, u_n and v_n. Define, for $(x,y) \in [0,\infty)^2$,

$$L_n^{(1)}(x) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\left\{U_i \leq \frac{k}{n}x\right\} \quad \text{and} \quad L_n^{(2)}(y) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\left\{V_i \leq \frac{k}{n}y\right\},$$

a rescaled version of the marginal empirical distribution functions of U and V. We now show that the $\mathbb{D}$-valued process

$$(x,y) \mapsto \left(H_n(x,y), \sqrt{m}\left(L_n^{(1)}(x) - x\right), \sqrt{m}\left(L_n^{(2)}(y) - y\right)\right) \tag{2.29}$$

converges in distribution to the Gaussian process $(W, W^{(1)}, W^{(2)})$ defined in Section 2.4.1 with covariance matrix Λ from (2.19), where $\mathbb{D} := \left(\ell^\infty([0,2T]^2)\right)^3$.

Again, let I denote the identity map on $\mathbb{R}$. The three processes H_n, $\sqrt{m}(L_n^{(1)} - I)$ and $\sqrt{m}(L_n^{(2)} - I)$ are individually tight (see Lemma 2.4) and hence it suffices to prove convergence of the marginal distributions. This in turn follows from convergence of the covariance function, by the multivariate Lindeberg-Feller theorem (see van der Vaart (2000), Theorem 2.27); verification of the Lindeberg condition is similar to condition (B) in the proof of Lemma 2.4. The convergence of $\mathbb{E}\left[H_n(x,y)H_n(x',y')\right]$ to $c(x \wedge x', y \wedge y')$ is already shown in Lemma 2.4. Using similar arguments and recalling that $m/k \to \chi > 0$, one easily deals with the other covariance terms and concludes that the processes in (2.29) weakly converge to $(W, W^{(1)}, W^{(2)})$ in $\mathbb{D}$.

Note that the random functions u_n and v_n are the generalized inverses of $L_n^{(1)} + 1/k$ and $L_n^{(2)} + 1/k$, respectively. Because $\sqrt{m}/k \to 0$, the term $1/k$ is negligible. Upon applying Vervaat's lemma (Vervaat (1972)), which states that the generalized inverse mapping is Hadamard differentiable around the identity function, we deduce that the

processes G_n, defined by

$$G_n(x,y) = (H_n(x,y), \sqrt{m}(u_n(x) - x), \sqrt{m}(v_n(y) - y)),$$

weakly converge to $(W, -W^{(1)}, -W^{(2)})$ in $\mathbb{D}$. For $t > 0$, define the sets

$$\mathcal{V}(t) := \{b \in \ell^\infty([0,2T]) : \forall x \in [0,T], x + tb(x) \in [0,2T]\}. \tag{2.30}$$

Let $\mathbb{D}_n \subset \mathbb{D}$ be the subset of functions $a = (a^{(0)}, a^{(1)}, a^{(2)})$ such that $a^{(1)}(x,y)$ is constant in y, $a^{(2)}(x,y)$ is constant in x and the functions $x \mapsto a^{(1)}(x,y)$ and $y \mapsto a^{(2)}(x,y)$ are elements of $\mathcal{V}(1/\sqrt{m})$. Let $\mathbb{E}$ be the space of equivalence classes $L^\infty([0,T]^2)$ equipped with the topology of hypi-convergence. Define the functionals $f_n : \mathbb{D}_n \to \mathbb{E}$ by

$$f_n(a)(x,y) := a^{(0)}(x,y) + \sqrt{m}\left(c\left(x + \frac{a^{(1)}(x,y)}{\sqrt{m}}, y + \frac{a^{(2)}(x,y)}{\sqrt{m}}\right) - c(x,y)\right).$$

(2.26) can be rephrased as $W_n = f_n(G_n) + o_{\mathbb{P}}(1)$, assuming that $G_n \in \mathbb{D}_n$, which is true with probability

$$\mathbb{P}(u_n(T) \leq 2T, v_n(T) \leq 2T) \longrightarrow 1.$$

Let $\mathbb{D}_0 \subset \mathbb{D}$ be the subset of continuous functions a such that $a(0) = 0$. As soon as $a_n \in \mathbb{D}_n$ converges uniformly to $a \in \mathbb{D}_0$, by Lemma 2.9, $f_n(a_n)$ hypi-converges to $f(a)$, where $f : \mathbb{D}_0 \to \mathbb{E}$ satisfies

$$f(a) := a^{(0)} + \dot{c}_1 a^{(1)} + \dot{c}_2 a^{(2)}.$$

Note that $(W, -W^{(1)}, -W^{(2)})$ concentrates on $\mathbb{D}_0$. Therefore, by the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1),

$$W_n = f_n(G_n) + o_{\mathbb{P}}(1) \rightsquigarrow f((W, -W^{(1)}, -W^{(2)})) = W - \dot{c}_1 W^{(1)} - \dot{c}_2 W^{(2)}$$

in $\mathbb{E}$. It remains to show that the result continues to hold if we replace the deterministic sequence $k = k_n$ by data-dependent $\widehat{k}$ as outlined in Section 2.3.1. This is established in Section 2.7.1. □

Proof that Theorems 2.1 and 2.2 continue to hold with $\widehat{k}$

Let $\widehat{c}_{n,\widehat{k}}$ be the estimator $\widehat{c}_n$ computed with the random quantity $\widehat{k}$ instead of k. We shall prove that $\sqrt{m}|\widehat{c}_{n,\widehat{k}} - \widehat{c}_n| \to 0$ in probability uniformly over $[0,T]^2$ (under asymptotic independence) or in the hypi semimetric (under asymptotic dependence).

Note that the definition of $\widehat{k}$ implies that $\widehat{c}_n(\widehat{k}/k, \widehat{k}/k) = 1$. By assumption, $\widehat{c}_n$

converges to c in probability uniformly in a neighborhood of $(1,1)$. Jointly with the fact that $c(x,x) = x^{1/\eta}$, this readily implies that $\widehat{k}/k \to 1$ in probability. Further note that

$$\widehat{c}_{n,\widehat{k}}(x,y) = \frac{q(k/n)}{q(\widehat{k}/n)}\widehat{c}_n(\widehat{k}x/k, \widehat{k}y/k).$$

We first discuss the case of asymptotic independence. By Theorem 2.1 and by Skorokhod's almost sure representation, we may assume that almost surely, $\widehat{c}_n = c + W/\sqrt{m} + o(1/\sqrt{m})$ and $\widehat{k}/k \to 1$. The object of interest is then equal, with probability one, to

$$\begin{aligned}
&\frac{q(k/n)}{q(\widehat{k}/n)}\sqrt{m}\left(\widehat{c}_n(\widehat{k}x/k, \widehat{k}y/k) - \frac{q(\widehat{k}/n)}{q(k/n)}\widehat{c}_n(x,y)\right) \\
&= \frac{q(k/n)}{q(\widehat{k}/n)}\Big\{\sqrt{m}\Big(c(\widehat{k}x/k, \widehat{k}y/k) - \frac{q(\widehat{k}/n)}{q(k/n)}c(x,y)\Big) + W(\widehat{k}x/k, \widehat{k}y/k) - W(x,y)\Big\} + o(1) \\
&= -\sqrt{m}c(x,y)\left(\frac{q(\widehat{k}/n)}{q(k/n)} - \left(\frac{\widehat{k}/n}{k/n}\right)^{1/\eta}\right)\frac{q(k/n)}{q(\widehat{k}/n)} + o(1), \qquad (2.31)
\end{aligned}$$

where we have used homogeneity of c, regular variation of q and the fact that almost surely, the sample paths of W are continuous, hence uniformly continuous on compact sets. The terms $o(1)$ are uniform over $[0,T]^2$. Finally, it is shown in Lemma 2.2 that uniformly over a in a neighborhood of 1, $q(at)/q(t) - a^{1/\eta} = O(q_1(t))$. Recalling that $\widehat{k}/k \to 1$ almost surely, the first term in (2.31) is then uniformly of the order of $\sqrt{m}q_1(k/n)$, which vanishes by Condition 2.1(ii).

In the case of asymptotic dependence, Theorem 2.2 ensures that $\widehat{c}_n = c + B/\sqrt{m} + o(1/\sqrt{m})$ in the hypi semimetric. We may apply the reasoning above except that, from the definition of the process B, we get the additional term

$$\begin{aligned}
&-\sum_{j=1}^{2}\Big(\dot{c}_j(\widehat{k}x/k, \widehat{k}y/k)W^{(j)}(\widehat{k}x/k, \widehat{k}y/k) - \dot{c}_j(x,y)W^{(j)}(x,y)\Big) \\
&\qquad\qquad = -\sum_{j=1}^{2}\dot{c}_j(x,y)\Big(W^{(j)}(\widehat{k}x/k, \widehat{k}y/k) - W^{(j)}(x,y)\Big); \qquad (2.32)
\end{aligned}$$

this follows from the fact that under asymptotic dependence, c is homogeneous of order 1 and the directional partial derivatives of such a function, when they exist, are constant along rays from the origin. The above term vanishes uniformly since $\dot{c}_j$ has to be locally bounded (only under asymptotic dependence) and since the sample paths of $W^{(j)}$ are almost surely continuous. We therefore obtain (2.31), except that this time the term $o(1)$ is understood in the hypi semimetric. From here on the proof

is completed in the same way as under asymptotic independence. □

Proof of Theorem 2.3

Recall the definition of Ψ_n from Section 2.3.2. Letting $\widehat{\sigma}_n = \frac{n}{m}\widehat{\zeta}_n$, the assumption that $(\widehat{\theta}_n, \widehat{\zeta}_n)$ minimizes the norm of Ψ_n^* becomes equivalent to $(\widehat{\theta}_n, \widehat{\sigma}_n)$ minimizing the norm of Ψ_n. The key is to note that for any θ, σ,

$$\Psi(\theta,\sigma) - \Psi_n(\theta,\sigma) = \int g(\widehat{c}_n - c)d\mu_{\mathrm{L}} = \frac{1}{\sqrt{m}}\int g W_n d\mu_{\mathrm{L}}, \tag{2.33}$$

with W_n defined as in Theorems 2.1 and 2.2. By the dominated convergence theorem, and because g is integrable, one easily sees that the functional $f \mapsto \int g f d\mu_{\mathrm{L}}$ is continuous in $\ell^\infty([0,T]^2)$. By Lemma 2.10, this is also true in the topology of hypi-convergence on $\ell^\infty([0,T]^2)$ at points f that are continuous Lebesgue-almost everywhere on $[0,T]^2$. It is the case of both limiting Gaussian processes appearing in Theorems 2.1 and 2.2: W, $W^{(1)}$ and $W^{(2)}$ have almost surely continuous sample paths and under asymptotic dependence, the directional derivatives $\dot{c}_j$ are almost everywhere continuous. Those two results and the continuous mapping theorem then imply that

$$\int g W_n d\mu_{\mathrm{L}} \rightsquigarrow N(0, A).$$

We may therefore apply Lemma 2.11 with $\phi = \Psi$, $x_0 = (\theta_0, 1)$, $Y_n = \frac{1}{\sqrt{m}}\int g W_n d\mu_{\mathrm{L}}$ and $a_n = 1/\sqrt{m}$, and as required we obtain

$$\sqrt{m}((\widehat{\theta}_n, \widehat{\sigma}_n) - (\theta_0, 1)) = (J^\top J)^{-1} J^\top \int g W_n d\mu_{\mathrm{L}} + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \Sigma).$$

□

2.7.2 Spatial estimation

For the proofs in the spatial setting, we assume the framework of Section 2.3.3, we define the transformed random variables $U^{(j)} = 1 - F^{(j)}(X^{(j)})$ and for a pair s, let $Q^{(s)}$ be the distribution function of the random vector $(U^{(s_1)}, U^{(s_2)})$. Define the transformed observations $U_i^{(j)} = 1 - F^{(j)}(X_i^{(j)})$ and denote by $U_{n,1}^{(j)}, \ldots, U_{n,n}^{(j)}$ the ordered versions thereof and define $U_{n,0}^{(j)} := 0$. For intermediate sequences $k^{(s)}$, we define the (weighted) empirical tail quantile functions $u_n^{(s,j)}$, $s \in \mathcal{P}, j \in \{1,2\}$, by

$$u_n^{(s,j)}(x) = \frac{n}{k^{(s)}} U_{n,\lfloor k^{(s)} x \rfloor}^{(s_j)}, \quad x \geq 0.$$

Recalling that $m^{(s)} = nq^{(s)}(k^{(s)}/n)$, it allows us to write

$$\widehat{c}_n^{(s)}(x,y) = \frac{n}{m^{(s)}} Q_n^{(s)} \left(\frac{k^{(s)}}{n} u_n^{(s,1)}(x), \frac{k^{(s)}}{n} u_n^{(s,2)}(y) \right).$$

where $Q_n^{(s)}$ denotes the empirical distribution function of $(U_1^{(s_1)}, U_1^{(s_2)}), \ldots, (U_n^{(s_1)}, U_n^{(s_2)})$. Following the discussion before the proof of Theorem 2.1, we may define

$$H_n^{(s)}(x,y) := \sqrt{m^{(s)}} \Big\{ \frac{1}{m^{(s)}} \sum_{i=1}^{n} \mathbb{I}\Big\{ U_i^{(s_1)} \le \frac{k^{(s)}}{n} x, U_i^{(s_2)} \le \frac{k^{(s)}}{n} y \Big\} - \frac{n}{m^{(s)}} \mathbb{P}\Big(U^{(s_1)} \le \frac{k^{(s)}}{n} x, U^{(s_2)} \le \frac{k^{(s)}}{n} y \Big) \Big\}.$$

and similarly obtain

$$W_n^{(s)}(x,y) = H_n^{(s)}(x,y) + \sqrt{m^{(s)}} \left(c^{(s)} \left(u_n^{(s,1)}(x), u_n^{(s,2)}(y) \right) - c^{(s)}(x,y) \right) + o_{\mathbb{P}}(1), \quad (2.34)$$

where $W_n^{(s)}$ is defined as in Theorem 2.4 and the term $o_{\mathbb{P}}(1)$ is uniform over compact sets.

Proof of Theorem 2.4

For asymptotically independent pairs, the second term of (2.34) vanishes uniformly, by the proof of Theorem 2.1. Define the $\mathbb{D}$-valued processes G_n by

$$G_n(x,y) := \left(\left(H_n^{(s)}(x,y) \right)_{s \in \mathcal{P}}, \left(\sqrt{m^{(s)}} \left(u_n^{(s,1)}(x) - x \right), \sqrt{m^{(s)}} \left(u_n^{(s,2)}(y) - y \right) \right)_{s \in \mathcal{P}_D} \right),$$

where $\mathbb{D} = (\ell^\infty([0,2T]^2))^{|\mathcal{P}|+2|\mathcal{P}_D|}$. The proof now proceeds similarly to that of Theorem 2.2; we show that G_n converges in distribution, that the processes of interest $W_n^{(s)}$ can be approximately represented as a transformation of G_n, and we conclude by applying a continuous mapping theorem.

For $s \in \mathcal{P}$, $j \in \{1,2\}$, let

$$L_n^{(s,j)}(x) = \frac{1}{k^{(s)}} \sum_{i=1}^{n} \mathbb{1}\left\{ U^{(s_j)} \le \frac{k^{(s)}}{n} x \right\}, \quad x \ge 0.$$

Recall that I denotes the identity mapping on $\mathbb{R}$. By standard arguments (see, e.g., the proofs of Theorems 2.1 and 2.2), we see that each of the processes $H_n^{(s)}$ and $\sqrt{m^{(s)}} \left(L_n^{(s,j)} - I \right)$ converge in distribution in $\ell^\infty([0,2T]^2)$, hence they are tight

random elements in that space. It follows that the sequence of processes

$$(x,y) \mapsto \left(\left(H_n^{(s)}(x,y)\right)_{s\in\mathcal{P}}, \left(\sqrt{m^{(s)}}\left(L_n^{(s,1)}(x)-x\right), \sqrt{m^{(s)}}\left(L_n^{(s,2)}(y)-y\right)\right)_{s\in\mathcal{P}_D}\right) \tag{2.35}$$

is tight in the product space $\mathbb{D}$. A Lindeberg-type condition (van der Vaart, 2000, Theorem 2.27) can easily be checked, so weak convergence of the process in (2.35) follows from convergence of $\mathbb{E}\left[G_n(x,y)G_n(x',y')^\top\right]$ to a suitable covariance matrix. This is simply a consequence of Condition 2.2; indeed, for suitable pairs $s, s' \in \mathcal{P}$, $j, j' \in \{1,2\}$ and $(x,y),(x',y') \in [0,\infty)^2$, this condition implies that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[H_n^{(s)}(x,y)H_n^{(s')}(x',y')\right] &= \Gamma^{(s,s')}((x,y),(x',y')),\\
\lim_{n\to\infty} \mathbb{E}\left[H_n^{(s)}(x,y)\sqrt{m^{(s')}}\left(L_n^{(s',j)}(x')-x'\right)\right] &= \Gamma^{(s,s',j)}((x,y),(x',y')),\\
\lim_{n\to\infty} \mathbb{E}\left[\sqrt{m^{(s)}}\left(L_n^{(s,j)}(x)-x\right)\sqrt{m^{(s')}}\left(L_n^{(s',j')}(x')-x'\right)\right] &= \Gamma^{(s,j,s',j')}((x,y),(x',y')).
\end{aligned}$$

We deduce that in $\mathbb{D}$, the processes in (2.35) weakly converge to the Gaussian process

$$\left((W^{(s)})_{s\in\mathcal{P}}, (W^{(s,j)})_{s\in\mathcal{P}_D, j\in\{1,2\}}\right)$$

as defined in Section 2.4.2. Noting that $u_n^{(s,j)}$ is the generalized inverse function of $L_n^{(s,j)} + 1/k^{(s)}$ and that $\sqrt{m^{(s)}}/k^{(s)} \to 0$, we apply Vervaat's lemma (Vervaat, 1972) to obtain that

$$G_n \rightsquigarrow G := \left((W^{(s)})_{s\in\mathcal{P}}, (-W^{(s,j)})_{s\in\mathcal{P}_D, j\in\{1,2\}}\right) \tag{2.36}$$

in $\mathbb{D}$.

Recall the definition of the sets $\mathcal{V}(t)$ in (2.30) and let $\mathbb{D}_n \subset \mathbb{D}$ be the subset of functions a of the form $\left((a^{(s)})_{s\in\mathcal{P}}, (a^{(s,j)})_{s\in\mathcal{P}_D, j\in\{1,2\}}\right)$ such that $a^{(s,1)}(x,y)$ is constant in y, $a^{(s,2)}(x,y)$ is constant in x and such that the functions $x \mapsto a^{(s,1)}(x,y)$ and $y \mapsto a^{(s,2)}(x,y)$ are elements of $\mathcal{V}\left(1/\sqrt{m^{(s)}}\right)$.

Defining $\mathbb{E}$ as the product space $(L^\infty([0,T]^2))^{|\mathcal{P}|}$, with $L^\infty([0,T]^2)$ equipped with the topology of hypi-convergence, consider the following functionals $f_n : \mathbb{D}_n \to \mathbb{E}$. For an element $a = \left((a^{(s)})_{s\in\mathcal{P}}, (a^{(s,j)})_{s\in\mathcal{P}_D, j\in\{1,2\}}\right) \in \mathbb{D}_n$, $f_n(a) = (f_n(a)^{(s)})_{s\in\mathcal{P}}$ is a function such that $f_n(a)^{(s)} = a^{(s)}$ if $s \in \mathcal{P}_I$, and

$$f_n(a)^{(s)}(x,y) = a^{(s)}(x,y) + \sqrt{m^{(s)}}\left(c^{(s)}\left(x + \frac{a^{(s,1)}(x,y)}{\sqrt{m^{(s)}}}, y + \frac{a^{(s,2)}(x,y)}{\sqrt{m^{(s)}}}\right) - c^{(s)}(x,y)\right)$$

if $s \in \mathcal{P}_D$. Referring to (2.34) and recalling that the second term thereof vanishes if $s \in \mathcal{P}_I$, we notice that for every pair s, $W_n^{(s)} = f_n(G_n)^{(s)} + o_{\mathbb{P}}(1)$. This representation,

of course, holds only if $G_n \in \mathbb{D}_n$; this is satisfied with probability at least

$$\mathbb{P}\left(\forall s \in \mathcal{P}_D, j \in \{1,2\}, u_n^{(s,j)}(T) \leq 2T\right) \longrightarrow 1$$

where the last convergence follows by Corollary 2.1 applied for each $s \in \mathcal{P}$. Define $f : \mathbb{D}_0 \to \mathbb{E}$, where $\mathbb{D}_0 \subset \mathbb{D}$ is the subset of continuous functions a such that $a(0) = 0$, as

$$f(a)^{(s)} = \begin{cases} a^{(s)}, & s \in \mathcal{P}_I \\ a^{(s)} + \dot{c}_1 a^{(s,1)} + \dot{c}_2 a^{(s,2)}, & s \in \mathcal{P}_D \end{cases}.$$

For a sequence $a_n \in \mathbb{D}_n$ that converges uniformly to a function $a \in \mathbb{D}_0$, $f_n(a_n) \to f(a)$ in $\mathbb{E}$. This can be seen by considering each pair separately; the result is obvious for asymptotically independent pairs, and for asymptotically dependent ones it follows from Lemma 2.9.

Finally, notice that the process G concentrates on $\mathbb{D}_0$. Therefore, by (2.36) and the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1),

$$\left(W_n^{(s)}\right)_{s\in\mathcal{P}} = f_n(G_n) + o_{\mathbb{P}}(1) \rightsquigarrow f(G) = \left(B^{(s)}\right)_{s\in\mathcal{P}}$$

in $\mathbb{E}$. □

Proof of Theorem 2.5

Similarly to the bivariate case, let

$$\Psi_n^{(s)}(\theta, \sigma) := (n/m)\Psi_n^{*(s)}(\theta, m\sigma/n).$$

As in the proof of Theorem 2.3, we may deduce that for every pair s, $\theta \in \widetilde{\Theta}$ and $\sigma > 0$,

$$\Psi^{(s)}(\theta,\sigma) - \Psi_n^{(s)}(\theta,\sigma) = \int g\left(\widehat{c}_n^{(s)} - c^{(s)}\right) d\mu_{\mathrm{L}} = \frac{1}{\sqrt{m}} \int g W_n^{(s)} d\mu_{\mathrm{L}},$$

with $W_n^{(s)}$ as defined in Theorem 2.4. By a similar argument to the bivariate case (involving the dominated convergence theorem and Lemma 2.10 to establish continuity of the mapping $f \mapsto \int g f d\mu_{\mathrm{L}}$, see the proof of Theorem 2.3 for the applicability of Lemma 2.10), Theorem 2.4 and the continuous mapping theorem yield

$$\left(\int g W_n^{(s)} d\mu_{\mathrm{L}}\right)_{s\in\mathcal{P}} \rightsquigarrow \left(\int g B^{(s)} d\mu_{\mathrm{L}}\right)_{s\in\mathcal{P}}. \tag{2.37}$$

The remaining proof consists of a number of successive applications of Lemma 2.11. We deal with each of the two estimators separately.

(i) For each pair s, applying Lemma 2.11 with $\phi = \Psi^{(s)}$, $x_0 = (h^{(s)}(\vartheta_0), 1)$, $a_n = 1/\sqrt{m}$ and $Y_n = \frac{1}{\sqrt{m}} \int g W_n^{(s)} d\mu_{\mathrm{L}}$ yields

$$\widehat{\theta}_n^{(s)} - h^{(s)}(\vartheta_0) = \frac{1}{\sqrt{m}} \mathcal{D}^{(s)} \int g W_n^{(s)} d\mu_{\mathrm{L}} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{m}}\right), \tag{2.38}$$

where $\mathcal{D}^{(s)}$ is the block corresponding to the pair s in the matrix $\mathcal{D}$ defined in (2.21); its existence, as well as the required smoothness of ϕ, are guaranteed by Condition 2.3. Now redefining ϕ as $\phi(\vartheta) = \left(h^{(s)}(\vartheta) - h^{(s)}(\vartheta_0)\right)_{s\in\mathcal{P}}$, we see that $\widehat{\vartheta}_n$ is in fact a minimizer of the norm of $\phi(\vartheta) - Y_n$, where Y_n is redefined as $\left(\widehat{\theta}_n^{(s)} - h^{(s)}(\vartheta_0)\right)_{s\in\mathcal{P}}$. Applying Lemma 2.11 again with ϕ and Y_n as above, $x_0 = \vartheta_0$ and $a_n = 1/\sqrt{m}$, we obtain

$$\begin{aligned}\widehat{\vartheta}_n - \vartheta_0 &= (J_1^\top J_1)^{-1} J_1^\top Y_n + o_{\mathbb{P}}\left(\frac{1}{\sqrt{m}}\right) \\ &= \frac{1}{\sqrt{m}} (J_1^\top J_1)^{-1} J_1^\top \left(\mathcal{D}^{(s)} \int g W_n^{(s)} d\mu_{\mathrm{L}}\right)_{s\in\mathcal{P}} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{m}}\right),\end{aligned}$$

where the last equality follows from (2.38) and J_1 is defined as in Section 2.4.2 in the paragraph below (2.21). The conclusion that $\sqrt{m}(\widehat{\vartheta}_n - \vartheta_0) \rightsquigarrow N(0, \Sigma_1)$ follows from this and (2.37).

(ii) Let $\widetilde{\sigma}_n = \frac{n}{m}\widetilde{\zeta}_n \in \mathbb{R}_+^{|\mathcal{P}|}$. Once more, we redefine

$$Y_n = \frac{1}{\sqrt{m}} \left(\int g W_n^{(s)} d\mu_{\mathrm{L}}\right)_{s\in\mathcal{P}} \quad \text{and} \quad \phi(\vartheta, \sigma) = \left(\Psi^{(s)}(h^{(s)}(\vartheta), \sigma^{(s)})\right)_{s\in\mathcal{P}}.$$

The estimator $(\widetilde{\vartheta}_n, \widetilde{\sigma}_n)$ can be seen to minimize the norm of $\phi - Y_n$. Therefore, applying Lemma 2.11 with $a_n = 1/\sqrt{m}$ and $x_0 = (\vartheta_0, 1, \ldots, 1)$, we obtain

$$(\widetilde{\vartheta}_n, \widetilde{\sigma}_n) - (\vartheta_0, 1, \cdots, 1) = \frac{1}{\sqrt{m}} (J_2^\top J_2)^{-1} J_2^\top \left(\int g W_n^{(s)} d\mu_{\mathrm{L}}\right)_{s\in\mathcal{P}} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{m}}\right),$$

which, combined with (2.37), implies $\sqrt{m}((\widetilde{\vartheta}_n, \widetilde{\sigma}_n) - (\vartheta_0, 1, \cdots, 1)) \rightsquigarrow N(0, \Sigma_2)$.

□

2.8 Auxiliary results

Throughout this chapter, particularly the proof of Lemma 2.2 below, we use (without reference when obvious) the following results on regularly varying functions at 0.

Lemma 2.1. *Suppose the functions f_1 and f_2 are regularly varying at 0 with indices ρ_1 and ρ_2, respectively.*

(i) If $\rho_1 > 0$ (respectively $\rho_1 < 0$), $\lim_{t\to 0} f_1(t) = 0$ (respectively ∞).

(ii) For any $\alpha \in \mathbb{R}$, f_1^α is $(\alpha\rho_1)$–RV at 0.

(iii) The product $f_1 f_2$ is $(\rho_1 + \rho_2)$–RV at 0.

(iv) If $\lim_{t\to 0} f_2(t) = 0$, then $f_1 \circ f_2$ is $(\rho_1\rho_2)$–RV at 0.

(v) If $\rho_1 > 0$, then f_1^{-1} is $(1/\rho_1)$–RV at 0, where we define the generalized inverse of f_1 as

$$f_1^{-1}(t) = \inf\{u > 0 : f_1(u) \geq t\}.$$

Proof. The assertions (ii) and (iii) are trivial consequences of the definition of regular variation. As for (i), (iv) and (v), analogue versions for regularly varying functions at ∞ are proved in Proposition 0.8 of Resnick (1987). The proof can readily be adapted, using the fact that f is ρ–RV at 0 if and only if $u \mapsto 1/f(1/u)$ is ρ–RV at ∞. □

Lemma 2.2. *(i) Assume (2.9). Then there exists $\eta \in (0, 1]$ such that q is a regularly varying (RV) function at 0 with index $1/\eta$ and c is $1/\eta$-homogeneous.*

(ii) Assume Condition 2.1(i) and suppose that q_1 is non-decreasing and that there exists $b > 1$ such that $q_1(bt) = O(q_1(t))$ as $t \to 0$. Then (2.9) holds locally uniformly on $[0, \infty)^2$.

Remark. In part (ii) of the previous result, the monotonicity condition on q_1 is artificial; it can be removed at the cost of replacing $q_1(t)$ by the non-decreasing function $\bar{q}_1(t) := \sup_{0<s\leq t} q_1(s)$. Indeed, if Condition 2.1 is satisfied with q_1, it is trivially satisfied with $\bar{q}_1$. Moreover, if $q_1(bt) = O(q_1(t))$, $\bar{q}_1$ also satisfies the same property.

Because q_1 is positive non-decreasing, that required property implies that $q_1(bt) = O(q_1(t))$ holds for every $b \geq 1$ (Bingham et al., 1987, Corollary 2.0.6). The function q_1 is then said to be O-regularly varying at 0.

Proof. (i) Recall that we assume $c(1, 1) = 1$. For any $x > 0$, (2.9) implies that $Q(tx, tx) = q(t)(c(x, x) + o(1))$ and $Q(tx, tx) = q(tx)(1 + o(1))$. This can be manipulated into

$$\frac{q(tx)}{q(t)} = \frac{c(x, x) + o(1)}{1 + o(1)} \longrightarrow c(x, x).$$

By Karamata's characterization theorem (Bingham et al., 1987, Theorem 1.4.1), q has to be ρ–RV and $c(x, x) = x^\rho$, for some $\rho \in \mathbb{R}$. However, since $q(t) \leq t$, we

must have $\rho \geq 1$. Moreover, for any $a, x, y > 0$,

$$c(ax, ay) = \lim_{t\to 0} \frac{Q(atx, aty)}{q(t)} = \lim_{t\to 0} \frac{Q(tx, ty)}{q(t/a)} = \lim_{t\to 0} \frac{Q(tx, ty)}{q(t)} \frac{q(t)}{q(t/a)} = a^\rho c(x, y).$$

Defining $\eta = 1/\rho$, this proves (i).

(ii) For arbitrary $(x, y) \in [0, \infty)^2$, we write $(x, y) = a(u, v)$. We will prove that (2.9) holds uniformly over all $(u, v) \in \mathcal{S}^+$ and over $a \in (0, b]$, for an arbitrary $b \in [1, \infty)$.

We have

$$\frac{Q(tx, ty)}{q(t)} = \frac{Q(atu, atv)}{q(t)} = \frac{q(at)}{q(t)} \frac{Q(atu, atv)}{q(at)}. \tag{2.39}$$

First, the term $Q(atu, atv)/q(at)$ is equal to $c(u, v) + O(q_1(at))$ uniformly in $(u, v) \in \mathcal{S}^+$. In order to control the term $q(at)/q(t)$, we note that since q is $1/\eta$-RV, there exists a slowly varying function ℓ such that for any $a > 0$,

$$\begin{aligned}
\frac{\ell(at)}{\ell(t)} - 1 &= a^{-1/\eta} \left(\frac{q(at)}{q(t)} - c(a, a) \right) \\
&= a^{-1/\eta} \left(\frac{Q(at, at)(1 + O(q_1(at)))}{q(t)} - c(a, a) \right) \\
&= a^{-1/\eta} \left(\frac{Q(at, at)}{q(t)} - c(a, a) + O(q_1(at)) \right) \\
&= O(q_1(t) + q_1(at)) = O(q_1(bt)) = O(q_1(t)),
\end{aligned}$$

where we have used the fact that $Q(at, at) = q(at)(1 + O(q_1(at)))$, which can be reversed into $q(at) = Q(at, at)(1 + O(q_1(at)))$. The function ℓ is thus said to be slowly varying with remainder (Bingham et al., 1987, Section 3.12). By theorem 3.12.1 of that book, the previous relation holds uniformly over all $a \in (1/2, b]$, so we henceforth focus on values $a \in (0, 1/2]$. Using Theorem 3.12.2 of the same book (which we adapt for slow variation at 0), we obtain that for some constants $C \in \mathbb{R}, T_0 \in (0, \infty)$ and for t small enough,

$$\ell(t) = \exp\left\{ C + \delta_1(t) + \int_t^{T_0} \frac{\delta_2(s)}{s} ds \right\},$$

where the functions δ_j are real-valued, measurable and satisfy $|\delta_j(t)| \leq K q_1(t)$ for some constant $K \in (0, \infty)$. The ratio $\ell(at)/\ell(t)$ becomes

$$\frac{\ell(at)}{\ell(t)} = \exp\left\{ \delta_1(at) - \delta_1(t) + \int_{at}^{t} \frac{\delta_2(s)}{s} ds \right\}.$$

As $t \to 0$, we can use the monotonicity of q_1 to control the integral in the previous display:

$$\left| \int_{at}^{t} \frac{\delta_2(s)}{s} ds \right| \leq K \int_{at}^{t} \frac{q_1(s)}{s} ds \leq K q_1(t) \int_{at}^{t} \frac{ds}{s} = K q_1(t) \log\left(\frac{1}{a}\right).$$

Because $a \leq 1/2$, $\log(1/a)$ is lower bounded, K can be chosen large enough so that $Kq_1(t)\log(1/a)$ also upper bounds the absolute value of $\delta_1(at) - \delta_1(t) + \int_{at}^{t} \frac{\delta_2(s)}{s} ds$. Therefore, using the fact that for every $h \in \mathbb{R}$, $|e^h - 1| \leq e^{|h|} - 1$, we obtain

$$\left| \frac{\ell(at)}{\ell(t)} - 1 \right| \leq \exp\left\{ K q_1(t) \log\left(\frac{1}{a}\right) \right\} - 1 = a^{-Kq_1(t)} - 1.$$

What we are interested in is bounding $q(at)/q(t) - a^{1/\eta}$. This can be done by recalling that

$$\left| \frac{q(at)}{q(t)} - a^{1/\eta} \right| = a^{1/\eta} \left| \frac{\ell(at)}{\ell(t)} - 1 \right| \leq a^{1/\eta} \left(a^{-Kq_1(t)} - 1 \right) =: \tau(a,t). \tag{2.40}$$

By simple differentiation, it is straightforward to see that for a fixed value of t small enough so that $Kq_1(t) < 1/\eta$, the function τ is differentiable in its first argument and that

$$\frac{\partial}{\partial a} \tau(a,t) = a^{1/\eta - 1} \left((1/\eta - Kq_1(t)) a^{-Kq_1(t)} - 1/\eta \right).$$

This suggests that the function attains its unique maximum at the point $a_{\max}(t) := (1 - \eta K q_1(t))^{1/(Kq_1(t))}$. Considering (2.40), we obtain that for all $a \in (0, 1/2]$,

$$\begin{aligned} \left| \frac{q(at)}{q(t)} - a^{1/\eta} \right| &\leq \tau(a_{\max}(t), t) \\ &= (1 - \eta K q_1(t))^{1/(\eta K q_1(t))} \left(\frac{1}{1 - \eta K q_1(t)} - 1 \right) \\ &= O(q_1(t)) \end{aligned}$$

as $t \to 0$, since $(1 - \eta K q_1(t))^{1/(\eta K q_1(t))} \to e^{-1}$ and since the function $x \mapsto 1/(1-x)$ is continuously differentiable at 0. Finally, this allows us to rewrite (2.39) as

$$\frac{Q(tx, ty)}{q(t)} = \left(a^{1/\eta} + O(q_1(t)) \right) \left(c(u,v) + O(q_1(at)) \right) = a^{1/\eta} c(u,v) + O(q_1(t)),$$

and the last equation holds uniformly over $a \in (0, b]$ and $(u, v) \in \mathcal{S}^+$. The proof is over since $a^{1/\eta} c(u,v) = c(x,y)$.

□

Lemma 2.3. *Let $\varphi : (0, T] \to (0, \infty)$ be a non-decreasing function such that $\varphi(t)/\sqrt{t} \to \infty$ as $t \to 0$ and assume there exists $c > 0$ such that*

$$\int_0^T \frac{1}{x} \exp\left\{-c\frac{\varphi^2(x)}{x}\right\} dx < \infty.$$

Then under the assumptions of Theorem 2.1, for every $\lambda \in (0, 1)$ we have

$$\sup_{\lambda/k \le x \le T} \frac{\sqrt{k}}{\varphi(x)} |u_n(x) - x| = O_{\mathbb{P}}(1),$$

where u_n is defined as in Section 2.7.1. In particular, note that $\varphi(x) := 1$, as well as any function that satisfies $\varphi(x) := \sqrt{x \log\log(1/x)}$ in a neighborhood of 0, are valid choices.

Proof. This is essentially proved in Csörgő and Horváth (1987), up to a slight difference between their definition of the quantiles and ours. We prove here that this difference does not change the result. More precisely, their Theorem 2.6 (ii) states that

$$\sup_{\lambda/k \le x \le T} \frac{|w_n(x)|}{\varphi(x)} = O_{\mathbb{P}}(1), \tag{2.41}$$

where we denote w_n what they call v_n (to avoid confusion with our definitions). From their definitions, one easily sees that

$$w_n(x) = \frac{n}{\sqrt{k}} \left(\frac{k}{n} x - U_{n,\lceil kx \rceil}\right) = \sqrt{k} \left(x - \frac{n}{k} U_{n,\lceil kx \rceil}\right).$$

Then, by the reverse triangle inequality,

$$\begin{aligned} |\sqrt{k}|u_n(x) - x| - |w_n(x)|| &\le |\sqrt{k}(u_n(x) - x) + w_n(x)| \\ &= \sqrt{k} \left| u_n(x) - \frac{n}{k} U_{n,\lceil kx \rceil} \right| = \frac{n}{\sqrt{k}} \left(U_{n,\lceil kx \rceil} - U_{n,\lfloor kx \rfloor}\right). \end{aligned}$$

Using this and the inequality $\lfloor x \rfloor \ge \lceil x \rceil - 1$, we have

$$\begin{aligned} &\left| \sup_{\lambda/k \le x \le T} \frac{\sqrt{k}}{\varphi(x)} |u_n(x) - x| - \sup_{\lambda/k \le x \le T} \frac{|w_n(x)|}{\varphi(x)} \right| \\ &\quad \le \frac{n}{\sqrt{k}} \sup_{\lambda/k \le x \le T} \frac{1}{\varphi(x)} \left(U_{n,\lceil kx \rceil} - U_{n,\lfloor kx \rfloor}\right) \\ &\quad \le \frac{n}{\sqrt{k}} \sup_{\lambda/k \le x \le T} \frac{1}{\varphi(x)} \left(U_{n,\lceil kx \rceil} - U_{n,\lceil kx \rceil - 1}\right) \end{aligned}$$

$$\leq \frac{n}{\sqrt{k}} \sup_{\lambda/k \leq x \leq (1+\lambda)/k} \frac{1}{\varphi(x)} \left(U_{n,\lceil kx \rceil} - U_{n,\lceil kx \rceil - 1} \right) + \frac{n}{\sqrt{k}} \sup_{(1+\lambda)/k \leq x \leq T} \frac{1}{\varphi(x)} \left(U_{n,\lceil kx \rceil} - U_{n,\lceil kx \rceil - 1} \right). \tag{2.42}$$

In the first term, since $\lambda/k \leq x \leq (1+\lambda)/k$ and $\lambda \in (0,1)$, we must have $\lceil kx \rceil \in \{1,2\}$. Therefore, we end up studying $U_{n,i} - U_{n,i-1}$, for some $i \in \{1,2\}$. It is a well known fact that those differences, regardless of the value of i, have a Beta distribution with parameters 1 and n. In particular, they are both $O_{\mathbb{P}}(1/n)$. It follows that the first supremum on the right hand side of (2.42) is asymptotically bounded in probability by

$$\frac{1}{\sqrt{k}} \sup_{\lambda/k \leq x \leq (1+\lambda)/k} \frac{1}{\varphi(x)} = \frac{1}{\sqrt{k}\varphi(\lambda/k)} \longrightarrow 0$$

by assumption on φ. As for the second term in (2.42), it is equal to

$$\frac{n}{\sqrt{k}} \sup_{(1+\lambda)/k \leq x \leq T} \frac{1}{\varphi(x)} \left(U_{n,\lceil kx \rceil} - U_{n,\lceil k(x-1/k) \rceil} \right) = \frac{n}{\sqrt{k}} \sup_{\lambda/k \leq x \leq T-1/k} \frac{1}{\varphi(x+1/k)} \left(U_{n,\lceil k(x+1/k) \rceil} - U_{n,\lceil kx \rceil} \right)$$

after shifting x to the right by $1/k$. Using (2.41), this is in turn equal to

$$\begin{aligned} &\frac{n}{\sqrt{k}} \sup_{\lambda/k \leq x \leq T-1/k} \frac{1}{\varphi(x+1/k)} \left(\frac{k}{n} \left(x + \frac{1}{k} \right) - \frac{k}{n} x \right) + \frac{n}{\sqrt{k}} O_{\mathbb{P}} \left(\frac{\sqrt{k}}{n} \right) \\ &= \frac{1}{\sqrt{k}} \sup_{\lambda/k \leq x \leq T-1/k} \frac{1}{\varphi(x+1/k)} + O_{\mathbb{P}}(1) \\ &= \frac{1}{\sqrt{k}\varphi((1+\lambda)/k)} + O_{\mathbb{P}}(1) \\ &= O_{\mathbb{P}}(1) \end{aligned}$$

once again by the properties of φ. We have shown that the difference between the quantity we are interested in and the term appearing in (2.41) is $O_{\mathbb{P}}(1)$. We may thus conclude, by (2.41), that

$$\sup_{\lambda/k \leq x \leq T} \frac{\sqrt{k}}{\varphi(x)} |u_n(x) - x| = \sup_{\lambda/k \leq x \leq T} \frac{|w_n(x)|}{\varphi(x)} + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1).$$

□

Corollary 2.1. *Define the random functions u_n and v_n as in Section 2.7.1. Then, as*

$n \to \infty$,

$$\sup_{0 \le x \le 2T} |u_n(x) - x| \quad \textit{and} \quad \sup_{0 \le y \le 2T} |v_n(y) - y|$$

are both $O_{\mathbb{P}}\left(1/\sqrt{k}\right)$.

Proof. Note that by definition, $u_n(z) = v_n(z) = 0$ whenever $z < 1/k$. It follows that

$$\begin{aligned}
\sup_{0 \le x \le 2T} |u_n(x) - x| &\le \sup_{0 \le x < 1/k} |u_n(x) - x| + \sup_{1/k \le x \le 2T} |u_n(x) - x| \\
&= \sup_{0 \le x < 1/k} x + \sup_{1/k \le x \le 2T} |u_n(x) - x| \\
&= \frac{1}{k} + \sup_{1/k \le x \le 2T} |u_n(x) - x|.
\end{aligned}$$

This is $O_{\mathbb{P}}\left(1/\sqrt{k}\right)$ by the preceding Lemma 2.3 with the function $\varphi(x) = 1$. The same proof holds with u_n replaced by v_n. □

Lemma 2.4. *Under Condition 2.1 the process H_n as defined in (2.24) converges to the process W from Theorem 2.1 in $\ell^\infty([0, 2T]^2)$.*

Proof. Denoting $f_{n,(x,y)}(u, v) := \sqrt{\frac{n}{m}} \mathbb{1}\left\{u \le \frac{k}{n}x, v \le \frac{k}{n}y\right\}$, we see that H_n can be written as

$$H_n(x, y) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_{n,(x,y)}(U_i, V_i) - \mathbb{E}\left[f_{n,(x,y)}(U, V)\right]\right).$$

Therefore, convergence of the process H_n to a Gaussian process in $\ell^\infty([0, 2T]^2)$ is equivalent to checking that the sequence of function classes

$$\mathcal{F}_n = \{f_{n,(x,y)} : (x, y) \in [0, 2T]^2\}$$

are Donsker classes for the distribution of (U, V). This is guaranteed by Theorem 11.20 of Kosorok (2008), provided that we can check the six conditions. Note that $\mathcal{F}_n$ admits the envelope function $F_n = f_{n,(2T,2T)}$.

(0) First, the AMS condition is trivially satisfied; by right continuity of indicator functions, for any $n \in \mathbb{N}$, $(x, y) \in [0, 2T]^2$ and $(u, v) \in [0, 1]^2$,

$$\inf_{(x',y') \in \mathbb{Q}^2} |f_{n,(x',y')}(u, v) - f_{n,(x,y)}(u, v)| = 0.$$

It follows that Equation (11.7) of Kosorok (2008) is satisfied with $T_n = \mathbb{Q}^2$, which is countable. Hence the classes $\mathcal{F}_n$ are AMS.

(A) For every n, it is easily checked that $\mathcal{F}_n$ is a VC class with VC-index 2. Therefore, condition (A) is a direct consequence of Lemma 11.21 of Kosorok (2008).

(B) For $(x,y),(x',y') \in [0,2T]^2$ arbitrary, it follows from the definition of H_n that

$$\begin{aligned}\mathbb{E}\left[H_n(x,y)H_n(x',y')\right] &= \mathbb{E}\left[f_{n,(x,y)}(U,V)f_{n,(x',y')}(U,V)\right] \\ &\quad - \mathbb{E}\left[f_{n,(x,y)}(U,V)\right]\mathbb{E}\left[f_{n,(x',y')}(U,V)\right] \\ &= \frac{n}{m}\mathbb{P}\left(U \le \frac{k}{n}(x\wedge x'), V \le \frac{k}{n}(y\wedge y')\right) \\ &\quad - \frac{n}{m}\mathbb{P}\left(U \le \frac{k}{n}x, V \le \frac{k}{n}y\right)\mathbb{P}\left(U \le \frac{k}{n}x', V \le \frac{k}{n}y'\right).\end{aligned}$$

Recall that $n/m = 1/q(k/n)$. Therefore, the first term of the last display converges to $c(x\wedge x', y\wedge y')$. The second term vanishes since both probabilities are of the order of m/n. The convariance functions of H_n thus converge pointwise to the covariance function of W.

(C) By definition of the envelope functions and by assumption, we have

$$\limsup_{n\to\infty}\mathbb{E}\left[F_n^2(U,V)\right] = \limsup_{n\to\infty}\frac{n}{m}\mathbb{P}\left(U \le \frac{k}{n}2T, V \le \frac{k}{n}2T\right) = c(2T,2T) < \infty.$$

(D) For every $\varepsilon > 0$,

$$\mathbb{E}\left[F_n^2(U,V)\mathbb{1}\left\{F_n(U,V) > \varepsilon\sqrt{n}\right\}\right] \le \frac{n}{m}\mathbb{1}\left\{\sqrt{\frac{n}{m}} > \varepsilon\sqrt{n}\right\},$$

which is equal to 0 as soon as $m \ge \varepsilon^{-2}$.

(E) We first recall that for arbitrary events A, B,

$$\mathbb{P}\left(\mathbb{1}_A \ne \mathbb{1}_B\right) = \mathbb{P}\left(A\backslash B\right) + \mathbb{P}\left(B\backslash A\right) = \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) - 2\mathbb{P}\left(A\cap B\right).$$

A direct application of this fact yields

$$\begin{aligned}\rho_n^2((x,y),(x',y')) &:= \mathbb{E}\left[\left(f_{n,(x,y)}(U,V) - f_{n,(x',y')}(U,V)\right)^2\right] \\ &= \frac{n}{m}\mathbb{P}\left(\mathbb{1}\left\{U \le \frac{k}{n}x, V \le \frac{k}{n}y\right\} \ne \mathbb{1}\left\{U \le \frac{k}{n}x', V \le \frac{k}{n}y'\right\}\right) \\ &= \frac{n}{m}\mathbb{P}\left(U \le \frac{k}{n}x, V \le \frac{k}{n}y\right) + \frac{n}{m}\mathbb{P}\left(U \le \frac{k}{n}x', V \le \frac{k}{n}y'\right) \\ &\quad - 2\frac{n}{m}\mathbb{P}\left(U \le \frac{k}{n}(x\wedge x'), V \le \frac{k}{n}(y\wedge y')\right)\end{aligned}$$

$$\begin{aligned}&\longrightarrow c(x,y)+c(x',y')-2c(x\wedge x',y\wedge y')\\&=:\rho^2((x,y),(x',y')).\end{aligned}$$

Moreover, by Lemma 2.2(ii), this convergence is uniform over $[0,2T]^4$. This means that for any sequences x_n, y_n, x'_n, y'_n in $[0,2T]$ such that $\rho((x_n,y_n),(x'_n,y'_n))\to 0$, $\rho_n((x_n,y_n),(x'_n,y'_n))$ is equal to

$$\begin{aligned}&\{\rho_n((x_n,y_n),(x'_n,y'_n))-\rho((x_n,y_n),(x'_n,y'_n))\}+\rho((x_n,y_n),(x'_n,y'_n))\\&\quad\le \sup_{(x,y,x',y')\in[0,2T]^4}|\rho_n((x,y),(x',y'))-\rho((x,y),(x',y'))|\\&\qquad+\rho((x_n,y_n),(x'_n,y'_n))\\&\quad\longrightarrow 0.\end{aligned}$$

Finally, the theorem implies that $H_n \rightsquigarrow W$ in $\ell^\infty([0,2T]^2)$. □

Lemma 2.5. *Let Q be a bivariate copula. If there exists a positive function q and a finite function c that is not everywhere 0 such that for every $(x,y)\in[0,\infty)^2$, as $n\to\infty$,*

$$\frac{Q(x/n,y/n)}{q(1/n)}\longrightarrow c(x,y),$$

then there exists a measure ν such that for every $(x,y)\in[0,\infty)^2$, $c(x,y)=\nu((0,x]\times(0,y])$. Note that (2.9) satisfies this setting.

Proof. Define the measures ν_n by

$$\nu_n((0,x]\times(0,y])=\frac{Q(x/n,y/n)}{q(1/n)}$$

and fix $a\in(0,\infty)$. Note that since c is not everywhere 0, $c(a,a)$ is eventually positive, so for n and a large enough, $\nu_n((0,a]^2)>0$. Then clearly

$$P_{n,a}:=\left(\nu_n((0,a]^2)\right)^{-1}\nu_n$$

is a probability measure on $[0,a]^2$. Since it is supported on the same compact set for every n, the sequence $\{P_{n,a}: n\in\mathbb{N}\}$ is tight. Thus, by Helly's selection theorem there exists a probability measure P_a also supported on $[0,a]^2$ and a subsequence $\{n_j: j\in\mathbb{N}\}$ such that $P_{n_j,a}\rightsquigarrow P_a$. However, by definition of ν_n, we have for every $(x,y)\in[0,a]^2$

$$P_{n_j,a}((0,x]\times(0,y])\longrightarrow\frac{c(x,y)}{c(a,a)}.$$

Therefore, we must have $P_a((0,x]\times(0,y])=c(x,y)/c(a,a)$, so choosing $\nu_a=$

$c(a,a)P_a$, the result holds for every $(x,y) \in [0,a]^2$. However, the value of $\nu_a((0,x] \times (0,y])$ is independent of a (as long as $x \vee y \leq a$), so ν_a can be uniquely extended to a measure ν on the bounded Borel sets of $[0,\infty)^2$. □

Lemma 2.6 (similar to Theorem 1 in Ramos and Ledford (2009))**.** *Define the function c as in (2.9). Then there exists a finite measure H on $[0,1]$ such that, for every $(x,y) \in [0,\infty)^2$,*

$$c(x,y) = \int_{[0,1]} \left(\frac{x}{1-w} \wedge \frac{y}{w} \right)^{1/\eta} H(dw).$$

It is also useful to note that this integral is equal to

$$\int_{[0,\frac{y}{x+y}]} \left(\frac{x}{1-w} \right)^{1/\eta} H(dw) + \int_{(\frac{y}{x+y},1]} \left(\frac{y}{w} \right)^{1/\eta} H(dw).$$

Proof. By Lemma 2.5, we can write

$$c(x,y) = \nu((0,x] \times (0,y]) = \int_{[0,\infty)^2} \mathbb{1}_{(0,x]\times(0,y]} d\nu = \int_{[0,\infty)^2 \setminus \{0\}} \mathbb{1}_{[0,x]\times[0,y]} d\nu. \tag{2.43}$$

In the last equality, nothing changed since $\nu((0,x] \times \{0\} \cup \{0\} \times (0,y]) \leq c(x,0) + c(0,y) = 0$. Then, through the mapping $f : [0,\infty)^2 \setminus \{0\} \to (0,\infty) \times [0,1]$ defined by $f(x,y) = (x+y, \frac{y}{x+y})$, define the push-forward measure $\mu = \nu \circ f^{-1}$. By homogeneity of ν, we see that μ is a product measure:

$$\mu((0,r] \times (0,w]) = r^{1/\eta} \mu((0,1] \times (0,w]) =: G((0,r])H((0,w]),$$

where G is a measure on $(0,\infty)$ and H is a measure on $[0,1]$. Finally, for any (x,y), define the function $g : (0,\infty) \times [0,1] \to \mathbb{R}$ as

$$g(r,w) = \mathbb{1}\left\{ r \leq \frac{x}{1-w} \wedge \frac{y}{w} \right\},$$

so that $g \circ f = \mathbb{1}_{[0,x]\times[0,y]}$. Using (2.43) and Theorem 9.15 from Teschl (1998), we have

$$\begin{aligned} c(x,y) &= \int_{[0,\infty)^2 \setminus \{0\}} g \circ f d\nu \\ &= \int_{(0,\infty)\times[0,1]} g d\mu \\ &= \int_{[0,1]} \int_{(0,\infty)} \mathbb{1}_{(0,\frac{x}{1-w} \wedge \frac{y}{w}]}(r) G(dr) H(dw) \\ &= \int_{[0,1]} \left(\frac{x}{1-w} \wedge \frac{y}{w} \right)^{1/\eta} H(dw), \end{aligned}$$

where we used Fubini's theorem to write the integral with respect to the product measure μ as a double integral. Moreover, note that H is finite since

$$H([0,1]) = \mu((0,1]\times[0,1]) = \nu\left(\left\{(x,y)\in[0,\infty)^2 : x+y\le 1\right\}\right) \le c(1,1) = 1.$$

□

Lemma 2.7. *Define the function c as in (2.9). Then for every $(x,y)\in[0,T]^2$ and $h>0$,*

$$c(x+h,y) - c(x,y) \le \frac{1}{\eta} h \frac{c(x+h,y)}{x+h}.$$

Proof. By Lemma 2.6, write

$$c(x,y) = \int_{[0,1]} \left(\frac{x}{1-w} \wedge \frac{y}{w}\right)^{1/\eta} H(dw) =: \int_{[0,1]} f(x,y,w) H(dw).$$

Clearly, it is sufficient to prove that for every x, y, h, w,

$$f(x+h,y,w) - f(x,y,w) \le \frac{1}{\eta} h \frac{f(x+h,y,w)}{x+h}, \tag{2.44}$$

because then the result follows by integrating both sides. To prove (2.44), first note that for any y, w,

$$f(x,y,w) = \begin{cases} \left(\frac{x}{1-w}\right)^{1/\eta} & , \quad x \le \frac{1-w}{w} y \\ \left(\frac{y}{w}\right)^{1/\eta} & , \quad x \ge \frac{1-w}{w} y \end{cases}.$$

As a function of x, this is continuously differentiable everywhere on $(0,T]$ except at the change point $x = \frac{1-w}{w} y$ and its derivative with respect to x, f', is equal to $f(x,y,h)/(\eta x)$ on the first part and 0 on the second. From here we consider three different cases, depending on the position of the change point with respect to x and $x+h$.

First, if $x + h \le \frac{1-w}{w} y$,

$$f(x+h,y,w) - f(x,y,w) = h f'(\xi,y,w) = h \frac{f(\xi,y,w)}{\eta\xi},$$

for some $\xi \in [x, x+h]$, by Taylor's theorem. By monotonicity, this is upper bounded by

$$\frac{1}{\eta} h \frac{f(x+h,y,w)}{x+h}.$$

Next, if $\frac{1-w}{w} y \le x$, $f(x+h,y,w) - f(x,y,w) = 0$ so the result is trivial.

Finally, if $x < \frac{1-w}{w}y < x+h$,

$$f(x+h,y,w)-f(x,y,w) = f\left(\frac{1-w}{w}y,y,w\right)-f(x,y,w) = \left(\frac{1-w}{w}y-x\right)\frac{f(\xi,y,w)}{\eta\xi},$$

for ξ between x and $\frac{1-w}{w}y$, once again by Taylor's theorem. By monotonicity, we have

$$\frac{f(\xi,y,w)}{\eta\xi} \le \frac{1}{\eta\frac{1-w}{w}y}\left(\frac{y}{w}\right)^{1/\eta} = \frac{1}{\eta\frac{1-w}{w}y}f(x+h,y,w).$$

Moreover,

$$\frac{\frac{1-w}{w}y-x}{\frac{1-w}{w}y} \le \frac{(x+h)-x}{(x+h)} = \frac{h}{x+h},$$

because the function $t \mapsto (t-x)/t$ is non-decreasing. Piecing everything together, we have

$$f(x+h,y,w)-f(x,y,w) \le \frac{1}{\eta}h\frac{f(x+h,y,w)}{x+h}.$$

We have proved that (2.44) holds for every $(x,y) \in [0,T]^2$, $h>0$ and $w \in [0,1]$. □

Lemma 2.8. *Define the function c as in (2.9) and assume Condition 2.1(i). Then there exists a constant $K := K_T < \infty$ such that for every $(x,y) \in [0,T]^2$,*

$$c(x,y) \le \frac{K}{\log(1/x)}.$$

Proof. We will prove that as $x \to 0$,

$$c(x,y) \lesssim \frac{1}{\log(1/x)}$$

uniformly for all $y \in [0,T]$. Since c is locally bounded, the result will follow.

Since Condition 2.1(i) is satisfied, we may assume it is satisfied with the function $q_1(t) = 1/\log(1/t)$. Recall that as $t \downarrow 0$, by Lemma 2.2,

$$Q(tx,ty) = q(t)c(x,y) + O(q(t)q_1(t))$$

uniformly over all $(x,y) \in [0,T]^2$. That is,

$$c(x,y) = \frac{Q(tx,ty)}{q(t)} + O(q_1(t)) \le \frac{tx}{q(t)} + O(q_1(t)) \tag{2.45}$$

uniformly, by Lipschitz continuity of the copula Q. The previous relation holds whenever $t \to 0$, and in particular it holds when t and x are related and both tend to 0.

Define $g(t) = q(t)q_1(t)/t \to 0$ as $t \to 0$. We argue, in the following, that for any x small enough, there exists $t(x) > 0$ such that $x \le g(t(x)) \le 2^{1/\eta}x$. Plugging $t(x)$ into (2.45), we find that as $x \to 0$,

$$c(x,y) \le \frac{t(x)x}{q(t(x))} + O(q_1(t(x))) = O(q_1(t(x))), \tag{2.46}$$

because, since we assume $x \le g(t(x))$,

$$\frac{t(x)x}{q(t(x))} \le \frac{t(x)}{q(t(x))} g(t(x)) = q_1(t(x)).$$

Moreover, since the function g is ρ-RV at 0, $\rho := 1/\eta - 1$, for small enough t we have $g(t) \ge t^\alpha$, as long as $\alpha > \rho$. This means that

$$q_1(t(x)) = \frac{1}{\log(1/t(x))} = \frac{\alpha}{\log(1/t(x)^\alpha)} \lesssim \frac{1}{\log(1/g(t(x)))}.$$

Finally, by the assumption that $g(t(x)) \le 2^{1/\eta}x$, we get

$$q_1(t(x)) \lesssim \frac{1}{\log(1/g(t(x)))} \lesssim \frac{1}{\log(1/x)}$$

which, in conjunction with (2.46), yields the desired bound for $c(x,y)$ as $x \to 0$, uniformly over bounded y.

The only thing left is to prove the existence of a point $t(x)$ such that $g(t(x)) \in [x, 2^{1/\eta}x]$ for every small enough x. This can be done by using the fact that the function g is ρ-RV at 0. Applying Theorem 1.5.6(iii) in Bingham et al. (1987) (adapted to functions of regular variation at 0) with any $\delta \in (0,1)$ and $A = 2^{1-\delta}$, we find that there exists $T_0 \in (0,\infty)$ such that for every $t \le T_0$,

$$\frac{g(t)}{g(t/2)} \le 2^{1-\delta}2^{\rho+\delta} = 2^{1/\eta}.$$

We now construct a non-increasing sequence the follwing way: take $t_0 = T_0$ and for $n \in \mathbb{N}$, define $t_n = t_{n-1}/2$ if $g(t_{n-1}/2) \le g(t_{n-1})$. Otherwise, $t_n = t_{n-1}/4$ if $g(t_{n-1}/4) \le g(t_{n-1})$. Otherwise, we try $t_{n-1}/8$, etc. In general

$$t_n = \max\left\{ \frac{t_{n-1}}{2^k} : k \in \mathbb{N}, g\left(\frac{t_{n-1}}{2^k}\right) \le g(t_{n-1}) \right\}.$$

Therefore, the sequence satisfies, for every natural n,

$$1 \le \frac{g(t_n)}{g(t_{n+1})} \le 2^{1/\eta}. \tag{2.47}$$

Now choose any $x \in (0, T_0/2]$ and let $t = \min_{n\in\mathbb{N}}\{t_n : g(t_n) \geq x\}$. Clearly, $g(t) \geq x$, and $g(t)$ has to be $\leq 2^{1/\eta}x$. Indeed, suppose the opposite. Then by (2.47), $g(t_{n+1}) \geq g(t)/2^{1/\eta} > x$, which contradicts the definition of t. We conclude that for every $x \in (0, T_0/2]$, the desired $t(x)$ exists. □

Lemma 2.9. *Assume the setting of Theorem 2.2. For arbitrary positive t and T, let*

$$\mathcal{V}(t) := \{b \in \ell^\infty([0, 2T]) : \forall x \in [0, T], x + tb(x) \in [0, 2T]\}.$$

Let $t_n \downarrow 0$ and assume that $b_n := (b_n^{(1)}, b_n^{(2)}) \in \mathcal{V}(t_n)^2$ converges uniformly to a continuous function $b = (b^{(1)}, b^{(2)})$ such that $b^{(1)}(0) = b^{(2)}(0) = 0$. Then, the functions $g_n : [0, T] \to \mathbb{R}$ defined by

$$g_n(x, y) := \frac{c\left(x + t_n b_n^{(1)}(x), y + t_n b_n^{(2)}(y)\right) - c(x, y)}{t_n}$$

hypi-converge to $\dot{c}_1(x, y)b^{(1)}(x) + \dot{c}_2(x, y)b^{(2)}(y)$, where $\dot{c}_1$ and $\dot{c}_2$ are defined as in Section 2.4.1.

Proof. Let L be the stable tail dependence function associated to the random vector (X, Y). Because we assume asymptotic dependence, we know that $\chi := \lim_{t\downarrow 0} q(t)/t > 0$ and that $c(x, y) = (x + y - L(x, y))/\chi$. Then,

$$g_n(x, y) = \chi^{-1}\left(b_n^{(1)}(x) + b_n^{(2)}(y) - \frac{L\left(x + t_n b_n^{(1)}(x), y + t_n b_n^{(2)}(y)\right) - L(x, y)}{t_n}\right).$$

By assumption, the sum of the first two terms converges uniformly to $b^{(1)}(x)+b^{(2)}(y)$. Let $\mathcal{S} \subset [0, \infty)^2$ be the set of points where L is differentiable. Since L is convex, the complement of $\mathcal{S}$ is Lebesgue-null and the gradient of L is continuous on $\mathcal{S}$ (Rockafellar, 1970, Theorem 25.5). By Lemma F.3 of Bücher et al. (2014), the last term hypi-converges to

$$\mathcal{L}_1(x, y) := \sup_{\varepsilon>0} \inf\left\{\dot{L}_1(x', y')b^{(1)}(x') + \dot{L}_2(x', y')b^{(2)}(y') : (x', y') \in \mathcal{S}, \|(x, y) - (x', y')\| < \varepsilon\right\},$$

where $\dot{L}_j$ are defined like $\dot{c}_j$: $\dot{L}_1(x, y)$ is the first partial derivative at (x, y) from the left, except if $x = 0$ in which case it is from the right, and $\dot{L}_2$ is always the second partial derivative from the right. We argue below that the hypi-distance between the functions $\mathcal{L}_1$ and $\mathcal{L}_2$, defined by $\mathcal{L}_2(x, y) = \dot{L}_1(x, y)b^{(1)}(x) + \dot{L}_2(x, y)b^{(2)}(y)$, is 0. That is, $\mathcal{L}_1$ and $\mathcal{L}_2$ belong to the same equivalence class in the space $L^\infty([0, 2T]^2)$ and

hypi-convergence to $\mathcal{L}_1$ is equivalent to hypi-convergence to $\mathcal{L}_2$. It follows that $g_n(x,y)$ hypi-converges to

$$\frac{b^{(1)}(x) + b^{(2)}(y) - \mathcal{L}_2(x,y)}{\chi} = \dot{c}_1(x,y)b^{(1)}(x) + \dot{c}_2(x,y)b^{(2)}(y), \tag{2.48}$$

where the last equality is a consequence of the relation $\dot{L}_j(x,y) = 1 - \chi \dot{c}_j(x,y)$, $j \in \{1,2\}$.

To prove the equivalence between $\mathcal{L}_1$ and $\mathcal{L}_2$, first note that by continuity of $b^{(1)}$ and $b^{(2)}$,

$$\mathcal{L}_1(x,y) := \sup_{\varepsilon>0} \inf \left\{ \dot{L}_1(x',y')b^{(1)}(x) + \dot{L}_2(x',y')b^{(2)}(y) : (x',y') \in \mathcal{S}, \|(x,y) - (x',y')\| < \varepsilon \right\}.$$

Let $\dot{L}_j^-$ and $\dot{L}_j^+$ denote the directional partial derivatives of L from the left and right, respectively. The function $\mathcal{L}_2$ can then be expressed the following way, and we analogously define $\mathcal{L}_3$:

$$\begin{aligned}\mathcal{L}_2(x,y) &= \dot{L}_1^-(x,y)b^{(1)}(x) + \dot{L}_2^+(x,y)b^{(2)}(y),\\ \mathcal{L}_3(x,y) &:= \dot{L}_1^+(x,y)b^{(1)}(x) + \dot{L}_2^-(x,y)b^{(2)}(y).\end{aligned}$$

The main tool is the homogeneity property of L ($L(ax,ay) = aL(x,y)$, $a \geq 0$). It implies that the directional derivatives $\dot{L}_j^\pm$ are constant along rays of the form $\{az : a > 0\}$, $z \in (0,\infty)^2$ and therefore that $\mathcal{S}$ consists exactly of a dense union of such rays.

Fix a point $(x,y) \in (0,\infty)^2$. For any sufficiently small $\varepsilon > 0$, the open ε-ball $B(\varepsilon)$ around (x,y) can be partitioned into the two open "half-balls"

$$B_1(\varepsilon) := \{(x',y') \in B(\varepsilon) : y'/x' > y/x\}, \quad B_2(\varepsilon) := \{(x',y') \in B(\varepsilon) : y'/x' < y/x\}$$

and the line $B_3(\varepsilon) := \{(x',y') \in B(\varepsilon) : y'/x' = y/x\}$. Provided that ε is sufficiently small, there exists a positive $\delta = \delta(\varepsilon)$ such that $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$, such that each point in $B_1(\varepsilon)$ is on the same ray as some $u \in (x-\delta, x] \times \{y\}$ and some $v \in \{x\} \times [y, y+\delta)$ and such that each point in $B_2(\varepsilon)$ is on the same ray as some $u \in (x, x+\delta) \times \{y\}$ and some $v \in \{x\} \times (y-\delta, y)$. By Rockafellar (1970), Theorem 24.1, we have

$$\begin{aligned}\lim_{\delta \downarrow 0} \dot{L}_1^\pm(x-\delta, y) &= \dot{L}_1^-(x,y), \quad \lim_{\delta \downarrow 0} \dot{L}_1^\pm(x+\delta, y) = \dot{L}_1^+(x,y),\\ \lim_{\delta \downarrow 0} \dot{L}_2^\pm(x, y-\delta) &= \dot{L}_2^-(x,y), \quad \lim_{\delta \downarrow 0} \dot{L}_2^\pm(x, y+\delta) = \dot{L}_2^+(x,y).\end{aligned}$$

Then, as $\varepsilon \to 0$, the vectors $(\dot{L}_1^{\pm}(x',y'), \dot{L}_2^{\pm}(x',y'))$ converge to $(\dot{L}_1^{-}(x,y), \dot{L}_2^{+}(x,y))$ for $(x',y') \in B_1(\varepsilon)$ and to $(\dot{L}_1^{+}(x,y), \dot{L}_2^{-}(x,y))$ for $(x',y') \in B_2(\varepsilon)$. It follows by continuity of b that for any sufficiently small $\varepsilon > 0$,

$$\lim_{(x',y')\to(x,y),(x',y')\in B_1(\varepsilon)} \mathcal{L}_2(x',y') = \lim_{(x',y')\to(x,y),(x',y')\in B_1(\varepsilon)} \mathcal{L}_3(x',y') = \mathcal{L}_2(x,y) \tag{2.49}$$

$$\lim_{(x',y')\to(x,y),(x',y')\in B_2(\varepsilon)} \mathcal{L}_2(x',y') = \lim_{(x',y')\to(x,y),(x',y')\in B_2(\varepsilon)} \mathcal{L}_3(x',y') = \mathcal{L}_3(x,y) \tag{2.50}$$

In particular, since $\dot{L}_j^{\pm}$ are constant on $B_3(\varepsilon)$, the semicontinuous hulls of $\mathcal{L}_2$ are

$$\mathcal{L}_{2,\wedge}(x,y) := \sup_{\varepsilon>0} \inf \{\mathcal{L}_2(x',y') : (x',y') \in B(\varepsilon)\} = \mathcal{L}_2(x,y) \wedge \mathcal{L}_3(x,y),$$
$$\mathcal{L}_{2,\vee}(x,y) := \inf_{\varepsilon>0} \sup \{\mathcal{L}_2(x',y') : (x',y') \in B(\varepsilon)\} = \mathcal{L}_2(x,y) \vee \mathcal{L}_3(x,y),$$

and since $B_1(\varepsilon) \cap \mathcal{S}$ and $B_2(\varepsilon) \cap \mathcal{S}$ are always nonempty, the preceding relations also hold if $B(\varepsilon)$ is intersected with $\mathcal{S}$, whence

$$\mathcal{L}_1(x,y) = \sup_{\varepsilon>0} \inf \{\mathcal{L}_2(x',y') : (x',y') \in B(\varepsilon) \cap \mathcal{S}\} = \mathcal{L}_{2,\wedge}(x,y).$$

One easily argues that $\mathcal{L}_1$ is lower semicontinuous, i.e. its lower semicontinuous hull is equal to $\mathcal{L}_1$ itself, which is also equal to the lower semicontinuous hull of $\mathcal{L}_2$.

Next observe that

$$\begin{aligned}\mathcal{L}_{1,\vee}(x,y) &= \inf_{\varepsilon>0} \sup \{\mathcal{L}_2(x',y') \wedge \mathcal{L}_3(x',y') : (x',y') \in B(\varepsilon)\} \\ &= \mathcal{L}_2(x,y) \vee \mathcal{L}_3(x,y) = \mathcal{L}_{2,\vee}(x,y).\end{aligned}$$

where the first equality follows from the definition of $\mathcal{L}_{1,\vee}$, the fact that $\mathcal{L}_1 = \mathcal{L}_{2,\wedge}$ as shown earlier and the representation for $\mathcal{L}_{2,\wedge}$ derived above while the second equality follows from (2.49) and (2.50).

The previous argument assumes $(x,y) \in (0,\infty)^2$. It remains to show that the semicontinuous hulls of $\mathcal{L}_1$ also correspond to those of $\mathcal{L}_2$ on the axes. For this, assume now that $x > 0, y = 0$. The ball $B(\varepsilon)$ around $(x,0)$ now becomes a "half-ball" (we intersect if with $[0,\infty)^2$). Let (x',y') be a point in that ball. Then (x',y') is on the same ray as (x,δ), for some $\delta \geq 0$ that can be made to converge to 0 as $\varepsilon \to 0$. We have $\dot{L}_2^{\pm}(x',y') = \dot{L}_2^{\pm}(x,\delta) \to \dot{L}_2^{+}(x,0)$ as $\varepsilon \to 0$. For the first derivative, the known bounds $x \vee y \leq L(x,y) \leq x + y$ imply that $x \leq L(x,\delta) \leq x + \delta$. The convexity and homogeneity properties then imply that

$$\dot{L}_1(x,0) = 1 \geq \dot{L}_1^{\pm}(x,\delta) \geq \frac{L(x,\delta) - L(0,\delta)}{x} \geq \frac{x-\delta}{x} \longrightarrow 1$$

as $\varepsilon \to 0$. By uniform boundedness of $\dot{L}_1^{\pm}, \dot{L}_2^{\pm}$ it follows easily that $\mathcal{L}_1$ and $\mathcal{L}_2$ are continuous at $(x, 0)$ and that $\mathcal{L}_1(x, 0) = \mathcal{L}_2(x, 0) = b^{(1)}(x)$, whence those two functions have the same semicontinuous hulls at that point.

Because $\dot{L}_1(0, y)$ was defined as the partial derivative from the right, one deals with a point $(0, y)$ in the same way.

Finally, note that since $b^{(1)}(0) = b^{(2)}(0) = 0$, and by uniform boundedness of $\dot{L}_1^{\pm}, \dot{L}_2^{\pm}$ the functions $\mathcal{L}_1$ and $\mathcal{L}_2$ are both continuous and take the value 0 at $(0, 0)$. Their semicontinuous hulls are therefore also equal at that point.

We have shown that everywhere on $[0, \infty)^2$, $\mathcal{L}_{1,\wedge} = \mathcal{L}_{2,\wedge}$ and $\mathcal{L}_{1,\vee} = \mathcal{L}_{2,\vee}$. By definition (see Bücher et al., 2014, Proposition 2.1), this means that $d_{\text{hypi}}(\mathcal{L}_1, \mathcal{L}_2) = 0$. □

Lemma 2.10. *Let $f : [0, T]^2 \to \mathbb{R}$ be continuous Lebesgue-almost everywhere, $g := (g_1, \ldots, g_q)^\top : [0, T]^2 \to \mathbb{R}^q$ be a vector of integrable functions and assume that f_n are measurable and hypi-converge to f on $[0, T]^2$. Then $\int g f_n d\mu_L \to \int g f d\mu_L$, where μ_L denotes the Lebesgue measure on $[0, T]^2$.*

Proof. For every $j \in \{1, \ldots, q\}$ and $M < \infty$, we have

$$
\begin{aligned}
&\int |g_j f_n - g_j f| d\mu_{\mathrm{L}} \\
&\quad = \int |g_j||f_n - f| \mathbb{1}\{|g_j| \le M\} \, d\mu_{\mathrm{L}} + \int |g_j||f_n - f| \mathbb{1}\{|g_j| > M\} \, d\mu_{\mathrm{L}} \\
&\quad \le M \int |f_n - f| d\mu_{\mathrm{L}} + \sup_{(x,y)\in[0,T]^2} |f_n(x, y) - f(x, y)| \int |g_j| \mathbb{1}\{|g_j| > M\} \, d\mu_{\mathrm{L}} \\
&\quad \le M \int |f_n - f| d\mu_{\mathrm{L}} \\
&\qquad + \left(\sup_{(x,y)\in[0,T]^2} |f_n(x, y)| + \sup_{(x,y)\in[0,T]^2} |f(x, y)| \right) \int |g_j| \mathbb{1}\{|g_j| > M\} \, d\mu_{\mathrm{L}}.
\end{aligned}
$$

The first term on the right hand side converges to 0 by Proposition 2.4 of Bücher et al. (2014) and since f is assumed continuous almost everywhere. By Proposition 2.3 of that paper, $\sup_{(x,y)\in[0,T]^2} |f_n(x, y)| \to \sup_{(x,y)\in[0,T]^2} |f(x, y)|$. Therefore, we have

$$
\lim_{n\to\infty} \int |g_j f_n - g_j f| d\mu_{\mathrm{L}} \le 2 \sup_{(x,y)\in[0,T]^2} |f(x, y)| \int |g_j| \mathbb{1}\{|g_j| > M\} \, d\mu_{\mathrm{L}},
$$

which can be made arbitrarily small by choosing M large enough, since g_j is integrable. The claim follows. □

Lemma 2.11. *Let $\phi : \mathbb{R}^p \to \mathbb{R}^q$, $p \le q$, have a unique, well separated zero at a point $x_0 \in \mathbb{R}^p$ and be continuously differentiable at x_0 with Jacobian matrix $J := J_\phi(x_0)$ of*

full rank p. Let Y_n be a random vector in $\mathbb{R}^q$ such that $a_n^{-1}Y_n$ weakly converges to a random vector Y, for some sequence $a_n \to 0$. Then if $X_n = \arg\min_x \|\phi(x) - Y_n\|$, we have

$$X_n - x_0 = (J^\top J)^{-1} J^\top Y_n + o_{\mathbb{P}}(a_n).$$

Proof. Let $h_n := a_n^{-1}(X_n - x_0 - (J^\top J)^{-1}J^\top Y_n)$. By definition of X_n, h_n is a minimizer of the random function $M_n : \mathbb{R}^p \to \mathbb{R}_+$ defined as

$$M_n(h) := a_n^{-1}\Big\|\phi\big(x_0 + (J^\top J)^{-1}J^\top Y_n + a_n h\big) - Y_n\Big\|.$$

By differentiability of ϕ, $M_n(h)$ is the norm of

$$\big(J(J^\top J)^{-1}J^\top - I\big)a_n^{-1}Y_n + Jh + o(1)$$

uniformly over bounded h, where I is the $q \times q$ identity matrix. The above display, seen as a function of h, weakly converges to

$$\big(J(J^\top J)^{-1}J^\top - I\big)Y + Jh$$

in $(\ell^\infty(\mathcal{K}))^q$, for any compact set $\mathcal{K}$. The mapping $f \mapsto \{h \mapsto \|f(h)\|\}$ being continuous from $(\ell^\infty(\mathcal{K}))^q$ onto $\ell^\infty(\mathcal{K})$, it follows that $M_n \rightsquigarrow M$ in $\ell^\infty(\mathcal{K})$, for

$$M(h) := \Big\|\big(J(J^\top J)^{-1}J^\top - I\big)Y + Jh\Big\|.$$

The function M^2 is strictly convex and has derivative $\partial(M^2(h))/\partial h = 2J^\top Jh$ which, since J has full rank, has a unique zero at $h = 0$. It follows that M^2, and thus M, has a unique minimizer at the point 0. Therefore, if we can show that the sequence $\{h_n\}$ is uniformly tight, Corollary 5.58 of van der Vaart (2000) will ensure that h_n converges in distribution (and hence in probability) to 0, which in turn implies the result.

It is known by Prohorov's theorem that $\{a_n^{-1}Y_n\}$ is uniformly tight. Therefore, it is sufficient to establish tightness of $\{a_n^{-1}(X_n - x_0)\}$. First, define for $\delta > 0$

$$\varepsilon(\delta) = \inf_{x \notin B(x_0,\delta)} \|\phi(x)\|,$$

where $B(x_0, \delta)$ denotes an open δ-ball around x_0. By assumption, $\varepsilon(\delta) > 0$ for every positive δ. Choose $\delta_0 > 0$ small enough so that for every $x \in B(x_0, \delta_0)$,

$$\|\phi(x) - J(x - x_0)\| < \frac{1}{2}\|J(x - x_0)\|,$$

which is possible by differentiability of ϕ (recall that J is the Jacobian at x_0). By

the reverse triangle inequality, this implies that $|\|\phi(x)\| - \|J(x - x_0)\|\|$ has the same upper bound. Then, for $\delta \leq \delta_0$,

$$\varepsilon(\delta) > \frac{1}{2} \inf_{x \in B(x_0,\delta)} \|J(x - x_0)\| = \frac{\sigma_1(J)}{2}\delta,$$

where $\sigma_1(J)$, the smallest singular value of J, is positive since J has full rank.

Now, fix an arbitrary $\eta > 0$. Because the sequence $\{a_n^{-1}Y_n\}$ is uniformly tight, there exists a finite $K = K(\eta)$ such that for $\delta_n := Ka_n$ and for n large enough so that $\delta_n \leq \delta_0$,

$$\mathbb{P}\left(\|Y_n\| \geq \frac{\varepsilon(\delta_n)}{2}\right) \leq \mathbb{P}\left(\|Y_n\| \geq \frac{K\sigma_1(J)}{4}a_n\right) \leq \eta$$

Hence with probability at least $1 - \eta$, $\|Y_n\| < \varepsilon(\delta_n)/2$. The last inequality implies two things. First, letting $\phi_n = \phi - Y_n$ and recalling that $\phi(x_0) = 0$, we have $\|\phi_n(x_0)\| = \|Y_n\| < \varepsilon(\delta_n)/2$. Second, for any $x \notin B(x_0, \delta_n)$, we have $\|\phi(x)\| \geq \varepsilon(\delta_n)$ so

$$\|\phi_n(x)\| = \|\phi(x) - Y_n\| \geq |\|\phi(x)\| - \|Y_n\|| > \frac{\varepsilon(\delta_n)}{2}.$$

That is, with probability at least $1 - \eta$, $X_n = \arg\min_x \|\phi_n(x)\| \in B(x_0, \delta_n)$. Since $\delta_n = O(a_n)$ and η was arbitrary, we conclude that $\{a_n^{-1}(X_n - x_0)\}$ is uniformly tight, and so is $\{h_n\}$. □

2.9 Proof of the claims in Examples 2.8, 2.11 and 2.12

2.9.1 Example 2.8

Recall that the random vector $Z := (1-X, 1-Y)$ is assumed max-stable with uniform margin and stable tail dependence function L, hence its distribution function is given by (2.3). Let $(x, y) \in (0, 1]^2$ (the result is trivial if x or y is zero). Note that we can without loss of generality focus on $(x, y) \in (0, 1]^2$ instead of general bounded sets since any bounded set can be rescaled to be contained in $[0, 1]^2$ at the cost of absorbing the scaling into t. The survival copula Q of (X, Y) satisfies

$$\begin{aligned}
Q(tx, ty) &:= \mathbb{P}\left(X \geq 1 - tx, Y \geq 1 - ty\right) \\
&= \mathbb{P}\left(1 - X \leq tx, 1 - Y \leq ty\right) \\
&= \exp\{-L(-\log(tx), -\log(ty))\} \\
&= \exp\left\{\log(t) L\left(1 + \frac{\log(x)}{\log(t)}, 1 + \frac{\log(y)}{\log(t)}\right)\right\},
\end{aligned}$$

where we have used the homogeneity property of L in the last line. By the assumed expansion of the function L,

$$L\left(1+\frac{\log(x)}{\log(t)}, 1+\frac{\log(y)}{\log(t)}\right) = L(1,1) + \dot{L}_1(1,1)\frac{\log(x)}{\log(t)} + \dot{L}_2(1,1)\frac{\log(y)}{\log(t)} + \delta(t,x,y),$$

where $\dot{L}_1$ and $\dot{L}_2$ are the right partial derivatives of L with respect to its first and second argument, respectively, and

$$\delta(t,x,y) \lesssim \left(\frac{\log(x)}{\log(t)}\right)^2 + \left(\frac{\log(y)}{\log(t)}\right)^2.$$

This is a linear approximation of the function L; since that function is convex, it lies above its sub gradient, so the error term $\delta(t,x,y)$ is non-negative. Plugging this in our expression for $Q(tx,ty)$ yields

$$Q(tx,ty) = t^{L(1,1)} x^{\dot{L}_1(1,1)} y^{\dot{L}_2(1,1)} e^{\delta'(t,x,y)},$$

where $\delta'(t,x,y) = \log(t)\delta(t,x,y)$ satisfies

$$\frac{\log(x)^2 + \log(y)^2}{\log(t)} \lesssim \delta'(t,x,y) \leq 0.$$

Letting $q(t) = t^{L(1,1)}$ and $c(x,y) = x^{\dot{L}_1(1,1)} y^{\dot{L}_2(1,1)}$, we obtain

$$\begin{aligned}\left|\frac{Q(tx,ty)}{q(t)} - c(x,y)\right| &= x^{\dot{L}_1(1,1)} y^{\dot{L}_2(1,1)} \left(1 - e^{\delta'(t,x,y)}\right) \\ &\leq x^{\dot{L}_1(1,1)} y^{\dot{L}_2(1,1)} |\delta'(t,x,y)| \\ &\lesssim \frac{x^{\dot{L}_1(1,1)} y^{\dot{L}_2(1,1)} (\log(x)^2 + \log(y)^2)}{\log(1/t)},\end{aligned}$$

where we used the fact that $0 \leq 1 - e^x \leq |x|$ for all $x \leq 0$. Since $\dot{L}_1(1,1)$ and $\dot{L}_2(1,1)$ are positive it follows that this upper bound is of order $1/\log(1/t)$ uniformly over x, y in bounded sets. The claim in Example 2.8 is proved. □

2.9.2 Example 2.11

Now, recall the setting of Example 2.11. The expression for $\Gamma^{(s,s)}$ is trivial. We shall treat the case where s and s' are two pairs that share an element, i.e. $s = (s_1, s_2)$ and $s' = (s_1, s_3)$. One similarly deals with different combinations of s, s', including the case where they are disjoint.

Let L be the stable tail dependence function of the max-stable, trivariate random

vector $(1-X^{(s_1)}, 1-X^{(s_2)}, 1-X^{(s_3)})$. By assumption and by the calculations above for the bivariate case, the pairs $(X^{(s_1)}, X^{(s_2)})$ and $(X^{(s_1)}, X^{(s_3)})$ satisfy Condition 2.1(i) with scaling functions $q^{(s)}(t) = t^{L(1,1,0)}$ and $q^{(s')}(t) = t^{L(1,0,1)}$, respectively. Since those functions are invertible, we may choose any diverging sequence $m = o(\log(n)^2)$ and invert them, setting $k^{(s)}/n = (m/n)^{1/L(1,1,0)}$ and $k^{(s')}/n = (m/n)^{1/L(1,0,1)}$. In fact, we may do so with every pair and obtain, as claimed, a universal sequence m.

Without loss of generality, let us assume that $L(1,1,0) \leq L(1,0,1)$ so that $k^{(s)} \leq k^{(s')}$. Let $t_n = k^{(s)}/n$ and $\alpha = L(1,1,0)/L(1,0,1) \in (0,1]$; observe that $k^{(s')}/n = t_n^{\alpha}$. By definition, for fixed $x^1, x^2 \in (0,1]^2$ (we can restrict our attention to this setting by similar arguments as in the bivariate case), we have

$$\Gamma^{(s,s')}(x^1, x^2) = \lim_{n\to\infty} \frac{n}{m} \mathbb{P}\left(1 - X^{(s_1)} \leq t_n x, 1 - X^{(s_2)} \leq t_n y, 1 - X^{(s_3)} \leq t_n^{\alpha} z\right), \quad (2.51)$$

where x is equal to $x_1^1 \wedge x_2^1$ if $\alpha = 1$ and to x_1^1 otherwise, $y = x_2^1$ and $z = x_2^2$. Using the same reasoning as in the bivariate case above (including the homogeneity property of L), the probability in (2.51) can be written as

$$\begin{aligned}
&\exp\left\{-L(-\log(t_n x), -\log(t_n y), -\log(t_n^{\alpha} z))\right\} \\
&\quad = \exp\left\{\log(t_n) L\left(1 + \frac{\log(x)}{\log(t_n)}, 1 + \frac{\log(y)}{\log(t_n)}, \alpha + \frac{\log(z)}{\log(t_n)}\right)\right\} \\
&\quad = t_n^{L(1,1,\alpha)} \exp\left\{\log(t_n)\left[L\left(1 + \frac{\log(x)}{\log(t_n)}, 1 + \frac{\log(y)}{\log(t_n)}, \alpha + \frac{\log(z)}{\log(t_n)}\right) - L(1,1,\alpha)\right]\right\}.
\end{aligned}$$

Eventually, $\log(t_n)$ is negative, which makes the difference in the square brackets non-negative by monotonicity of L. This eventually upper bounds the exponential by 1 and the entire expression by $t_n^{L(1,1,\alpha)}$, for any $x, y, z \in (0,1]$. Considering (2.51), it follows that for every fixed $x^1, x^2 \in (0,1]^2$,

$$\Gamma^{(s,s')}(x^1, x^2) \leq \lim_{n\to\infty} \frac{n}{m} t_n^{L(1,1,\alpha)} = \lim_{n\to\infty} \left(\frac{m}{n}\right)^{\frac{L(1,1,\alpha)}{L(1,1,0)} - 1} = 0,$$

since the assumption that L is component-wise strictly increasing means that $L(1,1,\alpha) > L(1,1,0)$. □

2.9.3 Example 2.12

We present here the main ideas, as most of the precise calculations are similar to the preceding section. As before, let $X^{(j)} = Y(u_j)$, and write $Z^{(j)}$ and $Z'^{(j)}$ for $Z(u_j)$ and $Z'(u_j)$. Consider a pair $s := (s_1, s_2)$ and let F be the distribution function of the unit

Fréchet distribution. Recall that $X^{(j)} = \max\{aZ^{(j)}, (1-a)Z'^{(j)}\}$. We have for $t \downarrow 0$

$$\begin{aligned}
&\mathbb{P}\left(F(X^{(s_1)}) \geq 1-tx, F(X^{(s_2)}) \geq 1-ty\right) \\
&= \mathbb{P}\left(F(Z^{(s_1)})^{1/a} \vee F(Z'^{(s_1)})^{1/(1-a)} \geq 1-tx, F(Z^{(s_2)})^{1/a} \vee F(Z'^{(s_2)})^{1/(1-a)} \geq 1-ty\right) \\
&= \mathbb{P}\left(F(Z^{(s_1)}) \geq (1-tx)^a, F(Z^{(s_2)}) \geq (1-ty)^a\right) \\
&\quad + \mathbb{P}\left(F(Z'^{(s_1)}) \geq (1-tx)^{1-a}, F(Z'^{(s_2)}) \geq (1-ty)^{1-a}\right) + O(t^2), \qquad (2.52)
\end{aligned}$$

where the term $O(t^2)$ is uniform over bounded x, y. Note that $(1-tx)^a = 1 - t(ax + O(tx^2))$. The first term of (2.52) is equal to

$$a\chi^{Z,(s)}t(x + y - L^{Z,(s)}(x,y)) + O(t^2)$$

uniformly over bounded x, y, where $\chi^{Z,(s)}$ and $L^{Z,(s)}$ are the extremal dependence coefficient and stable tail dependence function, respectively, corresponding to the random vector $(Z^{(s_1)}, Z^{(s_2)})$. From previous calculations, the second term of (2.52) is equal to

$$((1-a)t)^{L^{Z',(s)}(1,1)} x^{\dot{L}_1^{Z',(s)}(1,1)} y^{\dot{L}_2^{Z',(s)}(1,1)} + O\Big(t^{L^{Z',(s)}(1,1)}/\log(1/t)\Big),$$

uniformly over bounded x, y, where $L^{Z',(s)}$ is the stable tail dependence function corresponding to the max-stable random vector $(1/Z'^{(s_1)}, 1/Z'^{(s_2)})$. It follows that Condition 2.1(i) is satisfied for every pair of locations; depending on whether $(Z^{(s_1)}, Z^{(s_2)})$ is dependent or independent, either the first of the second of the last two expressions dominates. This determines that $q^{(s)}(t)$ is proportional to t for asymptotically dependent pairs and to $t^{1/\eta'^{(s)}}$ for asymptotically independent ones, where $\eta'^{(s)}$ is the coefficient of tail dependence of $(1/Z'^{(s_1)}, 1/Z'^{(s_2)})$, satisfying $1 < 1/\eta'^{(s)} < 2$ by assumption — for any inverted max-stable distribution, its coefficient of tail dependence η is always in $[1/2, 1)$, and can only be equal to $1/2$ under perfect independence. The coefficient of tail dependence $\eta^{(s)}$ of $(X^{(s_1)}, X^{(s_2)})$ is equal to 1 if $\chi^{Z,(s)} > 0$ and to $\eta'^{(s)}$ otherwise.

We now show how to obtain an expression for the functions $\Gamma^{(s,s')}$. First, since the functions $q^{(s)}$ are proportional to simple powers, for a sufficiently slow intermediate sequence m, we let $k^{(s)}/n$ be proportional to m/n if s is an asymptotically dependent pair and to $(m/n)^{\eta^{(s)}}$ otherwise, so that all $m^{(s)}$ are equal to m.

The case $s = s'$ follows trivially from the previous developments; $\Gamma^{(s,s)}$ can be derived from $c^{(s)}$. Next consider the case where s, s' share one element, i.e. $s = (s_1, s_2)$ and $s' = (s_1, s_3)$. Letting $t_n = k^{(s)}/n$ and $t'_n = k^{(s')}/n$, assume without loss of

generality that $t'_n \lesssim t_n$. The probability of interest is of the form

$$\begin{aligned}
&\mathbb{P}\left(F(X^{(s_1)}) \geq 1-(t_n x \wedge t'_n x'), F(X^{(s_2)}) \geq 1-t_n y, F(X^{(s_3)}) \geq 1-t'_n z\right) \\
&= \mathbb{P}\left(F(Z^{(s_1)}) \geq (1-(t_n x \wedge t'_n x'))^a, F(Z^{(s_2)}) \geq (1-t_n y)^a, F(Z^{(s_3)}) \geq (1-t'_n z)^a\right) \\
&\quad + \mathbb{P}\left(F(Z'^{(s_1)}) \geq (1-(t_n x \wedge t'_n x'))^{1-a}, F(Z'^{(s_2)}) \geq (1-t_n y)^{1-a},\right. \\
&\qquad \left. F(Z'^{(s_3)}) \geq (1-t'_n z)^{1-a}\right) + O(t_n^2).
\end{aligned}$$

Indeed, the third term above is the probability of a certain event that requires at least one of the Z and one of the Z' to be large, which has probability at most $O(t_n^2)$ since Z and Z' are assumed independent (recall that we assumed $t'_n = O(t_n)$). We note that the term in front of this probability in the definition of $\Gamma^{(s,s')}$ is equal to $q^{(s)}(t_n)^{-1} = t_n^{-1/\eta^{(s)}}$. However $t_n^2 = o(t_n^{1/\eta^{(s)}})$ since $\eta^{(s)} > 1/2$, and the second probability above is also $o(t_n^{1/\eta^{(s)}})$, following the calculations for Example 2.11. Therefore, in this case, $\Gamma^{(s,s')}((x,y),(x',z))$ is equal to the limit

$$\begin{aligned}
&\lim_{n\to\infty} t_n^{-1/\eta^{(s)}} \mathbb{P}\left(F(Z^{(s_1)}) \geq (1-(t_n x \wedge t'_n x'))^a,\right. \\
&\qquad\qquad \left. F(Z^{(s_2)}) \geq (1-t_n y)^a, F(Z^{(s_3)}) \geq (1-t'_n z)^a\right) \\
&= \lim_{n\to\infty} t_n^{-1/\eta^{(s)}} \mathbb{P}\left(F(Z^{(s_1)}) \geq 1-a(t_n x \wedge t'_n x'),\right. \\
&\qquad\qquad \left. F(Z^{(s_2)}) \geq 1-a t_n y, F(Z^{(s_3)}) \geq 1-a t'_n z\right)
\end{aligned}$$

which is non-zero if and only if $(Z^{(s_1)}, Z^{(s_2)}, Z^{(s_3)})$ is fully dependent (i.e., it contains no pairwise independence).

For the case where the pairs $s = (s_1, s_2)$ and $s' = (s_3, s_4)$ are disjoint, let $t_n = k^{(s)}/n$ and $t'_n = k^{(s')}/n$ and assume as before that $t'_n \lesssim t_n$. By similar arguments as above, one obtains that $\Gamma^{(s,s')}((x,y),(x',y'))$ is equal to the limit

$$\begin{aligned}
&\lim_{n\to\infty} t_n^{-1/\eta^{(s)}} \mathbb{P}\left(F(Z^{(s_1)}) \geq 1-a t_n x, F(Z^{(s_2)}) \geq 1-a t_n y,\right. \\
&\qquad\qquad \left. F(Z^{(s_3)}) \geq 1-a t'_n x', F(Z^{(s_4)}) \geq 1-a t'_n y'\right),
\end{aligned}$$

which is non-zero if and only if $(Z^{(s_1)}, Z^{(s_2)}, Z^{(s_3)}, Z^{(s_4)})$ has no independent pairs.

Using the same ideas and after straightforward computations, one may calculate the limits $\Gamma^{(s,s',j)}$, for $s' \in \mathcal{P}_D$. First, consider the case where $s = (s_1, s_2)$ and $s'_j = s_1$, that is the element s'_j is in the pair s. Defining t_n and t'_n as above, we still have $t'_n \lesssim t_n$ since s' is an asymptotically dependent pair. Then $\Gamma^{(s,s',j)}((x,y),(x',y'))$ is equal to

$$\chi^{Z,(s')} \lim_{n\to\infty} t_n^{-1/\eta^{(s)}} \mathbb{P}\left(F(Z^{(s_1)}) \geq 1-a(t_n x \wedge t'_n x'), F(Z^{(s_2)}) \geq 1-a t_n y\right),$$

which is non-zero if and only if $(Z^{(s_1)}, Z^{(s_2)})$ is dependent. Now if $s_3 := s'_j$ is not an element of s, $\Gamma^{(s,s',j)}((x,y),(x',y'))$ becomes

$$\chi^{Z,(s')} \lim_{n\to\infty} t_n^{-1/\eta^{(s)}} \mathbb{P}\left(F(Z^{(s_1)}) \geq 1 - at_n x, F(Z^{(s_2)}) \geq 1 - at_n y, F(Z^{(s_3)}) \geq 1 - at'_n x'\right),$$

which is non-zero if and only if $(Z^{(s_1)}, Z^{(s_2)}, Z^{(s_3)})$ is fully dependent.

Finally, for $s, s' \in \mathcal{P}_D$, again letting $t_n = k^{(s)}/n$ and $t'_n = k^{(s')}/n$, note that this time t'_n/t_n is constant. Without loss of generality, let $j = j' = 1$. Then $\Gamma^{(s,j,s',j')}((x,y),(x',y'))$ is equal to

$$\chi^{Z,(s)}\chi^{Z,(s')} \lim_{n\to\infty} t_n^{-1} \mathbb{P}\left(F(Z^{(s_1)}) \geq 1 - t_n x, F(Z^{(s'_1)}) \geq 1 - t'_n y'\right),$$

which is non-zero if and only if $(Z^{(s_1)}, Z^{(s'_1)})$ is dependent. □

2.10 Proof of the claims in Example 2.9

The multiplicative constant appearing in the scaling function q, as a function of λ, is given by

$$K_\lambda = \begin{cases} 2\frac{1-\lambda}{2-\lambda}, & \lambda \in (0,1) \\ 2, & \lambda = 1 \\ \left(1-\frac{1}{\lambda}\right)^{\lambda-1} \frac{2(\lambda-1)}{\lambda(2-\lambda)}, & \lambda \in (1,2) \\ \frac{1}{2}, & \lambda = 2 \\ \frac{\left(1-\frac{1}{\lambda}\right)^2}{1-\frac{2}{\lambda}}, & \lambda \in (2,\infty) \end{cases}; \tag{2.53}$$

it can be deduced from the proof.

The argument must be separated in two cases depending on whether $\lambda = 1$.

2.10.1 The case $\lambda \neq 1$

For now, assume that $\alpha_R \neq \alpha_W$. Let $\bar{F}_R$ denote the survival function of R. Then $\bar{F}_R(x) = x^{-\alpha_R}$ for $x > 1$, and $\bar{F}_R(x) = 1$ for $x \leq 1$. The first step in calculationg Q is to find an expression for the survival function $\bar{F}$ of X (and equivalently of Y) and its inverse. We have, for $x \geq 1$,

$$\begin{aligned} \bar{F}(x) &= \mathbb{P}\left(RW_1 > x\right) \\ &= \mathbb{P}\left(R > \frac{x}{W_1}\right) \\ &= \mathbb{E}\left[\bar{F}_R\left(\frac{x}{W_1}\right)\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(W_1 > x\right) + \int_1^x \left(\frac{w}{x}\right)^{\alpha_R} \frac{\alpha_W}{w^{\alpha_W+1}} dw \\
&= x^{-\alpha_W} + \alpha_W x^{-\alpha_R} \frac{x^{\alpha_R-\alpha_W} - 1}{\alpha_R - \alpha_W} \\
&= \frac{\alpha_R}{\alpha_R - \alpha_W} x^{-\alpha_W} - \frac{\alpha_W}{\alpha_R - \alpha_W} x^{-\alpha_R} \\
&= \frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge} x^{-\alpha_\wedge} \left(1 - \frac{\alpha_\wedge}{\alpha_\vee - \alpha_\wedge} x^{\alpha_\wedge - \alpha_\vee}\right),
\end{aligned}
$$

where $\alpha_\wedge$ and $\alpha_\vee$ denote the smallest and the largest of the two α's, respectively. Although not easily invertible, this function is close to $\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge} x^{-\alpha_\wedge}$, which has an analytical inverse. We now argue that this inverse is close to that of $\bar{F}$. First, for any $X \in (1, \infty)$, we have for $x \in [X, \infty)$

$$
\underbrace{\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge} x^{-\alpha_\wedge} \left(1 - \frac{\alpha_\wedge}{\alpha_\vee - \alpha_\wedge} X^{\alpha_\wedge - \alpha_\vee}\right)}_{f_1(x)} \leq \bar{F}(x) \leq \underbrace{\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge} x^{-\alpha_\wedge}}_{f_2(x)}.
$$

Now note that for two decreasing, invertible functions g_1 and g_2, $g_1 \leq g_2$ is equivalent to $g_1^{-1} \leq g_2^{-1}$. This means that as soon as $y \leq f_1(X)$, $f_1^{-1}(y) \leq \bar{F}^{-1}(y) \leq f_2^{-1}(y)$. In other words, for such y,

$$
\begin{aligned}
\left(1 - \frac{\alpha_\wedge}{\alpha_\vee - \alpha_\wedge} X^{\alpha_\wedge - \alpha_\vee}\right)^{1/\alpha_\wedge} \left(\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge}\right)^{1/\alpha_\wedge} y^{-1/\alpha_\wedge} &\leq \bar{F}^{-1}(y) \\
&\leq \left(\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge}\right)^{1/\alpha_\wedge} y^{-1/\alpha_\wedge}.
\end{aligned}
$$

Because these inequalities are true as soon as $y \leq f_1(X)$, they are true if $y = f_1(X)$. If y is small enough, choosing $X = \left(\frac{1}{2}\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge}\right)^{1/\alpha_\wedge} y^{-1/\alpha_\wedge}$ is sufficient to have $y \leq f_1(X)$. Therefore, if y is small enough, the first inequality in the last display becomes

$$
\bar{F}^{-1}(y) \geq \left(1 - O\left(y^{\frac{\alpha_\vee}{\alpha_\wedge} - 1}\right)\right) \left(\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge}\right)^{1/\alpha_\wedge} y^{-1/\alpha_\wedge}.
$$

Combining this with the upper bound (the second inequality) yields

$$
\bar{F}^{-1}(y) = \left(1 + O\left(y^\tau\right)\right) \left(\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge}\right)^{1/\alpha_\wedge} y^{-1/\alpha_\wedge}, \tag{2.54}
$$

where $\tau = \frac{\alpha_\vee}{\alpha_\wedge} - 1$.

The copula Q can now be expressed as

$$
Q(tx, ty) = \mathbb{P}\left(X \geq \bar{F}^{-1}(tx), Y \geq \bar{F}^{-1}(ty)\right)
$$

$$= \mathbb{P}\left(RW_1 \geq \bar{F}^{-1}(tx), RW_2 \geq \bar{F}^{-1}(ty)\right) = \mathbb{P}\left(R \geq Z\right) = \mathbb{E}\left[\bar{F}_R(Z)\right],$$

where

$$Z := Z(tx, ty) = \frac{\bar{F}^{-1}(tx)}{W_1} \vee \frac{\bar{F}^{-1}(ty)}{W_2}.$$

Recalling the definition of $\bar{F}_R$, we have

$$\begin{aligned}
Q(tx, ty) &= \mathbb{P}\left(Z \leq 1\right) + \mathbb{E}\left[Z^{-\alpha_R}; Z > 1\right] \\
&= \mathbb{P}\left(Z \leq 1\right) + \int_0^\infty \mathbb{P}\left(Z^{-\alpha_R} > a, Z > 1\right) da \\
&= \mathbb{P}\left(Z \leq 1\right) + \int_0^\infty \mathbb{P}\left(1 < Z \leq a^{-1/\alpha_R}\right) da \\
&= \mathbb{P}\left(Z \leq 1\right) + \int_0^1 \mathbb{P}\left(1 < Z \leq a^{-1/\alpha_R}\right) da \\
&= \mathbb{P}\left(Z \leq 1\right) + \int_0^1 \left(\mathbb{P}\left(Z \leq a^{-1/\alpha_R}\right) - \mathbb{P}\left(Z \leq 1\right)\right) da \\
&= \int_0^1 \mathbb{P}\left(Z \leq a^{-1/\alpha_R}\right) da.
\end{aligned}$$

In order to compute the previous integral, we need to derive the CDF of Z. From the definition of Z and by independence of W_1 and W_2, it is clear that, for any $z > 0$,

$$\mathbb{P}\left(Z \leq z\right) = \mathbb{P}\left(W_1 \geq \frac{\bar{F}^{-1}(tx)}{z}\right) \mathbb{P}\left(W_2 \geq \frac{\bar{F}^{-1}(ty)}{z}\right).$$

From now on, assume without loss of generality that $x \geq y$ since $c(x, y) = c(y, x)$ (because the random variables X and Y are exchangeable). Then $\bar{F}^{-1}(tx) \leq \bar{F}^{-1}(ty)$. The previous probability can take 3 different forms:

$$\mathbb{P}\left(Z \leq z\right) = \begin{cases} \left(\bar{F}^{-1}(tx)\bar{F}^{-1}(ty)\right)^{-\alpha_W} z^{2\alpha_W}, & \text{if } z \leq \bar{F}^{-1}(tx) \\ \left(\bar{F}^{-1}(ty)\right)^{-\alpha_W} z^{\alpha_W}, & \text{if } \bar{F}^{-1}(tx) < z \leq \bar{F}^{-1}(ty) \\ 1, & \text{if } z > \bar{F}^{-1}(ty) \end{cases}.$$

When substituting $z = a^{-1/\alpha_R}$, for $a \in (0, 1)$, notice that we are in the three preceding cases, respectively, when

$$\begin{cases} a \geq \left(\bar{F}^{-1}(tx)\right)^{-\alpha_R} \\ \left(\bar{F}^{-1}(ty)\right)^{-\alpha_R} \leq a < \left(\bar{F}^{-1}(tx)\right)^{-\alpha_R} \\ a < \left(\bar{F}^{-1}(ty)\right)^{-\alpha_R} \end{cases}.$$

This allows us to write

$$Q(tx,ty) = \int_0^{\left(\bar F^{-1}(ty)\right)^{-\alpha_R}} da + \left(\bar F^{-1}(ty)\right)^{-\alpha_W} \int_{\left(\bar F^{-1}(ty)\right)^{-\alpha_R}}^{\left(\bar F^{-1}(tx)\right)^{-\alpha_R}} a^{-\frac{\alpha_W}{\alpha_R}} da + \left(\bar F^{-1}(tx)\bar F^{-1}(ty)\right)^{-\alpha_W} \int_{\left(\bar F^{-1}(tx)\right)^{-\alpha_R}}^{1} a^{-2\frac{\alpha_W}{\alpha_R}} da. \quad (2.55)$$

Since we only need (2.9) to hold uniformly over a sphere, we may assume that $y \le x \le 1$. Then, (2.54) yields

$$\bar F^{-1}(tx) = (1+O(t^\tau)) \left(\frac{\alpha_\vee}{\alpha_\vee - \alpha_\wedge}\right)^{1/\alpha_\wedge} (tx)^{-1/\alpha_\wedge}$$

and the same for $\bar F^{-1}(ty)$. Moreover, the term $O(t^\tau)$ is uniform over all $(x,y) \in [0,1]^2$. The first term in (2.55) is then equal to

$$\left(\bar F^{-1}(ty)\right)^{-\alpha_R} = (1+O(t^\tau)) \left(1 - \frac{\alpha_\wedge}{\alpha_\vee}\right)^{\frac{\alpha_R}{\alpha_\wedge}} t^{\frac{\alpha_R}{\alpha_\wedge}} y^{\frac{\alpha_R}{\alpha_\wedge}} =: Q^{(1)}(tx,ty),$$

the second one is equal to

$$\begin{aligned}
&\left(\bar F^{-1}(ty)\right)^{-\alpha_W} \left.\frac{a^{1-\frac{\alpha_W}{\alpha_R}}}{1-\frac{\alpha_W}{\alpha_R}}\right|_{a=\left(\bar F^{-1}(ty)\right)^{-\alpha_R}}^{\left(\bar F^{-1}(tx)\right)^{-\alpha_R}} \\
&= \frac{1}{1-\frac{\alpha_W}{\alpha_R}} \left(\bar F^{-1}(ty)\right)^{-\alpha_W} \left(\bar F^{-1}(tx)^{-(\alpha_R-\alpha_W)} - \bar F^{-1}(ty)^{-(\alpha_R-\alpha_W)}\right) \\
&= (1+O(t^\tau)) \frac{\left(1-\frac{\alpha_\wedge}{\alpha_\vee}\right)^{\frac{\alpha_R}{\alpha_\wedge}}}{1-\frac{\alpha_W}{\alpha_R}} t^{\frac{\alpha_R}{\alpha_\wedge}} y^{\frac{\alpha_W}{\alpha_\wedge}} \left(x^{\frac{\alpha_R-\alpha_W}{\alpha_\wedge}} - y^{\frac{\alpha_R-\alpha_W}{\alpha_\wedge}}\right) \\
&=: Q^{(2)}(tx,ty)
\end{aligned}$$

and finally the third one is equal to

$$\begin{aligned}
&\left(\bar F^{-1}(tx)\bar F^{-1}(ty)\right)^{-\alpha_W} \left.\frac{a^{1-2\frac{\alpha_W}{\alpha_R}}}{1-2\frac{\alpha_W}{\alpha_R}}\right|_{a=\left(\bar F^{-1}(tx)\right)^{-\alpha_R}}^{1} \\
&= \frac{1}{1-2\frac{\alpha_W}{\alpha_R}} \left(\bar F^{-1}(tx)\bar F^{-1}(ty)\right)^{-\alpha_W} \left(1 - \left(\bar F^{-1}(tx)\right)^{2\alpha_W-\alpha_R}\right) \\
&= (1+O(t^\tau)) \frac{\left(1-\frac{\alpha_\wedge}{\alpha_\vee}\right)^{2\frac{\alpha_W}{\alpha_\wedge}}}{1-2\frac{\alpha_W}{\alpha_R}} t^{2\frac{\alpha_W}{\alpha_\wedge}} (xy)^{\frac{\alpha_W}{\alpha_\wedge}} \left(1 - \left(1-\frac{\alpha_\wedge}{\alpha_\vee}\right)^{\frac{\alpha_R-2\alpha_W}{\alpha_\wedge}} (tx)^{\frac{\alpha_R-2\alpha_W}{\alpha_\wedge}}\right) \\
&=: Q^{(3a)}(tx,ty)
\end{aligned}$$

in the case where $\alpha_R \neq 2\alpha_W$, and if $\alpha_R = 2\alpha_W$, it is equal to

$$\begin{aligned}
&- \left(\bar{F}^{-1}(tx)\bar{F}^{-1}(ty)\right)^{-\alpha_W} \log\left(\left(\bar{F}^{-1}(tx)\right)^{-\alpha_R}\right) \\
&= (1+O(t^\tau))\left(1-\frac{\alpha_\wedge}{\alpha_\vee}\right)^{2\frac{\alpha_W}{\alpha_\wedge}} t^{2\frac{\alpha_W}{\alpha_\wedge}} (xy)^{\frac{\alpha_W}{\alpha_\wedge}} \\
&\quad \times \left(-\log\left((1+O(t^\tau))\left(1-\frac{\alpha_\wedge}{\alpha_\vee}\right)^{\frac{\alpha_R}{\alpha_\wedge}}\right) + \frac{\alpha_R}{\alpha_\wedge}\left(\log(1/x)+\log(1/t)\right)\right) \\
&= \frac{1}{2}t^2\log(1/t)xy + O(t^2) \\
&=: Q^{(3b)}(tx,ty),
\end{aligned}$$

where the term $O(t^2)$ is uniform over $(x,y) \in [0,1]^2$. We now divide the possible values of $\lambda = \alpha_R/\alpha_W$ in four ranges and determine which of the three terms $Q^{(1)}$, $Q^{(2)}$ or $Q^{(3)}$ dominates.

If $\lambda \in (0,1)$

This is the case where we obtain asymptotic dependence. All three terms are of the order of t, so they all matter. In this case, $\alpha_\wedge = \alpha_R$, $\alpha_\vee = \alpha_W$ and $\tau = 1/\lambda - 1$. Therefore,

$$\begin{aligned}
Q^{(1)}(tx,ty) &= (1+O(t^\tau))\left(1-\frac{\alpha_R}{\alpha_W}\right)ty = (1-\lambda)ty + O\left(t^{1+\tau}\right), \\
Q^{(2)}(tx,ty) &= (1+O(t^\tau))\frac{1-\frac{\alpha_R}{\alpha_W}}{1-\frac{\alpha_W}{\alpha_R}} ty^{\frac{\alpha_W}{\alpha_R}}\left(x^{1-\frac{\alpha_W}{\alpha_R}} - y^{1-\frac{\alpha_W}{\alpha_R}}\right) \\
&= (1+O(t^\tau))\frac{\alpha_R}{\alpha_W} t\left(y - x^{1-\frac{\alpha_W}{\alpha_R}} y^{\frac{\alpha_W}{\alpha_R}}\right) \\
&= \lambda t\left(y - x^{1-1/\lambda}y^{1/\lambda}\right) + O\left(t^{1+\tau}\right), \\
Q^{(3a)}(tx,ty) &= (1+O(t^\tau))\frac{\left(1-\frac{\alpha_R}{\alpha_W}\right)^{2\frac{\alpha_W}{\alpha_R}}}{2\frac{\alpha_W}{\alpha_R}-1} t^{2\frac{\alpha_W}{\alpha_R}}(xy)^{\frac{\alpha_W}{\alpha_R}} \\
&\quad \times \left(\left(1-\frac{\alpha_R}{\alpha_W}\right)^{1-2\frac{\alpha_W}{\alpha_R}} (tx)^{1-2\frac{\alpha_W}{\alpha_R}} - 1\right) \\
&= (1+O(t^\tau))\frac{1-\frac{\alpha_R}{\alpha_W}}{2\frac{\alpha_W}{\alpha_R}-1} tx^{1-\frac{\alpha_W}{\alpha_R}} y^{\frac{\alpha_W}{\alpha_R}} + O\left(t^{2\frac{\alpha_W}{\alpha_R}}\right) \\
&= \lambda\frac{1-\lambda}{2-\lambda} tx^{1-1/\lambda}y^{1/\lambda} + O\left(t^{1+\tau}+t^{2/\lambda}\right) = \lambda\frac{1-\lambda}{2-\lambda} tx^{1-1/\lambda}y^{1/\lambda} + O\left(t^{1+\tau}\right),
\end{aligned}$$

where in the last line we have used $1+\tau = \alpha_\vee/\alpha_\wedge = 1/\lambda < 2/\lambda$. Therefore in this case we get

$$\begin{aligned}
Q(tx,ty) &= Q^{(1)}(tx,ty) + Q^{(2)}(tx,ty) + Q^{(3a)}(tx,ty) \\
&= (1-\lambda)ty + \lambda t\left(y - x^{1-1/\lambda}y^{1/\lambda}\right) + \lambda\frac{1-\lambda}{2-\lambda}tx^{1-1/\lambda}y^{1/\lambda} + O\left(t^{1+\tau}\right) \\
&= t\left(y + \left(-\lambda + \lambda\frac{1-\lambda}{2-\lambda}\right)x^{1-1/\lambda}y^{1/\lambda}\right) + O\left(t^{1+\tau}\right) \\
&= t\left(y - \frac{\lambda}{2-\lambda}x^{1-1/\lambda}y^{1/\lambda}\right) + O\left(t^{1+\tau}\right).
\end{aligned}$$

If $\lambda \in (1,2)$

Here again, all three terms are of the order of t^λ so they all matter. Note that here and in the next two cases, $\alpha_\wedge = \alpha_W$, $\alpha_\vee = \alpha_R$ and $\tau = \lambda - 1$. Through similar calculations as before, we obtain this time

$$\begin{aligned}
Q^{(1)}(tx,ty) &= (1+O(t^\tau))\left(1-\frac{\alpha_W}{\alpha_R}\right)^{\frac{\alpha_R}{\alpha_W}} t^{\frac{\alpha_R}{\alpha_W}} y^{\frac{\alpha_R}{\alpha_W}} = \left(1-\frac{1}{\lambda}\right)^\lambda t^\lambda y^\lambda + O\left(t^{\lambda+\tau}\right), \\
Q^{(2)}(tx,ty) &= (1+O(t^\tau))\frac{\left(1-\frac{\alpha_W}{\alpha_R}\right)^{\frac{\alpha_R}{\alpha_W}}}{1-\frac{\alpha_W}{\alpha_R}} t^{\frac{\alpha_R}{\alpha_W}} y\left(x^{\frac{\alpha_R}{\alpha_W}-1} - y^{\frac{\alpha_R}{\alpha_W}-1}\right) \\
&= \left(1-\frac{1}{\lambda}\right)^{\lambda-1} t^\lambda\left(x^{\lambda-1}y - y^\lambda\right) + O\left(t^{\lambda+\tau}\right), \\
Q^{(3a)}(tx,ty) &= (1+O(t^\tau))\frac{\left(1-\frac{\alpha_W}{\alpha_R}\right)^2}{2\frac{\alpha_W}{\alpha_R}-1} t^2xy\left(\left(1-\frac{\alpha_W}{\alpha_R}\right)^{\frac{\alpha_R}{\alpha_W}-2}(tx)^{\frac{\alpha_R}{\alpha_W}-2} - 1\right) \\
&= (1+O(t^\tau))\frac{\left(1-\frac{1}{\lambda}\right)^2}{\frac{2}{\lambda}-1} t^2xy\left(\left(1-\frac{1}{\lambda}\right)^{\lambda-2}(tx)^{\lambda-2} - 1\right) \\
&= (1+O(t^\tau))\frac{\left(1-\frac{1}{\lambda}\right)^\lambda}{\frac{2}{\lambda}-1} t^\lambda x^{\lambda-1}y + O\left(t^2\right) \\
&= \lambda\frac{\left(1-\frac{1}{\lambda}\right)^\lambda}{2-\lambda} t^\lambda x^{\lambda-1}y + O\left(t^{\lambda+\tau} + t^2\right) \\
&= \lambda\frac{\left(1-\frac{1}{\lambda}\right)^\lambda}{2-\lambda} t^\lambda x^{\lambda-1}y + O\left(t^{(2\lambda-1)\wedge 2}\right).
\end{aligned}$$

Therefore, Q can be calculated as

$$Q(tx,ty) = Q^{(1)}(tx,ty) + Q^{(2)}(tx,ty) + Q^{(3a)}(tx,ty)$$

$$
\begin{aligned}
&= \left(1-\frac{1}{\lambda}\right)^{\lambda-1} t^{\lambda}\left(\left(1-\frac{1}{\lambda}\right)y^{\lambda} + x^{\lambda-1}y - y^{\lambda} + \lambda\frac{1-\frac{1}{\lambda}}{2-\lambda}x^{\lambda-1}y\right) \\
&\quad + O\left(t^{(2\lambda-1)\wedge 2}\right) \\
&= \left(1-\frac{1}{\lambda}\right)^{\lambda-1} t^{\lambda}\left(-\frac{1}{\lambda}y^{\lambda} + \left(1+\lambda\frac{1-\frac{1}{\lambda}}{2-\lambda}\right)x^{\lambda-1}y\right) + O\left(t^{(2\lambda-1)\wedge 2}\right) \\
&= \left(1-\frac{1}{\lambda}\right)^{\lambda-1} t^{\lambda}\left(\frac{1}{2-\lambda}x^{\lambda-1}y - \frac{1}{\lambda}y^{\lambda}\right) + O\left(t^{(2\lambda-1)\wedge 2}\right).
\end{aligned}
$$

If $\lambda = 2$

In this case, $\alpha_R/\alpha_\wedge = 2$, so we easily see that both $Q^{(1)}(tx,ty)$ and $Q^{(2)}(tx,ty)$ are $O(t^2)$. Because the term $Q^{(3b)}$ is of the order of $t^2\log(1/t)$, it dominates the preceding two by a factor of $\log(1/t)$. Therefore,

$$
Q(tx,ty) = Q^{(3b)}(tx,ty) + O\left(t^2\right) = \frac{1}{2}t^2\log(1/t)xy + O\left(t^2\right).
$$

If $\lambda \in (2,\infty)$

Once again, the terms $Q^{(1)}$ and $Q^{(2)}$ are dominated by the third term; they are both of the order of t^{λ}, whereas the third term is of the order of t^2. Therefore,

$$
\begin{aligned}
Q(tx,ty) &= Q^{(3a)}(tx,ty) + O\left(t^{\frac{\alpha_R}{\alpha_W}}\right) \\
&= (1+O(t^{\tau}))\frac{\left(1-\frac{\alpha_W}{\alpha_R}\right)^2}{1-2\frac{\alpha_W}{\alpha_R}}t^2xy\left(1-\left(1-\frac{\alpha_W}{\alpha_R}\right)^{\frac{\alpha_R}{\alpha_W}-2}(tx)^{\frac{\alpha_R}{\alpha_W}-2}\right) \\
&\quad + O\left(t^{\frac{\alpha_R}{\alpha_W}}\right) \\
&= (1+O(t^{\tau}))\frac{\left(1-\frac{1}{\lambda}\right)^2}{1-\frac{2}{\lambda}}t^2xy + O\left(t^{\lambda}\right) \\
&= \frac{\left(1-\frac{1}{\lambda}\right)^2}{1-\frac{2}{\lambda}}t^2xy + O\left(t^{(2+\tau)\wedge\lambda}\right) \\
&= \frac{\left(1-\frac{1}{\lambda}\right)^2}{1-\frac{2}{\lambda}}t^2xy + O\left(t^{\lambda}\right),
\end{aligned}
$$

because, in the last line, $2+\tau = \lambda+1 > \lambda$.

2.10.2 The case $\lambda = 1$

From now on, we assume that $\alpha_R = \alpha_W = \alpha$. That is, R, W_1, W_2 are iid with a Pareto(α) distribution. Like before, we denote by $\bar{F}_R$ and $\bar{F}$ the survival functions of

R and of X (and equivalently Y), respectively. As before, we first find an expression for $\bar{F}$. For any $x \geq 1$,

$$\begin{aligned}
\bar{F}(x) &= \mathbb{P}\left(RW_1 > x\right) \\
&= \mathbb{P}\left(R > \frac{x}{W_1}\right) \\
&= \mathbb{E}\left[\bar{F}_R\left(\frac{x}{W_1}\right)\right] \\
&= \mathbb{P}\left(W_1 > x\right) + \int_1^x \left(\frac{w}{x}\right)^{\alpha} \frac{\alpha}{w^{\alpha+1}} dw \\
&= x^{-\alpha} + \alpha x^{-\alpha} \int_1^x \frac{dw}{w} \\
&= x^{-\alpha}\left(1 + \alpha \log(x)\right).
\end{aligned}$$

The inverse of this function is given by

$$\bar{F}^{-1}(y) = \left(\frac{-W_{-1}(-y/e)}{y}\right)^{1/\alpha},$$

where W_{-1} denotes the lower branch of the Lambert W function; for $y \in [-e^{-1}, 0)$, $W_{-1}(y)$ denotes the only solution in $x \in (-\infty, -1]$ of the equation $y = xe^x$. Indeed, it can be seen by a simple plug-in argument that for any $y \in (0, 1]$,

$$\bar{F}\left(\left(\frac{-W_{-1}(-y/e)}{y}\right)^{1/\alpha}\right) = y.$$

Repeating the steps leading to (2.55), we obtain the following similar integral representation for Q:

$$\begin{aligned}
Q(tx, ty) &= \int_0^{\left(\bar{F}^{-1}(ty)\right)^{-\alpha}} da + \left(\bar{F}^{-1}(ty)\right)^{-\alpha} \int_{\left(\bar{F}^{-1}(ty)\right)^{-\alpha}}^{\left(\bar{F}^{-1}(tx)\right)^{-\alpha}} a^{-1} da \\
&\quad + \left(\bar{F}^{-1}(tx)\bar{F}^{-1}(ty)\right)^{-\alpha} \int_{\left(\bar{F}^{-1}(tx)\right)^{-\alpha}}^{1} a^{-2} da \\
&= \left(\bar{F}^{-1}(ty)\right)^{-\alpha} + \left(\bar{F}^{-1}(ty)\right)^{-\alpha} \log\left(\frac{\left(\bar{F}^{-1}(tx)\right)^{-\alpha}}{\left(\bar{F}^{-1}(ty)\right)^{-\alpha}}\right) \\
&\quad + \left(\bar{F}^{-1}(tx)\bar{F}^{-1}(ty)\right)^{-\alpha} \left(\left(\bar{F}^{-1}(tx)\right)^{\alpha} - 1\right) \\
&= \left(\bar{F}^{-1}(ty)\right)^{-\alpha} \left(2 + \log\left(\frac{\left(\bar{F}^{-1}(tx)\right)^{-\alpha}}{\left(\bar{F}^{-1}(ty)\right)^{-\alpha}}\right)\right) - \left(\bar{F}^{-1}(tx)\bar{F}^{-1}(ty)\right)^{-\alpha}.
\end{aligned}$$

The last term in this expression is negligible, compared to the first one, by a factor

of at least $\left(\bar{F}^{-1}(ty)\right)^{-\alpha}$, which (we shall see) is small enough to be absorbed by the term $O(q_1(t))$.

Now, by Corless et al. (1996), Section 4, we may obtain the following expansion of $\left(\bar{F}^{-1}(t)\right)^{-\alpha}$ as $t \to 0$:

$$\begin{aligned}
\left(\bar{F}^{-1}(t)\right)^{-\alpha} &= \frac{t}{-W_{-1}(-t/e)} \\
&= \frac{t}{\log(e/t) + \log\log(e/t) + o(1)} \\
&= \frac{t}{\log(1/t) + \log\log(1/t) + O(1)} \\
&= \left(1 + O\left(\frac{1}{\log(1/t)}\right)\right) \frac{t}{\log(1/t) + \log\log(1/t)}.
\end{aligned}$$

Note that, since we are only interested in $(x, y) \in (0, 1]^2$ and since we assume $y \leq x$, $1/\log(1/ty) \leq 1/\log(1/tx) \leq 1/\log(1/t)$. Plugging the expansion in our expression for Q yields

$$\begin{aligned}
Q(tx, ty) = &\left\{1 + O\left(\frac{1}{\log(1/t)}\right)\right\} \frac{ty}{\log(1/ty) + \log\log(1/ty)} \\
&\times \left(2 + \log\left(\frac{\left\{1 + O\left(\frac{1}{\log(1/t)}\right)\right\} \frac{tx}{\log(1/tx) + \log\log(1/tx)}}{\left\{1 + O\left(\frac{1}{\log(1/t)}\right)\right\} \frac{ty}{\log(1/ty) + \log\log(1/ty)}}\right)\right) \\
&+ O\left(\left(\frac{t}{\log(1/t) + \log\log(1/t)}\right)^2\right) \\
= &\left\{1 + O\left(\frac{1}{\log(1/t)}\right)\right\} \frac{ty}{\log(1/t) + \log\log(1/t) + O(\log(1/y))} \\
&\times \left(2 + \log\left(\left\{1 + O\left(\frac{1}{\log(1/t)}\right)\right\}\right.\right. \\
&\left.\left. \times \frac{x}{y} \frac{\log(1/t) + \log\log(1/t) + O(\log(1/y))}{\log(1/t) + \log\log(1/t) + O(\log(1/x))}\right)\right) \\
&+ O\left(\left(\frac{t}{\log(1/t) + \log\log(1/t)}\right)^2\right)
\end{aligned} \tag{2.56}$$

Note that the first term thereof can be written as

$$\begin{aligned}
\frac{ty}{\log(1/t) + \log\log(1/t) + O(\log(1/y))} &= \frac{ty}{\log(1/t) + \log\log(1/t)} \left\{1 + O\left(\frac{\log(1/y)}{\log(1/t)}\right)\right\} \\
&= \frac{ty}{\log(1/t) + \log\log(1/t)} \left\{1 + O\left(\frac{1}{\log(1/t)}\right)\right\}
\end{aligned}$$

because as y approaches 0, the term $\log(1/y)$ gets absorbed by the term y on the numerator. Furthermore,

$$\begin{aligned}\frac{x}{y}\frac{\log(1/t)+\log\log(1/t)+O(\log(1/y))}{\log(1/t)+\log\log(1/t)+O(\log(1/x))} &= \frac{x}{y}\left\{1+O\left(\frac{\log(1/x)+\log(1/y)}{\log(1/t)}\right)\right\}\\ &= \frac{x}{y}\left\{1+O\left(\frac{\log(1/y)}{\log(1/t)}\right)\right\}.\end{aligned}$$

Thus the log term in (2.56) equals

$$\begin{aligned}\log\left(\frac{x}{y}+O\left(\frac{x}{y}\frac{\log(1/y)}{\log(1/t)}\right)\right) &= \log\left(\frac{x}{y}\right)+O\left(\frac{\frac{x}{y}\frac{\log(1/y)}{\log(1/t)}}{x/y}\right)\\ &= \log\left(\frac{x}{y}\right)+O\left(\frac{\log(1/y)}{\log(1/t)}\right),\end{aligned}$$

where we have used the fact that, for any $a \geq 1$ and $b \geq 0$, $\log(a+b) \leq \log(a) + b/a$ (recall that $x/y \geq 1$). Piecing everything together, (2.56) may be rewritten as

$$\begin{aligned}Q(tx,ty) &= \frac{ty}{\log(1/t)+\log\log(1/t)}\\ &\quad\times\left(2+\log\left(\frac{x}{y}\right)+O\left(\frac{\log(1/y)}{\log(1/t)}\right)\right)\left\{1+O\left(\frac{1}{\log(1/t)}\right)\right\}\\ &= \frac{ty}{\log(1/t)+\log\log(1/t)}\left(2+\log\left(\frac{x}{y}\right)\right)\left\{1+O\left(\frac{1}{\log(1/t)}\right)\right\},\end{aligned}$$

once again because the term $\log(1/y)$ is absorbed by y as y approaches 0. Recalling that we assumed $y \leq x$, the claim follows. □

2.11 A few words on the computational complexity of the method in spatial problems

Both estimators we propose in the spatial setting (defined in (2.16) and (2.17)) essentially rely on the evaluation of bivariate functions and as such are much faster than methods based on full likelihood (especially if the number of locations is large). A comparison with pairwise likelihood depends on the cost of likelihood evaluations in the particular model under consideration and the type of weight functions that we choose. For the sake of brevity we will focus on the estimator $\widehat{\vartheta}$ from (2.16); similar arguments apply to $\widetilde{\vartheta}$ from (2.17) with obvious modifications.

Typically, we expect that $\widehat{\vartheta}$ can be computed faster than a pairwise likelihood-based estimator. The main computational burden arises when computing the pairwise

empirical integrals $\int g(x,y)\widehat{Q}^{(s)}(kx/n, ky/n)dxdy$ and the corresponding estimators $\widehat{\theta}_n^{(s)}$. In computing those estimators, when finding the minimizer of

$$\left\| \int g(x,y)\widehat{Q}^{(s)}(kx/n, ky/n)dxdy - \zeta \int g(x,y)c_\theta(x,y)dxdy \right\|$$

through numerical optimization, only population level integrals $\int g(x,y)c_\theta(x,y)dxdy$ need to be re-computed for each optimization step. For specific models (such as the inverted Brown–Resnick process considered in our application) those integrals have simple analytic expressions, which additionally speeds up the computation. In comparison, the likelihood of a bivariate extreme value model may be substantially more costly to compute, and it needs to be evaluated at every optimization step.

The above procedure only needs to be completed once and can easily be parallelized by considering pairs independently. Once the estimators $\widehat{\theta}_n^{(s)}$ are available, the objective function in (2.16) only depends on evaluating the low-dimensional functions $h^{(s)}$. Again, in our example those are very simple analytic functions.

To give a rough idea of the computation times for the proposed methods in a specific example, we report below average computation times for the spatial simulation study in Section 2.5.2, with $d = 40$ locations (corresponding to 780 pairs), $n = 5000$ and a few different values of m. All computation times are for computing both spatial estimators simultaneously (but the time to compute only one is not so different since most of the "pairwise" steps leading to each estimator are the same). The values given are averaged based on 100 repetitions and the values in parenthesis are standard deviations. All computations were executed on a personal laptop with a 2.5GHz Intel Core i5-7200U processor without utilizing parallel computation.

m	25	100	250	500	1000
time (seconds)	9.6 (0.6)	9.5 (0.3)	9.6 (0.4)	9.8 (0.3)	9.8 (0.3)

Table 2.2: Computation time as a function of m

2.12 Additional simulation results

This section contains additional simulation results not included in Section 2.5.

2.12.1 Bivariate distributions

The following scatter plots represent data from each of the three bivariate models M1–M3 found in Section 2.5.1. For illustration purposes, there is no additive noise and the marginals are transformed to unit exponential.

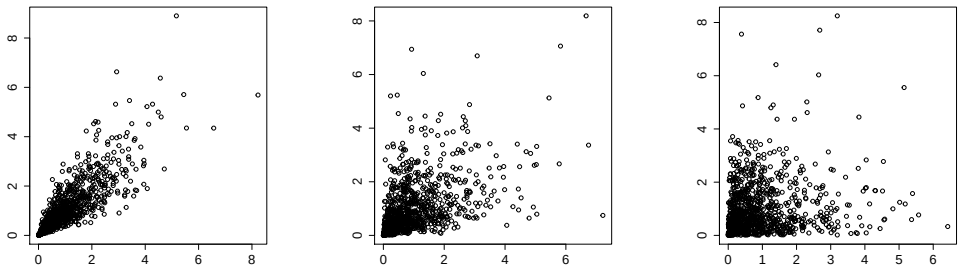

Figure 2.10: Samples of 1 000 data points from the inverted Hüsler–Reiss distribution with parameter θ equal to 0.6, 0.75 and 0.9, from left to right. The marginal distributions are scaled to unit exponential.

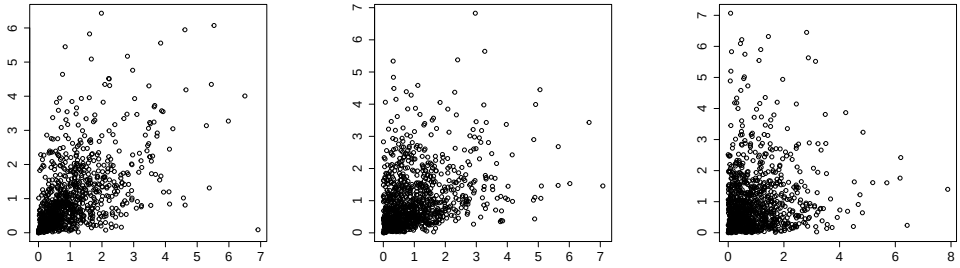

Figure 2.11: Samples of 1 000 data points from the inverted asymmetric logistic distribution with parameter θ equal to $(0.72, 0.72)$, $(0.75, 0.91)$ and $(0.91, 0.91)$, from left to right. The marginal distributions are unit exponential.

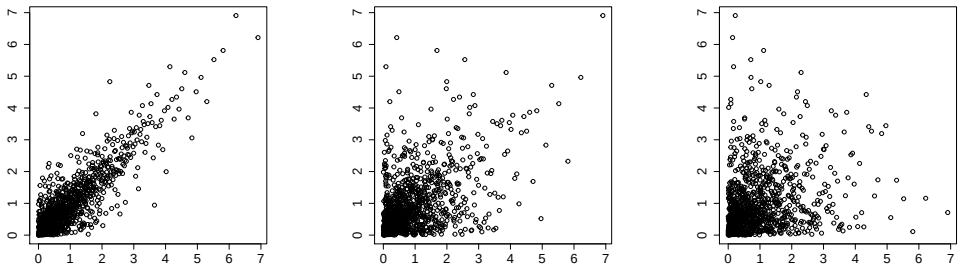

Figure 2.12: Samples of 1 000 data points from the Pareto random scale model with parameter λ equal to 0.4, 1 and 1.6, from left to right. The marginal distributions are approximately unit exponential.

Sensitivity with respect to the weight function

Recall the weight function in (2.22) that is used throughout Section 2.5. It is composed of the weighted indicator functions of the five rectangles $I_1 := [0,1]^2$, $I_2 := [0,2]^2$, $I_3 := [1/2, 3/2]^2$, $I_4 := [0,1] \times [0,3]$ and $I_5 := [0,3] \times [0,1]$. As explained in Section 2.5.1, those rectangles are chosen specifically to ensure identifiability in every model, so that a unique weight function may be used for all simulations.

We now consider different subsets of the five rectangles above and repeat the simulation study with each of the associated lower dimensional weight functions.

Precisely, we define $g^{(1)}$ as the function g in (2.22) and by the same principle we construct $g^{(2)}, \dots, g^{(7)}$, using the rectangles in Table 2.3.

Weight fct.	$g^{(1)}$	$g^{(2)}$	$g^{(3)}$	$g^{(4)}$	$g^{(5)}$	$g^{(6)}$	$g^{(7)}$
Rectangles	I_1, I_2, I_3, I_4, I_5	I_1, I_2	I_1, I_3	I_1, I_4, I_5	I_1, I_2, I_3	I_1, I_2, I_4, I_5	I_1, I_3, I_4, I_5

Table 2.3: Rectangles used to construct each weight function.

We repeat the simulation study from Section 2.5.1; 1 000 data sets of size $n = 5\,000$ are drawn from each of the three models, with the same noise mechanism as before, and from each data set seven estimators are computed based on the seven weight functions. We use the values k that were deemed good previously, that is 800 for the two inverted max-stable models (M1 and M2) and 400 for the Pareto random scale model (M3). For each model and each parameter value, we compare the weight functions based on the estimated RMSE of the M-estimator in Figure 2.13.

In the inverted Hüsler–Reiss model, the parameter has a one-to-one relation with the coefficient of homogeneity $1/\eta$ of c. In order to identify that coefficient, it is sufficient to compare the integral of c over the rectangles I_1 and I_2. It can moreover be deduced from the developments in Section 2.9 that in this model, the bias arising from the pre-asymptotic approximation of c is largest around the axes. Thus, as can be observed below, adding the non required rectangles I_4 and I_5, which contain a large portion of the axes, adds bias to the estimator. The best strategy for this model seems to be using I_1, I_2 and possibly I_3.

In contrast, the parameter in the inverted asymmetric logistic model is not identifiable if the rectangles used are all symmetric, since then (θ_1, θ_2) cannot be distinguished from (θ_2, θ_1). Therefore the estimator is not uniquely defined when neither I_4 nor I_5 is used, so the functions $g^{(2)}$, $g^{(3)}$ and $g^{(5)}$ were not included. It is to be noted that $g^{(4)}$ does not include either of I_2 and I_3, and as such is not able to estimate the homogeneity coefficient $\theta_1 + \theta_2$ well, even if it is able to recover the asymmetry. This explains the monotonic behavior of the error with respect to $\theta_1 + \theta_2$. The other three weight functions perform similarly to each other.

Finally, in the Pareto random scale model, the weight function $g^{(2)}$ only estimates the homogeneity and as such, it is unable to distinguish the parameters in the range $(0, 1)$, corresponding to asymptotic dependence. It was thus ignored. Among the other functions, the ones that use I_4 and I_5 ($g^{(1)}$, $g^{(4)}$, $g^{(6)}$, $g^{(7)}$) all have a similar performance whereas the other two ($g^{(3)}$ and $g^{(5)}$) incur a noticeably larger error. It seems that those rectangles help estimating characteristics that are strongly different from the coefficient of homogeneity, which explains why they significantly reduce the

RMSE under asymptotic dependence ($\lambda < 1$).

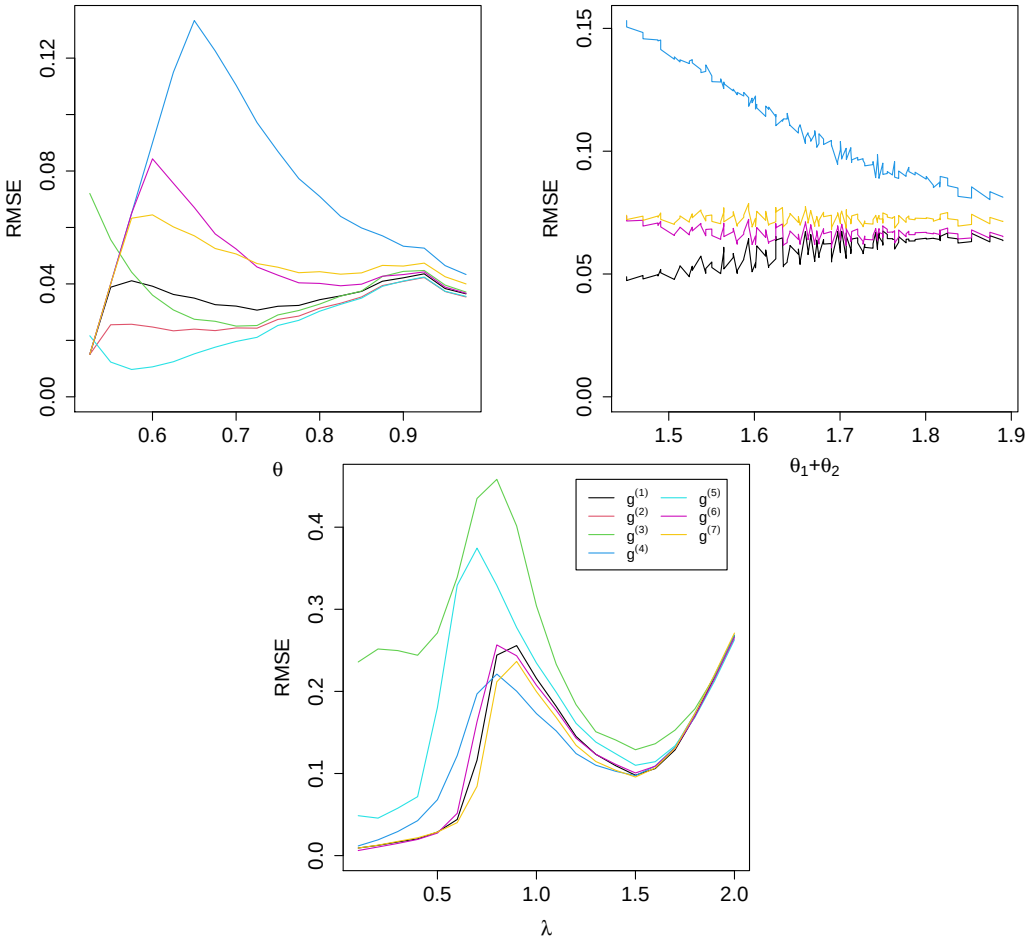

Figure 2.13: RMSE of the M-estimator in the models M1–M3 as a function of the parameter, based on 1 000 data sets of size $n = 5\,000$, $k = 800$ (for M1 and M2) and $k = 400$ (for M3). Colors represent the seven weight functions from Table 2.3.

2.12.2 Spatial models

Figure 2.14 shows the distribution of the distances of all the pairs that are used in the analysis in Section 2.5.2. Figures 2.15 and 2.16 present the same results as in Section 2.5.2 when the estimator (2.17) is used instead of (2.16).

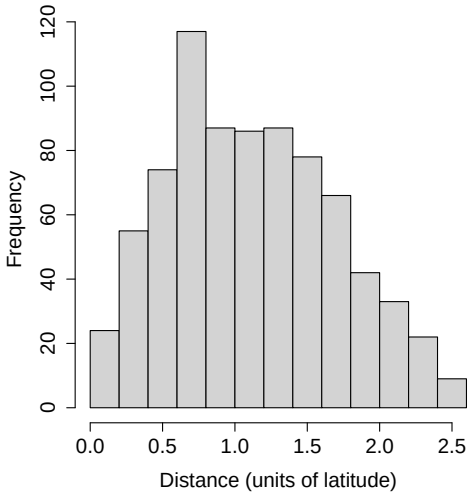

Figure 2.14: Distribution of the distances $\Delta^{(s)}$ for the 780 pairs used.

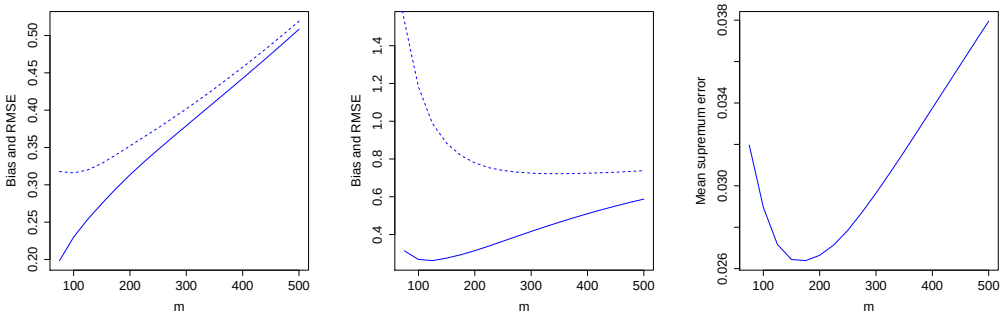

Figure 2.15: Left and middle columns: Bias (solid line) and RMSE (dotted line) of the estimators of the two spatial parameters α (left) and β (middle) as a function of m. Right: Mean of the supremum error $\sup_{0\leq\Delta\leq 3}|\theta(\Delta;\widehat{\alpha},\widehat{\beta})-\theta(\Delta;\alpha,\beta)|$ as a function of m.

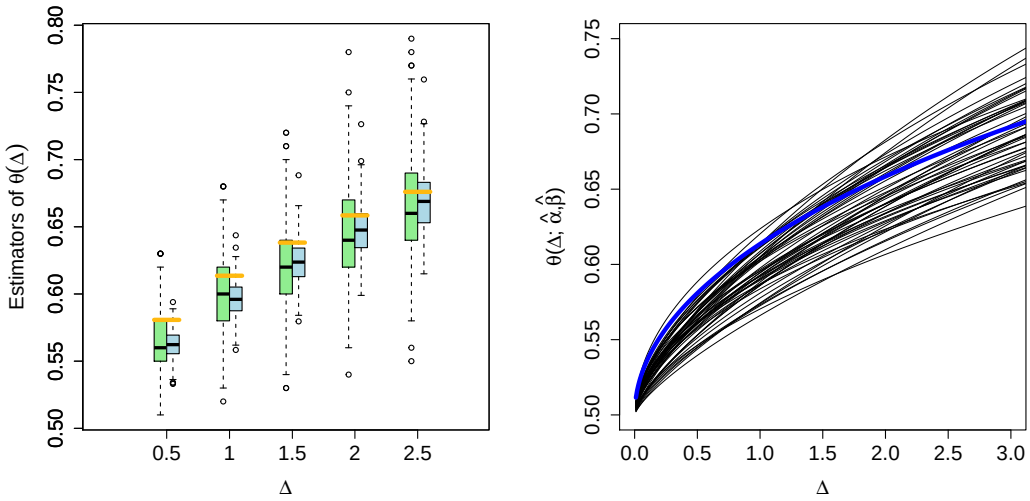

Figure 2.16: Left panel: Estimators of $\theta(\Delta)$ for 5 different distances. For each distance, bivariate M-estimator $\widehat{\theta}_n^{(s)}$ (green) and spatial estimator $\theta(\Delta^{(s)};\widehat{\alpha},\widehat{\beta})$ (blue) based on the $d=40$ locations. Right panel: 50 sampled curves $\theta(\cdot;\widehat{\alpha},\widehat{\beta})$. Blue represents the true curve $\theta(\cdot;\alpha,\beta)$.

Chapter 3

Learning extremal graphical models in high dimensions

3.1 Introduction

Extreme value theory plays an important role in risk quantification of rare events such as floods, heatwaves or financial crises (e.g., Katz et al., 2002; Poon et al., 2004; Engelke et al., 2019a). The univariate case is well understood; generalized extreme value and Pareto distributions allow for a parsimonious description of distributional tail of random variables. In dimension $d \geq 2$, the tail dependence between different components of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ does not follow a parametric model and can become arbitrarily complex. When the dimension is large, concepts such as sparsity or dimension reduction become crucial to obtain methods that are statistically sound and practically feasible; see Engelke and Ivanovs (2021) for a review of recent developments.

One popular approach to obtain interpretable dependence models for a random vectors $\boldsymbol{X}$ in high dimensions relies on graphical modeling (Lauritzen, 1996). Conditional independence relations between pairs of variables are described by the absence of edges in a graph $G = (V, E)$ with vertex set $V = \{1, \ldots, d\}$ and edges $E \subset V \times V$. Graphical models are particularly well studied for multivariate normal distributions where conditional independence relations are encoded as zeroes in the precision matrix (Lauritzen, 1996, Chapter 5).

For distributions arising in the setting of extreme value analysis, graphical modeling has been more challenging. Broadly speaking, there are two main approaches to modeling asymptotically dependent extremes. The first approach considers componentwise maxima of blocks of random vectors and leads to the notion of max-stable distributions; we refer the reader to Beirlant et al. (2004, Chapter 8) and de Haan and

Ferreira (2006, Chapter 6). For such distributions, Gissibl and Klüppelberg (2018), Klüppelberg and Lauritzen (2019) and Améndola et al. (2022) study max-linear models on directed acyclic graphs. The distributions considered in this line of work do not have densities, and a general result by Papastathopoulos and Strokorb (2016) shows that there exist no non-trivial density factorization of max-stable distributions on graphical structures.

The second approach relies on multivariate Pareto distributions, which arise as limits of conditional distributions of $\boldsymbol{X}$ given that at least one of its components is large (Rootzén and Tajvidi, 2006; Rootzén et al., 2018b). While multivariate Pareto distributions inherit certain structural properties such as homogeneity from their definition through limits, their class is still very flexible and too large to allow for efficient and interpretable inference in high dimensions. Classical conditional independence is not suited for these distributions since they are not supported on a product space. Engelke and Hitz (2020) overcome this by proposing new notions of conditional independence and extremal graphical models. They show that these definitions naturally link to density factorization and enable efficient inference on extremal graphical models; see also Asenova et al. (2021). Certain extremal graphical structures are known to arise as the extremes of regularly varying Markov trees (Segers, 2020) and Markov random fields on block graphs (Asenova and Segers, 2021).

The underlying graph plays a key role in graphical modeling. This graph is typically unknown and needs to be estimated in a data driven way. One may consider different levels of generality for the class of graph structures. Figure 3.1 shows four different graphs with increasingly complex structure: from left to right, a tree, a block graph, a general decomposable graph, and finally a non-decomposable graph. In general, estimating more complex graphs requires more assumptions on the underlying distribution. In the non-extreme world, two important cases correspond to graphical models on trees, which can be estimated non-parametrically (Liu et al., 2011), and multivariate normal distributions, for which general graphs can be estimated through the corresponding precision matrix; see Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Friedman et al. (2008) among many others.

To date, we are aware of only two approaches to estimating extremal graph structures, both of which are only applicable to fairly simple graphs or have other limitations. In their application to river discharge data, Engelke and Hitz (2020) use an ad-hoc forward selection method where edges are added one after the other to an initial tree to create a simple block graph (second from left in Figure 3.1). Block graphs are decomposable and the intersection of their cliques are only allowed to be singletons. As such, they are fairly close to tree structures and have rather limited flexibility for

general applications. This was also noted in several discussion contributions in Engelke and Hitz (2020), pointing out the need for estimation techniques for more general graphs. In addition to these limitations, the forward selection procedure is a heuristic method that does not guarantee consistent structure recovery. It moreover requires the choice of a prior tree, and is sensible to this choice. Engelke and Volgushev (2020) study structure estimation of extremal tree structures. They introduce the extremal variogram and show that it can be used in a minimum spanning tree algorithm for consistent tree recovery in a completely non-parametric way.

For more general graphs, such as the two graphs on the right of Figure 3.1, no methods that guarantee consistent graph recovery exist in the world of extremes. In this chapter, we propose a general methodology to learn arbitrary graphs for the class of Hüsler–Reiss distributions (Hüsler and Reiss, 1989). Those distributions share many attractive properties of multivariate normal distributions and have been widely used in modeling multivariate and spatial extremes (e.g., Davison et al., 2012b; Engelke et al., 2015). They are parametrized by a $d \times d$-dimensional variogram matrix Γ, which can be shown to contain the underlying graph structure. More precisely, Engelke and Hitz (2020) show that for any $m \in V$, the Farris transform $\Sigma^{(m)} \in \mathbb{R}^{d\times d}$ (Farris et al., 1970) with entries

$$\Sigma^{(m)}_{ij} = \frac{1}{2}\left(\Gamma_{im} + \Gamma_{jm} - \Gamma_{ij}\right), \quad i, j \in V,$$

encodes information about the extremal conditional independence structure through sparsity patterns in the entries and row-sums of a certain pseudo-inverse $\Theta^{(m)}$ of $\Sigma^{(m)}$ to be defined below.

In the present chapter, we develop a structure learning algorithm that leverages this information and leads to consistent recovery of arbitrary graphs. Due to the special nature of Hüsler–Reiss distributions, the entire graph structure cannot be recovered from the zero pattern of a single matrix $\Theta^{(m)}$, and estimators across all values of $m \in V$ need to be aggregated. Our approach therefore uses a majority voting algorithm to combine estimated sub-graphs for all $m \in V$, which are obtained from base learners derived from the theory of Gaussian graphical models. In principle, any base learner can be used, and we provide a thorough theoretical investigation for two of the most popular choices: neighborhood selection (Meinshausen and Bühlmann, 2006) and graphical lasso (Yuan and Lin, 2007; Friedman et al., 2008). We prove that consistent graph recovery is possible even when the dimension grows exponentially in the number of extreme samples.

A key difficulty in our analysis lies in the fact that, in practice, we typically only observe realizations whose tail can be approximated by a Hüsler–Reiss distribution, rather than from the latter, limiting model itself. Hence estimators of the variogram

matrix Γ use only the largest observations from a sample that need to be transformed marginally. This makes standard concentration results in the literature inapplicable, and a substantial part of our theoretical contribution is devoted to derive concentration bounds for an empirical version of Γ. Such inequalities provide a crucial ingredient for our proof of consistent graph recovery in increasing dimensions. The results are also of broader interest in multivariate extremes beyond Hüsler–Reiss distributions. For instance, the tail bounds we derive play a crucial role in the theoretical developments in Engelke and Volgushev (2020) and we expect that they can be leveraged elsewhere.

Necessary background information on multivariate extreme value distributions in general and Hüsler–Reiss models in particular is collected in Section 3.2. The estimation methodology is described in detail in Section 3.3. Section 3.4 contains all the theoretical results while finite-sample performance of the proposed methods is illustrated in a simulation study in Section 3.5. A data illustration is provided in Section 3.6 while potential extensions and directions for future work are described in Section 3.7. Section 3.8 contains additional numerical results that complement Sections 3.5 and 3.6. The rest of the chapter is dedicated to the proofs of all the theoretical results. We start with the theory related to graph recovery and precision matrix estimation (Sections 3.9 and 3.10), followed by the proof of Theorem 3.3 and related auxiliary results (Sections 3.11 and 3.12).

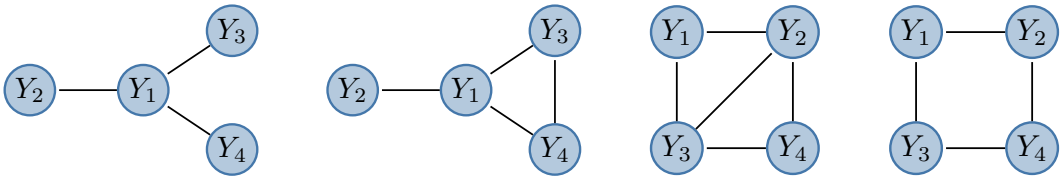

Figure 3.1: Four graph structures on the node set $V = \{1, \dots, 4\}$. From left to right: tree graph, block graph, decomposable graph, non-decomposable graph.

Notation Throughout the chapter we will use the following notation. For square matrices A with real eigenvalues, we let $\lambda_{\min}(A)$ denote the smallest eigenvalue of A. The notation $\|\boldsymbol{x}\|_\infty$ and $\|A\|_\infty$ will be used to denote the element-wise sup-norm of vectors $\boldsymbol{x}$ and matrices A. $|\!|\!|A|\!|\!|_p$ will denote the L^p/L^p operator norm of a (not necessarily square) matrix A. For a natural number $d \geq 1$ let $[d] := \{1, \dots, d\}$. For vectors $\boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ (or similarly for random vectors $\boldsymbol{X} = (X_1, \dots, X_d)^\top$) and subsets $J = \{j_1, \dots, j_k\} \subseteq [d]$, define the vector $\boldsymbol{x}_J := (x_{j_1}, \dots, x_{j_k})^\top$ where $j_1 < j_2 < \cdots < j_k$. Vectors with all entries 1 and 0 will be denoted with $\mathbf{1}$ and $\mathbf{0}$, respectively; the dimension will be clear form the context. Inequalities between vectors

are understood component-wise. The notation $\boldsymbol{x}_{\setminus J}$ will be used to denote the vector $\boldsymbol{x}_{V\setminus J}$, and if $J = \{j\}$ we will also write $\boldsymbol{x}_{\setminus j}$. We use similar notations to index rows and columns of matrices A.

3.2 Background

3.2.1 Multivariate Pareto distributions and domains of attraction

Let $\boldsymbol{Y} = (Y_j : j \in V)$ be a d-dimensional random vector indexed by $V = \{1, \dots, d\}$ with support contained in the space $\mathcal{L} = \{\boldsymbol{y} \geq 0 : \|\boldsymbol{y}\|_\infty > 1\}$. The random vector $\boldsymbol{Y}$ is said to have a multivariate Pareto distribution if $\mathbb{P}(Y_1 > 1) = \cdots = \mathbb{P}(Y_d > 1)$ and if it satisfies the homogeneity property

$$\mathbb{P}(\boldsymbol{Y} \in tA) = t^{-1}\mathbb{P}(\boldsymbol{Y} \in A), \qquad t \geq 1, \tag{3.1}$$

where for any Borel subset $A \subset \mathcal{L}$ we define $tA = \{t\boldsymbol{y} : \boldsymbol{y} \in A\}$ (Rootzén and Tajvidi, 2006). The homogeneity implies that for any $i \in V$ the univariate conditional margin satisfies $\mathbb{P}(Y_i \leq x \,|\, Y_i > 1) = 1 - 1/x$ for $x \geq 1$, that is, $Y_i \,|\, Y_i > 1$ follows a standard Pareto distribution. The class of multivariate Pareto distributions is very rich and contains parametric sub-families such as the extremal logistic (Tawn, 1990) and Dirichlet (Coles and Tawn, 1991) distributions. Of particular interest in the present chapter is the family of Hüsler–Reiss distributions (Hüsler and Reiss, 1989); they are often considered the "Gaussian distributions of extremes" and will be described in Section 3.2.3 in detail.

Multivariate Pareto distributions arise as natural models in the study of multivariate extreme events, since they are the only possible limits of so-called threshold exceedances. To formalize this, consider a random vector $\boldsymbol{X} = (X_j : j \in V)$ in $\mathbb{R}^d$ with eventually continuous marginal distributions F_i and define $F(\boldsymbol{x}) = (F_1(x_1), \dots, F_d(x_d))$. If for some random vector $\boldsymbol{Y}$ the limit relation

$$\lim_{q \downarrow 0} \mathbb{P}(F(\boldsymbol{X}) \leq 1 - q/\boldsymbol{x} \,|\, F(\boldsymbol{X}) \not\leq 1 - q) = \mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{x}), \tag{3.2}$$

holds at all continuity points $\boldsymbol{x} \in \mathcal{L}$ of $\mathbb{P}(\boldsymbol{Y} \leq \cdot)$, we say that $\boldsymbol{X}$ is in the domain of attraction of $\boldsymbol{Y}$.

In this case, the limit $\boldsymbol{Y}$ is necessarily a multivariate Pareto distribution, and conversely any multivariate Pareto distribution can appear as a limit in (3.2); see Engelke and Volgushev (2020, Proposition 6) for a more formal statement. Note that the margins of $\boldsymbol{X}$ are standardized via a marginal transformation, so this is really an assumption on the extremal dependence structure of $\boldsymbol{X}$. The notation $F(\boldsymbol{X}) \not\leq 1 - q$

means that at least one component of the vector $F(\boldsymbol{X})$ exceeds $1-q$, equivalently at least one component X_j of $\boldsymbol{X}$ exceeds its high marginal quantile $F_j^{-1}(1-q)$, $j \in V$. The above limit can thus be interpreted as a model for realizations of $\boldsymbol{X}$ with at least one extreme component.

There are several other popular objects that are routinely used to describe the tails of multivariate random vectors. For instance, one can show that (3.2) implies existence of the limit

$$L(\boldsymbol{x}) = \lim_{q \downarrow 0} q^{-1} \mathbb{P}(F(\boldsymbol{X}) \not\leq 1 - q\boldsymbol{x}), \quad \boldsymbol{x} \in [0, \infty)^d. \tag{3.3}$$

The function L is called the stable tail dependence function of $\boldsymbol{X}$ and is a popular object in the study of multivariate extremes (Huang, 1992; Einmahl et al., 2012; Fougères et al., 2015). The link between the distribution of $\boldsymbol{Y}$ and L is given by the relation

$$L(\boldsymbol{x}) = \frac{\mathbb{P}(\boldsymbol{Y} \not\leq 1/\boldsymbol{x})}{\mathbb{P}(Y_1 > 1)}, \quad \boldsymbol{x} \in \mathcal{L}.$$

3.2.2 Extremal graphical models

Since the support space $\mathcal{L}$ of a multivariate Pareto distribution $\boldsymbol{Y}$ is not a product space, the set of auxiliary random vectors $\boldsymbol{Y}^{(m)} = (\boldsymbol{Y} \,|\, Y_m > 1)$, $m \in V$, plays an important role in the analysis. In fact, they are the basis for the definition of conditional independence for the random vector $\boldsymbol{Y}$ in Engelke and Hitz (2020). For disjoint sets $A, B, C \subset V$, we say that $\boldsymbol{Y}_A$ is conditionally independent of $\boldsymbol{Y}_C$ given $\boldsymbol{Y}_B$ if the usual conditional independence holds for all auxiliary vectors:

$$\forall m \in V : \boldsymbol{Y}_A^{(m)} \perp \boldsymbol{Y}_C^{(m)} \,|\, \boldsymbol{Y}_B^{(m)}. \tag{3.4}$$

In this case, we speak of extremal conditional independence and denote it by $\boldsymbol{Y}_A \perp_e \boldsymbol{Y}_C \,|\, \boldsymbol{Y}_B$.

In graphical modeling, conditional independence is connected to graph structures to define sparse probabilistic models (Lauritzen, 1996). An undirected graph $G = (V, E)$ is a set of nodes $V = \{1, \ldots, d\}$ and a collection of edges $E \subset V \times V$ of unordered pairs of distinct nodes. With the notion of extremal conditional independence, we define an extremal graphical model as a multivariate Pareto distribution $\boldsymbol{Y}$ that satisfies the pairwise Markov property,

$$Y_i \perp_e Y_j \,|\, \boldsymbol{Y}_{\setminus\{i,j\}}, \text{ if } (i,j) \notin E. \tag{3.5}$$

When $\boldsymbol{Y}$ possesses a positive continuous density, the graph G is necessarily connected.

In this case, the pairwise Markov property is equivalent to the stronger global Markov property (Lauritzen, 1996, Chapter 3). If G is in addition decomposable, then the density factorizes on the graph into lower-dimensional densities, and inference is considerably more efficient thanks to the sparsity; see Engelke and Hitz (2020) for details.

A summary statistic for extremal dependence in $\boldsymbol{Y}$ that will turn out to be useful for graph structure learning is the extremal variogram (Engelke and Volgushev, 2020). The extremal variogram rooted at node $m \in V$ is defined as the matrix $\Gamma^{(m)}$ with entries

$$\Gamma_{ij}^{(m)} = \mathbb{V}\mathrm{ar}\left\{\log Y_i^{(m)} - \log Y_j^{(m)}\right\}, \quad i, j \in V, \tag{3.6}$$

whenever the right-hand side exists and is finite. For an extremal graphical model on a tree, for any $m \in V$, the minimum spanning tree with weights $\Gamma_{ij}^{(m)}$, $i, j \in V$, recovers the underlying tree structure corresponding to extremal conditional independence. This can be exploited for non-parametric consistent tree recovery without distributional assumptions (Engelke and Volgushev, 2020). Extremal variograms will also play a crucial role in learning general extremal graph structures.

3.2.3 Hüsler–Reiss distributions

Let $\mathcal{S}_0^d$ be the set of symmetric $d \times d$-matrices with zero diagonal and non-negative entries. A conditionally negative definite matrix $\Gamma \in \mathcal{S}_0^d$ is defined by the property that $\boldsymbol{x}^\top \Gamma \boldsymbol{x} \leq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$ with $\boldsymbol{x}^\top \mathbf{1} = 0$. If the inequality is strict except for $\boldsymbol{x} = \mathbf{0}$, then Γ is in the cone $\mathcal{C}^d \subset \mathcal{S}_0^d$ of strictly conditionally negative definite matrices, which we will also call variogram matrices. Let further $\mathcal{P}^{d-1}$ denote the space of symmetric, strictly positive definite $(d-1) \times (d-1)$-matrices.

The family of d-dimensional Hüsler–Reiss (Pareto) distributions consists of multivariate Pareto distributions parametrized by $\Gamma \in \mathcal{C}^d$ (Hüsler and Reiss, 1989). For each $m \in V$, the Farris transform $\varphi_m : \mathcal{C}^d \to \varphi_m(\mathcal{C}^d)$ (Farris et al., 1970) maps a given Γ to the matrix

$$\Sigma_{ij}^{(m)} := \frac{1}{2}\left(\Gamma_{im} + \Gamma_{jm} - \Gamma_{ij}\right), \quad i, j \in V. \tag{3.7}$$

The Farris transform is a bijection with inverse $\varphi_m^{-1} : \varphi_m(\mathcal{C}^d) \to \mathcal{C}^d$, where the image $\varphi_m(\mathcal{C}^d)$ consists of all matrices in $\mathcal{P}^{d-1}$ with a row and column of zeroes inserted in the mth position. Consequently, the submatrix of $\Sigma^{(m)}$ obtained by removing this row and column is invertible. Let $\Theta^{(m)}$ be obtained by inverting that submatrix and adding back the row and column of zeros in the mth position, i.e., $\Theta^{(m)} \in \mathbb{R}^{d \times d}$ is

defined by

$$\Theta^{(m)}_{\setminus m,\setminus m} = \big(\Sigma^{(m)}_{\setminus m,\setminus m}\big)^{-1}, \tag{3.8}$$

and $\Theta^{(m)}_{im} = \Theta^{(m)}_{mi} = 0$, $i \in V$. Then, the density of a Hüsler–Reiss distributed $\boldsymbol{Y}$ with parameter matrix Γ exists, and for any $m \in V$ it can be expressed as

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) \propto y_m^{-2}\Big(\prod_{i\neq m} y_i^{-1}\Big) \exp\left\{-\frac{1}{2}\widetilde{\boldsymbol{y}}^\top \Theta^{(m)} \widetilde{\boldsymbol{y}}\right\}, \tag{3.9}$$

where $\widetilde{\boldsymbol{y}} = \log \boldsymbol{y} - \mathbf{1}\log y_m + \Gamma_{Vm}/2$. The matrix $\Sigma^{(m)}$ is the covariance matrix of a transformation of the auxiliary random vector given by $(\log\{Y_i^{(m)}/Y_m^{(m)}\} : i \in V)$. From this it is easy to see that extremal variograms in (3.6) at nodes m are given by $\varphi_m^{-1}(\Sigma^{(m)})$, $m \in V$, and therefore they are all equal to the parameter matrix of the Hüsler–Reiss distribution, that is, $\Gamma = \Gamma^{(1)} = \cdots = \Gamma^{(d)}$.

The importance of the matrices $\Sigma^{(m)}$ and $\Theta^{(m)}$ comes from the fact that the latter contains the graphical structure of the Hüsler–Reiss distribution $\boldsymbol{Y}$ in its sparsity pattern. Indeed, for any fixed node $m \in V$ and $i \neq j$, Engelke and Hitz (2020, Proposition 3) show that

$$Y_i \perp_e Y_j \,|\, \boldsymbol{Y}_{\setminus\{i,j\}} \iff \begin{cases} \Theta^{(m)}_{ij} = 0, & \text{if } i,j \neq m, \\ \sum_{\ell=1}^d \Theta^{(m)}_{i\ell} = 0, & \text{if } j = m. \end{cases} \tag{3.10}$$

Relation (3.10) provides the key to learning general graphs for Hüsler–Reiss distributions through estimating sparse versions of $\Theta^{(m)}$.

The following equivalent parametrization of the Hüsler–Reiss distribution was proposed by Hentschel (2021); see also Röttger et al. (2021) for details. Consider the $d \times d$ matrix Θ with entries

$$\Theta_{ij} = \Theta^{(m)}_{ij}, \quad i,j \neq m, \tag{3.11}$$

and note that it is well-defined since $\Theta^{(m)}_{ij} = \Theta^{(m')}_{ij}$ for $i,j \notin \{m,m'\}$ by Engelke and Hitz (2020, Lemma 1). The matrix Θ is symmetric, positive semi-definite with rank $d-1$, and it has zero row sums: $\Theta\mathbf{1} = 0$. Let $P = I_d - \mathbf{1}\mathbf{1}^\top/d$. Then there is a one-to-one correspondence between the parameter matrix Γ and the matrix Θ given by $\Theta = (P(-\Gamma/2)P)^+$, where A^+ is the Moore–Penrose pseudoinverse of a matrix A. The matrix Θ uniquely defines the Hüsler–Reiss distribution since the parameter matrix can be recovered through the inverse Farris transform by $\Gamma = \varphi_m^{-1}(\Theta^+)$. Moreover, it

contains the graphical structure through the simple relation

$$Y_i \perp_e Y_j \,|\, \boldsymbol{Y}_{\setminus\{i,j\}} \iff \Theta_{ij} = 0. \tag{3.12}$$

The matrix Θ is therefore called the Hüsler–Reiss precision matrix of $\boldsymbol{Y}$.

3.3 Learning Hüsler–Reiss graphical models

3.3.1 `EGlearn`: a majority voting algorithm

For a Hüsler–Reiss distribution, in view of (3.10), a sparse estimate of $\Theta^{(m)}$ contains information on the conditional independence structure between nodes i, j where $i, j \neq m$. This is equivalent to the presence or absence of edges in the corresponding extremal graphical model $G = (V, E)$ that are not related to the mth node. This fact can be exploited by obtaining a sparse estimate of $\Theta^{(m)}$, which establishes a link to the problem of sparse precision matrix estimation and allows us to borrow tools from the Gaussian graphical models literature. In our case, it is natural to combine estimated sparsity patterns across different values of m to infer extremal conditional independence for all possible values of $i, j \in V$. We propose to do this through a majority voting algorithm.

More formally, for a given $m \in V$ and estimator $\widehat{\Gamma}$ of the variogram matrix Γ, consider an arbitrary algorithm $\mathcal{A}$, called base learner in what follows, that takes the submatrix $\widehat{\Sigma}^{(m)}_{\setminus m,\setminus m}$, $\widehat{\Sigma}^{(m)} := \varphi_m(\widehat{\Gamma})$, as input and returns an estimator of the set of non-zero entries of $\Theta^{(m)}_{\setminus m,\setminus m}$. The output of this algorithm, denoted by $\widehat{Z}^{(m)}$, will be represented as a $(d-1) \times (d-1)$ matrix with entries 1 in positions where $\Theta^{(m)}_{\setminus m,\setminus m}$ is estimated to be non-zero, and entries 0 elsewhere. Two examples of possible base learner algorithms are neighborhood selection (Meinshausen and Bühlmann, 2006) and graphical lasso (Yuan and Lin, 2007; Friedman et al., 2008); they are formally introduced in Section 3.3.2 below. The base learner $\mathcal{A}$ may require the choice of tuning parameters, as is the case for neighbourhood selection and the graphical lasso, which can be fixed or data-dependent.

Augmenting the matrix $\widehat{Z}^{(m)}$ with a row and column of zeros in the mth position, we obtain a $d \times d$ matrix $\widetilde{Z}^{(m)}$. The entries of $\widetilde{Z}^{(m)}$ outside its mth row and column are now considered as votes in favor or against the presence of certain edges in the graph G. Running the algorithm for each $m \in V$ results in d such matrices. Those are then combined into a final graph estimator $\widehat{G} = (V, \widehat{E})$ using majority voting: an edge (i, j), $i \neq j$, is included in the final graph if and only if a 1 appears in position (i, j) of more than half of the $d-2$ matrices $\widetilde{Z}^{(m)}$, $m \notin \{i, j\}$. The reason for excluding $\widetilde{Z}^{(i)}$

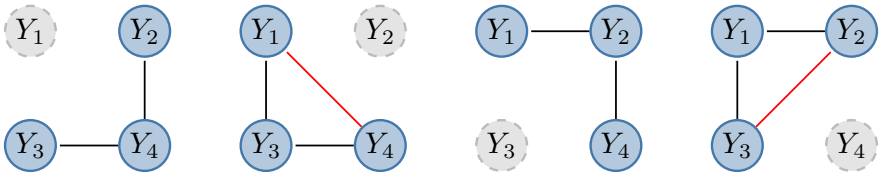

Figure 3.2: Illustration of the majority voting algorithm when the true underlying graph is the non-decomposable graph on the right-hand side of Figure 3.1. Left to right: graphical representation of the estimated matrices $\widetilde{Z}^{(m)}$, where the grey node Y_m is not considered in the mth step, $m = 1, \ldots, 4$; black and red edges indicate correctly and incorrectly estimated edges, respectively.

and $\widetilde{Z}^{(j)}$ from the voting for edge (i, j) is that by (3.10), zeroes in the mth row and column of $\Theta^{(m)}$ are not informative about conditional independence in $\boldsymbol{Y}$. The steps described above are summarized in the following algorithm.

Input: variogram estimate $\widehat{\Gamma}$, base learner algorithm $\mathcal{A}$
Output: extremal graph estimate $\widehat{G} = (V, \widehat{E})$

initialize $\widehat{G} := (V, \emptyset)$
for $m \in V$ **do**
  initialize $\widehat{\Sigma}^{(m)} := \varphi_m(\widehat{\Gamma})$
  obtain a $(d-1) \times (d-1)$ matrix $\widehat{Z}^{(m)}$ from algorithm $\mathcal{A}$ with $\widehat{\Sigma}^{(m)}_{\setminus m, \setminus m}$ as input
  obtain $\widetilde{Z}^{(m)}$ by augmenting $\widehat{Z}^{(m)}$ with a row and column of zeros in the mth position
for $i, j \in V$, $i \neq j$ **do**
  if $\frac{1}{d-2} \#\{m \in V \setminus \{i, j\} : \widetilde{Z}^{(m)}_{ij} = 1\} > \frac{1}{2}$ **then**
    add an edge in $\widehat{G}$ between nodes i and j

Algorithm 1: `EGlearn`: general algorithm for learning extremal graphical models.

Figure 3.2 shows an illustration of the majority voting algorithm where the true underlying graph $G = (V, E)$ is the non-decomposable graph on the right-hand side of Figure 3.1. In this example, the algorithm would output the true graph $\widehat{G} = G$ since exactly the true edges appear in the majority of the cases.

An alternative method that, based on (3.10), not only uses the information in $\Theta^{(m)}$ on nodes $i, j \in V$, but jointly enforces sparsity also on edges related to the mth node, turns out to be more involved and is discussed in Section 3.7.

3.3.2 Base learners for sparsity estimation

Two classical methods from Gaussian graphical modeling to obtain sparse estimators of precision matrices are neighborhood selection (Meinshausen and Bühlmann, 2006) and

graphical lasso (Yuan and Lin, 2007). The original theoretical guarantees for consistent recovery of the sparsity pattern rely on Gaussian data and empirical covariances as input. Since the input estimator $\widehat{\Sigma}^{(m)}_{\setminus m,\setminus m}$ for the base learner $\mathcal{A}$ in our `EGlearn` in Algorithm 1 uses neither Gaussian data nor the empirical covariance, we discuss in this section how the assumptions of sparse estimators for Gaussian distributions can be relaxed. Related observations were made in Liu et al. (2012) for data that have a Gaussian copula but are not marginally Gaussian and Loh and Wainwright (2013) for discrete graphical models.

Throughout this section, we let $A \in \mathbb{R}^{p\times p}$ denote a symmetric, positive definite matrix and we are interested in the sparsity pattern of its inverse $B = A^{-1}$. We aim to use neighborhood selection and graphical lasso as the base learner algorithm $\mathcal{A}$ in the framework of our `EGlearn` in Algorithm 1. In this case, in the mth step of the algorithm, the matrix $A = \Sigma^{(m)}_{\setminus m,\setminus m}$ with $p = d - 1$, and the interest is in the sparsity pattern of $B = \Theta^{(m)}_{\setminus m,\setminus m}$.

Neighborhood selection

Neighborhood selection was originally proposed by Meinshausen and Bühlmann (2006) for estimating Gaussian graphical models. Although the motivation in Meinshausen and Bühlmann (2006) relies on properties of multivariate normal distributions and their conditional independence, the underlying principle can be used to estimate the sparsity pattern of the inverse of a general symmetric positive definite matrix. Indeed, for A positive definite and B its inverse, we have the representation

$$\frac{-B_{\setminus\ell,\ell}}{B_{\ell\ell}} = (A_{\setminus\ell,\setminus\ell})^{-1}A_{\setminus\ell,\ell} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{p-1}} \big\{ -2A_{\ell,\setminus\ell}\beta + \beta^\top A_{\setminus\ell,\setminus\ell}\beta \big\}, \quad \ell = 1,\dots,p.$$

The first equation follows from matrix computations using block inversion formulae (see, for instance, Lauritzen, 1996, Equation (C.4)) and the second from computing the gradient of the minimization problem. Hence, given access to an estimator $\widehat{A}$, the sparsity pattern in $B_{\setminus\ell,\ell}$ can be estimated through the zero entries of

$$\widehat{\theta} := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{p-1}} \big\{ -2\widehat{A}_{\ell,\setminus\ell}\theta + \theta^\top \widehat{A}_{\setminus\ell,\setminus\ell}\theta + \rho_\ell\|\theta\|_1 \big\}, \tag{3.13}$$

where ρ_ℓ denotes a penalty parameter and the L^1 penalty is used for enforcing sparse solutions. The set $\widehat{\mathrm{ne}}(\ell)$ of indices of non-zero entries in $\widehat{\theta}$ is then taken as an estimate of the non-zero pattern in the ℓth row of B. The procedure is repeated for each variable ℓ. Since the matrix B is symmetric, pairs $(i,j),(j,i)$ are added to the estimated set of non-zero entries if and only if $i \in \widehat{\mathrm{ne}}(j)$ and $j \in \widehat{\mathrm{ne}}(i)$; see Algorithm 2 below.

To link the above approach to neighborhood selection as proposed in Meinshausen and Bühlmann (2006), assume that $\widehat{\Sigma}$ is the sample covariance matrix of $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$, a sample from a p-dimensional Gaussian distribution. Then regressing the variables $W_{1,\ell}, \ldots, W_{n,\ell}$ on $\boldsymbol{W}_{1,\setminus\ell}, \ldots, \boldsymbol{W}_{n,\setminus\ell}$ via the lasso amounts to solving the problem

$$\begin{aligned}&\arg\min_{\beta\in\mathbb{R}^{p-1}} \Big\{ \frac{1}{n}\sum_{i=1}^{n}(W_{i,\ell} - \beta^\top \boldsymbol{W}_{i,\setminus\ell})^2 + \rho_\ell \|\beta\|_1 \Big\} \\ &\quad = \arg\min_{\beta\in\mathbb{R}^{p-1}} \big\{ -2\widehat{\Sigma}_{\ell,\setminus\ell}\beta + \beta^\top \widehat{\Sigma}_{\setminus\ell,\setminus\ell}\beta + \rho_\ell\|\beta\|_1 \big\},\end{aligned}$$

where the left–hand side in the equation above was originally considered in Meinshausen and Bühlmann (2006).

Input: Matrix $\widehat{A} \in \mathbb{R}^{p\times p}$, penalty parameters $(\rho_\ell)_{\ell=1,\ldots,p}$
Output: Estimate $\widehat{Z}$ of sparsity pattern of A^{-1}

initialize $\widehat{Z}$ as a matrix of zeros
for $\ell = 1, \ldots, p$ **do**
  $\widehat{\theta} := \arg\min_{\theta\in\mathbb{R}^{p-1}} \big\{ -2\widehat{A}_{\ell,\setminus\ell}\theta + \theta^\top \widehat{A}_{\setminus\ell,\setminus\ell}\theta + \rho_\ell\|\theta\|_1 \big\}$
  $\widehat{\mathrm{ne}}(\ell) := \{j = 1, \ldots, p : \theta_j \neq 0\}$
for $i = 1, \ldots, p,\ j \neq i$ **do**
  if $i \in \widehat{\mathrm{ne}}(j)$ *and* $j \in \widehat{\mathrm{ne}}(i)$ **then**
    set $\widehat{Z}_{ij} = \widehat{Z}_{ji} = 1$

Algorithm 2: Sparsity pattern estimation through neighborhood selection.

Graphical lasso

As an alternative to running node-wise regressions as required for neighborhood selection, Yuan and Lin (2007) suggested to estimate the precision matrix $B = A^{-1}$ via penalized maximum likelihood, with a penalty on the off-diagonal, element-wise L^1 norm of the matrix:

$$\arg\min_{Q\in\mathcal{P}^p} \big\{ -\log\det Q + \mathrm{tr}(\widehat{A}Q) + \rho\|Q\|_{1,\mathrm{off}} \big\}, \quad \|Q\|_{1,\mathrm{off}} := \sum_{i\neq j}\sum |Q_{ij}|; \tag{3.14}$$

where $\widehat{A}$ denotes an estimator of the matrix A. The name "graphical lasso" was subsequently given to the algorithm that efficiently solves this problem (Friedman et al., 2008).

In the original procedure of Yuan and Lin (2007), $\widehat{A}$ denotes the sample covariance matrix of an iid sample of $N(0, A)$ random vectors. Subsequently, Ravikumar et al. (2011) proved that Gaussianity is not needed and the sparsity pattern of pre-

cision matrices can be recovered consistently as long as the data used to compute the empirical covariance matrix satisfy certain tail properties. A close look at the analysis of Ravikumar et al. (2011) further reveals that there is nothing special about empirical covariance matrices. More precisely, under certain technical conditions, the optimization in (3.14) yields an estimator of A^{-1} with the correct sparsity pattern, provided that $\widehat{A}$ is any estimator of A that is close to A in element-wise sup-norm. These claims will be made precise in Proposition 3.4.

3.3.3 The empirical extremal variogram

The extremal variogram matrix $\Gamma^{(m)}$ rooted at node m is defined in (3.6) for a general, not necessarily Hüsler–Reiss, multivariate Pareto distribution $\boldsymbol{Y}$. In typical applications we do not observe data from $\boldsymbol{Y}$ but rather from $\boldsymbol{X}$ in the domain of attraction of $\boldsymbol{Y}$ in the sense of (3.2). A simple computation then implies that

$$P(\boldsymbol{Y}^{(m)} \leq \boldsymbol{x}) = \lim_{q \downarrow 0} \mathbb{P}\Big(\frac{q}{1 - F(\boldsymbol{X})} \leq \boldsymbol{x} \,\Big|\, F_m(X_m) > 1 - q \Big).$$

Let $(\boldsymbol{X}_t := (X_{t1}, \ldots, X_{td}) : t \in [n])$ be a sample of independent copies of $\boldsymbol{X}$. Denoting by $\widetilde{F}_i$ the left-continuous empirical distribution function of $X_{1i}, \ldots, X_{ni}$, the sample

$$\Big\{ \frac{k}{n} \frac{1}{1 - \widetilde{F}(\boldsymbol{X}_t)} : \widetilde{F}_m(X_{tm}) > 1 - k/n \Big\},$$

with $\widetilde{F}(\boldsymbol{x}) := (\widetilde{F}_1(x_1), \ldots, \widetilde{F}_d(x_d))$, is an approximate sample from $\boldsymbol{Y}^{(m)}$ if k/n is sufficiently small. This motivates the empirical extremal variogram rooted at node m (Engelke and Volgushev, 2020), defined through the conditional variance

$$\widehat{\Gamma}^{(m)}_{ij} := \widehat{\mathbb{V}\mathrm{ar}}\Big(\log(1 - \widetilde{F}_i(X_{ti})) - \log(1 - \widetilde{F}_j(X_{tj})) \;\Big|\; \widetilde{F}_m(X_{tm}) > 1 - k/n \Big),$$

where $\widehat{\mathbb{V}\mathrm{ar}}$ denotes the empirical variance with scaling equal to k^{-1}, the inverse of the sample size.

When $\boldsymbol{Y}$ has a Hüsler–Reiss Pareto distribution, the population variograms $\Gamma^{(m)}$, $m \in V$, are all equal to the parameter matrix Γ. We therefore combine the estimators into the empirical (extremal) variogram

$$\widehat{\Gamma} := \frac{1}{d} \sum_{m \in V} \widehat{\Gamma}^{(m)}. \tag{3.15}$$

Concentration properties of this estimator will be derived in Section 3.4.2.

3.4 Consistent extremal graph recovery and concentration of empirical variograms

In this section we provide theoretical guarantees for Algorithm 1 to correctly recover the true extremal graph G when either neighborhood selection or graphical lasso is used as base learner. Our main results on consistent graph recovery are collected in Section 3.4.1.

A key technical ingredient of the corresponding proofs is a general concentration bound for the empirical extremal variogram; see Theorem 3.3. This result is obtained under general domain of attraction conditions and does not require the limiting multivariate Pareto distribution to be from the Hüsler–Reiss family. This finding is of independent interest. For instance, the theory of high-dimensional extremal tree recovery in Engelke and Volgushev (2020) relies on our Theorem 3.3.

As a second ingredient, we establish guarantees on the estimators provided by neighbourhood selection and graphical lasso when the input matrices are not empirical covariances of iid Gaussian data. This is a straightforward consequence of the analysis in Ravikumar et al. (2011) for the graphical lasso but requires more work for neighborhood selection, since classical arguments in Meinshausen and Bühlmann (2006) explicitly rely on properties of the multivariate normal distribution. We provide a formal statement in Proposition 3.3 and its proof in Section 3.10.

3.4.1 Consistent recovery of Hüsler–Reiss graphical models

We start by collecting technical assumptions which are needed for graph recovery in the Hüsler–Reiss case. In what follows, assume that $\boldsymbol{Y}$ has a Hüsler–Reiss distribution on $\mathbb{R}^d$ with parameter matrix $\Gamma \in \mathcal{C}^d$ that is an extremal grapical model on the connected graph $G = (V, E)$. Recall the definitions of the precision matrix Θ and the matrices $\Sigma^{(m)}$ and $\Theta^{(m)}$ in Section 3.2.3.

Our first assumption is a second order condition that essentially controls the speed of convergence in (3.3). For notational reasons, it turns out to me more convenient to work with slightly different versions of the probabilities appearing in (3.3).

Assumption 3.1 (Extended second order)**.** *The marginal distribution functions $F_1, \dots, F_d$ are continuous and there exist constants $\xi > 0$ and $K < \infty$ such that for all triples of distinct indices $J = (i, j, m)$ and $q \in (0, 1]$,*

$$\sup_{\boldsymbol{x} \in [0, q^{-1}]^2 \times [0,1]} \left| q^{-1} \mathbb{P}\big(F_J(\boldsymbol{X}_J) > 1 - q\boldsymbol{x}\big) - \frac{\mathbb{P}(\boldsymbol{Y}_J > 1/\boldsymbol{x})}{\mathbb{P}(Y_1 > 1)} \right| \leq K q^{\xi}, \tag{3.16}$$

where $F_J(\boldsymbol{x}_J) = (F_i(x_i), F_j(x_j), F_m(x_m))$.

This formulation differs from classical second order conditions in that the supremum is taken over sets that grow with q. For Hüsler–Reiss distributions it is implied by a standard second order condition on bounded sets (see Assumption 3.3 in the next section) which is routinely imposed in theoretical developments for multivariate extreme value theory (Einmahl et al., 2012; Fougères et al., 2015; Engelke and Volgushev, 2020); see Proposition 3.1 in Section 3.4.2 for details.

Depending on whether we use neighborhood selection or graphical lasso as base learners in Algorithm 1, additional assumptions are needed on the matrices $\Sigma^{(m)}$, $\Theta^{(m)}$ and Θ that parametrize the limiting model $\boldsymbol{Y}$. We start by discussing neighborhood selection.

Let s denote the maximal degree, that is, the largest number of edges connected to any node, of the graph G. For $m, \ell \in V$, $m \neq \ell$, define the set of all nodes, except for node m, that are connected to node ℓ as

$$S_{m,\ell} := \{i \in V \setminus \{m, \ell\} : \Theta_{i\ell} \neq 0\} \subset V \setminus \{m\}$$

and its complement $S^c_{m,\ell}$ taken in $V \setminus \{m\}$. Define the quantities

$$\begin{aligned}
\theta^{\text{ns}}_{\min} &:= \min_{i,\ell:\Theta_{i\ell}\neq 0} |\Theta_{i\ell}|/\Theta_{\ell\ell} \\
\lambda &:= \min_{m,\ell\in V, m\neq\ell} \lambda_{\min}(\Sigma^{(m)}_{S_{m,\ell},S_{m,\ell}}), \\
\kappa &:= \max_{m,\ell\in V, m\neq\ell} \left|\!\left|\!\left| \Sigma^{(m)}_{S^c_{m,\ell},S_{m,\ell}} \right|\!\right|\!\right|_\infty, \\
\vartheta &:= \max_{m,\ell\in V, m\neq\ell} \left|\!\left|\!\left| (\Sigma^{(m)}_{S_{m,\ell},S_{m,\ell}})^{-1} \right|\!\right|\!\right|_\infty.
\end{aligned}$$

Additionally, consider the neighborhood selection incoherence parameter

$$\eta^{\text{ns}} := \min_{m,\ell\in V, m\neq\ell} \eta^{\text{ns}}_{m,\ell}, \quad \eta^{\text{ns}}_{m,\ell} := 1 - \left|\!\left|\!\left| \Sigma^{(m)}_{S^c_{m,\ell},S_{m,\ell}} (\Sigma^{(m)}_{S_{m,\ell},S_{m,\ell}})^{-1} \right|\!\right|\!\right|_\infty. \tag{3.17}$$

Incoherence parameters of this sort are known to be a crucial ingredient for support recovery via the lasso (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006). In our theory below we will assume that η^{ns} is strictly positive; a technical relaxation to requiring sufficiently many $\eta^{\text{ns}}_{m,\ell}$ being positive is shown in the proof of Theorem 3.1.

The majority voting in Algorithm 1 applies the base learner algorithm to d distinct problems, namely for every $m \in V$. Using neighborhood selection as the base learner in turn requires the choice of d tuning parameters, resulting in a total of $d(d-1)$ tuning parameters $\rho^{\text{ns}}_{m,\ell}, m \in V, \ell \in [d-1]$, where $\rho^{\text{ns}}_{m,1}, \ldots, \rho^{\text{ns}}_{m,d-1}$ correspond to the tuning parameters in the mth step of Algorithm 1. Define $\rho^{\text{ns}}_{\min} := \min_{m,\ell} \rho^{\text{ns}}_{m,\ell}$ and $\rho^{\text{ns}}_{\max} := \max_{m,\ell} \rho^{\text{ns}}_{m,\ell}$.

Theorem 3.1. *Assume* (3.16) *and that* $\eta^{ns} > 0$. *Let* $G^{\text{ns}}(\widehat{\Gamma})$ *denote the estimated graph obtained through neighbourhood selection in Algorithm 1 with penalty parameters* $\rho^{\text{ns}}_{m,\ell}$. *Then, as soon as*

$$\left\|\widehat{\Gamma} - \Gamma\right\|_\infty < C^{\text{ns}} := \frac{2}{3} \min\left\{ \frac{\lambda}{2s}, \frac{\eta^{\text{ns}}}{4\vartheta(1+\kappa\vartheta)s}, \frac{\theta^{\text{ns}}_{\min} - \vartheta\rho^{\text{ns}}_{\max}}{2\vartheta(1+\kappa\vartheta)}, \frac{\rho^{\text{ns}}_{\min}\eta^{\text{ns}}}{8(1+\kappa\vartheta)^2} \right\},$$

we have $G^{\text{ns}}(\widehat{\Gamma}) = G$.

Assuming that $k \geq n^\zeta$ *for some* $\zeta > 0$, $\log d = o(k/(\log k)^8)$ *and the quantities* $\lambda, \kappa, \vartheta, \eta^{\text{ns}}$ *are bounded away form zero and infinity, we have*

$$\mathbb{P}\big(G^{\text{ns}}(\widehat{\Gamma}) = G\big) \to 1, \quad n \to \infty,$$

provided that $\rho^{\text{ns}}_{\max} < \theta^{\text{ns}}_{\min}/(2\vartheta)$ *and* $(k/n)^\xi(\log(n/k))^2 + \sqrt{(\log d)/k} = o(\min(\rho^{\text{ns}}_{\min}, s^{-1}))$.

In the statement of the second part of the above theorem, we sacrificed generality for the sake of simplicity. Combining the first statement with the general concentration bounds on $\max_{m\in V} \left\|\widehat{\Gamma}^{(m)} - \Gamma^{(m)}\right\|_\infty$, which we derive in Section 3.4.2, one can obtain lower bounds on the probability of correct graph recovery that are explicit in all constants appearing above. We have opted against providing such explicit expressions because the resulting terms are lengthy and do not add much in terms of interpretability. In the same vein, the quantities $\lambda, \kappa, \vartheta, s, \eta^{\text{ns}}$ are for simplicity taken as the worst case over m, ℓ. It is possible to introduce versions of $\lambda, \kappa, \vartheta, s, \eta^{\text{ns}}$ that depend on m, ℓ. This would allow for sharper but more complex results; in particular, the incoherence parameters $\eta^{\text{ns}}_{m,1}, \ldots, \eta^{\text{ns}}_{m,d-1}$ would need to be non-negative for only half of the values $m \in V$. The precise form of this statement is immediate from a close look at the proof in Section 3.9.

For interpreting the second part of the above theorem, note that the quantity $r_{k,n} := (k/n)^\xi(\log(n/k))^2 + (k^{-1}\log d)^{1/2}$ is the order at which $\|\widehat{\Gamma} - \Gamma\|_\infty$ concentrates with $(k/n)^\xi(\log(n/k))^2$ corresponding to the bias and $(k^{-1}\log d)^{1/2}$ to the stochastic part; see Theorem 3.3 and the discussion right after for additional details. The quantity $\theta^{\text{ns}}_{\min}$ can be interpreted as minimal signal strength among edges that are present in the graph. In order for such edges to be recovered with high probability, we need a minimal signal condition $\theta^{\text{ns}}_{\min} \gg r_{k,n}$. Similarly, the maximal edge degree s must satisfy $s^{-1} \gg r_{k,n}$. In general, such conditions on minimal signal and maximal edge degrees are unavoidable for consistent graph recovery. Conditions that are similar in spirit were also imposed in Meinshausen and Bühlmann (2006) for neighborhood selection and Ravikumar et al. (2011) for the graphical lasso.

We next discuss guarantees on structure recovery using the graphical lasso as base

learner for Algorithm 1. Define the set of edges in the graph excluding edges containing node m, and augmented by self loops, by

$$S_m := \{(i,j) : i,j \in V \setminus \{m\}, \Theta_{ij} \neq 0\},$$

as well as its complement S_m^c taken in $(V \setminus \{m\})^2$. Let $\Omega^{(m)} := \Sigma^{(m)} \otimes \Sigma^{(m)}$ and define the quantities $\theta_{\min}^{\mathrm{gl}} := \min_{i \neq j : \Theta_{ij} \neq 0} |\Theta_{ij}|$, $\kappa_\Sigma := \max_m |||\Sigma^{(m)}|||_\infty$ and $\kappa_\Omega := \max_m |||(\Omega^{(m)}_{S_m,S_m})^{-1}|||_\infty$. The incoherence parameter for graphical lasso is defined as

$$\eta^{\mathrm{gl}} := \min_{m \in V} \eta_m^{\mathrm{gl}}, \quad \eta_m^{\mathrm{gl}} := 1 - |||\Omega^{(m)}_{S_m^c,S_m} (\Omega^{(m)}_{S_m,S_m})^{-1}|||_\infty. \tag{3.18}$$

Similarly to neighborhood selection, such incoherence parameters play a crucial role in guarantees for consistent support recovery by the Gaussian graphical lasso (Ravikumar et al., 2011).

Using graphical lasso as the base learner in Algorithm 1 requires the choice of d tuning parameters $\rho_1^{\mathrm{gl}}, \ldots, \rho_d^{\mathrm{gl}}$, one for each step of the loop over m. Define $\rho_{\min}^{\mathrm{gl}} := \min_{m \in V} \rho_m^{\mathrm{gl}}$ and $\rho_{\max}^{\mathrm{gl}} := \max_{m \in V} \rho_m^{\mathrm{gl}}$.

Theorem 3.2. *Assume (3.16) and that $\eta^{\mathrm{gl}} > 0$. Let $G^{\mathrm{gl}}(\widehat{\Gamma})$ denote the estimated graph obtained through the graphical lasso as base learner in Algorithm 1 with penalty parameters $\rho_1^{\mathrm{gl}}, \ldots, \rho_d^{\mathrm{gl}}$. Then, as soon as*

$$\|\widehat{\Gamma} - \Gamma\|_\infty < C^{\mathrm{gl}} := \frac{2}{3} \min \left\{ \min_{i,m \in V, i \neq m} \Sigma_{ii}^{(m)}, \frac{\eta^{\mathrm{gl}} \rho_{\min}^{\mathrm{gl}}}{8}, \frac{1}{\chi_0 s} - \rho_{\max}^{\mathrm{gl}}, \frac{\theta_{\min}^{\mathrm{gl}}}{4\kappa_\Omega} - \rho_{\max}^{\mathrm{gl}} \right\},$$

for $\chi_0 := 6\kappa_\Sigma \kappa_\Omega (1 \vee (9\kappa_\Sigma^2 \kappa_\Omega / \eta^{\mathrm{gl}}))$, we have $G^{\mathrm{gl}}(\widehat{\Gamma}) = G$.

Assuming that $k \geq n^\zeta$ for some $\zeta > 0$, $\log d = o(k/(\log k)^8)$ and the quantities $\min_{i \neq m} \Sigma_{ii}^{(m)}, \kappa_\Sigma, \kappa_\Omega, \eta^{\mathrm{gl}}$ are bounded away form zero and infinity, we have

$$\mathbb{P}\big(G^{\mathrm{gl}}(\widehat{\Gamma}) = G\big) \to 1, \quad n \to \infty,$$

provided that $\rho_{\max}^{\mathrm{gl}} < (2\chi_0 s)^{-1} \wedge (\theta_{\min}^{\mathrm{gl}} / 8\kappa_\Omega)$ and $(k/n)^\xi (\log(n/k))^2 + \sqrt{(\log d)/k} = o(\rho_{\min}^{\mathrm{gl}})$.

Similarly to the statements in Theorem 3.1, we opted for simplicity over generality. It is possible to obtain sharper statements for the second part by combining the statement in the first part with concentration results on the empirical variogram. Moreover, it suffices if a majority of the incoherence parameters $\eta_1^{\mathrm{gl}}, \ldots, \eta_d^{\mathrm{gl}}$ are non-negative.

One important difference between the assumptions for consistent graph estimation via neighborhood selection and graphical lasso lies in the definition of corresponding

incoherence parameters. While there are no general results stating that one parameter is always smaller than the other, simulations indicate that in the models we considered, the neighborhood selection incoherence parameter is more likely to be positive; see Section 3.5 for additional details. This is in line with the discussion in Ravikumar et al. (2011, Sections 3.1.1, 3.1.2) who show that in two examples, conditions required for consistent graph recovery via neighborhood selection are weaker than those required by the graphical lasso.

Similarly to $\theta_{\min}^{\mathrm{ns}}$, the quantity $\theta_{\min}^{\mathrm{gl}}$ corresponds to a minimal signal strength condition. Both quantities are of the same order provided that all diagonal entries of $\Theta^{(m)}$ are bounded away from zero and infinity for all m. In Theorem 3.2, we find that assuming all other parameters fixed, $\theta_{\min}^{\mathrm{gl}} \gg r_{k,n}$ and $s^{-1} \gg r_{k,n}$ are required for consistent graph recovery with graphical lasso as the base learner. This matches the requirements when neighborhood selection is the base learner; see the discussion following Theorem 3.1.

3.4.2 Concentration of the empirical variogram

In this section we present concentration results on the empirical extremal variogram $\widehat{\Gamma}^{(m)}$ in (3.18) when the data $\boldsymbol{X}$ is in the domain of attraction of an arbitrary multivariate Pareto distribution $\boldsymbol{Y}$ with extremal variogram matrices $\Gamma^{(m)}$, $m \in V$. The bound holds simultaneously for all the estimators $\widehat{\Gamma}^{(m)}$. In the Hüsler–Reiss case, where all the population matrices $\Gamma^{(m)}$ are equal, the same bound holds trivially for the combined empirical variogram $\widehat{\Gamma}$.

We first introduce some additional notation and technical assumptions. Define the R-function by

$$R(\boldsymbol{x}) = \lim_{q \downarrow 0} q^{-1} \mathbb{P}(F(\boldsymbol{X}) > 1 - q\boldsymbol{x}), \quad \boldsymbol{x} \in [0, \infty)^d. \tag{3.19}$$

This function can be recovered from the stable tail dependence function L appearing in (3.3) through simple manipulations with the inclusion-exclusion formula; for $d = 2$, their relationship simplifies to $R(x, y) = x + y - L(x, y)$. Additionally, we have $R(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{Y} > 1/\boldsymbol{x})/\mathbb{P}(Y_1 > 1)$. The R-function is a popular object in describing multivariate extreme value distributions and turns out to be convenient for our theoretical analysis. In what follows, for $J \subset V$, let R_J denote the R-function corresponding to $\boldsymbol{Y}_J$. When J is a pair or a triple, we write R_{ij} and R_{ijm} for $R_{\{i,j\}}$ and $R_{\{i,j,m\}}$.

Assumption 3.1 is already sufficient to derive concentration bounds for the empirical extremal variogram, but sharper results are possible if the function R has pairwise

densities satisfying a certain bound.

Assumption 3.2 (Bounds on densities)**.** *For each $i, j \in V, i \neq j$ the functions R_{ij} have mixed partial derivatives r_{ij} satisfying*

$$r_{ij}(x,y) := \frac{\partial^2}{\partial x \partial y} R_{ij}(x,y) \leq \frac{K(\beta)}{x^{\beta} y^{1-\beta}}, \quad (x,y) \in (0,\infty)^2,$$

for constants $K(\beta)$ and every $\beta \in [-\varepsilon, 1+\varepsilon]$, for some some $\varepsilon > 0$.

Remark. Lemma 3.3 shows that Assumption 3.2 is satisfied by any non-degenerate Hüsler–Reiss distribution; the value ε therein can be chosen arbitrarily large and the constant $K(\beta)$ will additionally depend on Γ. In addition, it is trivial to check that the assumption holds if the functions r_{ij} satisfy

$$r_{ij}(x, 1-x) \leq K_r (x(1-x))^{\varepsilon}, \quad x \in (0,1),$$

for some positive constants K_r and ε.

We are now ready to state the main result in this section.

Theorem 3.3. *Let Assumption 3.1 hold and $\zeta \in (0,1]$ be arbitrary. There exist positive constants C, c and M only depending on K, ξ and ζ such that for any $k \geq n^{\zeta}$ and $\lambda \leq \sqrt{k}/(\log n)^4$,*

$$\mathbb{P}\left(\max_{m \in V} \left\| \widehat{\Gamma}^{(m)} - \Gamma^{(m)} \right\|_{\infty} > C \left\{ \left(\frac{k}{n}\right)^{\xi} (\log(n/k))^2 + \frac{(\log(n/k))^2 (1+\lambda)}{\sqrt{k}} \right\} \right) \leq M d^3 e^{-c\lambda^2}.$$

If in addition Assumption 3.2 holds, there exists a positive constant $\bar{C}$ only depending on K, ξ, ζ, ε and $K(\beta)$ such that for any k and λ as above,

$$\mathbb{P}\left(\max_{m \in V} \left\| \widehat{\Gamma}^{(m)} - \Gamma^{(m)} \right\|_{\infty} > \bar{C} \left\{ \left(\frac{k}{n}\right)^{\xi} (\log(n/k))^2 + \frac{1+\lambda}{\sqrt{k}} \right\} \right) \leq M d^3 e^{-c\lambda^2}.$$

This theorem is the main technical result of this chapter, and the proof turns out to be surprisingly involved, especially the part establishing the sharper bound under Assumption 3.2. A major difficulty stems from the use of empirical distribution functions to normalize the margins. As mentioned previously, this result is of general interest in structure learning for extremes. It provides a crucial ingredient in the analysis of tree learning in Engelke and Volgushev (2020) and should also prove useful in other settings such as estimation of extreme value distributions under total positivity constraints as in Röttger et al. (2021).

The term $(k/n)^{\xi}(\log(n/k))^2$ appearing in the upper bound above results from an upper bound on the bias in estimating $\Gamma^{(m)}$ due to the fact that we only observe data

from the domain of attraction of $\boldsymbol{Y}$, rather than from the distribution of $\boldsymbol{Y}$ directly. The second term involving λ in both cases arises from the stochastic error. We note that k corresponds to the effective number of observations used for estimation of $\Gamma^{(m)}$, and in that sense the term $\lambda/\sqrt{k}$ appearing in the second tail bound corresponds to the typical $\sqrt{k}$ convergence rates in extreme value theory.

We close this section by discussing the relation between Assumption 3.1 and standard second order conditions on bounded sets.

Assumption 3.3 (Second order)**.** *The marginal distribution functions $F_1, \dots, F_d$ are continuous and there exist constants $\xi' > 0$ and $K' < \infty$ such that for all $J \subset V$, $|J| \in \{2, 3\}$, and $q \in (0, 1]$,*

$$\sup_{\boldsymbol{x} \in [0,1]^{|J|}} \left| q^{-1} \mathbb{P}(F_J(\boldsymbol{X}_J) > 1 - q\boldsymbol{x}) - R_J(\boldsymbol{x}) \right| \leq K' q^{\xi'}. \tag{3.20}$$

As we show below, this condition together with an assumption on the tails of R_{ij} implies the stronger Assumption 3.1 and vice versa.

Assumption 3.4 (Tail)**.** *There exist constants $\xi_T > 0$ and $K_T < \infty$ such that for all $i \neq j \in V$ and $q \in (0, 1]$,*

$$1 - R_{ij}(q^{-1}, 1) \leq K_T q^{\xi_T}. \tag{3.21}$$

The relation between the above conditions is summarized in the following Proposition.

Proposition 3.1. *If Assumption 3.1 holds then Assumption 3.3 holds with $K' = 2K$, $\xi' = \xi$ and Assumption 3.4 holds with $K_T = K$, $\xi_T = \xi$. Conversely, if Assumption 3.3 holds with K', ξ' and Assumption 3.4 holds with K_T, ξ_T, then Assumption 3.1 holds with $K = (K' + 2K_T)$, $\xi = \xi' \xi_T / (1 + \xi' + \xi_T)$.*

Note that Hüsler–Reiss distributions satisfy Assumption 3.4 for any $\xi' > 0$ provided that all entries of the matrix Γ are bounded away from zero and infinity (see Lemma 3.3). Therefore, for Hüsler–Reiss distributions Assumption 3.1 holds as soon as Assumption 3.3 is satisfied for a strictly larger exponent ξ'. Theorems 3.1 and 3.2 thus hold under the more standard Assumption 3.3 only.

3.5 Simulations

3.5.1 Simulation setup

We conduct several simulation studies to compare the performance and properties of different structure learning methods for extremal graphs. Two classes of Hüsler–Reiss

distributions $\boldsymbol{Y}$ are chosen as the true extremal graphical model. They are described below by first sampling a random graph structure $G = (V, E)$ and then generating a random parameter matrix Γ that factorizes on that graph. Using the exact method of Dombry et al. (2016), we then simulate n samples of a max-stable random vector $\boldsymbol{X}$ associated to $\boldsymbol{Y}$ (cf., Rootzén et al., 2018b), whose copula is

$$\mathbb{P}(F(\boldsymbol{X}) \leq \boldsymbol{x}) = \exp\{-L(-\log \boldsymbol{x})\}, \quad \boldsymbol{x} \in [0, \infty)^d,$$

where L is the stable tail dependence function of $\boldsymbol{Y}$ in (3.3). It is shown in Section 3.12.7 that this distribution satisfies Assumption 3.3 with $\xi' = 1$. Hence by Proposition 3.1 and Lemma 3.3, it satisfies Assumption 3.1 with any $\xi < 1$. In particular, it is in the domain of attraction of $\boldsymbol{Y}$.

As the first random graph $G = (V, E)$ we consider the Barabasi–Albert model denoted by $\mathrm{BA}(d, q)$, which is a preferential attachment model with d nodes and a degree parameter $q \in \mathbb{N}$ (Albert and Barabási, 2002). Figure 3.3 shows two examples in dimension $d = 100$, one for degree $q = 1$, which is a tree, and one for degree $q = 2$. In order to randomly generate a valid Hüsler–Reiss parameter matrix Γ on G, we use the scheme in Ying et al. (2021) to sample a weighted graph Laplacian matrix. The latter can be used as a Hüsler–Reiss precision matrix Θ, which then corresponds uniquely to a variogram matrix; see Section 3.2.3. More precisely, we sample for every undirected edge $(i, j) \in E$ of G an independent uniform random variable $U_{ij} \sim \mathrm{Unif}[2, 5]$, and define the matrix $W \in \mathbb{R}^{d \times d}$ by

$$W_{ij} = W_{ji} := \begin{cases} U_{ij} \text{ if } (i, j) \in E, \\ 0 \text{ otherwise.} \end{cases}$$

Let D be the diagonal degree matrix with entry D_{ii} given by the ith row sum of W, $i \in V$. The matrix $\Theta = D - W$ is called a weighted Laplacian matrix over the graph G and is a valid Hüsler–Reiss precision matrix (Röttger et al., 2021).

We note that this construction always results in a precision matrix satisfying

$$\Theta_{ij} \leq 0, \quad i, j \in V, i \neq j. \tag{3.22}$$

By Röttger et al. (2021, Lemma 4.5) this implies that the corresponding Hüsler–Reiss distribution is EMTP_2, a notion of positive dependence for multivariate Pareto distributions. While encountered frequently in multivariate extreme value models (see Röttger et al., 2021, Section 4), such positive dependence is not present in all Hüsler–Reiss distributions.

As a second model for G and Γ, we therefore consider a setup where $\Theta_{ij} > 0$ for

some $i, j \in V$. Note that in this case there is no canonical construction similar to the Laplacian matrix above. Instead, we consider a graph with n_C fully connected cliques $C_1, \ldots, C_{n_C}$, each consisting of n_N nodes. We assume that the only intersections of these cliques are between C_j and C_{j+1}, $j = 1, \ldots, n_C - 1$, and that each intersection consists of a single node. This results in a block graph G and a block structure of the precision matrix Θ (Hentschel, 2021); see the right-hand side of Figure 3.3 for a block graph with $n_C = 10$ and $n_N = 4$. On this extremal graph structure, it suffices to specify Γ_{C_j,C_j} on each clique C_j, and the remaining entries are implied by the conditional independence structure (Engelke and Hitz, 2020, Proposition 4). Following Hentschel (2021), we can construct a valid Γ_{C_j,C_j} matrix by taking any $(n_N \times n_N)$-dimensional covariance matrix S and projecting it by PSP, where $P = I_d - \mathbf{1}\mathbf{1}^\top/d$; see Section 3.2.3. For each clique we generate independently a correlation matrix S following the method in Joe (2006), whose off-diagonal entries have marginal Beta$(\alpha - 1 + n_N/2, \alpha - 1 + n_N/2)$ distributions rescaled to $(-1, 1)$, where $\alpha > 0$ is a parameter. We denote this block model for Γ by BM(n_C, n_N, α). It has dimension $d = n_N + (n_C - 1)(n_N - 1)$ and is parametrized by the number of cliques n_C, the number of nodes n_N per clique, and the parameter α governing the dependence inside the cliques. Figure 3.7 shows the proportion of positive off-diagonal entries of the Hüsler–Reiss precision matrix Θ corresponding to the block graph model BM$(6, 4, \alpha)$ for different α values. It can be seen that for increasing α, less positive values appear and the model becomes closer to EMTP$_2$.

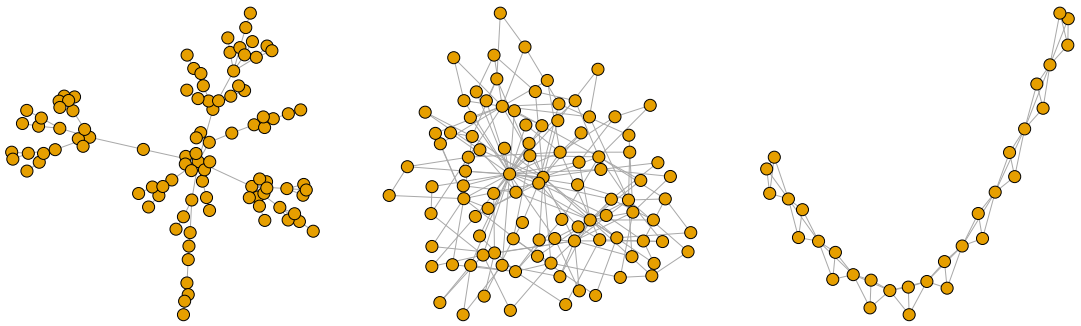

Figure 3.3: Realization of the Barabasi–Albert model of degree $q = 1$ (left) and $q = 2$ (center) in dimension $d = 100$, and of the block graph model in dimension $d = 31$.

3.5.2 Competing methods and evaluation

We apply several methods for structure estimation to the simulated data. All methods are based on the empirical extremal variogram $\widehat{\Gamma}$ defined in (3.15). For a given sample size n, this estimator uses only the k largest exceedances in each variable. Throughout

the simulation study, we choose $k = \lfloor n^{0.7} \rfloor$, which satisfies the assumptions for our theory. We refer to Engelke and Volgushev (2020, Section 5) for a detailed study of the choice of k in the framework of structure learning for extremal trees, which applies in the same manner here. In practice, k can be chosen based on stability plots of the estimated entries of $\widehat{\Gamma}$, similar to the Hill plot (e.g., Drees et al., 2000).

The first estimator of the extremal graph structure G is the extremal minimum spanning tree T_{mst} introduced in Engelke and Volgushev (2020). For a given estimate $\widehat{\Gamma}$ of the extremal variogram, it is defined as

$$T_{\text{mst}} = \underset{T=(V,E)}{\arg\min} \sum_{(i,j)\in E} \widehat{\Gamma}_{ij}, \tag{3.23}$$

where the minimum is taken over all tree structures T. In Engelke and Volgushev (2020) it is shown to consistently recover an underlying tree even in high dimensions. By construction, this always results in a tree and hence cannot be consistent for graphs that are not trees.

A second method is the estimator introduced in Röttger et al. (2021) who obtain an EMTP_2 estimator of Θ as the solution of the convex problem

$$-\log \det{}^*\Theta + \frac{1}{2} \operatorname{tr} \widehat{\Gamma}\Theta \tag{3.24}$$

over all Hüsler–Reiss precision matrices satisfying the EMTP_2 constraint (3.22). This method sometimes introduces sparsity, but it is important to note that it is not designed for structure estimation. While Röttger et al. (2021) show consistency of this estimator for the entries of Γ in a fixed–dimensional setting, there are no guarantees for consistent graph recovery of G, even if the true model is EMTP_2. The method should thus not be considered as a direct competitor but is included for comparison.

In Section 3.3.1 we introduced our `EGlearn` algorithm that uses majority voting for structure estimation of general extremal graphical models. It can either be combined with neighborhood selection or graphical lasso as the base learning method. Both methods depend on collections of tuning parameters, which are denoted as $\rho^{\text{ns}}_{m,\ell}$ and ρ^{gl}_{m}, respectively, $m \in V$, $\ell \in [d-1]$. We first set them all to the same value ρ to obtain a path of estimated graphs indexed by ρ, ranging from dense to sparse graphs for increasing values of ρ; see Figure 3.4 for typical paths for the two base learners. As a benchmark, we consider an oracle version of our estimator by selecting ρ to minimize the evaluation metric (3.25) below.

In practice, we need to select the amount of sparsity of the graph structure in a data driven way. We discuss automatic tuning of `EGlearn` with neighborhood selection as base learner, since it turns out to be superior to the graphical lasso alternative in

all the settings that we consider. At the mth step of the algorithm, d different lasso regressions are produced by solving (3.13) with $\widehat{A} := \widehat{\Sigma}^{(m)}_{\setminus m,\setminus m}$. For each m and ℓ, we define a surrogate for the deviance as

$$\mathrm{dev}_{m,\ell}(\theta) := k \log\left(\widehat{A}_{\ell\ell} - 2\widehat{A}_{\ell,\setminus\ell}\theta + \theta^\top \widehat{A}_{\setminus\ell,\setminus\ell}\theta\right) - k \log k.$$

Among a path of solutions $\widehat{\theta}$ of (3.13) indexed by the choice of $\rho^{\mathrm{ns}}_{m,\ell}$, the AIC, BIC and MBIC tuning strategies select the value $\widehat{\theta}$ minimizing

$$\mathrm{dev}_{m,\ell}(\widehat{\theta}) + 2\|\widehat{\theta}\|_0, \quad \mathrm{dev}_{m,\ell}(\widehat{\theta}) + (\log k)\|\widehat{\theta}\|_0, \quad \mathrm{dev}_{m,\ell}(\widehat{\theta}) + (\log k)(\log\log(d-1))\|\widehat{\theta}\|_0,$$

respectively, where $\|\widehat{\theta}\|_0$ is the number of non-zero elements in $\widehat{\theta}$. They are motivated by the traditional Akaike information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz, 1978), and on an extension of the latter to high-dimensional models developed in Wang et al. (2009).

In order to compare an estimated graph $\widehat{G} = (V, \widehat{E})$ with the true underlying graph $G = (V, E)$, we use as evaluation metric the F-score. It is defined as

$$F = \frac{|E \cap \widehat{E}|}{|E \cap \widehat{E}| + \frac{1}{2}(|E^c \cap \widehat{E}| + |E \cap \widehat{E}^c|)}, \tag{3.25}$$

where for a set of edges E, the set E^c denotes all possible undirected edges on $V \times V$ except for those in E. The F-score consists precisely of the harmonic mean between the precision and the recall.

3.5.3 Results

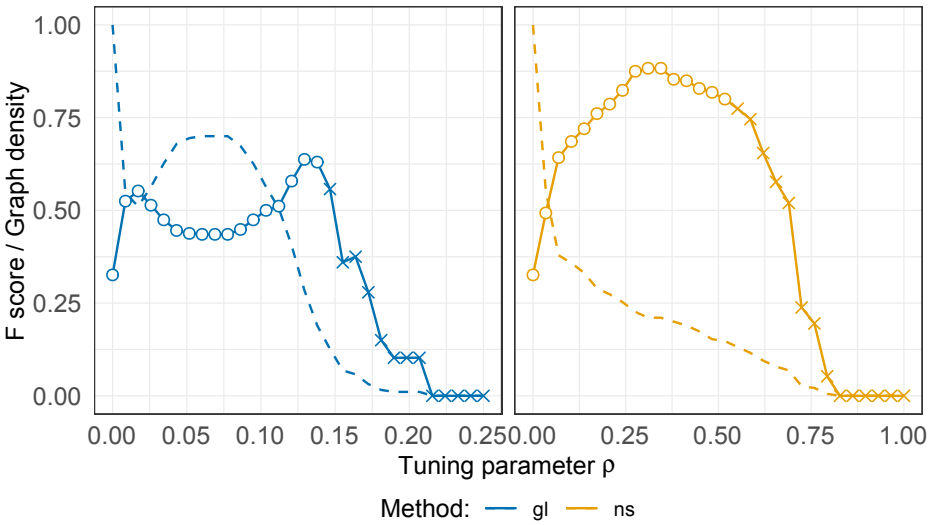

Figure 3.4: Paths for `EGlearn` with grahpical lasso (left) and neighborhood selection (right) as a function of the common tuning parameter ρ, fitted to data from the BA$(20, 2)$ model with $k = 5d$. Circles indicate that the estimated graph $\widehat{G}$ is connected, and crosses correspond to unconnected graphs. Dashed lines show the density of the graph, that is, the proportion of existing edges in $\widehat{G}$ out of all $d(d-1)/2$ possible edges.

The first set of simulations assesses the performance in terms of the F-score of the different methods described in the previous section. We simulate n samples from the Hüsler–Reiss distribution generated according to the Barabasi–Albert model BA(d, q) of degrees $q \in \{1, 2\}$ and in dimensions $d = \{20, 50, 100\}$. For a given dimension d, we simulate three different sample sizes n such that the number of exceedances satisfy $k/d \in \{0.5, 1, 2.5\}$ and $k/d \in \{0.5, 1, 5\}$ for the models with $q = 1$ and $q = 2$, respectively. The case of $k/d = 0.5$ is a high-dimensional setting since the number of effective samples is smaller than the dimension. The different methods are then applied based on the empirical variogram $\widehat{\Gamma}$ as described above using $k = \lfloor n^{0.7} \rfloor$ exceedances in each variable. Figure 3.4 shows typical paths of F-scores and densities of estimated graphs for `EGlearn` for the two base learners.

The results for dimensions 20 and 100 are shown in Figures 3.5 and 3.6 as boxplots of the F-scores for 100 repetitions of each experiment. Similar results are obtained in dimension 50 in Section 3.8.1. The minimum spanning tree in (3.23) outperforms the other methods when the underlying model is indeed a tree ($q = 1$). This is expected since the approach takes advantage of the special structure of a tree. For the more general Barabasi–Albert graph with degree $q = 2$ the minimum spanning tree is no longer consistent and even with larger sample sizes the F-scores stay bounded below

a certain level. The EMTP_2 estimator does not recover the graph structure well. As discussed in the previous section, this is not surprising, since there are no guarantees concerning structure estimation in this method, even if the true model is EMTP_2.

Turning to the methods that are designed to estimate extremal graphical structures for general graphs G, we first observe that `EGlearn` with grahpical lasso as the base learner does not perform well on any of the simulations, even with the oracle value for the penalty paramter ρ. This is surprising since the graphical lasso for Gaussian distributions is a well established method; we discuss this phenomenon in more detail below. On the other hand, `EGlearn` with neighborhood selection performs very well and seems to consistently recover the graph in all of the setups for large enough sample sizes.

Since the `EGlearn` with graphical lasso is generally not consistent even with oracle tuning parameter, we only consider data driven selection of the tuning parameters in the case of neighborhood selection. Figures 3.5 and 3.6 show the performance of this method for model selection based on the AIC, BIC and MBIC as described in Section 3.5.2. We observe that AIC does not work well for selection of the penalization parameter. This is to be expected since the AIC is too conservative to result in consistent model selection even in classical settings (Arlot and Celisse, 2010). On the other hand, BIC and MBIC behave similarly and both produce structure estimates that are fairly close to the oracle estimator.

We next run simulations with Hüsler–Reiss distributions generated according to the block model $\text{BM}(6, 4, \alpha)$, which results in $d = 19$ nodes in the graph. For the dependence parameter we choose a sequence of values $\alpha \in \{0.1, 1, 2, 10, 20\}$. As before we use $k = \lfloor n^{0.7} \rfloor$ exceedances and we simulate two different sample sizes n such that $k/d \in \{2, 10\}$. The results for 100 repititions are shown in Figure 3.7. The top right panel shows boxplots of the F-scores for the oracle `EGlearn` with graphical lasso and neighborhood selection as a functon of α. We observe that again, `EGlearn` with graphical lasso base learner does not seem to be consistent since even with the larger sample size the F-scores do not improve much. On the contrary, `EGlearn` with neighborhood selection performs well especially for larger samples sizes, suggesting consistency of the method. We further observe that in general, smaller values of α correspond to more difficult estimation problems. Note that this corresponds to the case of higher proportions of positive entries Θ_{ij} in the Hüsler–Reiss precision matrix (top left panel of Figure 3.7).

To understand this behaviour and the related phenomenon that the graphical lasso as base learner does not seem to work well, we take a closer look at the assumptions for consistent structure recovery by `EGlearn` in Theorems 3.1 and 3.2. For a given

parameter matrix Γ, a crucial requirement for both neighborhood selection and graphical lasso as base learner is the positivity of the incoherence parameters η^{ns} and η^{gl}, respectively. The bottom panels of Figure 3.7 show boxplots of these parameters for the generated block models. All incoherence parameters η^{ns} for neighborhood selection are positive and thus, Theorem 3.1 guarantees consistent graph recovery. We also note that as α increases, so does η^{ns}, and the graph recovery results improve. This is in line with our theory; the expression of C^{ns} in Theorem 3.1 suggests that a higher η^{ns} increases the probability of graph recovery. On the other hand, all incoherence parameters η^{gl} are negative and Theorem 3.2 is not applicable. More generally, for all the simulation settings we have considered, the neighborhood selection incoherence parameter η^{ns} is much more likely to be positive than its graphical lasso equivalent η^{gl}. It thus appears that the assumption of Theorem 3.1 is significantly weaker than that of Theorem 3.2. This is also consistent with results in the literature of Gaussian structure learning (Ravikumar et al., 2011, Sections 3.1.1, 3.1.2).

Since Hüsler–Reiss graphical models are not defined on disconnected graphs, subsequent inference for Γ on $\widehat{G}$ requires that the estimated graph is connected. Theorem 3.1 and 3.2 ensure that the paths obtained by `EGlearn` include the true, connected graph with high probability. In finite samples with data-driven penalty parameter, it can however happen that the selected graph is disconnected. Focusing again on the case of neihborhood selection, we observe that in all our simulations, the estimated paths are monotone in the sense that the graphs are nested in one another; no edge can enter the model when increasing the penalty parameter (see, e.g., the right panel of Figure 3.4). If the selected graph is disconnected, we therefore propose to use the connected graph with the largest penalty parameter ρ. Table 3.1 in Section 3.8.2 shows that the performance loss in terms of F-score due to this strategy is negligible, except for very sparse models such as trees, where sometimes disconnected graphs $\widehat{G}$ may be selected that are very similar to G. See the discussion and simulations therein for more details.

Note that the same strategy would not be sensible for `EGlearn` with the graphical lasso as base learner, since the path of the graph density is not monotone; see left panel of Figure 3.4.

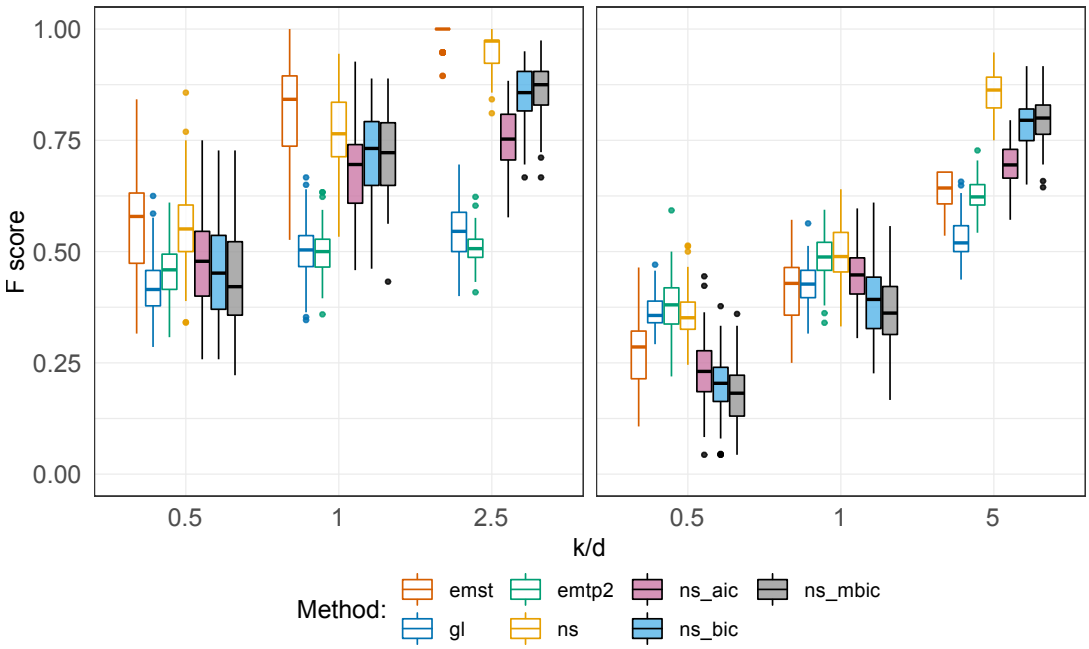

Figure 3.5: Results of 100 repetitions in dimension $d = 20$ and degree $q = 1$ (left) and $q = 2$ (right).

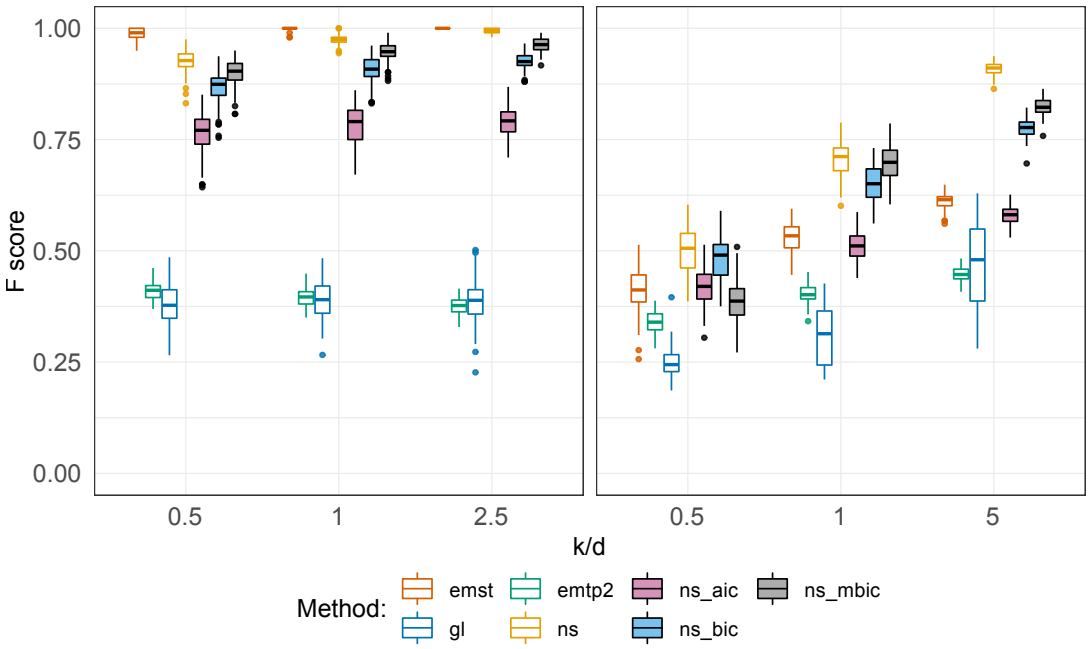

Figure 3.6: Results of 100 repetitions in dimension $d = 100$ and degree $q = 1$ (left) and $q = 2$ (right).

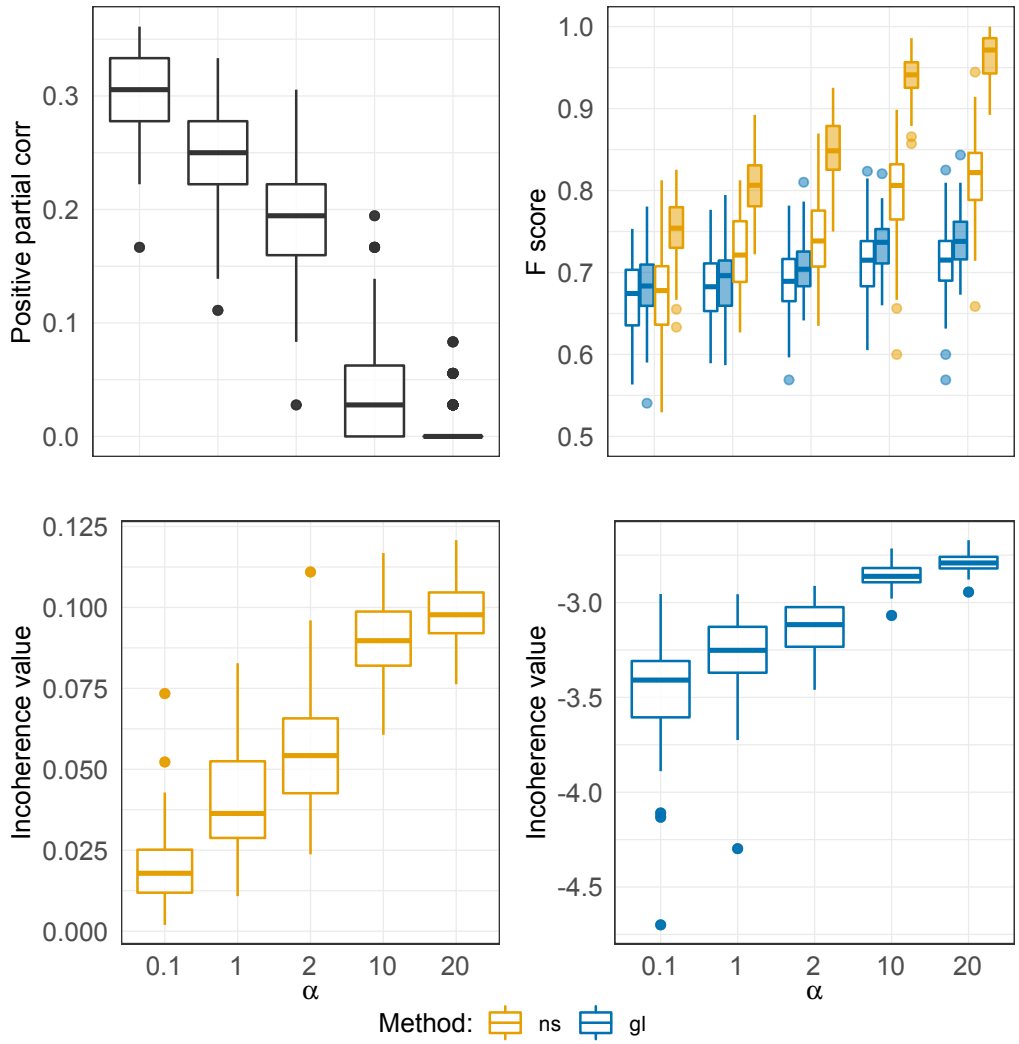

Figure 3.7: Results for the BM$(6, 4, \alpha)$ model for different α. Top left: proportion of positive off-diagonal entries of the Hüsler–Reiss precision matrix Θ; top right: F-score of `EGlearn` with graphical lasso (blue) and neighborhood selection (yellow) with $k = 2d$ and $k = 10d$ for empty and filled boxes, respectively; bottom left: parameter η^{ns}; bottom right: parameter η^{gl}.

3.6 Application

We use daily discharge data collected at $d = 31$ stations in the upper Danube basin (Asadi et al., 2015) to illustrate the proposed algorithm. The original data set spans over 50 years, but after removing seasonality, declustering and "aligning" the d time series, Asadi et al. (2015) are left with $n = 428$ observations. We use their preprocessed version of the data set. Considering the discharges as observations from a d-dimensional random vector $\boldsymbol{X}$, we are interested in the dependence structure between floodings along the basin, represented by the tail dependence of $\boldsymbol{X}$.

A certain amount of information can be deduced from the physical properties of the river network itself: the flow connections are known, and are graphically represented in the left panel of Figure 3.8. It is reasonable to expect that extremal dependence is strong along those connections. However, due to other geographical features (e.g., small Euclidean distance between some stations that are otherwise not directly flow-connected), the extremal dependence may be too complex to be represented by this simple tree structure.

We first fix the number k of tail observations to be used in each variable as $k = 42$, which corresponds to only using data where one component exceeds it's marginal 90th percentile. This choice coincides with that of Engelke and Hitz (2020), Section 6, who also fit a Hüsler–Reiss graphical model to this data, although their methods only allow for the limited structure of block graphs. Using the same threshold exceedances will allow us to compare our results with the aforementioned paper. Using `EGlearn` with neighborhood selection and a grid of 100 equidistant values of ρ between 0 and 0.5, we then compute a path of estimated extremal graphs for this data. We find that as soon as ρ exceeds 0.1, the estimated graph is disconnected. The center panel of Figure 3.8 shows the sparsest estimated graph that is connected, corresponding to $\rho = 0.1$. We also compute the AIC, BIC and MBIC graphs, as defined in Section 3.5. It turns out that those three graphs are disconnected, although the AIC graph is merely two edges away from being connected. See Figure 3.10.

The structure learning methods presented in this chapter output an estimated graph $\widehat{G} = (V, \widehat{E})$. They do not automatically provide an estimator of the Hüsler–Reiss parameter matrix Γ on this graph structure. A natural estimator $\widehat{\Gamma}^0$ that has the sparsity pattern of $\widehat{G}$ and agrees with the input estimator $\widehat{\Gamma}$ (here, the empirical variogram) on the edges of $\widehat{G}$ is the solution to the matrix completion problem

$$\begin{aligned} \widehat{\Gamma}^0_{ij} &= \widehat{\Gamma}_{ij}, \quad (i,j) \in \widehat{E}, \\ \widehat{\Theta}^0_{ij} &= 0, \quad (i,j) \notin \widehat{E}, \end{aligned}$$

where $\widehat{\Theta}^0$ is the Hüsler–Reiss precision matrix corresponding to $\widehat{\Gamma}^0$. Hentschel (2021) show that there is a unique solution to this completion problem, hence yielding a unique Hüsler–Reiss model fit along each estimate graph. Using this procedure, we are able to obtain estimated models for each connected graph in the estimated path and measure the model fit via, for instance, the Hüsler–Reiss log-likelihood. For means of comparison, we calculate the AIC of each such model, as defined in Engelke and Hitz (2020), and plot them in the right panel of Figure 3.8 as a function of the number of edges in the estimated graph. We obtain a path of increasingly complex estimated

models, seven of which outperform the block graph model obtained by Engelke and Hitz (2020) via forward selection (blue dashed line). The latter itself outperforms the hybrid spatial model of Asadi et al. (2015) (orange dashed line).

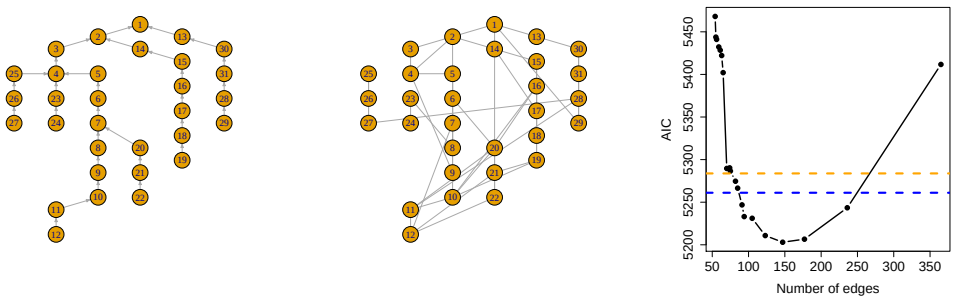

Figure 3.8: Left: graphical representation of the river flow across the 31 stations. Center: the most sparse connected graph in the estimation path ($\rho = 0.1$). Right: AIC of the estimated Hüsler–Reiss models as a function of the number of edges in the estimated graph. AIC of the best models estimated in Engelke and Hitz (2020) and Asadi et al. (2015) are in blue and orange, respectively.

3.7 Extensions and future work

In the present chapter, we have introduced a general methodology for estimating Hüsler–Reiss graphical models through `EGlearn` and provided a thorough theoretical analysis of the resulting procedure. This is the first principled approach for estimating extremal graphical models on arbitrary connected graphs, and there are many questions warranting further investigation.

A first direction is a systematic exploration of alternative base learners in `EGlearn`. We have focused on the two most popular and classical approaches, but many more possible choices exist; see Drton and Maathuis (2017, Chapter 3) and the references cited therein for a partial overview of the recent literature with a focus on graphical modeling.

Similarly, in this chapter we have used the empirical variogram, but different estimators of the variogram matrix Γ could be considered. For instance, one could consider method of moments or M-estimators for extremes (Einmahl et al., 2008, 2012). This could be especially interesting in the high-dimensional regime, where it might be possible to obtain estimators of the variogram matrix with better concentration properties than the empirical variogram.

Lastly, we discuss alternatives to the `EGlearn` algorithm. In view of the charaterization of extremal conditional independence for Hüsler–Reiss distributions in (3.10), a promising direction for future research is to consider methods that simultaneously

penalize the entries and row sums of $\Theta^{(m)}$. More precisely, for an arbitrary $m \in V$, one may attempt to estimate $\Theta^{(m)}$ by solving the doubly penalized graphical lasso problem

$$\underset{\substack{\Theta^{(m)} \in \mathcal{S}^d \\ \Theta^{(m)}_{mV} = \mathbf{0}}}{\arg\min} \Big\{ -\log\det{}^*\Theta^{(m)} + \operatorname{tr}(\widehat{\Sigma}^{(m)}\Theta^{(m)}) + \rho \sum_{i \neq j} \sum |\Theta^{(m)}_{ij}| + 2\rho \sum_i \Big| \sum_j \Theta^{(m)}_{ij} \Big| \Big\}, \quad (3.26)$$

where $\widehat{\Sigma}^{(m)}$ is defined as $\varphi_m(\widehat{\Gamma})$ for some estimate $\widehat{\Gamma}$ of the extremal variogram, where φ_m is as in (3.8). Here, $\det^*$ denotes the pseudo-determinant of a matrix (the product of its non-zero eigenvalues) and $\mathcal{S}^d$ is the space of symmetric positive semi-definite matrices in $\mathbb{R}^{d\times d}$. The second penalty term is used to impose sparsity of the row and column sums of $\Theta^{(m)}$, in addition to sparsity in the off-diagonal entries themselves. The motivation for such an approach is that zero row sums of $\Theta^{(m)}$ contain information on the absence of edges containing the node m; see (3.10).

Alternatively, in view of the charaterization in (3.12), one may consider a sparse estimate of the positive semi-definite matrix Θ defined in (3.11) by the modified graphical lasso problem

$$\underset{\Theta \in \mathcal{S}^d_{\mathbf{1}}}{\arg\min} \Big\{ -\log\det{}^*\Theta + \operatorname{tr}(\widehat{\Sigma}\Theta) + \rho \sum_{i \neq j} \sum |\Theta_{ij}| \Big\}, \quad (3.27)$$

where $\mathcal{S}^d_{\mathbf{1}}$ is the cone of symmetric, positive semi-definite matrices with rank $d-1$ and row sums equal to zero. The estimator $\widehat{\Sigma}$ is defined as a transformation of the extremal variogram estimator through $\widehat{\Sigma} = P(-\widehat{\Gamma}/2)P$, where $P = I_d - \mathbf{1}\mathbf{1}^\top/d$ as in Section 3.2.3. While this is a more symmetric approach, the difficulty is the semi-definiteness of Θ.

For any $m \in V$, the matrix $\Theta^{(m)}$ uniquely defines the variogram Γ, and hence characterizes the Hüsler–Reiss model. The solution to (3.26), an estimate of $\Theta^{(m)}$, can therefore be transformed to a variogram estimate $\widehat{\Gamma}^{(m)}_\rho$. Similarly, the solution to (3.27) is an estimate of Θ and can be uniquely transformed into a variogram estimate $\widehat{\Gamma}_\rho$. Interestingly, if the input $\widehat{\Gamma}$ of these optimization problems is the same (for example, the empirical variogram), then regardless of the choice of m, (3.26) and (3.27) result in the same estimated model.

Proposition 3.2. *Let $\widehat{\Gamma}$ be an arbitrary estimator of the extremal variogram matrix and suppose that (3.26) and (3.27) are solved with inputs $\widehat{\Sigma}^{(m)} = \varphi_m(\widehat{\Gamma})$ and $\widehat{\Sigma} = P(-\widehat{\Gamma}/2)P$, respectively, and with constant penalty parameter $\rho > 0$. Then the*

corresponding estimated models are all the same, i.e.,

$$\widehat{\Gamma}_\rho^{(1)} = \cdots = \widehat{\Gamma}_\rho^{(d)} = \widehat{\Gamma}_\rho.$$

While the above approaches are attractive, preliminary simulations indicate that it does in general not lead to consistent recovery of the sparsity structure of the matrices $\Theta^{(m)}$ and Θ. This is in line with theoretical results obtained by Ying et al. (2020a), who investigate Laplacian constrained degenerate Gaussian distributions. Replacing the vanilla L^1 penalties by adaptively weighted (Ying et al., 2021) or non-convex (Ying et al., 2020b) versions might provide a way out.

3.8 Additional numerical results

3.8.1 Simulation results for the $\mathrm{BA}(50, q)$ model

The results of our experiments on the Barabasi–Albert model of dimension 50 are presented in Figure 3.9. The conclusions are similar to the 100-dimensional case: with neighborhood selection as base learner, the algorithm has a seemingly consistent behavior as $k \to \infty$, as opposed to the graphical lasso case.

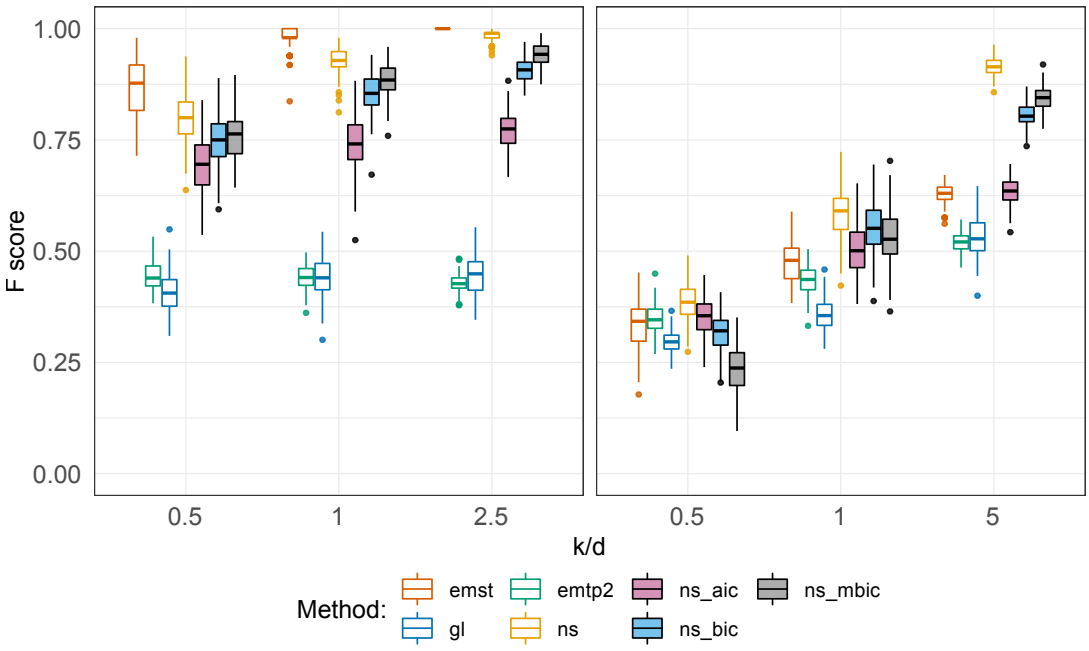

Figure 3.9: Results of 100 repetitions in dimension $d = 50$ and degree $q = 1$ (left) and $q = 2$ (right).

3.8.2 Connectedness

The graph estimated by the `EGlearn` algorithm with neighborhood selection as its base learner is not guaranteed to be connected in general. In Section 3.5, we suggested that when the selected graph is not connected, one can compute a path of solutions using $\rho^{\text{ns}}_{m,\ell} \equiv \rho$, for a grid of values ρ, and select the sparsest connected graph in the path.

Note that as illustrated in Section 3.6, each estimated graph in a path can be "completed" into a unique Hüsler–Reiss estimated model. A strategy to ensure connectedness, not investigated here, is to calculate the AIC for those models and maximize that measure, say, over all the connected graphs in the path.

To quantify the effect of the first strategy above, we study three specific scenarios: the models BA$(50, 1)$, BA$(50, 2)$ and BM$(10, 4, 2)$ (with dimension $d = 31$), which have been used above. For each model, we simulate 100 samples of size n such that $k := \lfloor n^{0.7} \rfloor$ is equal to $d/2$, d and to $5d$; we omit $k = d/2$ for the block model, since the small sample size $k = 15$ creates some degeneracies. For each model and sample, a path of estimated graphs is obtained according to `EGlearn` with neighborhood selection as base learner, and with 100 equidistant values of ρ between 0 and 1. We then calculate the ratio of the maximum F-score among connected graphs in the path to the unconstrained maximum F-score. The average ratio for each setup is presented in Table 3.1, with standard deviation. In addition, we calculate the proportion of samples where the oracle and the AIC, BIC and MBIC graphs (as defined in Section 3.5) are connected. These proportions are found in the last four columns of Table 3.1.

The BIC and MBIC graphs are frequently disconnected if the sample size is small, whereas the AIC, being more conservative, tends to avoid that problem. However, the performance of the algorithm is generally stable over a certain range of tuning parameters, and the loss incurred by only optimizing over the connected graphs is negligible. The largest loss of performance in terms of F-score is of the order of 12% in the BA$(50, 1)$ model. This graph is a tree, the most sparse connected graph structure, which explains why it is well approximated by certain disconnected estimates. For the other, denser models, the performance loss is less than 1%.

		F-score ratio	oracle	AIC	BIC	MBIC
BA(50, 1)	$k = d/2$	0.873 (0.066)	0.02	0.56	0.02	0
	$k = d$	0.894 (0.075)	0.09	0.87	0.41	0.13
	$k = 5d$	0.997 (0.018)	0.91	1	0.99	0.99
BA(50, 2)	$k = d/2$	0.992 (0.02)	0.72	0.86	0.04	0
	$k = d$	0.996 (0.01)	0.78	0.99	0.56	0.01
	$k = 5d$	1 (0)	1	1	1	1
BM(10, 4, 2)	$k = d$	0.996 (0.016)	0.87	0.91	0.66	0.42
	$k = 5d$	0.999 (0.003)	0.98	1	0.98	0.97

Table 3.1: Average ratio of the best F-score among connected graphs in the path obtained by `EGlearn` with neighborhood selection, to the best F-score in the whole path, based on 100 simulations. Standard deviations in parentheses. The last four columns contain the proportions of samples where the oracle, AIC, BIC and MBIC graphs are connected.

3.8.3 AIC and BIC estimated graphs from the Danube data

Figure 3.10 contains the graphs obtained by running `EGlearn` on the Danube data set, with neighborhood selection as the base learner. The empirical variogram is calculated with $k = 42$, and the AIC, BIC and MBIC are defined as in Section 3.5.

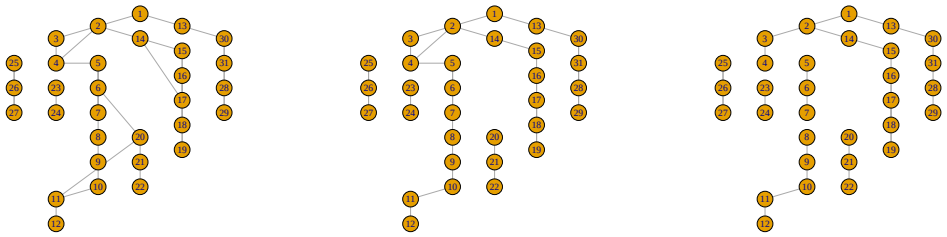

Figure 3.10: The AIC (left), BIC (center) and MBIC (right) graphs based on the Danube data.

3.9 Proofs of extremal graph recovery results

3.9.1 Proof of Theorem 3.1

When running the neighborhood selection algorithm on $\Sigma^{(m)}_{\backslash m,\backslash m}$, $m \in V$, recall that $\rho^{\mathrm{ns}}_{m,1}, \ldots, \rho^{\mathrm{ns}}_{m,d-1}$ are used as penalty parameters. Let

$$\lambda_{m,\ell} := \lambda_{\min}(\Sigma^{(m)}_{S_{m,\ell},S_{m,\ell}}),$$
$$\kappa_{m,\ell} := \left|\left|\left|\Sigma^{(m)}_{S^c_{m,\ell},S_{m,\ell}}\right|\right|\right|_\infty,$$

$$
\begin{aligned}
\vartheta_{m,\ell} &:= \left|\!\left|\!\left| (\Sigma^{(m)}_{S_{m,\ell},S_{m,\ell}})^{-1} \right|\!\right|\!\right|_{\infty}, \\
s_{m,\ell} &:= |S_{m,\ell}|, \\
C^{\mathrm{ns}}_{m,\ell} &:= \frac{2}{3} \min \left\{ \frac{\lambda_{m,\ell}}{2 s_{m,\ell}}, \frac{\eta^{\mathrm{ns}}_{m,\ell}}{4\vartheta_{m,\ell}(1+\kappa_{m,\ell}\vartheta_{m,\ell}) s_{m,\ell}}, \right. \\
&\qquad \left. \frac{\min_{i \in S_{m,\ell}} \frac{|B_{i\ell}|}{B_{\ell\ell}} - \vartheta_{m,\ell}\rho^{\mathrm{ns}}_{m,\ell}}{2\vartheta_{m,\ell}(1+\kappa_{m,\ell}\vartheta_{m,\ell})}, \frac{\rho^{\mathrm{ns}}_{m,\ell}\eta^{\mathrm{ns}}_{m,\ell}}{8(1+\kappa_{m,\ell}\vartheta_{m,\ell})^2} \right\}.
\end{aligned}
$$

By Proposition 3.3, we have

$$
\{(i,j) \in (V \setminus \{m\})^2 : \widetilde{Z}^{(m)} = 1\} = \{(i,j) \in (V \setminus \{m\})^2 \cap E\}
$$

provided that

$$
\|\widehat{\Sigma}^{(m)} - \Sigma^{(m)}\|_{\infty} < \frac{3}{2} \min_{\ell \neq m} C^{\mathrm{ns}}_{m,\ell}.
$$

From the definition of $\Sigma^m, \widehat{\Sigma}^{(m)}$ through $\Gamma, \widehat{\Gamma}$, it follows that

$$
\max_{m \in V} \|\widehat{\Sigma}^{(m)} - \Sigma^{(m)}\|_{\infty} \leq \frac{3}{2} \|\widehat{\Gamma} - \Gamma\|_{\infty}.
$$

The `EGlearn` algorithm correctly learns the presence (resp. absence) of an edge (i,j) if a 1 (resp. a 0) rightfully appears in position (i,j) of $\widetilde{Z}^{(m)}$ for at least $\lfloor d/2 \rfloor$ values of $m \notin \{i,j\}$. This is clearly the case for each pair (i,j) if at least $\lfloor d/2 \rfloor + 2$ of the neighborhood selections perfectly succeed, which is guaranteed if $\|\widehat{\Gamma} - \Gamma\|_{\infty} < \min_{\ell \neq m} C^{\mathrm{ns}}_{m,\ell}$ for at least $\lfloor d/2 \rfloor + 2$ values of m. Noting that the assumed bound C^{ns} lower bounds each $C^{\mathrm{ns}}_{m,\ell}$, the first statement of the theorem is proved.

For a proof of the second part, let

$$
\lambda := \sqrt{\tfrac{3}{c} \log d + \sigma}
$$

with c as in Theorem 3.3, and $\sigma \to \infty$ but $\sigma = o(k/(\log k)^8)$. Then, by assumption on d and k, we have eventually $\lambda \leq \sqrt{k}/(\log n)^4$, so that Theorem 3.3 applies. It states that with probability at least $1 - M \exp\{3 \log d - c\lambda^2\} = 1 - o(1)$,

$$
\|\widehat{\Gamma} - \Gamma\|_{\infty} \lesssim \left(\frac{k}{n}\right)^{\xi} (\log(n/k))^2 + \frac{\sqrt{\log d + \sigma}}{\sqrt{k}}.
$$

By assumption,

$$
\left(\frac{k}{n}\right)^{\xi} (\log(n/k))^2 + \frac{\sqrt{\log d}}{\sqrt{k}} = o(C^{\mathrm{ns}}),
$$

so if σ is chosen to diverge slowly enough, we have $\mathbb{P}(\|\widehat{\Gamma} - \Gamma\|_{\infty} < C^{\mathrm{ns}}) \to 1$. □

3.9.2 Proof of Theorem 3.2

When running the graphical lasso algorithm on $\Sigma^{(m)}_{\backslash m,\backslash m}$, $m \in V$, recall that ρ^{gl}_m is used as penalty parameter. Let

$$\kappa_{\Sigma,m} := |||\Sigma^{(m)}|||_\infty, \quad \kappa_{\Omega,m} := |||(\Omega^{(m)}_{S_m,S_m})^{-1}|||_\infty, \quad \chi_m := 6\kappa_{\Sigma,m}\kappa_{\Omega,m}\Big(1 \vee \frac{9\kappa^2_{\Sigma,m}\kappa_{\Omega,m}}{\eta^{\mathrm{gl}}_m}\Big),$$

$$s_m := \max_{\ell\in[d-1]} s_{m,\ell},$$

$$C^{\mathrm{gl}}_m := \frac{2}{3}\min\Big\{\min_{i\in[d-1]}\Sigma^{(m)}_{ii}, \frac{\eta^{\mathrm{gl}}_m\rho^{\mathrm{gl}}_m}{8}, \frac{1}{\chi_m s_m} - \rho^{\mathrm{gl}}_m, \frac{\theta^{\mathrm{gl}}_{\min}}{4\kappa_{\Omega,m}} - \rho^{\mathrm{gl}}_m\Big\}.$$

By Proposition 3.4, we have

$$\{(i,j) \in (V\setminus\{m\})^2 : \widetilde{Z}^{(m)} = 1\} = \{(i,j) \in (V\setminus\{m\})^2 \cap E\}$$

provided that

$$\|\widehat{\Sigma}^{(m)} - \Sigma^{(m)}\|_\infty < \frac{3}{2}C^{\mathrm{gl}}_m.$$

Similarly to the proof of Theorem 3.1, deduce that whenever $\|\widehat{\Gamma} - \Gamma\|_\infty < C^{\mathrm{gl}}_m$ for at least $\lfloor d/2\rfloor + 2$ values of m, `EGlearn` correctly recovers the extremal graph G. Moreover, $C^{\mathrm{gl}} \le C^{\mathrm{gl}}_m$ for each $m \in V$, which completes the proof of the first part. The proof of the second statement is identical to that of the second statement in Theorem 3.1. □

3.9.3 Consistency of neighborhood selection and graphical lasso

Let A be a p-dimensional covariance matrix, $B := A^{-1}$, $\widehat{A}$ be an estimator and $\varepsilon := \|\widehat{A} - A\|_\infty$. Recall the definition of the graph $G(B)$ associated to B, i.e., the graph with edge set $\{(i,j) : i \ne j, B_{ij} \ne 0\}$.

We start by discussing the neighborhood selection algorithm. Define the neighborhood of a node ℓ in the graph $G(B)$ by $\mathrm{ne}(\ell) := \{i \in [p]\setminus\ell : B_{i\ell} \ne 0\}$. Let

$$\begin{aligned}
\lambda_\ell &:= \lambda_{\min}(A_{\mathrm{ne}(\ell),\mathrm{ne}(\ell)}),\\
\kappa_\ell &:= |||A_{\mathrm{ne}(\ell)^c,\mathrm{ne}(\ell)}|||_\infty,\\
\vartheta_\ell &:= |||(A_{\mathrm{ne}(\ell),\mathrm{ne}(\ell)})^{-1}|||_\infty,\\
s_\ell &:= |\mathrm{ne}(\ell)|,\\
\eta_\ell &:= 1 - |||A_{\mathrm{ne}(\ell)^c,\mathrm{ne}(\ell)}(A_{\mathrm{ne}(\ell),\mathrm{ne}(\ell)})^{-1}|||_\infty.
\end{aligned}$$

Proposition 3.3. *Assume that* $\min_{\ell\in[p]}\eta_\ell > 0$ *and*

$$\varepsilon < \min_{\ell\in[p]} \min\left\{\frac{\lambda_\ell}{2s_\ell}, \frac{\eta_\ell}{4\vartheta_\ell(1+\kappa_\ell\vartheta_\ell)s_\ell}, \frac{\min_{i\in \mathrm{ne}(\ell)} \frac{|B_{i\ell}|}{B_{\ell\ell}} - \vartheta_\ell\rho_\ell}{2\vartheta_\ell(1+\kappa_\ell\vartheta_\ell)}, \frac{\rho_\ell\eta_\ell}{8(1+\kappa_\ell\vartheta_\ell)^2}\right\}.$$

Then the graph $G_{NS}(\widehat{A})$ *obtained through the neighborhood selection in Algorithm 2 with penalty parameters* $\rho_1,\dots,\rho_p$ *is equal to* $G(B)$.

We now discuss the graphical lasso. Define the maximal edge degree $s := \max_{\ell\in[p]} s_\ell$, with s_ℓ as earlier, S to be the augmented edge set $\{(i,j) : B_{ij} \neq 0\}$, including the self loops (i,i), and S^c to be its complement in V^2. Let $\Omega := A \otimes A$,

$$\kappa_A := |||A|||_\infty, \quad \kappa_\Omega := |||(\Omega_{SS})^{-1}|||_\infty, \quad \chi := 6\kappa_A\kappa_\Omega\Big(1 \vee \frac{9\kappa_A^2\kappa_\Omega}{\alpha}\Big),$$
$$\alpha := 1 - |||\Omega_{S^cS}(\Omega_{SS})^{-1}|||_\infty.$$

Proposition 3.4. *Assume that* $\alpha > 0$ *and*

$$\varepsilon < \min\left\{\min_{i\in[p]} A_{ii}, \frac{\alpha\rho}{8}, \frac{1}{\chi s} - \rho, \frac{1}{4\kappa_\Omega}\min_{i\neq j, B_{ij}\neq 0}|B_{ij}| - \rho\right\}.$$

Then the graph $G_{GL}(\widehat{A})$ *obtained through the graphical lasso with penalty parameter* ρ *is equal to* $G(B)$.

Remark. For both algorithms, $s\varepsilon \to 0$ is sufficient for model selection consistency as the sample size increases, if everything else is constant and the relevant incoherence condition is satisfied.

3.9.4 Proof of Proposition 3.2

Denote by f_m the bijective function that maps Θ to $\Theta^{(m)}$. It is proved in Röttger et al. (2021) that

$$-\log\det{}^*\Theta + \mathrm{tr}(\widehat{\Sigma}\Theta) = -\log\det f_m(\Theta) + \mathrm{tr}(\widehat{\Sigma}^{(m)} f_m(\Theta)) + d.$$

Moreover, for any i,j not equal to m, $f_m(\Theta)_{ij} = \Theta_{ij}$. This implies, first, that

$$\sum_{i\neq j}\sum |f_m(\Theta)_{ij}| = \sum_{i\neq m}\sum_{j\notin\{i,m\}} |\Theta_{ij}|.$$

It also implies that

$$2\sum_i \Big|\sum_j f_m(\Theta)_{ij}\Big| = 2\sum_{i\neq m}\Big|\sum_{j\neq m}\Theta_{ij}\Big| = 2\sum_{i\neq m}|-\Theta_{im}| = \sum_{i\neq m}|\Theta_{im}| + \sum_{j\neq m}|\Theta_{mj}|,$$

since Θ has row sums zero, that is $\Theta_{im} = -\sum_{j \neq m} \Theta_{ij}$. The two equations above combine into

$$\begin{aligned}\sum_{i \neq j} \sum |f_m(\Theta)_{ij}| + 2 \sum_i \left| \sum_j f_m(\Theta)_{ij} \right| &= \sum_{i \neq m} \left(\sum_{j \notin \{i,m\}} |\Theta_{ij}| + |\Theta_{im}| \right) + \sum_{j \neq m} |\Theta_{mj}| \\ &= \sum_{i \neq m} \sum_{j \neq i} |\Theta_{ij}| + \sum_{j \neq m} |\Theta_{mj}| \\ &= \sum_{i \neq j} \sum |\Theta_{ij}|.\end{aligned}$$

Conclude that since the objective functions in (3.26) and (3.27) differ only by an additive constant term, the solution $\widehat{\Theta}$ of (3.27) also minimizes

$$-\log\det f_m(\Theta) + \operatorname{tr}(\widehat{\Sigma}^{(m)} f_m(\Theta)) + \rho \sum_{i \neq j} \sum |f_m(\Theta)_{ij}| + 2\rho \sum_i \left| \sum_j f_m(\Theta)_{ij} \right|.$$

This completes the proof. □

3.10 Consistency of neighborhood selection and graphical lasso: proofs

3.10.1 Proof of Proposition 3.3

First note that if each of the lasso regressions therein succeeds in recovering the corresponding neighborhood, then clearly Algorithm 2 recovers the right graph.

We hereby fix an arbitrary index $\ell \in [p]$. Now, note that the assumed bound on ε implies

$$s_\ell \varepsilon \leq \lambda_\ell / 2, \tag{3.28}$$

$$\vartheta_\ell s_\ell \varepsilon \leq 1/2, \tag{3.29}$$

$$2\vartheta_\ell (1 + \kappa_\ell \vartheta_\ell) s_\ell \varepsilon \leq \eta_\ell / 2, \tag{3.30}$$

$$2\vartheta_\ell (1 + \kappa_\ell \vartheta_\ell)\varepsilon + \vartheta_\ell \rho_\ell < \min_{i \in \mathrm{ne}(\ell)} \frac{|B_{i\ell}|}{B_{\ell\ell}}, \tag{3.31}$$

$$2(1 + \kappa_\ell \vartheta_\ell)^2 \varepsilon < \frac{\rho_\ell \eta_\ell}{4}. \tag{3.32}$$

For the remainder of the proof, ℓ will be kept fix and hence will be partially removed from the notation. In particular, the subscripts of $\lambda_\ell, \kappa_\ell, \vartheta_\ell, s_\ell, \eta_\ell, \rho_\ell$ will be omitted.

Our goal is to show then that

$$\widehat{\theta} := \underset{\theta \in \mathbb{R}^{p-1}}{\arg\min} \big\{ -2\widehat{A}_{\ell,\setminus\ell}\theta + \theta^\top \widehat{A}_{\setminus\ell,\setminus\ell}\theta + \rho\|\theta\|_1 \big\} \tag{3.33}$$

has the same support (i.e., the set of indices where it is nonzero) as $B_{\setminus\ell,\ell}$.

Partition $[p]$ into the three subsets $\{\ell\}$, $\mathrm{ne}(\ell)$ and $\mathrm{ne}(\ell)^c \setminus \{\ell\}$. We shall index elements of vectors and rows/columns of matrices by ℓ, 1 and 2, respectively, to denote those subsets, e.g.,

$$\begin{aligned}
\widehat{A}_{11} &:= (\widehat{A}_{ij})_{i\in\mathrm{ne}(\ell), j\in\mathrm{ne}(\ell)} \\
\widehat{A}_{21} &:= (\widehat{A}_{ij})_{i\in\mathrm{ne}(\ell)^c\setminus\{\ell\}, j\in\mathrm{ne}(\ell)} \\
\widehat{A}_{1\ell} &:= (\widehat{A}_{i\ell})_{i\in\mathrm{ne}(\ell)} \\
\widehat{A}_{2\ell} &:= (\widehat{A}_{i\ell})_{i\in\mathrm{ne}(\ell)^c\setminus\{\ell\}};
\end{aligned}$$

similarly define the population versions A_{11}, A_{21}, $A_{1\ell}$ and $A_{2\ell}$. We use the same notation for partitioning the matrix B. We now show that (3.28) to (3.32) imply the bounds

$$\Big|\Big|\Big| \widehat{A}_{21}(\widehat{A}_{11})^{-1} \Big|\Big|\Big|_\infty \le 1 - \eta/2 \tag{3.34}$$

$$\Big\| (\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} + \frac{B_{1\ell}}{B_{\ell\ell}} \Big\|_\infty + \frac{\rho}{2}\Big|\Big|\Big| (\widehat{A}_{11})^{-1} \Big|\Big|\Big|_\infty < \min_{i\in\mathrm{ne}(\ell)} \frac{|B_{i\ell}|}{B_{\ell\ell}} \tag{3.35}$$

$$\Big\| \widehat{A}_{21}(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - \widehat{A}_{2\ell} \Big\|_\infty < \frac{\rho\eta}{4}. \tag{3.36}$$

Subsequently, by adapting the arguments of Zhao and Yu (2006), we will show that the invertibility of $\widehat{A}_{11}$, along with (3.34) to (3.36), imply the result.

Preliminaries to the proofs of (3.34) to (3.36): Let us start by obtaining a few useful bounds and identities.

We first prove that under the assumptions made, the matrix $\widehat{A}_{11}$ is invertible. Observe that

$$\lambda_{\min}(\widehat{A}_{11}) = \min_{b:\|b\|_2=1} b^\top \widehat{A}_{11} b \ge \lambda - \sup_{b:\|b\|_2=1} b^\top(\widehat{A}_{11} - A_{11})b \ge \lambda - s\|\widehat{A}_{11} - A_{11}\|_\infty = \lambda - s\varepsilon \ge \lambda/2,$$

by (3.28). In the second inequality, we have used the known result that the spectral norm of any square matrix is upper bounded by its dimension times its maximum norm. Most importantly, we have established that $\widehat{A}_{11}$ is invertible.

Now, by sub-multiplicativity of operator norms,

$$\Big|\Big|\Big| (\widehat{A}_{11})^{-1} - (A_{11})^{-1} \Big|\Big|\Big|_\infty = \Big|\Big|\Big| (\widehat{A}_{11})^{-1}(A_{11} - \widehat{A}_{11})(A_{11})^{-1} \Big|\Big|\Big|_\infty$$

$$\begin{aligned}
&\leq \left|\left|\left|(A_{11})^{-1}\right|\right|\right|_\infty \left|\left|\left|(\widehat{A}_{11})^{-1}\right|\right|\right|_\infty \left|\left|\left|\widehat{A}_{11} - A_{11}\right|\right|\right|_\infty \\
&\leq \vartheta\Big(\vartheta + \left|\left|\left|(\widehat{A}_{11})^{-1} - (A_{11})^{-1}\right|\right|\right|_\infty\Big) s\varepsilon.
\end{aligned}$$

Rearranging yields

$$\left|\left|\left|(\widehat{A}_{11})^{-1} - (A_{11})^{-1}\right|\right|\right|_\infty \leq \frac{\vartheta^2 s\varepsilon}{1 - \vartheta s\varepsilon} \leq 2\vartheta^2 s\varepsilon, \tag{3.37}$$

since the L^∞/L^∞-operator norm of a square matrix is also upper bounded by its dimension times its maximum norm. The last inequality is due to (3.29). For further use, note that this implies

$$\left|\left|\left|(\widehat{A}_{11})^{-1}\right|\right|\right|_\infty \leq \vartheta + \left|\left|\left|(\widehat{A}_{11})^{-1} - (A_{11})^{-1}\right|\right|\right|_\infty \leq \vartheta + 2\vartheta^2 s\varepsilon \leq 2\vartheta, \tag{3.38}$$

where in the last inequality we applied (3.29) again.

Using (3.37), it is possible to obtain a sharper bound on the maximum norm difference between $(A_{11})^{-1}$ and $(\widehat{A}_{11})^{-1}$. Indeed, note that for any matrices T_1, T_2, we have

$$\|T_1 T_2\|_\infty \leq \|T_1\|_\infty |||T_2|||_1, \tag{3.39}$$

which reduces to $\|T_1\|_\infty \|T_2\|_1$ if T_2 is a column vector. Similarly,

$$\|T_1 T_2\|_\infty \leq |||T_1|||_\infty \|T_2\|_\infty. \tag{3.40}$$

Repeatedly using those facts, along with symmetry,

$$\begin{aligned}
\left\|(\widehat{A}_{11})^{-1} - (A_{11})^{-1}\right\|_\infty &= \left\|(\widehat{A}_{11})^{-1}(A_{11} - \widehat{A}_{11})(A_{11})^{-1}\right\|_\infty \\
&\leq \left\|(\widehat{A}_{11})^{-1}(A_{11} - \widehat{A}_{11})\right\|_\infty \left|\left|\left|(A_{11})^{-1}\right|\right|\right|_1 \\
&\leq \left|\left|\left|(\widehat{A}_{11})^{-1}\right|\right|\right|_\infty \left\|A_{11} - \widehat{A}_{11}\right\|_\infty \left|\left|\left|(A_{11})^{-1}\right|\right|\right|_1 \\
&\leq 2\vartheta^2 \varepsilon,
\end{aligned} \tag{3.41}$$

where in the last step we used the fact that the L^1/L^1- and L^∞/L^∞-operator norms of a symmetric matrix (in this case, $(A_{11})^{-1}$) are equal, along with (3.38).

We now prove that

$$(A_{11})^{-1} A_{1\ell} = -\frac{B_{1\ell}}{B_{\ell\ell}}. \tag{3.42}$$

If $\boldsymbol{W} \sim N(0, A)$, it is known that $-B_{1\ell}/B_{\ell\ell}$ is the vector of non-zero coefficients for optimal linear prediction of W_ℓ using $\boldsymbol{W}_{\setminus\ell}$. Since removing the non-predictor variables

does not change the prediction, we have

$$-\frac{B_{1\ell}}{B_{\ell\ell}} = -\frac{B^*_{1\ell}}{B^*_{\ell\ell}}$$

where

$$B^* := \begin{bmatrix} A_{11} & A_{1\ell} \\ A_{\ell 1} & A_{\ell\ell} \end{bmatrix}^{-1}.$$

By the inversion formula for block matrices,

$$-B^*_{1\ell} = \frac{1}{A_{\ell\ell}}\Big(A_{11} - \frac{A_{1\ell}A_{\ell 1}}{A_{\ell\ell}}\Big)^{-1} A_{1\ell}, \quad B^*_{\ell\ell} = \frac{1}{A_{\ell\ell}}\Big(1 - \frac{A_{\ell 1}(A_{11})^{-1}A_{1\ell}}{A_{\ell\ell}}\Big)^{-1},$$

so letting $\lambda := \frac{A_{\ell 1}(A_{11})^{-1}A_{1\ell}}{A_{\ell\ell}} \in (0,1)$,

$$-\frac{B^*_{1\ell}}{B^*_{\ell\ell}} = (1-\lambda)\Big(A_{11} - \frac{A_{1\ell}A_{\ell 1}}{A_{\ell\ell}}\Big)^{-1} A_{1\ell}.$$

Applying Woodbury's matrix inversion formula, we find

$$\begin{aligned}\Big(A_{11} - \frac{A_{1\ell}A_{\ell 1}}{A_{\ell\ell}}\Big)^{-1} &= (A_{11})^{-1} - (A_{11})^{-1}A_{1\ell}\big(-A_{\ell\ell} + A_{\ell 1}(A_{11})^{-1}A_{1\ell}\big)^{-1}A_{\ell 1}(A_{11})^{-1} \\ &= (A_{11})^{-1} - (A_{11})^{-1}A_{1\ell}\frac{1}{A_{\ell\ell}(\lambda-1)}A_{\ell 1}(A_{11})^{-1}.\end{aligned}$$

Simple matrix algebra yields

$$\begin{aligned}-\frac{B^*_{1\ell}}{B^*_{\ell\ell}} &= (1-\lambda)(A_{11})^{-1}A_{1\ell}\Big(1 - \frac{1}{A_{\ell\ell}(\lambda-1)}A_{\ell 1}(A_{11})^{-1}A_{1\ell}\Big) \\ &= (1-\lambda)(A_{11})^{-1}A_{1\ell}\Big(1 - \frac{\lambda}{\lambda-1}\Big) \\ &= (A_{11})^{-1}A_{1\ell},\end{aligned}$$

which finally establishes (3.42).

Similarly, we prove that

$$A_{21}(A_{11})^{-1}A_{1\ell} = A_{2\ell}. \tag{3.43}$$

Indeed, notice that the Schur complement of A_{11} in A,

$$A_{(2,\ell),(2,\ell)\,|\,1} := \begin{bmatrix} A_{22} & A_{2\ell} \\ A_{\ell 2} & A_{\ell\ell} \end{bmatrix} - \begin{bmatrix} A_{21} \\ A_{\ell 1} \end{bmatrix} (A_{11})^{-1} \begin{bmatrix} A_{12} & A_{1\ell} \end{bmatrix},$$

is the conditional covariance matrix of the random vector $\boldsymbol{W}_{\mathrm{ne}(\ell)^c}$ given $\boldsymbol{W}_{\mathrm{ne}(\ell)}$. The off-diagonal block of $A_{(2,\ell),(2,\ell)\,|\,1}$, that is, $A_{2\ell} - A_{21}(A_{11})^{-1}A_{1\ell}$, is therefore the conditional

covariance between W_ℓ and the other variables not in its neighborhood, given the variables in its neighborhood. By definition of the neighborhood, this is zero, hence (3.43) holds.

Proof of (3.34): Using (3.37), we have

$$\begin{aligned}
&\left|\!\left|\!\left| \widehat{A}_{21}(\widehat{A}_{11})^{-1} - A_{21}(A_{11})^{-1} \right|\!\right|\!\right|_\infty \\
&\quad \le \left|\!\left|\!\left| \widehat{A}_{21}(\widehat{A}_{11})^{-1} - A_{21}(\widehat{A}_{11})^{-1} \right|\!\right|\!\right|_\infty + \left|\!\left|\!\left| A_{21}(\widehat{A}_{11})^{-1} - A_{21}(A_{11})^{-1} \right|\!\right|\!\right|_\infty \\
&\quad \le \left|\!\left|\!\left| (\widehat{A}_{11})^{-1} \right|\!\right|\!\right|_\infty \left|\!\left|\!\left| \widehat{A}_{21} - A_{21} \right|\!\right|\!\right|_\infty + |||A_{21}|||_\infty \left|\!\left|\!\left| (\widehat{A}_{11})^{-1} - (A_{11})^{-1} \right|\!\right|\!\right|_\infty \\
&\quad \le 2\vartheta s\varepsilon + \kappa 2\vartheta^2 s\varepsilon \\
&\quad = 2\vartheta(1+\kappa\vartheta)s\varepsilon,
\end{aligned}$$

where the third inequality follows by applying (3.37) and (3.38). Now by the reverse triangle inequality,

$$\begin{aligned}
\left|\!\left|\!\left| \widehat{A}_{21}(\widehat{A}_{11})^{-1} \right|\!\right|\!\right|_\infty &\le |||A_{21}(A_{11})^{-1}|||_\infty + \left|\!\left|\!\left| \widehat{A}_{21}(\widehat{A}_{11})^{-1} - A_{21}(A_{11})^{-1} \right|\!\right|\!\right|_\infty \\
&\le 1 - \eta + 2\vartheta(1+\kappa\vartheta)s\varepsilon.
\end{aligned}$$

(3.34) then follows from (3.30).

Proof of (3.35): First, by (3.42),

$$\begin{aligned}
\left\| (\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} + \frac{B_{1\ell}}{B_{\ell\ell}} \right\|_\infty &= \left\| (\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - (A_{11})^{-1}A_{1\ell} \right\|_\infty \\
&\le \left\| (\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - (\widehat{A}_{11})^{-1}A_{1\ell} \right\|_\infty + \left\| (\widehat{A}_{11})^{-1}A_{1\ell} - (A_{11})^{-1}A_{1\ell} \right\|_\infty \\
&\le \left|\!\left|\!\left| (\widehat{A}_{11})^{-1} \right|\!\right|\!\right|_\infty \left\| \widehat{A}_{1\ell} - A_{1\ell} \right\|_\infty + \left\| \left((\widehat{A}_{11})^{-1} - (A_{11})^{-1} \right) \right\|_\infty \|A_{1\ell}\|_1 \\
&\le 2\vartheta\varepsilon + 2\vartheta^2\varepsilon\kappa \\
&= 2\vartheta(1+\kappa\vartheta)\varepsilon,
\end{aligned}$$

where we first used (3.39) and (3.40), then (3.38) and (3.41). Noting that

$$\frac{\rho}{2}\left|\!\left|\!\left| (\widehat{A}_{11})^{-1} \right|\!\right|\!\right|_\infty \le \frac{\rho}{2} 2\vartheta = \vartheta\rho,$$

(3.35) follows from (3.31).

Proof of (3.36): First, by (3.43),

$$\left\| \widehat{A}_{21}(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - \widehat{A}_{2\ell} \right\|_\infty \le \left\| \widehat{A}_{21}(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - A_{21}(A_{11})^{-1}A_{1\ell} \right\|_\infty + \left\| \widehat{A}_{2\ell} - A_{2\ell} \right\|_\infty,$$

the second term of which is clearly upper bounded by ε. The first term above is upper

bounded by

$$\begin{aligned}
&\left\|\widehat{A}_{21}(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - \widehat{A}_{21}(A_{11})^{-1}A_{1\ell}\right\|_\infty + \left\|\widehat{A}_{21}(A_{11})^{-1}A_{1\ell} - A_{21}(A_{11})^{-1}A_{1\ell}\right\|_\infty \\
&\quad \le \left|\!\left|\!\left|\widehat{A}_{21}\right|\!\right|\!\right|_\infty \left\|(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - (A_{11})^{-1}A_{1\ell}\right\|_\infty + \left\|\widehat{A}_{21} - A_{21}\right\|_\infty \left|\!\left|\!\left|(A_{11})^{-1}A_{1\ell}\right|\!\right|\!\right|_1 \\
&\quad \le (\kappa + s\varepsilon)2\vartheta(1+\kappa\vartheta)\varepsilon + \kappa\vartheta\varepsilon \\
&\quad \le 2(1+\kappa\vartheta)^2\varepsilon - \varepsilon,
\end{aligned}$$

where we used (3.39) and (3.40) and the result in the proof of (3.35) above where we bound $\|(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - (A_{11})^{-1}A_{1\ell}\|_\infty$. The final bound was obtained by applying (3.29) and rearranging. Hence (3.36) follows from (3.32).

Proof that (3.33) recovers the support of $B_{\setminus\ell,\ell}$: Note that

$$\nabla_\theta\Big\{ -2\widehat{A}_{\ell,\setminus\ell}\theta + \theta^\top \widehat{A}_{\setminus\ell,\setminus\ell}\theta\Big\} = 2\widehat{A}_{\setminus\ell,\setminus\ell}\theta - 2\widehat{A}_{\setminus\ell,\ell}$$

and that the subdifferential of the 1-norm at a point $\theta \in \mathbb{R}^{p-1}$ is given by the set of all $x \in [-1,1]^{p-1}$ such that

$$\theta_j \neq 0 \Longrightarrow x_j = \operatorname{sign}(\theta_j).$$

Considering the optimization problem in (3.33), the KKT conditions state that any point $\widehat{\theta}$ satisfying

$$\big(2\widehat{A}_{\setminus\ell,\setminus\ell}\widehat{\theta} - 2\widehat{A}_{\setminus\ell,\ell}\big)_{\widehat{J}} = -\rho\operatorname{sign}(\widehat{\theta}_{\widehat{J}}) \tag{3.44}$$

$$\big\|\big(2\widehat{A}_{\setminus\ell,\setminus\ell}\widehat{\theta} - 2\widehat{A}_{\setminus\ell,\ell}\big)_{\widehat{J}^c}\big\|_\infty \le \rho \tag{3.45}$$

where $\widehat{J} := \{j \in [p-1] : \widehat{\theta}_j \neq 0\}$, is a solution. Following the arguments in the proof of Proposition 1 in Zhao and Yu (2006), we shall identify one such solution that has the right support. We will subsequently show that it is unique, utilizing arguments similar to (but not implied by) what is found in Section 2.1 of Tibshirani (2013).

Let $\theta^* := -B_{\setminus\ell,\ell}/B_{\ell\ell}$ and θ_1^*, θ_2^* denote its subvectors indexed by $\mathrm{ne}(\ell)$ and $\mathrm{ne}(\ell)^c\setminus\{\ell\}$, respectively[1]; we will use the same notation for $\widehat{\theta}$ defined below. The candidate solution $\widehat{\theta}$ is defined by $\widehat{\theta}_2 = 0$ and

$$\widehat{\theta}_1 := (\widehat{A}_{11})^{-1}\Big(\widehat{A}_{1\ell} - \frac{\rho}{2}\operatorname{sign}(\theta_1^*)\Big),$$

[1]For simplicity, we slightly abuse the notation here. In fact, θ_1^* and θ_2^* are the subvectors that, in the product $A_{\setminus\ell,\setminus\ell}\theta^*$, are multiplied by the columns of A corresponding to variables in $\mathrm{ne}(\ell)$ and in $\mathrm{ne}(\ell)^c\setminus\{\ell\}$, respectively.

where the sign function is applied to θ_1^* coordinate-wise. First note that by (3.35),

$$\|\widehat{\theta}_1 - \theta_1^*\|_\infty < \min_{i \in \mathrm{ne}(\ell)} |\theta_i^*|,$$

hence $\mathrm{sign}(\theta_1^*) = \mathrm{sign}(\widehat{\theta}_1)$. Thus, we find

$$2\widehat{A}_{\mathrm{ne}(\ell),\setminus\ell}\widehat{\theta} - 2\widehat{A}_{1\ell} = 2\widehat{A}_{11}\widehat{\theta}_1 - 2\widehat{A}_{1\ell} = -\rho\,\mathrm{sign}(\theta_1^*) = -\rho\,\mathrm{sign}(\widehat{\theta}_1),$$

i.e., (3.44) is satisfied. Next, by (3.34) and (3.36) we have

$$\begin{aligned}
\|2\widehat{A}_{\mathrm{ne}(\ell)^c\setminus\{\ell\},\setminus\ell}\widehat{\theta} - 2\widehat{A}_{2\ell}\|_\infty &= \|2\widehat{A}_{21}\widehat{\theta}_1 - 2\widehat{A}_{2\ell}\|_\infty \\
&= \Big\|2\widehat{A}_{21}(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - 2\widehat{A}_{2\ell} - \rho\widehat{A}_{21}(\widehat{A}_{11})^{-1}\,\mathrm{sign}(\theta_1^*)\Big\|_\infty \\
&\le 2\Big\|\widehat{A}_{21}(\widehat{A}_{11})^{-1}\widehat{A}_{1\ell} - \widehat{A}_{2\ell}\Big\|_\infty + \rho\Big|\Big|\Big|\widehat{A}_{21}(\widehat{A}_{11})^{-1}\Big|\Big|\Big|_\infty \\
&< \frac{\rho\eta}{2} + \rho\Big(1 - \frac{\eta}{2}\Big) \\
&= \rho,
\end{aligned}$$

i.e., (3.45) is satisfied. Therefore, we have proved the existence of a solution $\widehat{\theta}$ to (3.33) which has the same sign pattern (hence the same support) as θ^*.

It remains to show that this solution is unique. We first prove a weaker statement: $\widehat{A}_{\setminus\ell,\setminus\ell}\widetilde{\theta}$ is unique across all solutions $\widetilde{\theta}$ to (3.33). Indeed, suppose that $\widetilde{\theta}^{(1)}$ and $\widetilde{\theta}^{(2)}$ are two distinct solutions such that $\widehat{A}_{\setminus\ell,\setminus\ell}\widetilde{\theta}^{(1)} \neq \widehat{A}_{\setminus\ell,\setminus\ell}\widetilde{\theta}^{(2)}$. Let W be a (possibly rectangular) matrix such that $\widehat{A}_{\setminus\ell,\setminus\ell} = W^\top W$; for instance, W can be obtained from a Cholesky decomposition. By assumption, $W\widetilde{\theta}^{(1)} \neq W\widetilde{\theta}(2)$. Both points being a solution means that the objective function in (3.33) attains its minimum value at both points: if

$$Q(\theta) := -2\widehat{A}_{\ell,\setminus\ell}\theta + \|W\theta\|_2^2 + \rho\|\theta\|_1,$$

then $Q(\widetilde{\theta}^{(1)}) = Q(\widetilde{\theta}^{(2)}) = \min_\theta Q(\theta) =: v_{\min}$, say. Then, evaluating Q at a point $\alpha\widetilde{\theta}^{(1)} + (1-\alpha)\widetilde{\theta}^{(2)}$, for some $\alpha \in (0,1)$, yields

$$\begin{aligned}
&-2\widehat{A}_{\ell,\setminus\ell}\big(\alpha\widetilde{\theta}^{(1)} + (1-\alpha)\widetilde{\theta}^{(2)}\big) + \big\|\alpha W\widetilde{\theta}^{(1)} + (1-\alpha)W\widetilde{\theta}^{(2)}\big\|_2^2 + \rho\big\|\alpha\widetilde{\theta}^{(1)} + (1-\alpha)\widetilde{\theta}^{(2)}\big\|_1 \\
&\quad< \alpha\big\{-2\widehat{A}_{\ell,\setminus\ell}\widetilde{\theta}^{(1)} + \|W\widetilde{\theta}^{(1)}\|_2^2 + \rho\|\widetilde{\theta}^{(1)}\|_1\big\} \\
&\qquad + (1-\alpha)\big\{-2\widehat{A}_{\ell,\setminus\ell}\widetilde{\theta}^{(2)} + \|W\widetilde{\theta}^{(2)}\|_2^2 + \rho\|\widetilde{\theta}^{(2)}\|_1\big\} \\
&\quad= \alpha Q(\widetilde{\theta}^{(1)}) + (1-\alpha)Q(\widetilde{\theta}^{(2)}) \\
&\quad= v_{\min},
\end{aligned}$$

where the strict inequality is a consequence of the strict convexity of the squared

Euclidean norm and the convexity of the 1-norm. This is a contradiction since $v_{\min}$ was assumed to be the minimum value. Hence $\widehat{A}_{\setminus\ell,\setminus\ell}\widetilde{\theta}$ must be unique across all solutions $\widetilde{\theta}$ of (3.33).

Now, using this preliminary uniqueness result, we notice that for any solution $\widetilde{\theta}$, we have

$$\left\|2\widehat{A}_{\mathrm{ne}(\ell)^c\setminus\{\ell\},\setminus\ell}\widetilde{\theta} - 2\widehat{A}_{2\ell}\right\|_\infty = \left\|2\widehat{A}_{\mathrm{ne}(\ell)^c\setminus\{\ell\},\setminus\ell}\widehat{\theta} - 2\widehat{A}_{2\ell}\right\|_\infty < \rho,$$

hence $\widetilde{\theta}_2 = 0$. Therefore,

$$2\widehat{A}_{11}\widetilde{\theta}_1 - 2\widehat{A}_{1\ell} = 2\widehat{A}_{\mathrm{ne}(\ell),\setminus\ell}\widetilde{\theta} - 2\widehat{A}_{1\ell} = 2\widehat{A}_{\mathrm{ne}(\ell),\setminus\ell}\widehat{\theta} - 2\widehat{A}_{1\ell} = -\rho\,\mathrm{sign}(\theta_1^*),$$

which uniquely defines $\widetilde{\theta}_1$ by the invertibility of $\widehat{A}_{11}$. Deduce that $\widetilde{\theta} = \widehat{\theta}$. □

3.10.2 Proof of Proposition 3.4

First note that the assumed bound on ε implies

$$\varepsilon \le \frac{\alpha\rho}{8}, \tag{3.46}$$

$$2\kappa_\Omega(\varepsilon+\rho) \le \min\left\{\frac{1}{3\kappa_A s}, \frac{1}{3\kappa_A^3\kappa_\Omega s}\right\}, \tag{3.47}$$

$$6\kappa_A^3\kappa_\Omega^2 s(\varepsilon+\rho) \le \frac{\alpha}{9} \le \frac{1}{1+8/\alpha}, \tag{3.48}$$

$$2\kappa_\Omega(\varepsilon+\rho) \le \frac{1}{2}\min_{i\neq j, B_{ij}\neq 0}|B_{ij}|. \tag{3.49}$$

Without loss of generality, assume that $\rho > 0$. Otherwise, (3.46) implies that $\widehat{A} = A$, in which case the result is trivial.

The proof is heavily based on that of Theorems 1 and 2 of Ravikumar et al. (2011). Note that by assumption, each diagonal element of $\widehat{A}$ satisfies $\widehat{A}_{ii} \ge A_{ii} - \varepsilon > 0$, $i \in [p]$. Then by Lemma 3 of that paper, the positivity of ρ ensures that the solution $\widehat{B}$ exists, is unique and satisfies

$$-\widehat{B}^{-1} + \widehat{A} + \rho Z = 0,$$

for some matrix Z in the sub-differential of the off-diagonal norm at the point $\widehat{B}$, as defined in Ravikumar et al. (2011). The strategy is now to consider the solution $\widetilde{B}$ of the graphical lasso optimization problem with the additional constraint that $B_{S^c} = 0$, which is also guaranteed to exist and to be unique by (3.46). Define $\Delta := \|\widetilde{B} - B\|_\infty$. By (3.47), the condition of Lemma 6 of Ravikumar et al. (2011) is satisfied. It then follows from that result that $\|\Delta\|_\infty \le 2\kappa_\Omega(\varepsilon+\rho)$. Lemma 5 from that paper now

implies that the matrix $R(\Delta)$, as defined therein, satisfies

$$\|R(\Delta)\|_\infty \le \frac{3}{2}\kappa_A^3 s\|\Delta\|_\infty^2 \le 6\kappa_A^3\kappa_\Omega^2 s(\varepsilon+\rho)^2 \le \frac{\varepsilon+\rho}{1+8/\alpha} \le \frac{(\alpha/8+1)\rho}{1+8/\alpha} = \frac{\alpha\rho}{8},$$

where the last three inequalities are due to the previously obtained bound on $\|\Delta\|_\infty$, to (3.48) and to (3.46), respectively. Now that $\varepsilon \vee \|R(\Delta)\|_\infty \le \alpha\rho/8$, we may apply Lemma 4 of Ravikumar et al. (2011) and find that in fact, $\widehat{B} = \widetilde{B}$. It follows, by definition of $\widetilde{B}$, that $\widehat{B}_{ij} = 0$ for all $(i,j) \in S^c$ and that for $(i,j) \in S$, $i \neq j$,

$$|\widehat{B}_{ij} - B_{ij}| \le \|\Delta\|_\infty \le \frac{|B_{ij}|}{2}$$

by (3.49). That is, the element-wise error is too small for $\widehat{B}_{ij}$ to reach 0. We have therefore guaranteed that the sparsity pattern of $\widehat{B}$ is the same as that of B. □

3.11 Proof of Theorem 3.3

We start by introducing useful additional notation and auxiliary variables that will be useful throughout this proof. For $\ell \in \{1,2\}$, let $e_i^{(m),\ell} := \mathbb{E}[(\log Y_i^{(m)})^\ell]$ and $e_{ij}^{(m)} := \mathbb{E}[(\log Y_i^{(m)})(\log Y_j^{(m)})]$. Then we have

$$\Gamma_{ij}^{(m)} = e_i^{(m),2} + e_j^{(m),2} - 2e_{ij}^{(m)} - \big(e_i^{(m),1} - e_j^{(m),1}\big)^2, \quad i \neq j, m \in V. \tag{3.50}$$

Similarly

$$\widehat{\Gamma}_{ij}^{(m)} = \widehat{e}_i^{(m),2} + \widehat{e}_j^{(m),2} - 2\widehat{e}_{ij}^{(m)} - \big(\widehat{e}_i^{(m),1} - \widehat{e}_j^{(m),1}\big)^2, \quad i \neq j, m \in V,$$

where

$$\widehat{e}_i^{(m),\ell} := \frac{1}{k}\sum_{t=1}^n \left\{\log\Big(\frac{k}{n\widehat{F}_i(U_{ti})}\Big)\right\}^\ell \mathbb{1}\left\{\widehat{F}_m(U_{tm}) \le k/n\right\},$$

$$\widehat{e}_{ij}^{(m)} := \frac{1}{k}\sum_{t=1}^n \log\Big(\frac{k}{n\widehat{F}_j(U_{tj})}\Big)\log\Big(\frac{k}{n\widehat{F}_i(U_{ti})}\Big)\mathbb{1}\left\{\widehat{F}_m(U_{tm}) \le k/n\right\},$$

$U_{ti} := 1 - F_i(X_{ti})$, $1 \le t \le n$, are independent and uniformly distributed, and $\widehat{F}_i$ is the (right-continuous) empirical distribution function of $(U_{ti})_{1\le t\le n}$, satisfying $\widehat{F}_i(U_{ti}) = 1 - \widetilde{F}_i(X_{ti})$.

Let $i \neq j$ and m be arbitrary. An expression for the estimation error is given by

$$\widehat{\Gamma}_{ij}^{(m)} - \Gamma_{ij}^{(m)} = \big(\widehat{e}_i^{(m),2} - e_i^{(m),2}\big) + \big(\widehat{e}_j^{(m),2} - e_j^{(m),2}\big) - 2\big(\widehat{e}_{ij}^{(m)} - e_{ij}^{(m)}\big)$$

$$- 2\big(e_i^{(m),1} - e_j^{(m),1}\big)\Big(\big(\widehat{e}_i^{(m),1} - e_i^{(m),1}\big) - \big(\widehat{e}_j^{(m),1} - e_j^{(m),1}\big)\Big)$$
$$- \Big(\big(\widehat{e}_i^{(m),1} - e_i^{(m),1}\big) - \big(\widehat{e}_j^{(m),1} - e_j^{(m),1}\big)\Big)^2, \tag{3.51}$$

the last two terms stemming from the identity $y^2 - x^2 = 2x(y-x) + (y-x)^2$. In order to prove the result, it is sufficient to bound the differences

$$\widehat{e}_i^{(m),\ell} - e_i^{(m),\ell}, \quad \widehat{e}_m^{(m),\ell} - e_m^{(m),\ell}, \quad \widehat{e}_{im}^{(m)} - e_{im}^{(m)}, \quad \widehat{e}_{ij}^{(m)} - e_{ij}^{(m)}$$

for all distinct triples (i, j, m) and $\ell \in \{1, 2\}$. The terms $\widehat{e}_m^{(m),\ell} - e_m^{(m),\ell}$ are entirely deterministic, since it is known that the observations X_{tm} that are used for the estimator $\widehat{\Gamma}^{(m)}$ have ranks $n - k + 1, \ldots, n$ (by continuity, it can be assumed that there are no ties in the data). They are on the order of $(\log k)^\ell / k$, as is proved in Section 3.12.3. The rest of the proof thus focuses on the other three differences.

3.11.1 Preliminaries, additional notation and structure of the proof

Recall that $\widehat{F}_i$ is the empirical distribution function of $(U_{ti})_{1 \le t \le n}$ and denote its left-continuous inverse by $\widehat{F}_i^-$, where $f^-(t) := \inf\{x : f(x) \ge t\}$. Consider the rescaled tail quantile processes

$$u_n^{(i)}(x) := \frac{n}{k} \widehat{F}_i^-(kx/n). \tag{3.52}$$

Similarly to R and its margins R_J, $J \subset V$, let

$$\widehat{R}_J^0(\boldsymbol{x}_J) := \frac{1}{k} \sum_{t=1}^n \mathbb{1}\left\{U_{ti} \le \frac{k}{n} x_i, i \in J\right\}, \quad \widehat{R}_J(\boldsymbol{x}_J) := \widehat{R}_J^0(\widehat{\boldsymbol{x}}_J), \quad \boldsymbol{x}_J \in [0, \infty)^{|J|}, \tag{3.53}$$

where $\widehat{\boldsymbol{x}}_J := (u_n^{(i)}(x_i))_{i \in J}$. The function $\widehat{R}_J$ can be seen as the tail empirical copula of the random vector $\boldsymbol{U}_J$. As an intermediate between R_J and $\widehat{R}_J$, let

$$R_{J,n}(\boldsymbol{x}_J) := \frac{n}{k} \mathbb{P}\Big(\boldsymbol{U}_J \le \frac{k}{n} \boldsymbol{x}_J\Big) = \frac{n}{k} \mathbb{P}\Big(F_J(\boldsymbol{X}_J) > 1 - \frac{k}{n} \boldsymbol{x}_J\Big),$$

the pre-asymptotic version of R_J.

The function R_J can be seen as a measure on $[0, \infty)^{|J|}$ and for measurable sets $A_i \subset \mathbb{R}$, we will write $R_J((A_i)_{i \in J})$ to denote $R_J(\otimes_{i \in J} A_i)$. If in place of one of the A_i there is a number a_i, it will be understood that $A_i = [0, a_i]$. For example, $R_{ij}([x, \infty), y) = R_{ij}([x, \infty) \times [0, y])$. We use the same conventions for the functions $R_{J,n}$, $\widehat{R}_J^0$ and $\widehat{R}_J$, as well as $G_{J,n}$ and $\bar{R}_J$ to be defined later.

The functions R_J, $R_{J,n}$, $\widehat{R}_J^0$ and $\widehat{R}_J$ enjoy certain properties which will be used throughout the proof, as well as throughout Section 3.12: they are non-decreasing in each component and are upper bounded by their minimum argument. Beyond

component-wise monotonicity, the induced measures, as described above, are non-negative. The functions R_J moreover inherit the homogeneity property of multivariate Pareto distributions ($R_J(q\boldsymbol{x}_J) = qR_J(\boldsymbol{x}_J)$, $q \geq 0$).

From now on, fix a number $a \in (0,1)$. It is proved in Lemma 3.7 that for all distinct triples (i,j,m) and $\ell \in \{1,2\}$, we have

$$\begin{aligned}\widehat{e}_i^{(m),\ell} - e_i^{(m),\ell} = &\int_a^1 \frac{\big(\widehat{R}_{im}(x,1) - R_{im}(x,1)\big)(-2\log x)^{\ell-1}}{x}dx \\ &- \int_1^{n/k} \frac{\big(\widehat{R}_{im}([x,\infty),1) - R_{im}([x,\infty),1)\big)(-2\log x)^{\ell-1}}{x}dx \\ &+ O\bigg(\Big(\frac{k}{n}\Big)^{\xi}\log(n/k) + \frac{(\log(n/k)+\log(1/a))^2}{k} \\ &\quad + a(\log(n/k)+\log(1/a))\bigg),\end{aligned} \tag{3.54}$$

$$\begin{aligned}\widehat{e}_{im}^{(m)} - e_{im}^{(m)} = &\int_a^1\int_a^1 \frac{\widehat{R}_{im}(x,y) - R_{im}(x,y)}{xy}dxdy \\ &- \int_a^1\int_1^{n/k} \frac{\widehat{R}_{im}([x,\infty),y) - R_{im}([x,\infty),y)}{xy}dxdy \\ &+ O\bigg(\Big(\frac{k}{n}\Big)^{\xi}\log(n/k) + \frac{(\log(n/k)+\log(1/a))^2}{k} \\ &\quad + a(\log(n/k)+\log(1/a))\bigg),\end{aligned} \tag{3.55}$$

$$\begin{aligned}\widehat{e}_{ij}^{(m)} - e_{ij}^{(m)} = &\int_a^1\int_a^1 \frac{\widehat{R}_{ijm}(x,y,1) - R_{ijm}(x,y,1)}{xy}dxdy \\ &- \int_a^1\int_1^{n/k} \frac{\widehat{R}_{ijm}([x,\infty),y,1) - R_{ijm}([x,\infty),y,1)}{xy}dxdy \\ &- \int_1^{n/k}\int_a^1 \frac{\widehat{R}_{ijm}(x,[y,\infty),1) - R_{ijm}(x,[y,\infty),1)}{xy}dxdy \\ &+ \int_1^{n/k}\int_1^{n/k} \frac{\widehat{R}_{ijm}([x,\infty),[y,\infty),1) - R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy \\ &+ O\bigg(\Big(\frac{k}{n}\Big)^{\xi}\log(n/k) + \frac{(\log(n/k)+\log(1/a))^2}{k} \\ &\quad + a(\log(n/k)+\log(1/a))\bigg)\end{aligned} \tag{3.56}$$

almost surely, where the error terms are not stochastic. We shall separately bound

each of the eight integrals above. Denote these integrals, in order of appearance in (3.54) to (3.56), by $\mathcal{I}_i^{(m),\ell,-}$, $\mathcal{I}_i^{(m),\ell,+}$, $\mathcal{I}_{im}^{(m),--}$, $\mathcal{I}_{im}^{(m),+-}$, $\mathcal{I}_{ij}^{(m),--}$, $\mathcal{I}_{ij}^{(m),+-}$, $\mathcal{I}_{ij}^{(m),-+}$ and $\mathcal{I}_{ij}^{(m),++}$.

The processes $\widehat{R}_J - R_J$ can be decomposed into the stochastic error $\widehat{R}_J - R_{J,n}$ and the difference $R_{J,n} - R_J$ between the tail at finite and infinite levels. Replacing $\widehat{R}_J - R_J$ by $(\widehat{R}_J - R_{J,n}) + (R_{J,n} - R_J)$, each integral $\mathcal{I}_\cdot^{(m),\cdot}$ is written as

$$\mathcal{I}_\cdot^{(m),\cdot} =: A_\cdot^{(m),\cdot} + B_\cdot^{(m),\cdot},$$

where the A terms are stochastic and the B terms represent deterministic bias.

We proceed as follows. In Section 3.11.2, it is shown that Assumption 3.1 is sufficient to bound all the bias terms, up to a constant, by $(k/n)^\xi(\log(n/k)+\log(1/a))^2$. Subsequently, we prove in Section 3.11.3 concentration results for the stochastic terms which are then leveraged in Section 3.11.4 to complete the proof.

Before moving on, we highlight some consequences of Assumption 3.1 and Proposition 3.1 in terms of the functions $R_{J,n}$ and R_J. For any distinct triple (i,j,m) and $q \in (0,1]$,

$$\sup_{x \le n/k, y \le 1} |R_{ij,n}(x,y) - R_{ij}(x,y)| \le 2K\Big(\frac{k}{n}\Big)^\xi, \tag{3.57}$$

$$\sup_{x \le n/k, y \le n/k, z \le 1} |R_{ijm,n}(x,y,z) - R_{ijm}(x,y,z)| \le K\Big(\frac{k}{n}\Big)^\xi, \tag{3.58}$$

$$1 - R_{ijm}(q^{-1},q^{-1},1) \le 1 - R_{im}(q^{-1},1) + 1 - R_{jm}(q^{-1},1) \le 2Kq^\xi; \tag{3.59}$$

to obtain the first inequality in (3.59), apply the inequality

$$P([0,q^{-1}] \times [0,q^{-1}]) \ge P([0,q^{-1}] \times [0,\infty)) + P([0,\infty) \times [0,q^{-1}]) - 1,$$

valid for probability measures P on $[0,\infty)^2$, with $P = R_{ijm}(\cdot \times [0,1])$.

3.11.2 The bias terms B

(3.57) directly implies

$$\begin{aligned}
\big|B_i^{(m),\ell,-}\big| &\le \int_a^1 \frac{\big|R_{im,n}(x,1) - R_{im}(x,1)\big|(-2\log x)^{\ell-1}}{x}dx \\
&\le 2K\Big(\frac{k}{n}\Big)^\xi \int_a^1 \frac{(-2\log x)^{\ell-1}}{x}dx \\
&\lesssim \Big(\frac{k}{n}\Big)^\xi (\log(1/a))^\ell,
\end{aligned}$$

$$\begin{aligned}\left|B_i^{(m),\ell,+}\right| &\le \int_1^{n/k} \frac{\left|R_{im,n}([x,\infty),1) - R_{im}([x,\infty),1)\right|(2\log x)^{\ell-1}}{x}dx \\ &\le 2K\Big(\frac{k}{n}\Big)^{\xi}\int_1^{n/k}\frac{(2\log x)^{\ell-1}}{x}dx \\ &\lesssim \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^{\ell},\end{aligned}$$

$$\begin{aligned}\left|B_{im}^{(m),--}\right| &\le \int_a^1\int_a^1 \frac{\left|R_{im,n}(x,y) - R_{im}(x,y)\right|}{xy}dxdy \\ &\le 2K\Big(\frac{k}{n}\Big)^{\xi}\int_a^1\int_a^1\frac{1}{xy}dxdy \\ &\lesssim \Big(\frac{k}{n}\Big)^{\xi}(\log(1/a))^2,\end{aligned}$$

$$\begin{aligned}\left|B_{im}^{(m),+-}\right| &\le \int_a^1\int_1^{n/k} \frac{\left|R_{im,n}([x,\infty),y) - R_{im}([x,\infty),y)\right|}{xy}dxdy \\ &\le 2K\Big(\frac{k}{n}\Big)^{\xi}\int_a^1\int_1^{n/k}\frac{1}{xy}dxdy \\ &\lesssim \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))(\log(1/a)).\end{aligned}$$

Note further that by (3.58), $B_{ij}^{(m),--}$ admits the same bound as $B_{im}^{(m),--}$. Similarly, $B_{ij}^{(m),+-}$, and by symmetry $B_{ij}^{(m),-+}$, admit the same bound as $B_{im}^{(m),+-}$. Finally, since

$$R_{ijm}([x,\infty),[y,\infty),1) = 1 - R_{jm}(y,1) - R_{im}(x,1) + R_{ijm}(x,y,1)$$

and the same relation holds for $R_{ijm,n}$, (3.57) and (3.58) also imply that

$$\sup_{x\le n/k, y\le n/k}|R_{ijm,n}([x,\infty),[y,\infty),1) - R_{ijm}([x,\infty),[y,\infty),1)| \le 5K\Big(\frac{k}{n}\Big)^{\xi}.$$

Deduce that

$$\begin{aligned}\left|B_{ij}^{(m),++}\right| &\le \int_1^{n/k}\int_1^{n/k} \frac{\left|R_{ijm,n}([x,\infty),[y,\infty),1) - R_{ijm}([x,\infty),[y,\infty),1)\right|}{xy}dxdy \\ &\le 5K\Big(\frac{k}{n}\Big)^{\xi}\int_1^{n/k}\int_1^{n/k}\frac{1}{xy}dxdy \\ &\lesssim \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2.\end{aligned}$$

3.11.3 The stochastic error terms A

It remains to bound the stochastic error terms $A^{\cdot}_{\cdot}$, which entirely depend on the processes $\widehat{R}_J - R_{J,n}$. Recall how, for $\boldsymbol{x}_J \in [0,\infty)^{|J|}$, we define $\widehat{\boldsymbol{x}}_J$ in (3.53). Consider further the relation $\widehat{R}_J(\boldsymbol{x}_J) = \widehat{R}^0_J(\widehat{\boldsymbol{x}}_J)$. We shall rely on the decomposition

$$\begin{aligned}\widehat{R}_J(\boldsymbol{x}_J) - R_{J,n}(\boldsymbol{x}_J) &= \big(\widehat{R}^0_J(\widehat{\boldsymbol{x}}_J) - R_{J,n}(\widehat{\boldsymbol{x}}_J)\big) + \big(R_{J,n}(\widehat{\boldsymbol{x}}_J) - R_{J,n}(\boldsymbol{x}_J)\big) \\ &= \big(G_{J,n}(\widehat{\boldsymbol{x}}_J) - G_{J,n}(\boldsymbol{x}_J)\big) + G_{J,n}(\boldsymbol{x}_J) + \big(R_{J,n}(\widehat{\boldsymbol{x}}_J) - R_{J,n}(\boldsymbol{x}_J)\big),\end{aligned} \tag{3.60}$$

where

$$G_{J,n} := \widehat{R}^0_J - R_{J,n}. \tag{3.61}$$

Accordingly, each $A^{\cdot}_{\cdot}$ term is further decomposed into three integrals $A^{\cdot}_{\cdot,1}$, $A^{\cdot}_{\cdot,2}$ and $A^{\cdot}_{\cdot,3}$. For instance,

$$\begin{aligned}A^{(m),--}_{ij} &= A^{(m),--}_{ij,1} + A^{(m),--}_{ij,2} + A^{(m),--}_{ij,3} \\ &:= \int_a^1\int_a^1 \frac{G_{ijm,n}(u^{(i)}_n(x), u^{(j)}_n(y), u^{(m)}_n(1)) - G_{ijm,n}(x,y,1)}{xy}dxdy \\ &\quad + \int_a^1\int_a^1 \frac{G_{ijm,n}(x,y,1)}{xy}dxdy \\ &\quad + \int_a^1\int_a^1 \frac{R_{ijm,n}(u^{(i)}_n(x), u^{(j)}_n(y), u^{(m)}_n(1)) - R_{ijm,n}(x,y,1)}{xy}dxdy.\end{aligned} \tag{3.62}$$

The first of the three terms in (3.60) is proportional to a standard empirical process evaluated at a set corresponding to the difference between $\boldsymbol{x}_J$ and $\widehat{\boldsymbol{x}}_J$. It will be uniformly bounded by using well known concentration inequalities for empirical processes appearing in Koltchinskii (2006) and Massart (2000). The second term, $G_{J,n}$, is now a rescaled sum of n independent and identically distributed (iid) processes which, when integrated as in (3.62), becomes a sum of iid, bounded random variables. We will be able to control this sum via Bernstein's inequality. Finally, the third term relates to the difference between $\boldsymbol{x}_J$ and $\widehat{\boldsymbol{x}}_J$. It can be controlled by weighted approximation results on the uniform quantile processes given in Lemmas 3.1 and 3.2. Two bounds are then proved on the corresponding term in (3.62), the stronger of which only holds under Assumption 3.2. This gives rise to the two desired results.

Technical preliminaries on uniform quantile processes

Before tackling each term in (3.62), we prove a few properties of the rescaled quantile functions $u^{(i)}_n$ which will be used throughout. In Lemma 3.1, we first prove an

approximation of $u_n^{(i)}$ by a certain Gaussian process. We then establish various properties of $u_n^{(i)}$ and its approximation in Corollary 3.1 and Lemma 3.2.

Lemma 3.1. *For any fixed $i \in V$, define the random function $u_n^{(i)}$ as in (3.52) and let $0 < \nu < 1/2$. There exist random functions $w_n^{(i)}$ defined on a possibly enriched probability space and universal constants $A, B, C \in (0, \infty)$ such that for any $z > 0$,*

$$\mathbb{P}\Big(\sup_{x \in [0, n/k]} |u_n^{(i)}(x) - w_n^{(i)}(x)| > k^{-1}(A \log n + z) \Big) \leq B e^{-Cz}.$$

Moreover, the functions $w_n^{(i)}$, along with constants $\widetilde{A}, \widetilde{B}, \widetilde{C}$ possibly depending on ν, can be chosen such that for any $z > 0$,

$$\mathbb{P}\Big(\max\Big\{ \sup_{0 \leq x \leq 1} \frac{|w_n^{(i)}(x) - x|}{x^{\nu}}, \sup_{1 \leq x \leq n/k} \frac{|w_n^{(i)}(x) - x|}{x^{1-\nu}} \Big\} > k^{-1/2}(\widetilde{A} + z) \Big) \leq \widetilde{B} e^{-\widetilde{C} z^2}.$$

Proof. By definition, the quantile function $\widehat{F}_i^-$ appearing in $u_n^{(i)}(x)$ is the right-continuous function

$$\widehat{F}_i^-(x) = U_{ni, \lfloor nx \rfloor},$$

where $U_{ni,j}$ is the jth order statistic from the sample $U_{1i}, \ldots, U_{ni}$. We use the convention $U_{ni,0} = 0$. Similarly define the left-continuous quantile function

$$\widehat{F}_i^+(x) = U_{ni, \lceil nx \rceil}.$$

Theorem 1 from Csorgo and Revesz (1978) states that for every $z > 0$,

$$\mathbb{P}\Big(\sup_{x \in [0,1]} |(\widehat{F}_i^+(x) - x) - n^{-1/2} B_n(x)| > n^{-1}(A^+ \log n + z) \Big) \leq B^+ e^{-C^+ z},$$

for positive constants A^+, B^+, C^+ and a sequence of Brownian bridges B_n. We first establish a similar tail bound for the right-continuous quantile function. Using $\lfloor y \rfloor \geq \lceil y \rceil - 1 = \lceil y - 1 \rceil$, note that

$$\widehat{F}_i^+(x) \geq \widehat{F}_i^-(x) \geq U_{ni, \lceil nx - 1 \rceil} = \widehat{F}_i^+(x - 1/n),$$

so for every $x \in [0, 1]$, using the convention $\widehat{F}_i^+(x) = \widehat{F}_i^+(0)$ if $x < 0$,

$$\begin{aligned} 0 &\leq \widehat{F}_i^+(x) - \widehat{F}_i^-(x) \leq \widehat{F}_i^+(x) - \widehat{F}_i^+(x - 1/n) \\ &\leq 2 \sup_{x \in [0,1]} |(\widehat{F}_i^+(x) - x) - n^{-1/2} B_n(x)| \\ &\quad + \Big| \Big(x + n^{-1/2} B_n(x) \Big) - \Big(x - \frac{1}{n} + n^{-1/2} B_n(x - 1/n) \Big) \Big| \end{aligned}$$

$$\leq 2 \sup_{x\in[0,1]} |(\widehat{F}_i^+(x) - x) - n^{-1/2}B_n(x)| + \frac{1}{n} + n^{-1/2}|B_n(x) - B_n(x - 1/n)|.$$

Thus

$$\sup_{x\in[0,1]} |(\widehat{F}_i^-(x) - x) - n^{-1/2}B_n(x)|$$
$$\leq 3 \sup_{x\in[0,1]} |(\widehat{F}_i^+(x) - x) - n^{-1/2}B_n(x)| + \frac{1}{n} + n^{-1/2}|B_n(x) - B_n(x - 1/n)|.$$

Using the covariance function of the standard Brownian bridge, $n^{-1/2}(B_n(x) - B_n(x - 1/n))$ is normally distributed with mean 0 and variance upper bounded by $4/n^2$. Thus, its absolute value is upper bounded by $\sqrt{z}/n$ with probability greater than $1 - e^{-z/4}$. Therefore,

$$\mathbb{P}\Big(\sup_{x\in[0,1]} |(\widehat{F}_i^-(x) - x) - n^{-1/2}B_n(x)| > n^{-1}(3A^+ \log n + 3z + 1 + \sqrt{z})\Big) \leq B^+ e^{-C^+ z} + e^{-z/4}$$

which implies, for the right choice of A, B, C,

$$\mathbb{P}\Big(\sup_{x\in[0,1]} |(\widehat{F}_i^-(x) - x) - n^{-1/2}B_n(x)| > n^{-1}(A \log n + z)\Big) \leq Be^{-Cz}. \tag{3.63}$$

Now define

$$w_n^{(i)}(x) := x + n^{-1/2}\frac{n}{k}B_n(kx/n).$$

Then

$$\begin{aligned} u_n^{(i)}(x) - w_n^{(i)}(x) &= \frac{n}{k}\widehat{F}_i^-(kx/n) - x - n^{-1/2}\frac{n}{k}B_n(kx/n) \\ &= \frac{n}{k}\Big(\widehat{F}_i^-(kx/n) - \frac{kx}{n} - n^{-1/2}B_n(kx/n)\Big). \end{aligned}$$

Thus for $w_n^{(i)}$ defined above

$$\begin{aligned} &\mathbb{P}\Big(\sup_{x\in[0,n/k]} |u_n^{(i)}(x) - w_n^{(i)}(x)| > k^{-1}(A \log n + z)\Big) \\ &= \mathbb{P}\Big(\sup_{x\in[0,n/k]} \frac{n}{k}\Big|\widehat{F}_i^-(kx/n) - \frac{kx}{n} - n^{-1/2}B_n(kx/n)\Big| > k^{-1}(A \log n + z)\Big) \\ &= \mathbb{P}\Big(\sup_{x\in[0,1]} \Big|\widehat{F}_i^-(x) - x - n^{-1/2}B_n(x)\Big| > n^{-1}(A \log n + z)\Big) \\ &\leq Be^{-Cz} \end{aligned}$$

by (3.63), which proves the first claim.

We now prove the second claim. Let

$$Z_n(x) := n^{-1/2}\frac{n}{k}B_n(kx/n).$$

Observe that

$$\begin{aligned}\{Z_n(x)\}_{x\in[0,n/k]} &\overset{\mathcal{D}}{=} \left\{n^{-1/2}\frac{n}{k}W_n(kx/n) - n^{-1/2}xW_n(1)\right\}_{x\in[0,n/k]}\\ &\overset{\mathcal{D}}{=} \left\{k^{-1/2}W_n(x) - k^{-1/2}kx/nW_n(n/k)\right\}_{x\in[0,n/k]},\end{aligned}$$

where W_n are standard Wiener processes on $[0,\infty)$. If the sequences of suprema $\sup_{0<x\leq 1} k^{1/2}|Z_n(x)|/x^\nu$ and $\sup_{1\leq x\leq n/k} k^{1/2}|Z_n(x)|/x^{1-\nu}$ are uniformly tight, their distributions have finite medians independent of n. Hence by Proposition A.2.1 of van der Vaart and Wellner (1996), there exist constants $\widetilde{A},\widetilde{B},\widetilde{C}$, depending only on those medians, such that

$$\mathbb{P}\Big(\max\Big\{\sup_{0<x\leq 1}\frac{k^{1/2}|Z_n(x)|}{x^\nu}, \sup_{1\leq x\leq \frac{n}{k}}\frac{k^{1/2}|Z_n(x)|}{x^{1-\nu}}\Big\} > \widetilde{A}+z\Big) \leq \widetilde{B}e^{-\widetilde{C}z^2}.$$

and the result follows.

To establish tightness, note that since $\nu < 1/2$,

$$\sup_{0<x\leq 1}\frac{k^{1/2}|Z_n(x)|}{x^\nu} \leq \sup_{0<x\leq 1}\frac{|W_n(x)|}{x^\nu} + \frac{k}{n}|W_n(n/k)| = O_{\mathbb{P}}\left(1+\sqrt{\frac{k}{n}}\right) = O_{\mathbb{P}}(1)$$

and

$$\begin{aligned}\sup_{1\leq x\leq \frac{n}{k}}\frac{k^{1/2}|Z_n(x)|}{x^{1-\nu}} &\leq \sup_{1\leq x<\infty}\frac{|W_n(x)|}{x^{1-\nu}} + \left(\frac{k}{n}\right)^{1-\nu}|W_n(n/k)|\\ &= O_{\mathbb{P}}\left(1+\left(\frac{k}{n}\right)^{1/2-\nu}\sqrt{\log\log\frac{n}{k}}\right)\\ &= O_{\mathbb{P}}(1),\end{aligned}$$

where we used the law of the iterated logarithm at 0 and at ∞, respectively, for Wiener processes (see for instance Durrett, 2010, Section 8.11). Note that the probability bounds on the suprema above, and hence the bounds on their medians, are uniform in n but may depend on ν, hence the dependence of the constants $\widetilde{A},\widetilde{B},\widetilde{C}$ on this parameter. □

Let $w_n^{(i)}$ be as in the proof of Lemma 3.1. For $a \in (0,1)$ and $\nu_1, \nu_2 \in [0,1]$, let

$$\widetilde{\Delta}_n^{(i)}(a, \nu_1, \nu_2) = \max\left\{ \sup_{a \le x \le 1} \frac{|w_n^{(i)}(x) - x|}{x^{\nu_1}}, \sup_{1 \le x \le n/k} \frac{|w_n^{(i)}(x) - x|}{x^{1-\nu_2}} \right\}$$

and

$$\widetilde{\Delta}_n^{(i)}(a, \nu) = \widetilde{\Delta}_n^{(i)}(a, \nu, \nu).$$

Similarly, let

$$\widehat{\Delta}_n^{(i)}(a, \nu_1, \nu_2) = \max\left\{ \sup_{a \le x \le 1} \frac{|u_n^{(i)}(x) - x|}{x^{\nu_1}}, \sup_{1 \le x \le n/k} \frac{|u_n^{(i)}(x) - x|}{x^{1-\nu_2}} \right\}$$

and

$$\widehat{\Delta}_n^{(i)}(a, \nu) = \widehat{\Delta}_n^{(i)}(a, \nu, \nu).$$

It is established in Lemma 3.1 that there are constants $\widetilde{A}, \widetilde{B}, \widetilde{C}$ only depending on $\nu \in (0, 1/2)$ such that

$$\mathbb{P}\Big(\widetilde{\Delta}_n^{(i)}(0, \nu) > k^{-1/2}(\widetilde{A} + z)\Big) \le \widetilde{B} e^{-\widetilde{C} z^2}.$$

Lemma 3.1 also allows to obtain a certain bound for the terms $\widehat{\Delta}_n^{(i)}$.

Corollary 3.1. *Let $\widehat{\Delta}_n^{(i)}$ be as above. There exist constants $\widehat{A}, \widehat{B}, \widehat{C}$ depending only on $\nu \in (1/2)$ such that for all $z \ge 0$,*

$$\mathbb{P}\left(\widehat{\Delta}_n^{(i)}(0, 0, \nu) > \widehat{A}\Big(\frac{1}{\sqrt{k}} + \frac{\log n}{k}\Big) + \sqrt{\frac{z}{k}} + \frac{z}{k}\right) \le \widehat{B} e^{-\widehat{C} z}.$$

Proof. Write

$$\begin{aligned}
\widehat{\Delta}_n^{(i)}(0, 0, \nu) &\le \widetilde{\Delta}_n^{(i)}(0, 0, \nu) + \big|\widehat{\Delta}_n^{(i)}(0, 0, \nu) - \widetilde{\Delta}_n^{(i)}(0, 0, \nu)\big| \\
&\le \widetilde{\Delta}_n^{(i)}(0, \nu) + \max\left\{ \sup_{0 \le x \le 1} |u_n^{(i)}(x) - w_n^{(i)}(x)|, \sup_{1 \le x \le n/k} \frac{|u_n^{(i)}(x) - w_n^{(i)}(x)|}{x^{1-\nu}} \right\} \\
&\le \widetilde{\Delta}_n^{(i)}(0, \nu) + \sup_{0 \le x \le n/k} |u_n^{(i)}(x) - w_n^{(i)}(x)|,
\end{aligned}$$

where the second inequality follows form the fact that for two functions f and g defined on the same domain, $|\sup_x f(x) - \sup_x g(x)| \le \sup_x |f(x) - g(x)|$.

By Lemma 3.1, the first term above is larger than $(\widetilde{A} + \sqrt{z})/\sqrt{k}$ with probability at most $\widetilde{B} e^{-\widetilde{C} z}$ and the second one is larger than $(A \log n + z)/k$ with probability at most $B e^{-C z}$. The result follows by the right choice of $\widehat{A}, \widehat{B}, \widehat{C}$. □

Lemma 3.2. *With $w_n^{(i)}$ as above, there exists a universal positive constant c' such that for all $a \in (0,1)$,*

$$\mathbb{P}\Big(\sup_{a \le x \le n/k} \frac{|w_n^{(i)}(x) - x|}{x} > 1/2 \Big) \le 6 \exp\Big\{ - c' k \Big(1 \wedge \frac{a}{\log\log(1/a)} \Big) \Big\}.$$

Proof. From the proof of Lemma 3.1, there is a standard Brownian motion W such that $\{w_n^{(i)}(x) - x\}_{x \in [a,n/k]}$ is equal in distribution to the zero-mean Gaussian process

$$\{Z_n(x)\}_{x \in [a,n/k]} := \Big\{ W(x)/\sqrt{k} - \sqrt{k} x/n W(n/k) \Big\}_{x \in [a,n/k]}.$$

Assume first that $a \le e^{-2}$. We are therefore interested in

$$\begin{aligned} &\sup_{x \in [a,n/k]} \frac{|Z_n(x)|}{x} \qquad (3.64) \\ &\le \frac{1}{\sqrt{k}} \Big(\sup_{x \in [a,n/k]} \frac{|W(x)|}{x} + \frac{k}{n} W(n/k) \Big) \\ &\le \frac{1}{\sqrt{k}} \Big(\sup_{x \in [a,e^{-2}]} \sqrt{\frac{\log\log(1/x)}{x}} \frac{|W(x)|}{\sqrt{x \log\log(1/x)}} + \sup_{x \in [e^{-2},n/k]} \frac{|W(x)|}{x} + \frac{k}{n} W(n/k) \Big) \\ &\le \frac{1}{\sqrt{k}} \Big(\sqrt{\frac{\log\log(1/a)}{a}} \sup_{x \in [0,e^{-2}]} \frac{|W(x)|}{\sqrt{x \log\log(1/x)}} + \sup_{x \in [e^{-2},\infty)} \frac{|W(x)|}{x} + \frac{k}{n} W(n/k) \Big). \qquad (3.65) \end{aligned}$$

By the laws of the iterated logarithm at 0 and at infinity, respectively, the above two suprema of Gaussian processes are tight random variables. It follows that they have finite medians and hence, by Proposition A.2.1 of van der Vaart and Wellner (1996), that they have sub-Gaussian tails. The same can be said of the uniformly (in n) tight random variable $\frac{k}{n} W(n/k)$. Therefore,

$$\begin{aligned} \mathbb{P}\Big(\sup_{x \in [a,n/k]} \frac{|Z_n(x)|}{x} > 1/2 \Big) &\le \mathbb{P}\Big(\sup_{x \in [0,e^{-2}]} \frac{|W(x)|}{\sqrt{x \log\log(1/x)}} > \frac{\sqrt{ka}}{6\sqrt{\log\log(1/a)}} \Big) \\ &\quad + \mathbb{P}\Big(\sup_{x \in [e^{-2},\infty)} \frac{|W(x)|}{x} > \frac{\sqrt{k}}{6} \Big) + \mathbb{P}\Big(\frac{k}{n} W(n/k) > \frac{\sqrt{k}}{6} \Big) \\ &\le 2 \exp\Big\{ - c_1 \frac{ka}{\log\log(1/a)} \Big\} + 2e^{-c_2 k} + 2e^{-c_3 k}, \end{aligned}$$

for some universal positive constants c_1, c_2, c_3. The result follows for some c' depending on those three constants only.

If, instead, $a > e^{-2}$, then the sum of the last two terms of (3.65) is a valid bound in itself. The rest of the proof goes through and we obtain, instead, the upper bound

$4e^{-c'k}$. □

In particular, Lemma 3.2 lower bounds the probability that for all $x \in [a, n/k]$, $x/2 \leq w_n^{(i)}(x) \leq 2x$, which will be repeatedly used in Section 3.11.3.

The following sections are respectively dedicated to each of the three terms of the decomposition introduced in (3.60).

Increments of empirical processes

We first consider the terms $A_{\cdot,1}^{\cdot}$. In this section we prove that for any $\nu \in (0, 1/2)$ there exists a constant $C_1 < \infty$ (which can also depend on the constant K from Assumption 3.1) such that for any $\varepsilon \leq 1$,

$$\begin{aligned}
&\mathbb{P}\Big(\max \big|A_{\cdot,1}^{\cdot}\big| > C_1(\log(n/k) + \log(1/a))^2 \Big\{ \Big(\frac{\log(n/k)}{k}\Big)^{1/2} \Big((k/n)^{\xi} + \varepsilon\Big)^{1/2} \\
&\qquad + \frac{\log(n/k)}{k} + \frac{\lambda}{\sqrt{k}} \Big((k/n)^{\xi} + \varepsilon\Big)^{1/2} + \frac{\lambda^2}{k} \Big\} \Big) \\
&\leq d^3 e^{-\lambda^2} + \mathbb{P}\Big(\max_{i \in V} \widehat{\Delta}_n^{(i)}(a, 0, \nu) > \varepsilon \Big).
\end{aligned}$$

Consider the following decompositions. For all $x, y \in [a, 1]$, the numerator in the integral $A_{ij,1}^{(m),--}$ satisfies

$$\begin{aligned}
&\big|G_{ijm,n}(u_n^{(i)}(x), u_n^{(j)}(y), u_n^{(m)}(1)) - G_{ijm,n}(x, y, 1)\big| \\
&\quad \leq \big|G_{ijm,n}(u_n^{(i)}(x), u_n^{(j)}(y), [1 \wedge u_n^{(m)}(1), 1 \vee u_n^{(m)}(1)])\big| \\
&\qquad + \big|G_{ijm,n}(u_n^{(i)}(x), [y \wedge u_n^{(j)}(y), y \vee u_n^{(j)}(y)], 1)\big| \\
&\qquad + \big|G_{ijm,n}([x \wedge u_n^{(i)}(x), x \vee u_n^{(i)}(x)], y, 1)\big|.
\end{aligned}$$

The numerators in $A_{im,1}^{(m),--}$ and $A_{i,1}^{(m),\ell,-}$ satisfy a similar bound with only the first two terms, up to a logarithmic factor that is everywhere bounded by $\log(1/a)$ in the case of $A_{i,1}^{(m),2,-}$.

For all $x \in [1, n/k]$, $y \in [a, 1]$, the numerator in the integral $A_{ij,1}^{(m),+-}$ satisfies

$$\begin{aligned}
&\big|G_{ijm,n}([u_n^{(i)}(x), \infty), u_n^{(j)}(y), u_n^{(m)}(1)) - G_{ijm,n}([x, \infty), y, 1)\big| \\
&\quad \leq \big|G_{ijm,n}([u_n^{(i)}(x), \infty), u_n^{(j)}(y), [1 \wedge u_n^{(m)}(1), 1 \vee u_n^{(m)}(1)])\big| \\
&\qquad + \big|G_{ijm,n}([u_n^{(i)}(x), \infty), [y \wedge u_n^{(j)}(y), y \vee u_n^{(j)}(y)], 1)\big| \\
&\qquad + \big|G_{ijm,n}([x \wedge u_n^{(i)}(x), x \vee u_n^{(i)}(x)], y, 1)\big|.
\end{aligned}$$

The numerators in $A_{im,1}^{(m),+-}$ and $A_{i,1}^{(m),\ell,+}$ satisfy a similar bound with only the first two terms, up to a logarithmic factor that is everywhere bounded by $\log(n/k)$ in the case

of $A_{i,1}^{(m),2,+}$, as well as the numerators in $A_{ij,1}^{(m),-+}$ by symmetry.

For all $x, y \in [1, n/k]$, the numerator in the integral $A_{ij,1}^{(m),++}$ satisfies

$$\begin{aligned}
&\big|G_{ijm,n}([u_n^{(i)}(x),\infty),[u_n^{(j)}(y),\infty),u_n^{(m)}(1)) - G_{ijm,n}([x,\infty),[y,\infty),1)\big| \\
&\quad\le \big|G_{ijm,n}([u_n^{(i)}(x),\infty),[u_n^{(j)}(y),\infty),[1\wedge u_n^{(m)}(1), 1\vee u_n^{(m)}(1)])\big| \\
&\qquad + \big|G_{ijm,n}([u_n^{(i)}(x),\infty),[y\wedge u_n^{(j)}(y), y\vee u_n^{(j)}(y)],1)\big| \\
&\qquad + \big|G_{ijm,n}([x\wedge u_n^{(i)}(x), x\vee u_n^{(i)}(x)],[y,\infty),1)\big|.
\end{aligned}$$

Define, for any $\varepsilon \in (0,1]$, $\mathcal{F}(\varepsilon) := \cup_{i,j,m}\mathcal{F}_{ijm}(\varepsilon)$ where

$$\mathcal{F}_{ijm}(\varepsilon) := \Big\{\frac{n}{k}\mathbb{1}_{\frac{k}{n}S} : S \in \mathcal{S}_{ijm}^- \cup \mathcal{S}_{ijm}^+\Big\},$$

and where the classes of sets $\mathcal{S}_{ijm}^-$ and $\mathcal{S}_{ijm}^+$ are defined as

$$\begin{aligned}
\mathcal{S}_{ijm}^- := \big\{\{w\in[0,\infty)^d : x-\varepsilon \le w_i \le x+\varepsilon,\, a_j \le w_j \le b_j, a_m \le w_m \le b_m\} : \\
a \le x \le 1, a_j, b_j, a_m, b_m \in [0,\infty]\big\},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{S}_{ijm}^+ := \big\{\{w\in[0,\infty)^d : x-x^{1-\nu}\varepsilon \le w_i \le x+x^{1-\nu}\varepsilon,\, a_j \le w_j \le b_j, 0 \le w_m \le 1\} : \\
1 \le x \le n/k, a_j, b_j, \in [0,\infty]\big\}.
\end{aligned}$$

Recalling the definition of $\widehat{\Delta}_n^{(i)}(a,0,\nu)$, it follows from the definition of the class $\mathcal{F}(\varepsilon)$ that whenever

$$\widehat{\Delta}_n^{(i)}(a,0,\nu) = \max_{i\in V}\max\Big\{\sup_{a\le x\le 1}|u_n^{(i)}(x)-x|, \sup_{1\le x\le n/k}\frac{|u_n^{(i)}(x)-x|}{x^{1-\nu}}\Big\} \le \varepsilon,$$

the numerator inside any of the integrals $A_{\cdot,1}^{\cdot}$ can be expressed as a sum of at most three terms of the form $(P_nf_1 - Pf_1) + (P_nf_2 - Pf_2) + (P_nf_3 - Pf_3)$, for functions $f_1, f_2, f_3 \in \mathcal{F}(\varepsilon)$. Here, P_n is the empirical distribution of the random vectors $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ whereas P is their true distribution. In the case of the terms $A_{i,1}^{(m),2,\pm}$, the sum is multiplied by a logarithmic term. In this case, all the integrals $A_{\cdot,1}^{\cdot}$ are upper bounded, in absolute value, by

$$3(\log(n/k) + \log(1/a))^2 \sup_{f\in\mathcal{F}(\varepsilon)}|P_nf - Pf|.$$

What we have established so far is that each such term $A_{\cdot,1}^{\cdot}$ satisfies, for any $t > 0$,

$$\mathbb{P}\Big(|A_{\cdot,1}^{\cdot}| \geq t\Big) \leq \mathbb{P}\Big(3(\log(n/k) + \log(1/a))^2 \sup_{f \in \mathcal{F}(\varepsilon)} |P_n f - Pf| \geq t\Big) + \mathbb{P}\Big(\max_{i \in V} \max\Big\{ \sup_{a \leq x \leq 1} |u_n^{(i)}(x) - x|, \sup_{1 \leq x \leq n/k} \frac{|u_n^{(i)}(x) - x|}{x^{1-\nu}} \Big\} > \varepsilon\Big).$$

For any triple (i, j, m), $\mathcal{F}_{ijm}(\varepsilon)$ clearly admits the constant envelope function of the form n/k. Moreover it is a VC-subgraph class that satisfies (3.92) with universal constants A and V (see for instance van der Vaart and Wellner (1996), Theorem 2.6.7). Moreover, the variance of any single function f in $\mathcal{F}_{ijm}(\varepsilon)$ is bounded by

$$\begin{aligned} Pf^2 &\leq \Big(\frac{n}{k}\Big)^2 \max\Big\{ \sup_{a \leq x \leq 1} \mathbb{P}\Big(U_i \in \frac{k}{n}[x - \varepsilon, x + \varepsilon]\Big), \\ &\qquad \sup_{1 \leq x \leq n/k} \mathbb{P}\Big(U_i \in \frac{k}{n}[x - x^{1-\nu}\varepsilon, x + x^{1-\nu}\varepsilon], U_m \leq \frac{k}{n}\Big)\Big\} \\ &\leq \frac{n}{k}\Big\{2\varepsilon \vee \sup_{1 \leq x \leq n/k} R_{im,n}([x - x^{1-\nu}\varepsilon, x + x^{1-\nu}\varepsilon], 1)\Big\} \\ &\leq \frac{n}{k}\Big\{2\varepsilon + \sup_{1 \leq x \leq n/k} R_{im}([x - x^{1-\nu}\varepsilon, x + x^{1-\nu}\varepsilon], 1) + 2K\Big(\frac{k}{n}\Big)^{\xi}\Big\} \\ &\lesssim \frac{n}{k}\Big\{\varepsilon + \Big(\frac{k}{n}\Big)^{\xi}\Big\} \end{aligned}$$

where the last two inequalities follow from (3.57) and Lemma 3.9, respectively. By (3.93) we therefore have

$$\begin{aligned} \mathbb{E}\Big[\sup_{f \in \mathcal{F}_{ijm}(\varepsilon)} |P_n f - Pf|\Big] &\lesssim k^{-1/2}\Big((k/n)^{\xi} + \varepsilon\Big)^{1/2} \log\Big((n/k)^{1/2}\Big((k/n)^{\xi} + \varepsilon\Big)^{-1/2}\Big)^{1/2} \\ &\quad + k^{-1} \log\Big((n/k)^{1/2}\Big((k/n)^{\xi} + \varepsilon\Big)^{-1/2}\Big) \\ &\lesssim \Big(\frac{\log(n/k)}{k}\Big)^{1/2}\Big((k/n)^{\xi} + \varepsilon\Big)^{1/2} + \frac{\log(n/k)}{k}. \end{aligned}$$

It follows from (3.94) that there exists a constant c such that for each triple (i, j, m) and each $\lambda > 0$,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{f \in \mathcal{F}_{ijm}(\varepsilon)} |P_n f - Pf| \geq c\Big\{\Big(\frac{\log(n/k)}{k}\Big)^{1/2}\Big((k/n)^{\xi} + \varepsilon\Big)^{1/2} + \frac{\log(n/k)}{k} \\ &\quad + \frac{\lambda}{\sqrt{k}}\Big((k/n)^{\xi} + \varepsilon\Big)^{1/2} + \frac{\lambda^2}{k}\Big\}\Big) \\ &\leq e^{-\lambda^2}. \end{aligned}$$

Combined with the union bound, this completes the proof.

Sums of iid processes

We now deal with the terms $A^{\cdot}_{\cdot,2}$ involving integrated empirical processes, such as in (3.62). In this section, we show that there exists a positive constant c_2 such that for all $\lambda > 0$,

$$\begin{aligned}\mathbb{P}\Bigg(\max \left|A^{\cdot}_{\cdot,2}\right| > \left(1 + \left(\frac{k}{n}\right)^{\xi} (\log(n/k))^2 (\log(n/k) + \log(1/a))^2\right)^{1/2} \frac{\lambda}{\sqrt{k}} \\ + (\log(n/k) + \log(1/a))^2 \frac{\lambda^2}{k}\Bigg) \\ \leq 16 d^3 e^{-c_2 \lambda^2/2}.\end{aligned}$$

Starting with $A^{(m),--}_{ij,2}$, we have by definition of $G_{ijm,n}$

$$A^{(m),--}_{ij,2} = \int_a^1 \int_a^1 \frac{G_{ijm,n}(x,y,1)}{xy} dxdy =: \sum_{t=1}^{n} \left(V^{(m),--}_{t,ijm} - \mathbb{E}\big[V^{(m),--}_{t,ijm}\big]\right),$$

where $V^{(m),--}_{t,ijm}$, $1 \leq t \leq n$, are independent copies of the random variable

$$\begin{aligned}V^{(m),--}_{ijm} &:= \frac{1}{k} \int_a^1 \int_a^1 \frac{\mathbb{1}\left\{U_i \leq \frac{k}{n}x, U_j \leq \frac{k}{n}y, U_m \leq \frac{k}{n}\right\}}{xy} dxdy \\ &= \frac{1}{k} \log\left(\frac{k}{nU_i} \wedge a^{-1}\right) \log\left(\frac{k}{nU_j} \wedge a^{-1}\right) \mathbb{1}\left\{U_i \leq \frac{k}{n}, U_j \leq \frac{k}{n}, U_m \leq \frac{k}{n}\right\}.\end{aligned}$$

Recall that by assumption $a < 1$. We may then write

$$V^{(m),--}_{ijm} = \frac{1}{k} \log(W_i) \log(W_j) \mathbb{1}\left\{W_i, W_j > 1\right\},$$

with

$$W_i := \left(\frac{k}{nU_i} \wedge a^{-1}\right)^{\mathbb{1}\left\{U_m \leq \frac{k}{n}\right\}}$$

and W_j defined the same way. We easily notice that $0 \leq V^{(m),--}_{ijm} \leq (\log(1/a))^2/k$. Moreover, an application of Lemma 3.6 (particularly (3.83)) gives

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big(V^{(m),--}_{ijm}\big) &\leq \mathbb{E}\big[\big(V^{(m),--}_{ijm}\big)^2\big] \\ &= \frac{4}{k^2} \int_a^1 \int_a^1 \frac{\frac{k}{n} R_{ijm,n}(x,y,1) |(\log x)(\log y)|}{xy} dxdy \\ &\leq \frac{4}{kn} \int_0^1 \int_0^1 \frac{|(\log x)(\log y)|}{\sqrt{xy}} dxdy\end{aligned}$$

$$= \frac{64}{kn},$$

where we used once again that $R_{ijm,n}(x,y,1) \le x \wedge y \le \sqrt{xy}$, along with the formula $\int_0^1 \log(x)/\sqrt{x}\, dx = -4$. We may therefore apply Bernstein's inequality for bounded random variables (van der Vaart and Wellner, 1996, Lemma 2.2.9) with $v = 64/k$, $M = (\log(1/a))^2/k$, which yields

$$\mathbb{P}\big(|A_{ij,2}^{(m),--}| > \lambda\big) \le 2\exp\Big\{-\frac{k\lambda^2}{2(64 + \lambda(\log(1/a))^2/3)}\Big\}.$$

Now considering $A_{ij,2}^{(m),+-}$, we use the same approach and see that

$$A_{ij,2}^{(m),+-} = \sum_{t=1}^{n} \big(V_{t,ijm}^{(m),+-} - \mathbb{E}[V_{t,ijm}^{(m),+-}]\big),$$

where

$$V_{ijm}^{(m),+-} = -\frac{1}{k}\log(W_i)\log(W_j)\mathbb{1}\{W_i < 1, W_j > 1\}$$

and W_i, W_j are as before. This time, $0 \le V_{ijm}^{(m),+-} \le (\log(n/k))\log(1/a)/k$. An application of Lemma 3.6 (this time, (3.84)) gives

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\big(V_{ijm}^{(m),+-}\big) &\le \mathbb{E}\big[\big(V_{ijm}^{(m),+-}\big)^2\big] \\
&= \frac{4}{k^2}\int_a^1\int_1^{n/k} \frac{\frac{k}{n}R_{ijm,n}([x,\infty),y,1)|(\log x)(\log y)|}{xy}dxdy \\
&\le \frac{4}{kn}\int_a^1\int_1^{n/k} \frac{\big(R_{ijm}([x,\infty),y,1) + 3K(k/n)^{\xi}\big)|(\log x)(\log y)|}{xy}dxdy,
\end{aligned}$$

by (3.57) and (3.58). By (3.59), $R_{ijm}([x,\infty),y,1) \le R_{im}([x,\infty),1) \wedge R_{jm}(y,1) \le Kx^{-\xi} \wedge y \le Kx^{-\xi/2}y^{1/2}$. The integral above is thus bounded by

$$\begin{aligned}
&K\int_0^1\int_1^{\infty} \frac{(\log x)(-\log y)}{x^{1+\xi/2}y^{1/2}}dxdy + 3K\Big(\frac{k}{n}\Big)^{\xi}\int_a^1\int_1^{n/k}\frac{(\log x)(-\log y)}{xy}dxdy \\
&\le \frac{16K}{\xi^2} + 3K\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2(\log(1/a))^2 \le C_2\Big(1 + \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2(\log(1/a))^2\Big),
\end{aligned}$$

for a suitably chosen constant C_2 depending on K and ξ only. Bernstein's inequality, with

$$v = \frac{4C_2}{k}\Big(1 + \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2(\log(1/a))^2\Big)$$

and $M = (\log(n/k))\log(1/a)/k$, therefore implies that for a positive constant c_2

depending on C_2 only,

$$\mathbb{P}\big(\big|A_{ij,2}^{(m),+-}\big| > \lambda\big) \\ \leq 2\exp\Big\{-c_2\frac{k\lambda^2}{1+(k/n)^\xi(\log(n/k))^2(\log(1/a))^2+\lambda(\log(n/k))\log(1/a)}\Big\}.$$

By symmetry, $A_{ij,2}^{(m),-+}$ admits the same bound.

As for $A_{ij,2}^{(m),++}$, we write it as

$$A_{ij,2}^{(m),++} = \sum_{t=1}^{n}\big(V_{t,ijm}^{(m),++} - \mathbb{E}\big[V_{t,ijm}^{(m),++}\big]\big),$$

where

$$V_{ijm}^{(m),++} = \frac{1}{k}\log(W_i)\log(W_j)\mathbb{1}\{W_i, W_j < 1\}$$

and W_i, W_j are as before. Again, $0 \leq V_{ijm}^{(m),++} \leq (\log(n/k))^2/k$. An application of Lemma 3.6 (this time, (3.85)) gives

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big(V_{ijm}^{(m),++}\big) &\leq \mathbb{E}\big[\big(V_{ijm}^{(m),++}\big)^2\big] \\ &= \frac{4}{k^2}\int_1^{n/k}\int_1^{n/k}\frac{\frac{k}{n}R_{ijm,n}([x,\infty),[y,\infty),1)(\log x)(\log y)}{xy}dxdy \\ &\leq \frac{4}{kn}\int_1^{n/k}\int_1^{n/k}\frac{\big(R_{ijm}([x,\infty),[y,\infty),1)+5K(k/n)^\xi\big)(\log x)(\log y)}{xy}dxdy,\end{aligned}$$

by (3.57) and (3.58). By (3.59), $R_{ijm}([x,\infty),[y,\infty),1) \leq R_{im}([x,\infty),1)\wedge R_{jm}([y,\infty),1) \leq Kx^{-\xi}\wedge Ky^{-\xi} \leq Kx^{-\xi/2}y^{-\xi/2}$. The integral above is thus bounded by

$$\begin{aligned}&K\int_1^\infty\int_1^\infty\frac{(\log x)(\log y)}{(xy)^{1+\xi/2}}dxdy + 5K\Big(\frac{k}{n}\Big)^\xi\int_1^{n/k}\int_1^{n/k}\frac{(\log x)(\log y)}{xy}dxdy \\ &\leq \frac{16K}{\xi^4} + 5K\Big(\frac{k}{n}\Big)^\xi(\log(n/k))^4 \leq C_2\Big(1+\Big(\frac{k}{n}\Big)^\xi(\log(n/k))^4\Big),\end{aligned}$$

after possibly enlarging the constant C_2. Hence by a similar application of Bernstein's inequality for bounded random variables as before,

$$\mathbb{P}\big(\big|A_{ij,2}^{(m),++}\big| > \lambda\big) \leq 2\exp\Big\{-c_2\frac{k\lambda^2}{1+(k/n)^\xi(\log(n/k))^4+\lambda(\log(n/k))^2}\Big\},$$

after possibly decreasing the (still positive) constant c_2.

Finally, the terms $A_{i,2}^{(m),\ell,-}$, $A_{i,2}^{(m),\ell,+}$, $A_{im,2}^{(m),--}$ and $A_{im,2}^{(m),+-}$ can be shown to satisfy similar tail bounds by the same strategy, using (3.81) and (3.82) instead of (3.83) to (3.85) for $A_{i,2}^{(m),\ell,-}$ and $A_{i,2}^{(m),\ell,+}$.

The conclusion of this section is that the positive constant c_2 can be chosen sufficiently small (only depending on K and ξ) such that for each term $A^{\cdot}_{\cdot,2}$ and for all $\lambda > 0$, $\mathbb{P}\big(|A^{\cdot}_{\cdot,2}| > \lambda\big)$ is bounded above by

$$2\exp\left\{ - c_2 \frac{k\lambda^2}{1 + (k/n)^{\xi}(\log(n/k))^2(\log(n/k) + \log(1/a))^2 + \lambda(\log(n/k) + \log(1/a))^2} \right\}.$$

The denominator in the exponential above is clearly upper bounded by

$$2\max\Big\{1 + (k/n)^{\xi}(\log(n/k))^2(\log(n/k) + \log(1/a))^2, \lambda(\log(n/k) + \log(1/a))^2\Big\},$$

so the whole exponential is upper bounded by

$$\max\left\{2\exp\left\{ - c_2 \frac{k\lambda^2}{2(1 + (k/n)^{\xi}(\log(n/k))^2(\log(n/k) + \log(1/a))^2)} \right\}, \right.$$
$$\left. 2\exp\left\{ - c_2 \frac{k\lambda}{2(\log(n/k) + \log(1/a))^2} \right\}.\right.$$

Deduce that at least one of

$$\mathbb{P}\left(|A^{\cdot}_{\cdot,2}| > \left(1 + \left(\frac{k}{n}\right)^{\xi}(\log(n/k))^2(\log(n/k) + \log(1/a))^2\right)^{1/2} \frac{\lambda}{\sqrt{k}} \right) \leq 2e^{-c_2\lambda^2/2}$$

or

$$\mathbb{P}\left(|A^{\cdot}_{\cdot,2}| > (\log(n/k) + \log(1/a))^2 \frac{\lambda^2}{k} \right) \leq 2e^{-c_2\lambda^2/2}$$

holds. Therefore

$$\begin{aligned}\mathbb{P}\Bigg(|A^{\cdot}_{\cdot,2}| > &\left(1 + \left(\frac{k}{n}\right)^{\xi}(\log(n/k))^2(\log(n/k) + \log(1/a))^2\right)^{1/2} \frac{\lambda}{\sqrt{k}} \\ &+ (\log(n/k) + \log(1/a))^2 \frac{\lambda^2}{k} \Bigg) \\ &\leq 2e^{-c_2\lambda^2/2},\end{aligned}$$

and a union bound allows to conclude.

Increments of rescaled copulae

It remains to bound the terms $A^{\cdot}_{\cdot,3}$, corresponding to increments of the measures $R_{J,n}$ when the rescaled quantile functions $u_n^{(i)}$ are applied to their arguments. In this section, we prove that under Assumption 3.1, there exists a constant C_3 such that for any $\lambda, \tau > 0$,

$$\mathbb{P}\Bigg(\max |A^{\cdot}_{\cdot,3}| > 3\tau(\log(n/k) + \log(1/a))^2$$

$$
\begin{aligned}
&\quad + C_3(\log(n/k) + \log(1/a))^2\Big(\frac{\widetilde{A}+\lambda}{\sqrt{k}} + \Big(\frac{k}{n}\Big)^{\xi}\Big)\Big) \\
&\leq \mathbb{P}\Big(\max_i \widetilde{\Delta}_n^{(i)}(a,\nu) > \frac{\widetilde{A}+\lambda}{\sqrt{k}}\Big) + \mathbb{P}\Big(\max_i \sup_{x\in[0,n/k]} |u_n^{(i)}(x) - w_n^{(i)}(x)| > \tau\Big) \\
&\quad + \mathbb{P}\Big(\max_i \sup_{a\leq x\leq n/k} \frac{|w_n^{(i)}(x) - x|}{x} > 1/2\Big),
\end{aligned}
$$

where the constant $\widetilde{A}$ is as in Lemma 3.1. Moreover, if Assumption 3.2 is also satisfied, C_3 can be modified in a way that only depends on ε and $K(\beta)$ and such that the slightly larger probability

$$
\mathbb{P}\Big(\max\big|A^{\cdot}_{\cdot,3}\big| > 3\tau(\log(n/k)+\log(1/a))^2 + C_3\Big(\frac{\widetilde{A}+\lambda}{\sqrt{k}} + \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k)+\log(1/a))^2\Big)\Big)
$$

admits the same upper bound.

By Assumption 3.1, the measure $R_{J,n}$ in any integrand can be replaced by R_J at the cost of adding a deterministic error of the order of $(k/n)^{\xi}$. After being integrated, such an error is of order at most $(k/n)^{\xi}(\log(n/k) + \log(1/a))^2$. We will use this fact on multiple occasions by bounding the increments of R_J instead of $R_{J,n}$.

Next we observe that by Lipschitz continuity of $R_{J,n}$ the quantities $u_n^{(i)}(x)$, $u_n^{(j)}(y)$, and $u_n^{(m)}(1)$ appearing in the arguments of R_J inside $A^{\cdot}_{\cdot,3}$ can be replaced by $w_n^{(i)}(x)$, $w_n^{(j)}(y)$, and $w_n^{(m)}(1)$ with an error that is controlled by Lemma 3.1, uniformly over x, y, i, j, m. For example

$$
\begin{aligned}
&\max_{i,j,m} \sup_{x,y\in[0,n/k]} \Big|R_{ijm,n}(u_n^{(i)}(x), u_n^{(j)}(y), u_n^{(m)}(1)) - R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), w_n^{(m)}(1))\Big| \\
&\quad \leq 3 \max_i \sup_{x\in[0,n/k]} |u_n^{(i)}(x) - w_n^{(i)}(x)|.
\end{aligned}
$$

Define

$$
\widetilde{A}_{ij,3}^{(m),--} := \int_a^1 \int_a^1 \frac{|R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}(x,y,1)|}{xy} dxdy,
$$

and similarly define other terms $\widetilde{A}^{\cdot}_{\cdot,3}$ replacing the different $A^{\cdot}_{\cdot,3}$. The difference between those quantities and the original $A^{\cdot}_{\cdot,3}$ terms that they replace can be uniformly controlled as

$$
\max\big|A^{\cdot}_{\cdot,3} - \widetilde{A}^{\cdot}_{\cdot,3}\big| \leq 3(\log(n/k) + \log(1/a))^2 \max_i \sup_{x\in[0,n/k]} |u_n^{(i)}(x) - w_n^{(i)}(x)|,
$$

some of those bounds using the fact that

$$\int \frac{(\log x)^{\ell-1}}{x} dx = \frac{(\log x)^{\ell}}{\ell} + \text{constant}. \tag{3.66}$$

We then obtain that for all $\tau > 0$,

$$\begin{aligned}&\mathbb{P}\Big(\max \left| A^{\cdot}_{\cdot,3} - \widetilde{A}^{\cdot}_{\cdot,3} \right| > 3\tau(\log(n/k) + \log(1/a))^2 \Big) \\ &\quad \leq \mathbb{P}\Big(\max_i \sup_{x \in [0,n/k]} |u_n^{(i)}(x) - w_n^{(i)}(x)| > \tau \Big).\end{aligned}$$

Hence it remains to bound the terms $\widetilde{A}^{\cdot}_{\cdot,3}$ which are defined in the same way as $A^{\cdot}_{\cdot,3}$ but with $w_n^{(i)}(x), w_n^{(j)}(y), w_n^{(m)}(1)$ replacing $u_n^{(i)}(x), u_n^{(j)}(y), u_n^{(m)}(1)$.

Note that whenever

$$\max_i \sup_{a \leq x \leq n/k} \frac{|w_n^{(i)}(x) - x|}{x} \leq 1/2,$$

we have for all i and $x \in [a, n/k]$ that $x/2 \leq u_n^{(i)}(x) \leq 2x$. We will assume, for the remainder of the section, that this is realized. Lemma 3.2 allows to lower bound the probability of that event.

Finally, recall the quantities

$$\widetilde{\Delta}_n^{(i)}(a, \nu) = \max \left\{ \sup_{x \in [a,1]} \frac{|w_n^{(i)}(x) - x|}{x^{\nu}}, \sup_{x \in [1,n/k]} \frac{|w_n^{(i)}(x) - x|}{x^{1-\nu}} \right\}$$

from the discussion after Lemma 3.1.

The general case. We first prove the weaker bound that does not rely on Assumption 3.2.

Firstly, for $x, y \in [a, 1]$ and every triple (i, j, m),

$$\begin{aligned}&R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}(x, y, 1) \\ &\quad = R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), 1) \\ &\qquad + R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), 1) - R_{ijm,n}(w_n^{(i)}(x), y, 1) \\ &\qquad + R_{ijm,n}(w_n^{(i)}(x), y, 1) - R_{ijm,n}(x, y, 1).\end{aligned} \tag{3.67}$$

Each of the three differences above, by Lipschitz continuity of $R_{ij,m}$, is of course

bounded by $\max_i \widetilde{\Delta}_n^{(i)}(a,0) \leq \max_i \widetilde{\Delta}_n^{(i)}(a,\nu)$, for arbitrary $1/2 > \nu > 0$. Deduce that

$$\begin{aligned}|\widetilde{A}_{ij,3}^{(m),--}| &\leq \int_a^1 \int_a^1 \frac{|R_{ijm,n}(w_n^{(i)}(x), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}(x,y,1)|}{xy} dxdy \\ &\leq (\log(1/a))^2 \max_i \widetilde{\Delta}_n^{(i)}(0,\nu).\end{aligned}$$

The term $\widetilde{A}_{im,3}^{(m),--}$ is bounded using the same strategy, following an expansion similar to (3.67) but without the first term. As for the term $\widetilde{A}_{i,3}^{(m),\ell,-}$, it is also bounded after an expansion similar to the third term in (3.67), using the indefinite integral (3.66).

Secondly, for all $x \in [1, n/k]$, $y \in [a,1]$ and every triple (i,j,m),

$$\begin{aligned}&R_{ijm,n}([w_n^{(i)}(x),\infty), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}([x,\infty), y, 1) \\ &\quad = R_{ijm,n}([w_n^{(i)}(x),\infty), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}([w_n^{(i)}(x),\infty), w_n^{(j)}(y), 1) \\ &\qquad + R_{ijm,n}([w_n^{(i)}(x),\infty), w_n^{(j)}(y), 1) - R_{ijm,n}([w_n^{(i)}(x),\infty), y, 1) \\ &\qquad + R_{ijm,n}([w_n^{(i)}(x),\infty), y, 1) - R_{ijm,n}([x,\infty), y, 1). \end{aligned} \tag{3.68}$$

The first two differences are again uniformly bounded by $\max_i \widetilde{\Delta}_n^{(i)}(a,\nu)$ by Lipshitz continuity. As for the third difference in (3.68), let us replace $R_{ijm,n}$ by R_{ijm} as described at the beginning of this section. We are then left with

$$R_{ijm}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)), y, 1) \leq y \frac{|w_n^{(i)}(x) - x|}{x \wedge w_n^{(i)}(x)} \leq 2x^{-\nu} y \max_i \widetilde{\Delta}_n^{(i)}(a,\nu), \tag{3.69}$$

by Lemma 3.9 and the fact that we are on the event $w_n^{(i)}(x) \geq x/2$. Deduce that

$$\begin{aligned}|\widetilde{A}_{ij,3}^{(m),+-}| &\leq \int_1^{n/k} \int_a^1 \frac{|R_{ijm,n}([w_n^{(i)}(x),\infty), w_n^{(j)}(y), w_n^{(m)}(1)) - R_{ijm,n}([x,\infty), y, 1)|}{xy} dxdy \\ &\lesssim (\log(n/k))(\log(1/a)) \max_i \widetilde{\Delta}_n^{(i)}(a,\nu) + \Big(\frac{k}{n}\Big)^{\xi} (\log(n/k))(\log(1/a)),\end{aligned}$$

where the last term comes from the approximation of $R_{ijm,n}$ by R_{ijm}. By symmetry, $A_{ij,3}^{(m),-+}$ enjoys the same bound. Moreover, $A_{im,3}^{(m),+-}$ is bounded using the same strategy, following an expansion similar to (3.68) but without the first term. As for the term $A_{i,3}^{(m),\ell,+}$, it is also bounded after an expansion similar to the third term in (3.68), using (3.66).

Thirdly, for all $x, y \in [1, n/k]$ and every triple (i,j,m),

$$\begin{aligned}&R_{ijm,n}([w_n^{(i)}(x),\infty), [w_n^{(j)}(y),\infty), w_n^{(m)}(1)) - R_{ijm,n}([x,\infty), [y,\infty), 1) \\ &\quad = R_{ijm,n}([w_n^{(i)}(x),\infty), [w_n^{(j)}(y),\infty), w_n^{(m)}(1)) - R_{ijm,n}([w_n^{(i)}(x),\infty), [w_n^{(j)}(y),\infty), 1)\end{aligned}$$

$$+ R_{ijm,n}([w_n^{(i)}(x),\infty),[w_n^{(j)}(y),\infty),1) - R_{ijm,n}([w_n^{(i)}(x),\infty),[y,\infty),1)$$
$$+ R_{ijm,n}([w_n^{(i)}(x),\infty),[y,\infty),1) - R_{ijm,n}([x,\infty),[y,\infty),1). \tag{3.70}$$

The first difference is upper bounded similarly to before by using the Lipschitz continuity of $R_{ijm,n}$. As for the second term, we once again replace $R_{ijm,n}$ by its limit R_{ijm} and obtain

$$\begin{aligned} R_{ijm}([w_n^{(i)}(x),\infty),[y\wedge w_n^{(j)}(y), y\vee w_n^{(j)}(y)),1) &\le R_{jm}([y\wedge w_n^{(j)}(y), y\vee w_n^{(j)}(y)),1) \\ &\le \frac{|w_n^{(j)}(y)-y|}{y\wedge w_n^{(j)}(y)} \le 2y^{-\nu}\max_i \widetilde{\Delta}_n^{(i)}(a,\nu). \end{aligned} \tag{3.71}$$

The third term of (3.70) admits the same bound with x replacing y. Deduce that $\big|\widetilde{A}_{ij,3}^{(m),++}\big|$ is bounded above by

$$\int_1^{n/k}\int_1^{n/k} \frac{|R_{ijm,n}([w_n^{(i)}(x),\infty),[w_n^{(j)}(y),\infty),w_n^{(m)}(1)) - R_{ijm,n}([x,\infty),[y,\infty),1)|}{xy}dxdy$$
$$\lesssim (\log(n/k))^2\max_i \widetilde{\Delta}_n^{(i)}(a,\nu) + \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2,$$

the last term again appearing by replacing $R_{ijm,n}$ by R_{ijm}.

We have therefore proved that, for any $1/2 > \nu > 0$, each term $A_{\cdot,3}^{\cdot}$ is upper bounded by a constant multiple of

$$(\log(n/k)+\log(1/a))^2\max_i \widetilde{\Delta}_n^{(i)}(a,\nu) + \Big(\frac{k}{n}\Big)^{\xi}(\log(n/k)+\log(1/a))^2.$$

Assuming bounded densities. Let us now suppose that Assumption 3.2 is satisfied with a certain $\varepsilon \in (0,4)$. While in the general case above $\nu \in (0,1/2)$ was arbitrary, let now $\nu = 1/2 - \varepsilon/8$.

The various bounds above on the numerators in the integrals $\widetilde{A}_{\cdot,3}^{\cdot}$ were for the most part uniform in the integrands x and y. By integrating them over a growing domain, a polylogarithmic factor was paid. We shall now derive more subtle bounds that are proportional to functions $f(x,y)$ such that $f(x,y)/xy$ is integrable over the infinite domain, thus allowing us to remove the extra polylogarithmic factors.

Firstly, for $x,y \in [a,1]$, consider the three terms in (3.67) and in each one, replace $R_{ijm,n}$ by R_{ijm}. By Lemma 3.10 with $\beta = \nu/2$, the third term is then bounded by

$$R_{ij}([x\wedge w_n^{(i)}(x), x\vee w_n^{(i)}(x)], y) \lesssim y^{\nu/2}\frac{|w_n^{(i)}(x)-x|}{x^{\nu/2}} \le (xy)^{\nu/2}\max_i \widetilde{\Delta}_n^{(i)}(a,\nu).$$

The second term admits the same bound up to a factor of $2^{\nu/2}$, since by assumption $w_n^{(i)}(x) \le 2x$. As for the first term, now using Lemma 3.10 with $\beta = \nu$, it is upper bounded by both

$$R_{im}(2x, [1 \wedge w_n^{(m)}(1), 1 \vee w_n^{(m)}(1)]) \lesssim x^\nu \max_i \widetilde{\Delta}_n^{(i)}(a, \nu)$$

and

$$R_{jm}(2y, [1 \wedge w_n^{(m)}(1), 1 \vee w_n^{(m)}(1)]) \lesssim y^\nu \max_i \widetilde{\Delta}_n^{(i)}(a, \nu),$$

hence by

$$(xy)^{\nu/2} \max_i \widetilde{\Delta}_n^{(i)}(a, \nu)$$

up to a constant. It then follows that

$$\begin{aligned} \big|\widetilde{A}_{ij,3}^{(m),--}\big| &\lesssim \max_i \widetilde{\Delta}_n^{(i)}(a, \nu) \int_0^1 \int_0^1 (xy)^{\nu/2-1} dxdy + \Big(\frac{k}{n}\Big)^\xi (\log(1/a))^2 \\ &\lesssim \max_i \widetilde{\Delta}_n^{(i)}(a, \nu) + \Big(\frac{k}{n}\Big)^\xi (\log(1/a))^2. \end{aligned}$$

The bounds on $A_{im,3}^{(m),--}$ and $A_{i,3}^{(m),\ell,-}$ follow from the same argument, noting for the latter that

$$\int_0^1 \frac{(\log x)^{\ell-1}}{x^{1-\nu/2}} dx < \infty.$$

Secondly, for $x \in [1, n/k]$, $y \in [a, 1]$ and every triple (i, j, m), consider the three terms in (3.68) and in each one, replace $R_{ijm,n}$ by R_{ijm}. It was already proved in the general case that by Lemma 3.9, the third term satisfies (see (3.69))

$$\begin{aligned} R_{ijm}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], y, 1) &\le R_{ij}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], y) \\ &\le 2x^{-\nu} y \max_i \widetilde{\Delta}_n^{(i)}(a, \nu). \end{aligned}$$

The second term, by an application of Lemma 3.11 with $\beta = -\varepsilon$, is upper bounded by

$$\begin{aligned} R_{ij}([w_n^{(i)}(x), \infty), [y \wedge w_n^{(j)}(y), y \vee w_n^{(j)}(y)]) &\lesssim w_n^{(i)}(x)^{-\varepsilon} (y \vee w_n^{(j)}(y))^\varepsilon |w_n^{(j)}(y) - y| \\ &\lesssim x^{-\varepsilon} y^{\varepsilon+\nu} \max_i \widetilde{\Delta}_n^{(i)}(a, \nu). \end{aligned}$$

The first term of (3.68) is upper bounded by both

$$R_{jm}(w_n^{(j)}(y), [1 \wedge w_n^{(m)}(1), 1 \vee w_n^{(m)}(1)]) \le w_n^{(j)}(y) |w_n^{(m)}(1) - 1| \lesssim y \max_i \widetilde{\Delta}_n^{(i)}(a, \nu),$$

by Lemma 3.10, and

$$R_{im}([w_n^{(i)}(x), \infty), [1 \wedge w_n^{(m)}(1), 1 \vee w_n^{(m)}(1)]) \lesssim w_n^{(i)}(x)^{-\varepsilon} |w_n^{(m)}(1) - 1|$$

$$\lesssim x^{-\varepsilon} \max_i \widetilde{\Delta}_n^{(i)}(a,\nu),$$

again by Lemma 3.11 with $\beta = -\varepsilon$. Hence the first term is in fact bounded by

$$x^{-\varepsilon/2} y^{1/2} \max_i \widetilde{\Delta}_n^{(i)}(a,\nu)$$

up to a constant. It then follows that

$$\begin{aligned}\big|\widetilde{A}_{ij,3}^{(m),+-}\big| &\lesssim \max_i \widetilde{\Delta}_n^{(i)}(a,\nu) \int_0^1 \int_1^\infty \big(x^{-1-\nu} + x^{-1-\varepsilon} y^{\varepsilon+\nu-1} + x^{-1-\varepsilon/2} y^{-1/2}\big) dx dy \\ &\quad + \Big(\frac{k}{n}\Big)^{\xi} (\log(n/k))(\log(1/a)) \\ &\lesssim \max_i \widetilde{\Delta}_n^{(i)}(a,\nu) + \Big(\frac{k}{n}\Big)^{\xi} (\log(n/k))(\log(1/a)).\end{aligned}$$

The same holds for $A_{ij,3}^{(m),-+}$ by symmetry. The bounds on $A_{im,3}^{(m),+-}$ and $A_{i,3}^{(m),\ell,+}$ follow from the same argument, noting for the latter that

$$\int_1^\infty \frac{(\log x)^{\ell-1}}{x^{1+\zeta}} dx < \infty$$

for any positive ζ.

Finally, for $x, y \in [1, n/k]$ and every triple (i,j,m), consider the three terms in (3.70) and in each one, replace $R_{ijm,n}$ by R_{ijm}. By Lemma 3.10 with $\beta = 1+\varepsilon$, the third term of (3.70) satisfies

$$\begin{aligned}R_{ijm}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], [y,\infty), 1) &\le R_{im}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], 1) \\ &\lesssim \frac{|w_n^{(i)}(x) - x|}{x^{1+\varepsilon}} \\ &\lesssim x^{-\nu-\varepsilon} \max_i \widetilde{\Delta}_n^{(i)}(a,\nu)\end{aligned}$$

but at the same time, Lemma 3.11 with $\beta = -\varepsilon/2$ yields

$$\begin{aligned}R_{ijm}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], [y,\infty), 1) &\le R_{ij}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], [y,\infty)) \\ &\lesssim y^{-\varepsilon/2} x^{\varepsilon/2} |w_n^{(i)}(x) - x| \\ &\lesssim x^{1-\nu+\varepsilon/2} y^{-\varepsilon/2} \max_i \widetilde{\Delta}_n^{(i)}(a,\nu).\end{aligned}$$

The minimum between the bounds above being smaller than their geometric mean, we then have

$$R_{ijm}([x \wedge w_n^{(i)}(x), x \vee w_n^{(i)}(x)], [y,\infty), 1) \lesssim x^{-\varepsilon/8} y^{-\varepsilon/4} \max_i \widetilde{\Delta}_n^{(i)}(a,\nu),$$

recalling that $\nu = 1/2 - \varepsilon/8$, so that $(-\nu - \varepsilon) + (1 - \nu + \varepsilon/2) = 1 - 2\nu - \varepsilon/2 = -\varepsilon/4$. The second term of (3.70) admits a similar bound since by assumption $w_n^{(i)}(x) \geq x/2$. As for the first term, by Lemma 3.11 with $\beta = -\varepsilon$ it is bounded by

$$R_{im}([x/2, \infty), [1 \wedge w_n^{(m)}(1), 1 \vee w_n^{(m)}(1)]) \leq x^{-\varepsilon} |w_n^{(m)}(1) - 1| \lesssim x^{-\varepsilon} \max_i \widetilde{\Delta}_n^{(i)}(a, \nu),$$

but also by

$$R_{jm}([y/2, \infty), [1 \wedge w_n^{(m)}(1), 1 \vee w_n^{(m)}(1)]) \leq y^{-\varepsilon} |w_n^{(m)}(1) - 1| \lesssim y^{-\varepsilon} \max_i \widetilde{\Delta}_n^{(i)}(a, \nu),$$

hence by

$$(xy)^{-\varepsilon/2} \max_i \widetilde{\Delta}_n^{(i)}(a, \nu)$$

up to a constant. It then follows that

$$\begin{aligned} \left|\widetilde{A}_{ij,3}^{(m),++}\right| &\lesssim \max_i \widetilde{\Delta}_n^{(i)}(a, \nu) \int_1^\infty \int_1^\infty \left(x^{-1-\varepsilon/8} y^{-1-\varepsilon/4} + (xy)^{-1-\varepsilon/2}\right) dx dy \\ &\quad + \left(\frac{k}{n}\right)^\xi (\log(n/k))^2 \\ &\lesssim \max_i \widetilde{\Delta}_n^{(i)}(a, \nu) + \left(\frac{k}{n}\right)^\xi (\log(n/k))^2. \end{aligned}$$

We have therefore proved that each term $A_{\cdot,3}^{\cdot}$ is upper bounded, up to a constant, by

$$\max_i \widetilde{\Delta}_n^{(i)}(a, \nu) + \left(\frac{k}{n}\right)^\xi (\log(n/k) + \log(1/a))^2.$$

□

3.11.4 Proof of Theorem 3.3

As per the statement of the theorem, pick an arbitrary $\zeta \in (0, 1)$ and assume that $k \geq n^\zeta$. Since the statement is trivial for $\lambda < 1$ (with the right choice of constants), suppose that $1 \leq \lambda \leq \sqrt{k}/(\log n)^4$. Moreover let a satisfy

$$\max\left\{\frac{\lambda^2 \log n}{k}, \left(\frac{k}{n}\right)^\xi\right\} \leq a \leq \max\left\{\frac{\lambda}{\sqrt{k} \log n}, \left(\frac{k}{n}\right)^\xi\right\}. \tag{3.72}$$

Note that by our choice of λ, the interval above is always non-empty. Introduce the notation $l_{n,a} := \log(n/k) + \log(1/a)$ and note that by (3.72), $l_{n,a} \lesssim \log(n/k)$. Consider the results of Section 3.11.3 with

$$\varepsilon := \widehat{A}\left(\frac{1}{\sqrt{k}} + \frac{\log n}{k}\right) + \frac{\lambda}{\sqrt{k}} + \frac{\lambda^2}{k} \lesssim \frac{\lambda}{\sqrt{k}} \leq \frac{1}{(\log n)^4}$$

and

$$\tau := \frac{A \log n + \lambda^2}{k},$$

for $\widehat{A}$ and A as in Corollary 3.1 and Lemma 3.1, respectively. Combining these results with those of Section 3.11.2, we obtain the following simultaneous upper bound on each integral $\mathcal{I}_\cdot$ in (3.54) to (3.56):

$$\begin{aligned} & C_1 l_{n,a}^2 \Big\{ \Big(\frac{\log(n/k)}{k} \Big)^{1/2} \Big(\Big(\frac{k}{n} \Big)^{\xi} + \varepsilon \Big)^{1/2} + \frac{\log(n/k)}{k} + \frac{\lambda}{\sqrt{k}} \Big(\Big(\frac{k}{n} \Big)^{\xi} + \varepsilon \Big)^{1/2} + \frac{\lambda^2}{k} \Big\} \\ & + \Big(1 + \Big(\frac{k}{n} \Big)^{\xi} (\log(n/k))^2 l_{n,a}^2 \Big)^{1/2} \frac{\lambda}{\sqrt{k}} + l_{n,a}^2 \frac{\lambda^2}{k} \\ & + 3\tau l_{n,a}^2 + C_3 l_{n,a}^2 \Big(\frac{\widetilde{A} + \lambda}{\sqrt{k}} + \Big(\frac{k}{n} \Big)^{\xi} \Big) \\ & + O\Big(\Big(\frac{k}{n} \Big)^{\xi} l_{n,a}^2 \Big). \end{aligned} \tag{3.73}$$

Note that $(k/n)^{\xi} (\log(n/k))^2 l_{n,a}^2 \lesssim (k/n)^{\xi} (\log(n/k))^4$ can be upper bounded by a constant only depending on ξ. Using this and the fact that $(x+y)^{1/2} \leq x^{1/2} + y^{1/2}$ for $x, y \geq 0$, we find

$$\begin{aligned} l_{n,a}^2 \Big(\frac{\log(n/k)}{k} \Big)^{1/2} \Big(\Big(\frac{k}{n} \Big)^{\xi} + \varepsilon \Big)^{1/2} &\lesssim l_{n,a}^2 \Big(\frac{k}{n} \Big)^{\xi/2} \Big(\frac{\log(n/k)}{k} \Big)^{1/2} + l_{n,a}^2 \Big(\frac{\log(n/k)}{k} \Big)^{1/2} \varepsilon^{1/2} \\ &\lesssim \frac{1}{\sqrt{k}} + l_{n,a}^2 \Big(\frac{\log(n/k)}{k} \Big)^{1/2} \lambda^{1/2} k^{-1/4} \lesssim \frac{\lambda}{\sqrt{k}} \end{aligned}$$

since $(\log n)^{5/2} \lesssim k^{1/4}$. By similar arguments using that $\varepsilon \lesssim 1/(\log n)^4$,

$$l_{n,a}^2 \Big(\Big(\frac{k}{n} \Big)^{\xi} + \varepsilon \Big)^{1/2} \lesssim 1.$$

Moreover,

$$\tau l_{n,a}^2 \lesssim \frac{(\log n)^3}{k} + l_{n,a}^2 \frac{\lambda^2}{k}.$$

In addition, notice that by our choice of λ,

$$\frac{\lambda^2}{k} \leq l_{n,a}^2 \frac{\lambda^2}{k} \lesssim \frac{\lambda (\log n)^{-2} \sqrt{k}}{k} \leq \frac{\lambda}{\sqrt{k}}$$

and that since $k \geq n^{\zeta}$,

$$\frac{(\log n)^3}{k} \lesssim \frac{1}{\sqrt{k}}.$$

Piecing those results together, (3.73) can be bounded by

$$C'\Big\{\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2+\frac{(\log(n/k))^2(1+\lambda)}{\sqrt{k}}\Big\}, \tag{3.74}$$

for the right constant C'. If Assumption 3.2 is made, the same strategy yields the sharper bound

$$\begin{aligned}
&C_1 l_{n,a}^2\Big\{\Big(\frac{\log(n/k)}{k}\Big)^{1/2}\Big(\Big(\frac{k}{n}\Big)^{\xi}+\varepsilon\Big)^{1/2}+\frac{\log(n/k)}{k}+\frac{\lambda}{\sqrt{k}}\Big(\Big(\frac{k}{n}\Big)^{\xi}+\varepsilon\Big)^{1/2}+\frac{\lambda^2}{k}\Big\}\\
&\quad+\Big(1+\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2 l_{n,a}^2\Big)^{1/2}\frac{\lambda}{\sqrt{k}}+l_{n,a}^2\frac{\lambda^2}{k}\\
&\quad+3\tau l_{n,a}^2+C_3\Big(\frac{\widetilde{A}+\lambda}{\sqrt{k}}+\Big(\frac{k}{n}\Big)^{\xi}l_{n,a}^2\Big)\\
&\quad+O\Big(\Big(\frac{k}{n}\Big)^{\xi}l_{n,a}^2\Big)\\
&\le \bar{C}'\Big\{\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))^2+\frac{1+\lambda}{\sqrt{k}}\Big\},
\end{aligned} \tag{3.75}$$

for the right constant $\bar{C}'$. It is left to control the deterministic error terms in (3.54) to (3.56) arising from the truncation of the integrals. Those terms are upper bounded by a constant multiple of

$$\begin{aligned}
&\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))+l_{n,a}^2k^{-1}+al_{n,a}\\
&\quad\lesssim\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))+\frac{1}{\sqrt{k}}+\max\Big\{\frac{\lambda}{\sqrt{k}},\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))\Big\}\\
&\quad\lesssim\Big(\frac{k}{n}\Big)^{\xi}(\log(n/k))+\frac{1+\lambda}{\sqrt{k}}
\end{aligned}$$

so they are absorbed into the bounds above. Note that this time we have used the upper bound on a in (3.72) in order to bound $al_{n,a}$.

The probability that each of the two bounds in (3.74) and (3.75) holds is at least

$$\begin{aligned}
&1-d^3e^{-\lambda^2}-\mathbb{P}\Big(\max_{i\in V}\widehat{\Delta}_n^{(i)}(a,0,\nu)>\widehat{A}\Big(\frac{1}{\sqrt{k}}+\frac{\log n}{k}\Big)+\frac{\lambda}{\sqrt{k}}+\frac{\lambda^2}{k}\Big)-16d^3e^{-c_2\lambda^2/2}\\
&\quad-\mathbb{P}\Big(\max_i\widetilde{\Delta}_n^{(i)}(a,\nu)>\frac{\widetilde{A}+\lambda}{\sqrt{k}}\Big)-\mathbb{P}\Big(\max_i\sup_{x\in[0,n/k]}|u_n^{(i)}(x)-w_n^{(i)}(x)|>\frac{A\log n+\lambda^2}{k}\Big)\\
&\quad-\mathbb{P}\Big(\max_i\sup_{a\le x\le n/k}\frac{|w_n^{(i)}(x)-x|}{x}>1/2\Big)\\
&\ge 1-d^3e^{-\lambda^2}-\widehat{B}de^{-\widehat{C}\lambda^2}-16d^3e^{-c_2\lambda^2/2}-\widetilde{B}de^{-\widetilde{C}\lambda^2}-Bde^{-C\lambda^2}
\end{aligned}$$

$$
\begin{aligned}
&- 6d \exp\Big\{ - c'k\Big(1 \wedge \frac{a}{\log\log(1/a)}\Big)\Big\} \\
&\geq 1 - Md^3 \exp\Big\{ - c \min\Big\{\lambda^2, \frac{ka}{\log\log(1/a)}\Big\}\Big\},
\end{aligned}
$$

for suitable constants M and c, where we have used Corollary 3.1 and Lemmas 3.1 and 3.2. By (3.72), since $a \geq \lambda^2(\log n)/k$, we find

$$
\frac{ka}{\log\log(1/a)} \geq \frac{\lambda^2 \log n}{\log\log k} \geq \lambda^2,
$$

so that the probability above is equal to

$$
1 - Md^3 e^{-c\lambda^2}.
$$

Combining this with (3.51) and (3.54) to (3.56) finally concludes the proof, upon noting that the factor $e_i^{(m),1} - e_j^{(m),1}$ appearing in (3.51) is upper bounded by $1 + K/\xi$ (see the proof of Lemma 3.8) and properly choosing the constants C and $\bar{C}$ in terms of C' and $\bar{C}'$. □

3.12 Auxiliary results and proofs

3.12.1 Proof of Proposition 3.1

We first show that (3.16) is sufficient for Assumptions 3.3 and 3.4, and subsequently prove the converse which turns out to be more involved.

Assumption 3.1 implies Assumptions 3.3 and 3.4: Assume that (3.16) holds for all $q \in (0, 1]$ and all sets $J \subset V$ of size 3. We then have, for any i, j, m,

$$
\begin{aligned}
R_{im}(q^{-1}, 1) &= R_{ijm}(q^{-1}, \infty, 1) \\
&\geq R_{ijm}(q^{-1}, q^{-1}, 1) \\
&\geq q^{-1}\mathbb{P}(F_i(X_i) > 0, F_j(X_j) > 0, F_m(X_m) > 1 - q) - Kq^{\xi} \\
&= 1 - Kq^{\xi},
\end{aligned}
$$

since the marginal distribution of $F_m(X_m)$ is uniform on $(0, 1)$. Thus (3.21) holds with $K_T = K$, $\xi_T = \xi$.

Now, that (3.20) follows from (3.16) when $|J| = 3$ is trivial. For the case where J is a pair, say (i, m), let $x \leq q^{-1}$, $z \leq 1$. We have

$$
\Big| q^{-1}\mathbb{P}\big(F_i(X_i) > 1 - qx, F_m(X_m) > 1 - qz\big) - R_{im}(x, z) \Big|
$$

$$
\begin{aligned}
&= \left| q^{-1}\mathbb{P}\big(F_i(X_i) > 1 - qx, F_j(X_j) > 1 - qq^{-1}, F_m(X_m) > 1 - qz\big) - R_{im}(x,z) \right| \\
&\le \left| R_{ijm}(x, q^{-1}, z) - R_{im}(x,z) \right| + Kq^{\xi} \\
&= R_{ijm}([0,x], [q^{-1}, \infty), [0,z]) + Kq^{\xi}, \qquad (3.76)
\end{aligned}
$$

using (3.16) and the representation of R_{ijm} as a non-negative measure. Then, (3.21) implies that the first term above is upper bounded by $R_{jm}([q^{-1}, \infty), 1) \le Kq^{\xi}$. Hence for pairs J, (3.20) (in fact, a stronger version thereof where one component of $\boldsymbol{x}$ is allowed to grow) holds with $K' = 2K$, $\xi' = \xi$.

Assumptions 3.3 and 3.4 imply Assumption 3.1: Assume that (3.20) and (3.21) hold for all $q \in (0,1]$ and all pairs and triples $J \subset V$ of indices. As in the statement of the result, let $\xi := \xi'\xi_T/(1 + \xi' + \xi_T)$. Let

$$
\psi := \frac{\xi'}{1 + \xi' + \xi_T} \in (0,1),
$$

and note that both $-\psi + (1-\psi)\xi'$ and $\psi\xi_T$ are equal to ξ.

We wish to bound

$$
\left| q^{-1}\mathbb{P}(F_i(X_i) > 1 - qx, F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) - R_{ijm}(x,y,z) \right| \qquad (3.77)
$$

uniformly over all $x, y \in [0, q^{-1}]$, $z \in [0,1]$. Let us divide the square $[0, q^{-1}]^2$ of possible values of (x,y) into four quadrants defined by the axes $x = q^{-\psi}$ and $y = q^{-\psi}$. First, for all $x, y, z \le q^{-\psi}$,

$$
\begin{aligned}
&\left| q^{-1}\mathbb{P}(F_i(X_i) > 1 - qx, F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) - R_{ijm}(x,y,z) \right| \\
&= \Big| q^{-1}\mathbb{P}(F_i(X_i) > 1 - q^{1-\psi}q^{\psi}x, F_j(X_j) > 1 - q^{1-\psi}q^{\psi}y, F_m(X_m) > 1 - q^{1-\psi}q^{\psi}z) \\
&\quad - q^{-\psi}R_{ijm}(q^{\psi}x, q^{\psi}y, q^{\psi}z) \Big| \\
&= q^{-\psi} \Big| q^{\psi-1}\mathbb{P}(F_i(X_i) > 1 - q^{1-\psi}q^{\psi}x, F_j(X_j) > 1 - q^{1-\psi}q^{\psi}y, F_m(X_m) > 1 - q^{1-\psi}q^{\psi}z) \\
&\quad - R_{ijm}(q^{\psi}x, q^{\psi}y, q^{\psi}z) \Big| \\
&\le K' q^{-\psi + (1-\psi)\xi'} = K' q^{\xi}, \qquad (3.78)
\end{aligned}
$$

where we applied (3.20) with q replaced by $q^{1-\psi}$, since $q^{\psi}(x,y,z) \in [0,1]^3$. This bounds (3.77) for $x, y \le q^{-\psi}$.

Second, for $q^{-\psi} \le x, y \le q^{-1}$, $z \le 1$,

$$
\begin{aligned}
z \ge R_{ijm}(x,y,z) &= zR_{ijm}(x/z, y/z, 1) \ge zR_{ijm}(q^{-\psi}, q^{-\psi}, 1) \\
&= zR_{ijm}([0, q^{-\psi}] \times [0, q^{-\psi}] \times [0,1])
\end{aligned}
$$

$$\begin{aligned}
&\geq z(R_{im}([0, q^{-\psi}], [0,1]) + R_{jm}([0, q^{-\psi}], [0,1]) - 1) \\
&\geq z\big(1 - 2K_T(q^{\psi})^{\xi_T}\big) \\
&\geq z - 2K_T q^{\xi}, \qquad (3.79)
\end{aligned}$$

using (3.21) to lower bound R_{im} and R_{jm}. Similarly,

$$\begin{aligned}
z &\geq q^{-1}\mathbb{P}(F_i(X_i) > 1 - qx, F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) \\
&\geq q^{-1}\mathbb{P}(F_i(X_i) > 1 - qq^{-\psi}, F_j(X_j) > 1 - qq^{-\psi}, F_m(X_m) > 1 - qz) \\
&\geq R_{ijm}(q^{-\psi}, q^{-\psi}, z) - K'q^{\xi} \\
&= zR_{ijm}(q^{-\psi}/z, q^{-\psi}/z, 1) - K'q^{\xi} \\
&\geq zR_{ijm}(q^{-\psi}, q^{-\psi}, 1) - K'q^{\xi},
\end{aligned}$$

where the third inequality follows from (3.78). Using the developments leading to (3.79), this lower bound is itself lower bounded by

$$z - (K' + 2K_T)q^{\xi}.$$

Deduce that (3.77) is bounded by $(K' + 2K_T)q^{\xi}$ for $q^{-\psi} \leq x, y \leq q^{-1}$.

Third, let $q^{-\psi} \leq x \leq q^{-1}$, $y \leq q^{-\psi}$, $z \leq 1$; the case where $q^{-\psi} \leq y \leq q^{-1}$ and $x \leq q^{-\psi}$ is handled symmetrically. We will again sandwich the two terms in (3.77). We first have

$$\begin{aligned}
R_{jm}(y,z) &\geq R_{ijm}(x,y,z) \\
&\geq R_{ijm}(q^{-\psi}, y, z) \\
&= R_{jm}(y,z) - (R_{jm}(y,z) - R_{ijm}(q^{-\psi}, y, z)) \\
&\geq R_{jm}(y,z) - (z - R_{im}(q^{-\psi}, z)) \\
&\geq R_{jm}(y,z) - (1 - R_{im}(q^{-\psi}, 1)) \\
&\geq R_{jm}(y,z) - K_T q^{\xi},
\end{aligned}$$

where in the last step we use (3.21). The other term in (3.77) enjoys similar upper and lower bounds: by (3.78) and by the preceding lower bound on $R_{ijm}(q^{-\psi}, y, z)$,

$$\begin{aligned}
&q^{-1}\mathbb{P}(F_i(X_i) > 1 - qx, F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) \\
&\quad \geq q^{-1}\mathbb{P}(F_i(X_i) > 1 - qq^{-\psi}, F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) \\
&\quad \geq R_{ijm}(q^{-\psi}, y, z) - K'q^{\xi} \\
&\quad \geq R_{jm}(y,z) - (K' + K_T)q^{\xi}.
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
&q^{-1}\mathbb{P}(F_i(X_i) > 1 - qx, F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) \\
&\quad \le q^{-1}\mathbb{P}(F_j(X_j) > 1 - qy, F_m(X_m) > 1 - qz) \\
&\quad \le q^{-\psi}q^{\psi-1}\mathbb{P}(F_j(X_j) > 1 - q^{1-\psi}q^{\psi}y, F_m(X_m) > 1 - q^{1-\psi}q^{\psi}z) \\
&\quad \le q^{-\psi}\big(R_{jm}(q^{\psi}y, q^{\psi}z) + K'q^{(1-\psi)\xi'}\big) \\
&\quad = R_{jm}(y, z) + K'q^{\xi},
\end{aligned}$$

where we have used (3.20) with q replaced by $q^{1-\psi}$, since $q^{\psi}(y, z) \in [0, 1]^2$. Deduce that (3.77) is bounded by $(K' + K_T)q^{\xi}$ for $y \le q^{-\psi} \le x \le q^{-1}$ (and also for $y \le q^{-\psi} \le x \le q^{-1}$ by symmetry).

We have therefore established that for all $x, y \le q^{-1}$, $z \le 1$, (3.77) is upper bounded by $(K' + 2K_T)q^{\xi}$, i.e., Assumption 3.1 is satisfied with the desired values K and ξ. □

3.12.2 Densities of Hüsler–Reiss Pareto distributions

Using the known expression for the stable tail dependence function of the bivariate Hüsler–Reiss distribution, we now show that any such distribution satisfies Assumptions 3.2 and 3.4.

Lemma 3.3. *Suppose that $\boldsymbol{Y}$ has a Hüsler–Reiss distribution with parameter matrix Γ. For any distinct pair (i, j), as long as $\lambda := \sqrt{\Gamma_{ij}} > 0$, its bivariate R-function R_{ij} satisfies the following.*

(i) For any positive ξ, there exists a finite constant K_ξ (which also depends on λ) such that

$$1 - R_{ij}(q^{-1}, 1) \le K_\xi q^{\xi}, \quad q \in (0, 1].$$

(ii) The function R_{ij} has density

$$r_{ij}(x, y) = \frac{1}{2\sqrt{2\pi}\lambda\sqrt{xy}} \exp\left\{-\frac{\lambda^2}{2} - \frac{(\log x - \log y)^2}{8\lambda^2}\right\}, \quad (x, y) \in (0, \infty)^2.$$

For any $\beta \in \mathbb{R}$, this density enjoys the upper bound

$$r_{ij}(x, y) \le \frac{K(\beta)}{x^{\beta}y^{1-\beta}}, \quad K(\beta) := \frac{\exp\{\lambda^2(2(\beta - 1/2)^2 - 1/2)\}}{2\sqrt{2\pi}\lambda}.$$

Proof. The pair (Y_i, Y_j) has a bivariate Hüsler–Reiss distribution with dependence

parameter λ^2, so its stable tail dependence function is

$$L_{ij}(x,y) = x\Phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) + y\Phi\bigg(\lambda + \frac{\log y - \log x}{2\lambda}\bigg),$$

so R_{ij} is given by

$$R_{ij}(x,y) = x\Phi^c\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) + y\Phi^c\bigg(\lambda + \frac{\log y - \log x}{2\lambda}\bigg),$$

where Φ^c denotes the standard Gaussian survival function.

Proof of *(i)***:** Fix a number $\xi > 0$. First note that if $q \geq e^{-4\lambda^2}$, we trivially have that

$$1 - R_{ij}(q^{-1}, 1) \leq 1 \leq e^{4\lambda^2\xi} q^{\xi},$$

so we shall assume without loss of generality that $q \leq e^{-4\lambda^2}$. This implies that $\log q^{-1} \geq 4\lambda^2$, or equivalently

$$\frac{\log q^{-1}}{2\lambda} - \lambda \geq \frac{\log q^{-1}}{4\lambda}.$$

We then have

$$\begin{aligned} 1 - R_{ij}(q^{-1}, 1) &\leq 1 - \Phi^c\bigg(\lambda + \frac{\log q}{2\lambda}\bigg) = \Phi^c\bigg(\frac{\log q^{-1}}{2\lambda} - \lambda\bigg) \leq \Phi^c\bigg(\frac{\log q^{-1}}{4\lambda}\bigg) \\ &\leq \frac{4\lambda}{\sqrt{2\pi}\log q^{-1}} \exp\bigg\{ - \frac{1}{32\lambda^2}\big(\log q^{-1}\big)^2 \bigg\}, \end{aligned}$$

the last inequality following from well known bounds on the Gaussian tails (Durrett, 2010, Theorem 1.2.3). This is in turn upper bounded by

$$\frac{1}{\sqrt{2\pi}\lambda} q^{(\log q^{-1})/32\lambda^2},$$

which is of smaller order than any power of q since the exponent diverges as $q \downarrow 0$. We can therefore upper bound it by any power q^{ξ}, up to a multiplicative constant depending on both ξ and λ.

Proof of *(ii)***:** The density of R_{ij} is defined as

$$r_{ij}(x,y) := \frac{\partial^2}{\partial x \partial y} R_{ij}(x,y) = -\frac{\partial^2}{\partial x \partial y} L_{ij}(x,y).$$

First, we have

$$\frac{\partial}{\partial x} x\Phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) = \Phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) + x\phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg)\frac{1}{2\lambda x}$$

$$= \Phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) + \frac{1}{2\lambda}\phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg),$$

so

$$\begin{aligned}
&\frac{\partial^2}{\partial x \partial y} x\Phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) \\
&\quad = -\frac{1}{2\lambda y}\phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) - \frac{1}{4\lambda^2 y}\phi'\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg) \\
&\quad = -\frac{1}{4\lambda^2 y}\bigg(\lambda + \frac{\log y - \log x}{2\lambda}\bigg)\phi\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg),
\end{aligned}$$

where we used the expression $\phi'(t) = -t\phi(t)$ for the derivative of the standard Gaussian density ϕ. Now by definition of ϕ, this is equal to

$$\begin{aligned}
&-\frac{1}{4\sqrt{2\pi}\lambda^2 y}\bigg(\lambda + \frac{\log y - \log x}{2\lambda}\bigg)\exp\bigg\{-\frac{\lambda^2}{2} - \frac{(\log x - \log y)^2}{8\lambda^2} + \frac{\log y - \log x}{2}\bigg\} \\
&\quad = -\frac{1}{4\sqrt{2\pi}\lambda^2\sqrt{xy}}\bigg(\lambda + \frac{\log y - \log x}{2\lambda}\bigg)\exp\bigg\{-\frac{\lambda^2}{2} - \frac{(\log x - \log y)^2}{8\lambda^2}\bigg\}.
\end{aligned}$$

Adding this to

$$\begin{aligned}
&\frac{\partial^2}{\partial x \partial y} y\Phi\bigg(\lambda + \frac{\log y - \log x}{2\lambda}\bigg) \\
&\quad = -\frac{1}{4\sqrt{2\pi}\lambda^2\sqrt{xy}}\bigg(\lambda + \frac{\log x - \log y}{2\lambda}\bigg)\exp\bigg\{-\frac{\lambda^2}{2} - \frac{(\log x - \log y)^2}{8\lambda^2}\bigg\},
\end{aligned}$$

obtained by a symmetric argument, yields the desired density. As for the upper bound, note that for any $\beta \in \mathbb{R}$,

$$r_{ij}(x,y) = \frac{1}{2\sqrt{2\pi}\lambda x^\beta \sqrt{y}}\exp\bigg\{-\frac{\lambda^2}{2} - \frac{(\log x - \log y)^2}{8\lambda^2} + (\beta - 1/2)\log x\bigg\}.$$

Writing u and v for $\log x$ and $\log y$, the exponent above is

$$-\frac{\lambda^2}{2} - \frac{(u-v)^2}{8\lambda^2} + (\beta - 1/2)u = \frac{-u^2}{8\lambda^2} + \Big((\beta - 1/2) + \frac{v}{4\lambda^2}\Big)u - \frac{v^2}{8\lambda^2}$$

which is maximized (in u) at $u = v + 4\lambda^2(\beta - 1/2)$, hence

$$\begin{aligned}
-\frac{\lambda^2}{2} - \frac{(u-v)^2}{8\lambda^2} + (\beta - 1/2)u &\le -\frac{\lambda^2}{2} - 2\lambda^2(\beta - 1/2)^2 + (\beta - 1/2)v + 4\lambda^2(\beta - 1/2)^2 \\
&= (\beta - 1/2)v + \lambda^2(2(\beta - 1/2)^2 - 1/2).
\end{aligned}$$

Conclude that

$$r_{ij}(x,y) \leq \frac{1}{2\sqrt{2\pi}\lambda x^{\beta} y^{1-\beta}} \exp\{\lambda^2(2(\beta-1/2)^2 - 1/2)\}.$$

□

3.12.3 The moments $e_m^{(m),\ell}$

Recalling that for any m, $Y_m^{(m)}$ has a unit Pareto distribution, and thus that $\log Y_m^{(m)}$ has a unit exponential distribution, it is evident that $e_m^{(m),1} = 1$ and $e_m^{(m),2} = 2$. As for the empirical versions $\widehat{e}_m^{(m),\ell}$ of those moments, they are in fact deterministic, since the terms $\widehat{F}_m(U_{tm})$ appearing in the sum are exactly the k smallest such terms $\{1/n, \ldots, k/n\}$. Precisely, we have the following result.

Lemma 3.4. *As long as $k \geq 3$, we have*

$$\left|\widehat{e}_m^{(m),1} - 1\right| \leq \frac{3\log k}{k}, \quad \left|\widehat{e}_m^{(m),2} - 2\right| \leq \frac{8(\log k)^2}{k}.$$

Proof. By definition, we have

$$\widehat{e}_m^{(m),\ell} = \frac{1}{k}\sum_{j=1}^{k} \{\log(k/j)\}^{\ell} = \begin{cases} \log k - \frac{1}{k}\sum_{j=1}^{k} \log j, & \ell = 1 \\ (\log k)^2 - \frac{2\log k}{k}\sum_{j=1}^{k} \log j + \frac{1}{k}\sum_{j=1}^{k} (\log j)^2, & \ell = 2 \end{cases}. \tag{3.80}$$

Note that

$$\sum_{j=1}^{k} (\log j)^{\ell} = \sum_{j=2}^{k} \int_{j}^{j+1} (\log j)^{\ell} dt \in \left[\int_{1}^{k} (\log t)^{\ell} dt, \int_{2}^{k+1} (\log t)^{\ell} dt \right].$$

Evaluating those integrals yields

$$k\{\log k - 1\} + 1 \leq \sum_{j=1}^{k} \log j \leq (k+1)\{\log(k+1) - 1\} - 2(\log 2 - 1)$$

and

$$\begin{aligned} k\big\{(\log k)^2 - 2\log k + 2\big\} - 2 &\leq \sum_{j=1}^{k} (\log j)^2 \\ &\leq (k+1)\big\{(\log(k+1))^2 - 2\log(k+1) + 2\big\} - 2\big\{(\log 2)^2 - 2\log 2 + 2\big\}. \end{aligned}$$

Denote by a_ℓ and b_ℓ the lower and upper bound on $\sum_{j=1}^{k} (\log j)^{\ell}$ above, $\ell \in \{1, 2\}$. As

long as $k \geq 3$, we have by (3.80) and by simple computations

$$k\left|\widehat{e}_m^{(m),1} - 1\right| \leq |a_1 - k\log k + k| \vee |b_1 - k\log k + k| \leq 3\log k$$

and

$$\begin{aligned} k\left|\widehat{e}_m^{(m),2} - 2\right| &\leq |a_2 - 2(\log k)b_1 + k(\log k)^2 - 2k| \vee |b_2 - 2(\log k)a_1 + k(\log k)^2 - 2k| \\ &\leq 8(\log k)^2, \end{aligned}$$

which is the desired result. □

3.12.4 Verifying the integral representations of different moments

We start by deriving general expressions for the moments of logarithms of random vectors which will lead to proving the representations in (3.54) to (3.56). The following result is a multivariate version of the so-called "Darth Vader rule".

Lemma 3.5. *Let $X_1, \ldots, X_d$ be non-negative random variables and $p_1, \ldots, p_d > 0$. Then*

$$\mathbb{E}\left[\prod_{j=1}^d X_j^{p_j}\right] = \int_{[0,\infty)^d} \prod_{j=1}^d p_j x_j^{p_j - 1} \mathbb{P}(X_1 \geq x_1, \ldots, X_d \geq x_d) dx_1 \ldots dx_d.$$

Moreover, any number of "$\geq$" can be replaced by "$>$", as this changes the value of the probability, at most, on a Lebesgue-null set.

Proof. Letting $(\Omega, \mathcal{F}, P)$ be the underlying probability space containing all the random variables, we have

$$\begin{aligned} \mathbb{E}\left[\prod_{j=1}^d X_j^{p_j}\right] &= \int_\Omega \prod_{j=1}^d X_j(\omega)^{p_j} P(d\omega) \\ &= \int_\Omega \int_{[0, X_1(\omega)^{p_1}] \times \cdots \times [0, X_d(\omega)^{p_d}]} du_1 \ldots du_d P(d\omega) \\ &= \int_\Omega \int_{[0,\infty)^d} \mathbb{1}\left\{X_1(\omega) \geq u_1^{1/p_1}, \ldots, X_d(\omega) \geq u_d^{1/p_d}\right\} du_1 \ldots du_d P(d\omega) \\ &= \int_{[0,\infty)^d} \left(\int_\Omega \mathbb{1}\left\{X_1(\omega) \geq u_1^{1/p_1}, \ldots, X_d(\omega) \geq u_d^{1/p_d}\right\} P(d\omega) \right) du_1 \ldots du_d \\ &= \int_{[0,\infty)^d} \mathbb{P}\big(X_1 \geq u_1^{1/p_1}, \ldots, X_d \geq u_d^{1/p_d}\big) du_1 \ldots du_d, \end{aligned}$$

where we have use the fact that $X_j(\omega) \geq 0$ for almost every ω to justify the second equality. The change in the order of integration was allowed by Tonelli's theorem.

Finally, applying the change of variable $x_j = u_j^{1/p_j}$, $du_j/dx_j = p_j x_j^{p_j-1}$ produces the desired result. □

Lemma 3.6. *Let X and Y be almost surely positive random variables and let S be the distribution function of $(1/X, 1/Y)$, so that for positive x, y, $\mathbb{P}(X \geq x, Y \geq y) = S(1/x, 1/y)$. Then for any $p \in \{1, 2, \dots\}$,*

$$\mathbb{E}\big[(\log X)^p \mathbb{1}\{X > 1\}\big] = p \int_0^1 \frac{S(x,\infty)|\log x|^{p-1}}{x} dx, \tag{3.81}$$

$$\mathbb{E}\big[(-\log X)^p \mathbb{1}\{X < 1\}\big] = p \int_1^\infty \frac{S([x,\infty),\infty)|\log x|^{p-1}}{x} dx, \tag{3.82}$$

$$\mathbb{E}\big[((\log X)(\log Y))^p \mathbb{1}\{X, Y > 1\}\big] = p^2 \int_0^1 \int_0^1 \frac{S(x,y)|(\log x)(\log y)|^{p-1}}{xy} dxdy, \tag{3.83}$$

$$\mathbb{E}\big[(-(\log X)(\log Y))^p \mathbb{1}\{X < 1, Y > 1\}\big] = p^2 \int_0^1 \int_1^\infty \frac{S([x,\infty),y)|(\log x)(\log y)|^{p-1}}{xy} dxdy, \tag{3.84}$$

$$\mathbb{E}\big[((\log X)(\log Y))^p \mathbb{1}\{X, Y < 1\}\big] = p^2 \int_1^\infty \int_1^\infty \frac{S([x,\infty),[y,\infty))|(\log x)(\log y)|^{p-1}}{xy} dxdy, \tag{3.85}$$

where $S([x,\infty),y)$ and $S([x,\infty),[y,\infty))$ are shorthand for $S(\infty,y) - S(x,y)$ and $1 - S(x,\infty) - S(\infty,y) + S(x,y)$, respectively.

Proof. First, by Lemma 3.5 with $d = 1$, $p_1 = p$,

$$\begin{aligned}\mathbb{E}\big[(\log X)^p \mathbb{1}\{X > 1\}\big] &= p \int_0^\infty u^{p-1} \mathbb{P}(\log X \geq u) du \\ &= p \int_0^\infty u^{p-1} S(e^{-u}, \infty) du \\ &= p \int_0^1 \frac{S(x,\infty)(-\log x)^{p-1}}{x} dx,\end{aligned}$$

by the change of variable $x = e^{-u}$. Similarly,

$$\begin{aligned}\mathbb{E}\big[(-\log X)^p \mathbb{1}\{X < 1\}\big] &= p \int_0^\infty u^{p-1} \mathbb{P}(\log X \leq -u) du \\ &= p \int_0^\infty u^{p-1} S([e^u,\infty),\infty) du \\ &= p \int_1^\infty \frac{S([x,\infty),\infty)(\log x)^{p-1}}{x} dx,\end{aligned}$$

by the change of variable $x = e^u$. This establishes (3.81) and (3.82).

(3.83) to (3.85) are proved in a similar fashion by using Lemma 3.5 with $d = 2$, $p_1 = p_2 = p$. First,

$$\begin{aligned}\mathbb{E}[((\log X)(\log Y))^p \mathbb{1}\{X, Y > 1\}] &= p^2 \int_0^\infty \int_0^\infty (uv)^{p-1} \mathbb{P}(\log X \geq u, \log Y \geq v) du dv \\ &= p^2 \int_0^\infty \int_0^\infty (uv)^{p-1} S(e^{-u}, e^{-v}) du dv \\ &= p^2 \int_0^1 \int_0^1 \frac{S(x, y)((\log x)(\log y))^{p-1}}{xy} dx dy,\end{aligned}$$

using the change of variable $x = e^{-u}$, $y = e^{-v}$. Second,

$$\begin{aligned}&\mathbb{E}[((-\log X)(\log Y))^p \mathbb{1}\{X < 1, Y > 1\}] \\ &= p^2 \int_0^\infty \int_0^\infty (uv)^{p-1} \mathbb{P}(\log X \leq -u, \log Y \geq v) du dv \\ &= p^2 \int_0^\infty \int_0^\infty (uv)^{p-1} S([e^u, \infty), e^{-v}) du dv \\ &= p^2 \int_0^1 \int_1^\infty \frac{S([x, \infty), y)(-(\log x)(\log y))^{p-1}}{xy} dx dy,\end{aligned}$$

using the change of variable $x = e^u$, $y = e^{-v}$. Third,

$$\begin{aligned}&\mathbb{E}[((\log X)(\log Y))^p \mathbb{1}\{X, Y < 1\}] \\ &= p^2 \int_0^\infty \int_0^\infty (uv)^{p-1} \mathbb{P}(\log X \leq -u, \log Y \leq -v) du dv \\ &= p^2 \int_0^\infty \int_0^\infty (uv)^{p-1} S([e^u, \infty), [e^v, \infty)) du dv \\ &= p^2 \int_1^\infty \int_1^\infty \frac{S([x, \infty), [y, \infty))((\log x)(\log y))^{p-1}}{xy} dx dy,\end{aligned}$$

using the change of variable $x = e^u$, $y = e^v$. This establishes (3.83) to (3.85). □

Lemma 3.7. *Under Assumption 3.1,* (3.54) *to* (3.56) *hold for any* $a \in (0, 1)$.

Proof. Recall that i, j, m are assumed to be distinct indices. It is already proved in (Engelke and Volgushev, 2020, Section S.7) that the moments of interest satisfy

$$e_i^{(m),\ell} = \int_0^1 \frac{R_{im}(x, 1)(-2 \log x)^{\ell-1}}{x} dx - \int_1^\infty \frac{R_{im}([x, \infty), 1)(-2 \log x)^{\ell-1}}{x} dx, \quad (3.86)$$

$$e_{im}^{(m)} = \int_0^1 \int_0^1 \frac{R_{im}(x, y)}{xy} dx dy - \int_0^1 \int_1^\infty \frac{R_{im}([x, \infty), y)}{xy} dx dy, \quad (3.87)$$

$$e_{ij}^{(m)} = \int_0^1\int_0^1 \frac{R_{ijm}(x,y,1)}{xy}dxdy - \int_0^1\int_1^\infty \frac{R_{ijm}([x,\infty),y,1)}{xy}dxdy - \int_1^\infty\int_0^1 \frac{R_{ijm}(x,[y,\infty),1)}{xy}dxdy + \int_1^\infty\int_1^\infty \frac{R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy \tag{3.88}$$

and that their empirical versions satisfy

$$\widehat{e}_i^{(m),\ell} = \int_{1/k}^1 \frac{\bar{R}_{im}(x,1)(-2\log x)^{\ell-1}}{x}dx - \int_1^{n/k} \frac{\bar{R}_{im}([x,\infty),1)(-2\log x)^{\ell-1}}{x}dx, \tag{3.89}$$

$$\widehat{e}_{im}^{(m)} = \int_{1/k}^1\int_{1/k}^1 \frac{\bar{R}_{im}(x,y)}{xy}dxdy - \int_{1/k}^1\int_1^{n/k} \frac{\bar{R}_{im}([x,\infty),y)}{xy}dxdy, \tag{3.90}$$

$$\widehat{e}_{ij}^{(m)} = \int_{1/k}^1\int_{1/k}^1 \frac{\bar{R}_{ijm}(x,y,1)}{xy}dxdy - \int_{1/k}^1\int_1^{n/k} \frac{\bar{R}_{ijm}([x,\infty),y,1)}{xy}dxdy - \int_1^{n/k}\int_{1/k}^1 \frac{\bar{R}_{ijm}(x,[y,\infty),1)}{xy}dxdy + \int_1^{n/k}\int_1^{n/k} \frac{\bar{R}_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy, \tag{3.91}$$

where

$$\bar{R}_J(\boldsymbol{x}_J) := \frac{1}{k}\sum_{t=1}^n \mathbb{1}\left\{\widehat{F}_i(U_{ti}) \le \frac{k}{n}x_i, i \in J\right\}, \quad \boldsymbol{x}_J := (x_i)_{i\in J} \in [0,\infty)^{|J|}.$$

The integrals in (3.86) to (3.88) can be truncated above by using (3.59), which allows to upper bound the tails of the functions R_J. In particular, we have

$$\int_{n/k}^\infty \frac{R_{im}([x,\infty),1)(2\log x)^{\ell-1}}{x}dx \lesssim \int_{n/k}^\infty \frac{(\log x)^{\ell-1}}{x^{1+\xi}}dx \lesssim \left(\frac{k}{n}\right)^\xi \log(n/k),$$

$$\begin{aligned}\int_0^1\int_{n/k}^\infty \frac{R_{im}([x,\infty),y)}{xy}dxdy &= \int_0^1\int_{n/k}^\infty \frac{R_{im}([x/y,\infty),1)}{x}dxdy \\ &\lesssim \int_0^1\int_{n/k}^\infty \frac{(x/y)^{-\xi}}{x}dxdy \\ &\lesssim \left(\frac{k}{n}\right)^\xi,\end{aligned}$$

and

$$\begin{aligned}&\iint_{[1,\infty)^2\setminus[1,n/k]^2} \frac{R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy \\ &= \int_{n/k}^\infty\int_{n/k}^\infty \frac{R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy + \int_1^{n/k}\int_{n/k}^\infty \frac{R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy\end{aligned}$$

$$
\begin{aligned}
&+ \int_{n/k}^{\infty}\int_{1}^{n/k} \frac{R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy \\
&\leq \int_{n/k}^{\infty}\int_{n/k}^{\infty} \frac{R_{ijm}([x,\infty),[y,\infty),1)}{xy}dxdy + \int_{1}^{n/k}\int_{n/k}^{\infty} \frac{R_{im}([x,\infty),1)}{xy}dxdy \\
&+ \int_{n/k}^{\infty}\int_{1}^{n/k} \frac{R_{jm}([y,\infty),1)}{xy}dxdy \\
&\lesssim \int_{n/k}^{\infty}\int_{n/k}^{\infty} \frac{x^{-\xi}\wedge y^{-\xi}}{xy}dxdy + 2\int_{1}^{n/k}\int_{n/k}^{\infty} \frac{x^{-\xi}}{xy}dxdy \\
&\lesssim \left(\frac{k}{n}\right)^{\xi}\log(n/k).
\end{aligned}
$$

Hence we proved that all integral can be truncated above at n/k while incurring an error of at most $O((k/n)^{\xi}\log(n/k))$. Next we show that the integrals can as well be truncated below.

Recall that $a \in (0,1)$. Since by their definitions, R_J and $\bar{R}_J$ are both upper bounded by the minimum component of their argument, so is $|\bar{R}_J - R_J|$. We then have for $\ell \in \{1,2\}$

$$
\int_0^a \frac{|\bar{R}_{im}(x,1) - R_{im}(x,1)|(-2\log x)^{\ell-1}}{x}dx \leq \int_0^a (-2\log x)^{\ell-1}dx \lesssim a(1+\log(1/a)),
$$

$$
\begin{aligned}
&\iint_{[0,1]^2\setminus[a,1]^2} \frac{|\bar{R}_{im}(x,y) - R_{im}(x,y)|}{xy}dxdy \\
&\quad\leq \int_0^a\int_0^a \frac{x\wedge y}{xy}dxdy + 2\int_a^1\int_0^a \frac{1}{y}dxdy \lesssim a(1+\log(1/a)),
\end{aligned}
$$

$$
\int_0^a\int_1^{n/k} \frac{|\bar{R}_{im}([x,\infty),y) - R_{im}([x,\infty),y)|}{xy}dxdy \leq \int_0^a\int_1^{n/k}\frac{1}{x}dxdy \leq a\log(n/k),
$$

and by symmetry

$$
\int_1^{n/k}\int_0^a \frac{|\bar{R}_{im}(x,[y,\infty)) - R_{im}(x,[y,\infty))|}{xy}dxdy
$$

admits the same bound as the latter integral. Finally,

$$
\iint_{[0,1]^2\setminus[a,1]^2} \frac{|\bar{R}_{ijm}(x,y,1) - R_{ijm}(x,y,1)|}{xy}dxdy
$$

is handled similarly as

$$\iint_{[0,1]^2\setminus[a,1]^2} \frac{|\bar{R}_{im}(x,y) - R_{im}(x,y)|}{xy} dxdy.$$

We have therefore proved that each of the integrals in (3.86) to (3.91) can be truncated below at a point a and above at n/k, up to a deterministic additive error which satisfies the bound $\lesssim (k/n)^\xi \log(n/k) + a(\log(n/k) + \log(1/a))$. It follows that with probability 1,

$$\begin{aligned}
\widehat{e}_i^{(m),\ell} - e_i^{(m),\ell} &= \int_a^1 \frac{\big(\bar{R}_{im}(x,1) - R_{im}(x,1)\big)(-2\log x)^{\ell-1}}{x} dx \\
&\quad - \int_1^{n/k} \frac{\big(\bar{R}_{im}([x,\infty),1) - R_{im}([x,\infty),1)\big)(-2\log x)^{\ell-1}}{x} dx \\
&\quad + O\bigg(\Big(\frac{k}{n}\Big)^\xi \log(n/k) + a(\log(n/k) + \log(1/a))\bigg), \\
\widehat{e}_{im}^{(m)} - e_{im}^{(m)} &= \int_a^1 \int_a^1 \frac{\bar{R}_{im}(x,y) - R_{im}(x,y)}{xy} dxdy \\
&\quad - \int_a^1 \int_1^{n/k} \frac{\bar{R}_{im}([x,\infty),y) - R_{im}([x,\infty),y)}{xy} dxdy \\
&\quad + O\bigg(\Big(\frac{k}{n}\Big)^\xi \log(n/k) + a(\log(n/k) + \log(1/a))\bigg), \\
\widehat{e}_{ij}^{(m)} - e_{ij}^{(m)} &= \int_a^1 \int_a^1 \frac{\bar{R}_{ijm}(x,y,1) - R_{ijm}(x,y,1)}{xy} dxdy \\
&\quad - \int_a^1 \int_1^{n/k} \frac{\bar{R}_{ijm}([x,\infty),y,1) - R_{ijm}([x,\infty),y,1)}{xy} dxdy \\
&\quad - \int_1^{n/k} \int_a^1 \frac{\bar{R}_{ijm}(x,[y,\infty),1) - R_{ijm}(x,[y,\infty),1)}{xy} dxdy \\
&\quad + \int_1^{n/k} \int_1^{n/k} \frac{\bar{R}_{ijm}([x,\infty),[y,\infty),1) - R_{ijm}([x,\infty),[y,\infty),1)}{xy} dxdy \\
&\quad + O\bigg(\Big(\frac{k}{n}\Big)^\xi \log(n/k) + a(\log(n/k) + \log(1/a))\bigg),
\end{aligned}$$

where the error terms are deterministic. All that remains to obtain the desired result is to replace the functions $\bar{R}_J$ above by $\widehat{R}_J$, which amounts to comparing the left- and right-continuous versions of an empirical tail copula. By the result in Appendix C.1 of Radulović et al. (2017), we have

$$\max_{J:|J|\le 3} \sup_{\boldsymbol{x}_J \in [0,\infty)^{|J|}} |\bar{R}_J(\boldsymbol{x}_J) - \widehat{R}_J(\boldsymbol{x}_J)| \le \frac{3}{k}$$

almost surely, so replacing $\bar{R}_J$ by $\widehat{R}_J$ in the integrals above adds an error that is at most of the order of $(\log(n/k) + \log(1/a))^2/k$. □

Lemma 3.8. *Under Assumption 3.1,* $\max_{m\in V} \left\|\Gamma^{(m)}\right\|_\infty$ *admits an upper bound that depends only on* K *and* ξ.

Proof. First, as is pointed out in Section 3.12.3, for $\ell \in \{1, 2\}$, $e_m^{(m),\ell} = \ell$.

The remaining arguments are based on (3.59), which holds by assumption and states that for all distinct triples (i, j, m),

$$R_{ij}([x,\infty), 1) \le Kx^{-\xi}, \quad R_{ijm}([x,\infty), [y,\infty), 1) \le K(x \wedge y)^{-\xi}, \quad x, y \ge 1.$$

Equally important is the fact that every function R_J is upper bounded by its minimum argument. The proof consists of plugging those different bounds in (3.86) to (3.88) above, which provided expressions for the moments $e_i^{(m),\ell}$, $e_{im}^{(m)}$ and $e_{ij}^{(m)}$. Repeatedly using the inequality $a \wedge b \le (ab)^{1/2}$ for positive a, b, deduce that

$$\begin{aligned}
\left|e_i^{(m),\ell}\right| &\le \int_0^1 (-2\log x)^{\ell-1} dx + K \int_1^\infty \frac{(-2\log x)^{\ell-1}}{x^{1+\xi}} dx, \\
\left|e_{im}^{(m)}\right| &\le \int_0^1 \int_0^1 (xy)^{-1/2} dxdy + \sqrt{K} \int_0^1 \int_1^\infty x^{-1-\xi/2} y^{-1/2} dxdy, \\
\left|e_{ij}^{(m)}\right| &\le \int_0^1 \int_0^1 (xy)^{-1/2} dxdy + 2\sqrt{K} \int_0^1 \int_1^\infty x^{-1-\xi/2} y^{-1/2} dxdy \\
&\quad + K \int_1^\infty \int_1^\infty (xy)^{-1-\xi/2} dxdy.
\end{aligned}$$

Simply plugging those bounds in (3.50) yields the result. □

3.12.5 Bounds on the measures R_{ij}

Recall the representation of R_{ij} as a non-negative measure, for an arbitrary pair $i \ne j$. The following bounds necessarily hold.

Lemma 3.9. *Let* $0 < a \le b$ *and* $y > 0$. *Then for every distinct pair* (i, j),

$$R_{ij}([a, b], y) \le y \frac{b-a}{a}.$$

Proof. The idea is that the rectangle $[a, b] \times [0, y]$ is included in the trapezoid $\{(u, v) \in [0,\infty)^2 : a \le u \le b, v \le yu/a\} = S(b) \backslash S(a)$, where

$$S(x) := \{(u, v) \in [0,\infty)^2 : u \le x, v \le yu/a\}.$$

By homogeneity of R_{ij},

$$R_{ij}(S(b)\backslash S(a)) = (b-a)R_{ij}(S(1)) \leq (b-a)R_{ij}(1, y/a) \leq y\frac{b-a}{a},$$

since R_{ij} is always upper bounded by its smallest argument. □

The following bound assumes more but is considerably more flexible, as β can be both smaller and larger than 1.

Lemma 3.10. *Under Assumption 3.2, for every $\beta \in (0, 1+\varepsilon]$ there exists $K(\beta) < \infty$ such that for any $0 < a \leq b$, $y > 0$ and every distinct pair (i, j),*

$$R_{ij}([a, b], y) \leq \frac{K(\beta)}{\beta} y^{\beta} \frac{b-a}{a^{\beta}}.$$

Proof. The bound in Assumption 3.2 gives

$$R_{ij}([a, b], y) = \int_0^y \int_a^b r_{ij}(u, v)dudv \leq K(\beta) \int_0^y v^{\beta-1}dv \int_a^b u^{-\beta}du \leq \frac{K(\beta)}{\beta} y^{\beta} \frac{b-a}{a^{\beta}}.$$

□

Lemma 3.11. *Under Assumption 3.2, for every $\beta \in [-\varepsilon, 0)$ there exists $K(\beta) < \infty$ such that for any $0 < a \leq b$, $y > 0$ and every distinct pair (i, j),*

$$R_{ij}([a, b], [y, \infty)) \leq \frac{K(\beta)}{-\beta} y^{\beta} b^{-\beta}(b-a).$$

Proof. Following the proof of Lemma 3.10,

$$R_{ij}([a, b], [y, \infty)) \leq K(\beta) \int_y^{\infty} v^{\beta-1}dv \int_a^b u^{-\beta}du \leq \frac{K(\beta)}{-\beta} y^{\beta} b^{-\beta}(b-a).$$

□

3.12.6 Technical results from empirical process theory

We collect here two fundamental inequalities from empirical process theory that are used in Section 3.11.3. Denote by $\mathcal{G}$ a class of real-valued functions that satisfies $|f(x)| \leq F(x) \leq U$ for every $f \in \mathcal{G}$ and let $\sigma^2 \geq \sup_{f \in \mathcal{G}} Pf^2$. Additionally, suppose that for some positive A, V and for all $\varepsilon > 0$,

$$N(\varepsilon, \mathcal{G}, L_2(\mathbb{P}_n)) \leq \left(\frac{A\|F\|_{L^2(\mathbb{P}_n)}}{\varepsilon}\right)^V \tag{3.92}$$

almost surely. In that case, the symmetrization inequality and inequality (2.2) from Koltchinskii (2006) yield

$$\mathbb{E}[\|\mathbb{P}_n - P\|_{\mathcal{G}}] \le c_0 \Big[\sigma\Big(\frac{V}{n}\log\frac{A\|F\|_{L^2(P)}}{\sigma}\Big)^{1/2} + \frac{VU}{n}\log\frac{A\|F\|_{L^2(P)}}{\sigma}\Big] \tag{3.93}$$

for a universal constant $c_0 > 0$ provided that $1 \ge \sigma^2 > \text{const} \times n^{-1}$. In fact, the inequality in Koltchinskii (2006) is for $\sigma^2 = \sup_{f\in\mathcal{G}} Pf^2$. However, this is not a problem since we can replace $\mathcal{G}$ by $\mathcal{G}\sigma/(\sup_{f\in\mathcal{G}} Pf^2)^{1/2}$.

The second inequality (a refined version of Talagrand's concentration inequality) states that for any countable class of measurable functions $\mathcal{F}$ with elements mapping into $[-M, M]$,

$$\mathbb{P}\Big(\|\mathbb{P}_n - P\|_{\mathcal{F}} \ge 2\mathbb{E}[\|\mathbb{P}_n - P\|_{\mathcal{F}}] + c_1 n^{-1/2}\Big(\sup_{f\in\mathcal{F}} Pf^2\Big)^{1/2}\sqrt{t} + n^{-1}c_2 Mt\Big) \le e^{-t}, \tag{3.94}$$

for all $t > 0$ and some universal constants $c_1, c_2 > 0$. This is a special case of Theorem 3 in Massart (2000) (in the notation of that paper, set $\varepsilon = 1$).

3.12.7 Discussion of max-stable distributions

In this section, we take $\boldsymbol{X}$ to be distributed according to the max-stable distribution associated to an arbitrary multivariate Pareto $\boldsymbol{Y}$ with stable dependence function L. That is, the copula of $\boldsymbol{X}$ is given by

$$\mathbb{P}(F(\boldsymbol{X}) \le \boldsymbol{x}) = \exp\{-L(-\log \boldsymbol{x})\}.$$

We shall demonstrate the following result. Note that the constant $K' = 48$ therein is not particularly sharp, and can be improved at the cost of more detailed calculations.

Proposition 3.5. *The max-stable random vector $\boldsymbol{X}$, assuming that its marginal distributions are continuous, satisfies Assumption 3.3 with $K' = 48$ and $\xi' = 1$.*

Proof. First note that for $q > 1/2$ and $\boldsymbol{x} \in [0,1]^{|J|}$, we have

$$\Big|q^{-1}\mathbb{P}(F_J(\boldsymbol{X}_J) > 1 - q\boldsymbol{x}) - R_J(\boldsymbol{x})\Big| \le 1 \le 2q,$$

so we may without loss of generality consider only $q \le 1/2$.

Now let J be a subset of size 2 or 3. For $q \le 1/2$, we have

$$\begin{aligned}\mathbb{P}(F_J(\boldsymbol{X}_J) \not\le 1 - q\boldsymbol{x}) &= 1 - \exp\{-L_J(-\log(1 - q\boldsymbol{x}))\} \\ &\in \Big[L_J(-\log(1-q\boldsymbol{x})) - \frac{1}{2}L_J(-\log(1-q\boldsymbol{x}))^2, L_J(-\log(1-q\boldsymbol{x}))\Big],\end{aligned} \tag{3.95}$$

using a Taylor expansion of the exponential around the origin: for $z > 0$, $e^{-z} = 1 - z + \bar{z}^2/2$ for some $0 \le \bar{z} \le z$. Here L_J is the stable tail dependence function of the subvector $\boldsymbol{X}_J$; it is obtained by evaluating L at a point the components of which in positions J^c are zero. Using a similar decomposition of $-\log(1-z)$ around $z = 0$, we find that

$$z \le -\log(1-z) \le z + 2z^2.$$

Now recal that $\boldsymbol{x} \in [0,1]^{|J|}$. Using the properties of stable tail dependence functions, namely that L (and L_J) is component-wise monotone, convex, homogeneous and upper bounded by the sum of its arguments, we find

$$\begin{aligned} L_J(q\boldsymbol{x}) &\le L_J(-\log(1-q\boldsymbol{x})) \\ &\le L_J(q\boldsymbol{x} + 2q^2\mathbf{1}) \\ &= 2L_J\big(\tfrac{1}{2}q\boldsymbol{x} + \tfrac{1}{2}2q^2\mathbf{1}\big) \\ &\le L_J(q\boldsymbol{x}) + L_J(2q^2\mathbf{1}) \\ &\le L_J(q\boldsymbol{x}) + 2|J|q^2. \end{aligned}$$

Finally, deduce from (3.95) that

$$q^{-1}\mathbb{P}(F_J(\boldsymbol{X}_J) \not\le 1 - q\boldsymbol{x}) \le L_J(\boldsymbol{x}) + 6q,$$

and that

$$\begin{aligned} &q^{-1}\mathbb{P}(F_J(\boldsymbol{X}_J) \not\le 1 - q\boldsymbol{x}) \\ &\quad \ge \min\Big\{L_J(\boldsymbol{x}) - q^{-1}\frac{1}{2}L_J(q\boldsymbol{x})^2, L_J(\boldsymbol{x}) + 6q - q^{-1}\frac{1}{2}\big(L_J(q\boldsymbol{x}) + 6q^2\big)^2\Big\} \\ &\quad \ge \min\Big\{L_J(\boldsymbol{x}) - \frac{9}{2}q, L_J(\boldsymbol{x}) + 6q - q^{-1}\frac{1}{2}(3q + 3q)^2\Big\} \\ &\quad = L_J(\boldsymbol{x}) - 12q. \end{aligned}$$

We have established an approximation similar to what is desired, but for the probabilities $\mathbb{P}(F_J(\boldsymbol{X}_J) \not\le 1 - q\boldsymbol{x})$ by the functions L_J:

$$\Big|q^{-1}\mathbb{P}(F_J(\boldsymbol{X}_J) \not\le 1 - q\boldsymbol{x}) - L_J(\boldsymbol{x})\Big| \le 12q. \tag{3.96}$$

We shall use this result to complete the proof.

For $i \in V$, let $E_i = \{F_i(X_i) > 1 - qx_i\}$. Now, if $J = (i,j)$ has size 2, then

$$\mathbb{P}(F_J(\boldsymbol{X}_J) > 1 - q\boldsymbol{x}) = \mathbb{P}(E_i \cap E_J) = qx_i + qx_j - \mathbb{P}(E_i \cup E_j)$$

and

$$R_J(\boldsymbol{x}) = x_i + x_j - L_J(\boldsymbol{x}),$$

so the result follows from (3.96). If $J = (i, j, m)$ has size 3,

$$\begin{aligned}
\mathbb{P}(F_J(\boldsymbol{X}_J) > 1 - q\boldsymbol{x}) &= \mathbb{P}(E_i \cap E_j \cap E_m) \\
&= \mathbb{P}(E_i \cup E_j \cup E_m) - \mathbb{P}(E_i \cup E_j) - \mathbb{P}(E_i \cup E_m) - \mathbb{P}(E_j \cup E_m) \\
&\quad + \mathbb{P}(E_i) + \mathbb{P}(E_j) + \mathbb{P}(E_m) \\
&= \mathbb{P}(E_i \cup E_j \cup E_m) - \mathbb{P}(E_i \cup E_j) - \mathbb{P}(E_i \cup E_m) - \mathbb{P}(E_j \cup E_m) \\
&\quad + qx_i + qx_j + qx_m,
\end{aligned}$$

and similarly

$$R_J(\boldsymbol{x}) = L_J(\boldsymbol{x}) - L_{ij}(x_i, x_j) - L_{im}(x_i, x_m) - L_{jm}(x_j, x_m) + x_i + x_j + x_m,$$

so the result again follows by aplying (3.96) to approximate each of the four probabilities above by the corresponding L terms. □

Conclusion

"Wondering when I will return to the world
of the living again
Do I even want to leave?"

Gabriel Lucas Riccio

In this thesis, two general methodologies were introduced for the estimation of tail dependence structures. They both add to existing work in various ways by allowing more complex distributions to be modeled and inferred. The methods in Chapter 2 adapt those in Einmahl et al. (2012, 2016) by allowing for asymptotic independence, while however being constrained to the setting of bivariate distributions or pairwise identifiable processes. The algorithm presented in Chapter 3 extends the recent contributions of Engelke and Hitz (2020) and Engelke and Volgushev (2020) to extremal graphical models by allowing for arbitrary (albeit connected) graph structures to be learned.

Some of the most important contributions of the thesis are strong theoretical results about the probabilistic behavior of certain non-parametric estimators, namely Theorems 2.1, 2.2, 2.4 and 3.3. While they have been developed here with the goal of supporting our parametric methodologies, they are applicable on a wider scope. The behavior of any inference based on the function c or, in the spatial context, the functions $c^{(s)}$ introduced in Chapter 2 would likely be explained by Theorems 2.1, 2.2 and 2.4. In Chapter 3, it was already hinted that Theorem 3.3 is applicable to EMTP2-constrained inference for Hüsler–Reiss models (Röttger et al., 2021) and to extremal tree learning (in fact, the result is used in Engelke and Volgushev (2020)).

Nevertheless, there are numerous open questions that would warrant further investigation, a few of which are now touched upon.

The most obvious question to ask about the work in Chapter 2 is: (how) can it be extended to dimensions higher than two? Modeling extreme values in a way that allows for non-trivial asymptotic independence becomes tricky in three or more dimensions, since asymptotic independence is a fundamentally pairwise notion. Whereas in the bivariate case, the accepted solution is to model a certain "higher order" tail

dependence by considering joint extremes, in d dimensions there can be exponentially many "layers" of tail dependence depending on which subset of variables are required to simultaneously be large. In principle, one can formulate a tail model similar to (2.2) for each of the $2^d - d - 1$ subsets of variables. It would be interesting to understand how those models relate to each other, if at all; given the function c_I that arises in the limit for a certain subset I of variables, does that constrain the form of the functions c_J for other subsets J which contain/are contained in I? Inference for those different functions c would probably become infeasible for subsets of more than a few variables, since observations with a high number of simultaneous extremes are usually exceptionally rare. This idea could however be coupled with a preprocessing analysis meant to detect the small groups of variables that are most susceptible to be asymptotically dependent, exploiting for instance the work of Chiapino et al. (2019), Simpson et al. (2020) or Meyer and Wintenberger (2020).

Chapter 3 opens up several avenues that were briefly discussed in Section 3.7. An obvious one is a different choice of the base learner $\mathcal{A}$ used in `EGlearn`, for which options abound; the tuning-free method of Lederer and Müller (2022) is an attractive one. The relationship between the extremal graph structure and the matrices $\Theta^{(m)}$ in Hüsler–Reiss models is very similar to that between the precision matrix and graphical structure in Gaussian models. Using this similarity, some of the innumerable algorithms for Gaussian graphical model selection could certainly be adapted to perform extremal graph selection. One example is the adaptive Laplacian constrained optimization considered in Ying et al. (2021) discussed in Section 3.7. Another possibility would be to adapt the idea of neighborhood selection to directly estimate the neighborhood of each variable Y_j in the extremal graph itself.

While Hüsler–Reiss distributions have the enjoyable property that their conditional independence relations are encoded in transformations of certain moments, it would be interesting to move away from this parametric assumption to other multivariate Pareto distributions which share a similar property. Of particular interest, can the "generalized score matching for graphical models" of Yu et al. (2019) be adapted to recover extremal graphical models?

In a different vein, we are working towards a parallel result to Theorem 3.3 where we obtain the asymptotic distribution of the empirical variogram in fixed-dimensional settings. This could be instrumental to achieving uncertainty quantification for extremal graphical models, perhaps through "confidence sets" of trees (as discussed in Willis, 2019) or generalizations thereof to non-tree graphs.

Finally, a fascinating research avenue that combines the topics of Chapters 2 and 3 would be to devise a notion of extremal graphical model that allows for asymptotically

independent groups of variables. While there have been some recent progress in classifying variables with respect to their asymptotic dependence and independence relations (Nolde and Wadsworth, 2020), how to integrate this structure in a graphical modeling approach is not yet understood. It would be exciting if the tools developed in Chapter 2 could help to estimate such models.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Améndola, C., Klüppelberg, C., Lauritzen, S., and Tran, N. M. (2022). Conditional independence in max-linear Bayesian networks. *Ann. Appl. Probab.*, 32(1):1–45.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79.

Asadi, P., Davison, A., and Engelke, S. (2015). Extremes on river networks. *Ann. Appl. Stat.*, 9:2023–2050.

Asenova, S., Mazo, G., and Segers, J. (2021). Inference on extremal dependence in the domain of attraction of a structured Hüsler–Reiss distribution motivated by a Markov tree with latent variables. *Extremes*, 24:461–500.

Asenova, S. and Segers, J. (2021). Extremes of markov random fields on block graphs. *arXiv preprint arXiv:2112.04847*.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *Ann. Probability*, 2:792–804.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.

Bingham, N., Goldie, C., and Teugels, J. (1987). *Regular Variation*. Cambridge University Press.

Brown, B. M. and Resnick, S. I. (1977). Extreme values of independent stochastic processes. *Journal of Applied Probability*, 14(4):732–739.

Bücher, A. and Dette, H. (2013). Multiplier bootstrap of tail copulas with applications. *Bernoulli*, 19(5A):1655–1687.

Bücher, A., Segers, J., and Volgushev, S. (2014). When Uniform Weak Convergence Fails: Empirical Processes for Dependence Functions and Residuals via Epi- and Hypographs. *Ann. Stat.*, 42:1598–1634.

Bücher, A., Volgushev, S., and Zou, N. (2019). On second order conditions in the multivariate block maxima and peak over threshold method. *Journal of Multivariate Analysis*, 173:604–619.

Chiapino, M., Sabourin, A., and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):377–392.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. (1996). On the lambertw function. *Advances in Computational mathematics*, 5(1):329–359.

Csörgő, M. and Horváth, L. (1987). Approximation of intermediate quantile processes. *Journal of multivariate analysis*, 21(2):250–262.

Csorgo, M. and Revesz, P. (1978). Strong approximations of the quantile process. *The Annals of Statistics*, pages 882–894.

Davison, A. C. and Gholamrezaee, M. M. (2012). Geostatistics of extremes. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 468:581–608.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012a). Statistical modeling of spatial extremes. *Statistical science*, 27(2):161–186.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012b). Statistical modeling of spatial extremes. *Statist. Sci.*, 27:161–186.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*. Springer.

de Haan, L., Lin, T., et al. (2001). On convergence toward an extreme value distribution in c [0, 1]. *The Annals of Probability*, 29(1):467–483.

de Haan, L., Neves, C., and Peng, L. (2008). Parametric tail copula estimation and model testing. *Journal of Multivariate Analysis*, 99(6):1260–1275.

de Haan, L. and Resnick, S. I. (1977). Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 40(4):317–337.

Dombry, C., Engelke, S., and Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika*, 103(2):303–317.

Dombry, C., Engelke, S., and Oesting, M. (2017). Asymptotic properties of the maximum likelihood estimator for multivariate extreme value distributions. Available from `https://arxiv.org/abs/1612.05178`.

Draisma, G., Drees, H., Ferreira, A., and de Haan, L. (2004). Bivariate Tail Estimation: Dependence in Asymptotic Independence. *Bernoulli*, 10:251–280.

Drees, H. and Huang, X. (1998). Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–46.

Drees, H., Resnick, S., and de Haan, L. (2000). How to make a Hill plot. *The Annals of Statistics*, 28(1):254 – 274.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.

Durrett, R. (2010). *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition.

Einmahl, J.H.J.and Kiriliouk, A., Krajina, A., and Segers, J. (2016). An M-estimator of spatial tail dependence. *J. R. Stat. Soc. B*, 78:275–298.

Einmahl, J., Krajina, A., and Segers, J. (2008). An method of moments estimator of tail dependence. *Bernoulli*, 14:1003–1026.

Einmahl, J., Krajina, A., and Segers, J. (2012). An M-Estimator for Tail Dependence in Arbitrary Dimensions. *Ann. Stat.*, 40:1764–1793.

Einmahl, J. H. J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Statist.*, 37:2953–2989.

Engelke, S., de Fondeville, R., and Oesting, M. (2019a). Extremal behaviour of aggregated data with an application to downscaling. *Biometrika*, 106:127–144.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 82:871–932.

Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8. To appear.

Engelke, S., Lalancette, M., and Volgushev, S. (2021). Concentration bounds for the extremal variogram. *arXiv preprint arXiv:2111.00840*.

Engelke, S., Lalancette, M., and Volgushev, S. (2022). Learning extremal graphical models in high dimensions. In preparation.

Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 77(1):239–265.

Engelke, S., Opitz, T., and Wadsworth, J. (2019b). Extremal dependence of random scale constructions. *Extremes*, 22(4):623–666.

Engelke, S. and Volgushev, S. (2020). Structure learning for extremal tree models. *arXiv preprint arXiv:2012.06179*.

Farris, J. S., Kluge, A. G., and Eckardt, M. J. (1970). A numerical approach to phylogenetic systematics. *Systematic Zoology*, 19(2):172–189.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Fougères, A.-L., De Haan, L., Mercadier, C., et al. (2015). Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2):903–934.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math.*, 6:93–116.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24:2693–2720.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of mathematics*, pages 423–453.

Goix, N., Sabourin, A., and Clémençon, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory*, pages 843–860. PMLR.

Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric bayesian inference on bivariate extremes. *J. R. Stat. Soc. B*, 73(3):377–406.

Gumbel, E. J. (1960). Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Heffernan, J. E. and Resnick, S. I. (2007). Limit laws for random vectors with an extreme component. *Ann. Appl. Probab.*, 17(2):537–571.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66:497–546.

Hentschel, M. (2021). Statistical inference for Hüsler–Reiss graphical models. Master's thesis, Universität Mannheim.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174.

Huang, X. (1992). *Statistics of Bivariate Extreme Value Theory*. PhD thesis, Erasmus University Rotterdam.

Huser, R., Opitz, T., and Thibaud, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using gaussian scale mixtures. *Spatial Statistics*, 21:166 – 186.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189.

Kabluchko, Z., Schlather, M., De Haan, L., et al. (2009). Stationary max-stable fields associated to negative definite functions. *The Annals of Probability*, 37(5):2042–2065.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25:1287–1304.

Keef, C., Tawn, J., and Svensson, C. (2009). Spatial risk assessment for extreme river flows. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 58:601–618.

Klüppelberg, C. and Lauritzen, S. (2019). Bayesian networks for max-linear models. In *Network science*, pages 79–97. Springer, Cham.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Lalancette, M., Engelke, S., and Volgushev, S. (2021). Rank-based estimation under asymptotic dependence and independence, with applications to spatial extremes. *The Annals of Statistics*, 49(5):2552–2576.

Lauritzen, S. (1996). *Graphical models*. Clarendon Press.

Le, P. D., Davison, A. C., Engelke, S., Leonard, M., and Westra, S. (2018). Dependence properties of spatial rainfall extremes and areal reduction factors. *J. Hydrol.*, 565:711–719.

Lederer, J. and Müller, C. L. (2022). Topology adaptive graph estimation in high dimensions. *Mathematics*, 10(8):1244.

Ledford, A. and Tawn, J. (1997). Modelling dependence within joint tail regions. *J. R. Stat. Soc. B*, 59(2):475–499.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., and Wasserman, L. (2011). Forest density estimation. *The Journal of Machine Learning Research*, 12:907–951.

Loh, P.-L. and Wainwright, M. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Ann. Statist.*, 41:3022–3049.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M., editors (2019). *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, pages 863–884.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meyer, N. and Wintenberger, O. (2020). Multivariate sparse clustering for extremes. *arXiv preprint arXiv:2007.11848*.

Nolde, N. and Wadsworth, J. L. (2020). Linking representations for multivariate extremes via a limit set. *arXiv preprint arXiv:2012.00990*.

Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.*, 105:263–277.

Papastathopoulos, I. and Strokorb, K. (2016). Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15.

Peng, L. (1999). Estimation of the coefficient of tail dependence in bivariate extremes. *Stat. & Probab. Lett.*, 43(4):399–409.

Peng, L. and Qi, Y. (2008). Bootstrap approximation of tail dependence function. *Journal of Multivariate Analysis*, 99(8):1807–1824.

Pickands, J. (1981). Multivariate Extreme Value Distributions. *Bull. Int. Stat. Inst.*, 49:859–878.

Pickands, III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, 3:119–131.

Poon, S.-H., Rockinger, M., and Tawn, J. (2004). Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *Rev. Financ. Stud.*, 17:581–610.

Radulović, D., Wegkamp, M., Zhao, Y., et al. (2017). Weak convergence of empirical copula processes indexed by functions. *Bernoulli*, 23(4B):3346–3384.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *J. R. Stat. Soc. B*, 71:219–241.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5:303–336.

Resnick, S. I. (1987). *Extreme values, regular variation and point processes*. Springer.

Rockafellar, R. T. (1970). *Convex Analysis*, volume 28. Princeton university press.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018a). Multivariate generalized Pareto distributions: parametrizations, representations, and properties. *J. Multivariate Anal.*, 165:117–131.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018b). Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145.

Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12:917–930.

Röttger, F., Engelke, S., and Zwiernik, P. (2021). Total positivity in multivariate extremes. *arXiv preprint arXiv:2112.14727*.

Schlather, M. (2002). Models for Stationary Max-Stable Random Fields. *Extremes*, 5:33–44.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Segers, J. (2020). One- versus multi-component regular variation and extremes of Markov trees. *Advances in Applied Probability*, 52:855–878.

Sibuya, M. (1960). Bivariate extreme statistics. I. *Ann. Inst. Statist. Math. Tokyo*, 11:195–210.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Smith, R. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

Tawn, J. (1988). Bivariate Extreme Value Theory: Models and Estimation. *Biometrika*, 75:397–415.

Tawn, J. (1990). Modelling Multivariate Extreme Value Distributions. *Biometrika*, 77:245–253.

Teschl, G. (1998). *Topics in Real and Functional Analysis*. Unpublished, available online at https://www.mat.univie.ac.at/gerald/ftp/book-fa.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer.

Vervaat, W. (1972). Functional Central Limit Theorems for Processes with Positive Drift and Their Inverses. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, 23:245–253.

von Mises, R. (1936). La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, 1:141–160.

Wadsworth, J. and Tawn, J. (2012). Dependence modelling for spatial extremes. *Biometrika*, 99(2):253–272.

Wadsworth, J., Tawn, J. A., Davison, A., and Elton, D. M. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):149–175.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Weller, G. B. and Cooley, D. (2014). A sum characterization of hidden regular variation with likelihood inference via expectation-maximization. *Biometrika*, 101:17–36.

Westra, S. and Sisson, S. A. (2011). Detection of non-stationarity in precipitation extremes using a max-stable process model. *Journal of Hydrology*, 406:119 – 128.

Willis, A. (2019). Confidence sets for phylogenetic trees. *J. Amer. Statist. Assoc.*, 114(525):235–244.

Ying, J., de Miranda Cardoso, J. V., and Palomar, D. P. (2020a). Does the ℓ_1-norm learn a sparse graph under laplacian constrained graphical models? *arXiv preprint arXiv:2006.14925*.

Ying, J., de Miranda Cardoso, J. V., and Palomar, D. P. (2020b). Nonconvex sparse graph learning under Laplacian constrained graphical model. *Advances in Neural Information Processing Systems*, 33:7101–7113.

Ying, J., de Miranda Cardoso, J. V., and Palomar, D. P. (2021). Minimax estimation of Laplacian constrained precision matrices. In *International Conference on Artificial Intelligence and Statistics*, pages 3736–3744. PMLR.

Yu, S., Drton, M., and Shojaie, A. (2019). Generalized score matching for non-negative data. *The Journal of Machine Learning Research*, 20(1):2779–2848.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhou, C. (2010). Dependence structure of risk factors and diversification effects. *Insur. Math. Econ.*, 46:531–540.

Zscheischler, J. and Seneviratne, S. I. (2017). Dependence of drivers affects risks associated with compound events. *Science Advances*, 3.