Statistical Modeling and Inference for Spatio-Temporal Extremes

THÈSE N^{O} 5946 (2013)

PRÉSENTÉE LE 25 SEPTEMBRE 2013

À LA FACULTÉ DES SCIENCES DE BASE

CHAIRE DE STATISTIQUE

PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Raphaël HUSER

acceptée sur proposition du jury:

Prof. T. Mountford, président du jury
Prof. A. C. Davison, directeur de thèse
Prof. S. Morgenthaler, rapporteur
Prof. H. Rootzén, rapporteur
Prof. J. Tawn, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

"Wisdom is the principal thing; therefore get wisdom:
and with all thy getting get understanding."
— Proverbs, The Bible

To Franceline

Acknowledgements

I have been very fortunate to be supervised by Anthony Davison. In addition to his kindness and his great patience in guiding my research, Anthony has always been available when I needed to. I have learnt a lot from his help, advice and guidance, and I thank him for the knowledge he has transmitted to me. I thank him also very much for his flexibility, openness, and support for letting me go abroad to many interesting conferences. I will definitely miss his birds-clock, which gave rhythm to my work.

I would like also to thank Stephan Morgenthaler, Thomas Mountford, Holger Rootzén and Jonathan Tawn, who accepted to serve on my thesis committee, and to read carefully my thesis, making helpful suggestions and corrections.

Many thanks also to all my colleagues at EPFL, in particular Jacques Ferrez, Linda Frossard, Alix Leboucq, Claudio Semadeni and Emeric Thibaud, who proofread quickly and carefully my thesis and made a lot of comments and suggestions. Emeric deserves a special thank for having put up with me during three and a half years in the same office. I thank him for the many discussions that we had about work... but not only! And also for the bike rides.

I have also benefitted from various discussions with other current or former members of STAT/STAP/SMAT groups, including Mathieu Ribatet, Mehdi Gholamrezaee, Simone Padoan, Miguel Bras de Carvalho, Juliette Blanchet, Jenny Wadsworth, Kjell Konis, Andrea Kvitkovičová, Shahin Tavakoli and Yoav Zemel.

I acknowledge the Swiss National Fund, who primarily funded my research, and also MeteoSwiss, who provided the rainfall and temperature data for statistical analysis.

I thank also Marc-André Dupertuis, with whom I discussed the opportunity to do a PhD in statistics after my MSc, and who convinced me to do it. Although research is not everyday a bed of roses, I have never regretted this choice, so I thank him for his advice.

I'm very grateful to all my friends who studied mathematics with me at EPFL (Carol,

Daniele, Gwenol, Pirmin, Sam, Stéphane) and those from outside EPFL —from high school (Edgar, Lucile, Mathias, Matthieu, Robin, Sybille), from church and from youth camps; in particular, I want to thank a lot Séb & Déb and Amanda, as well as the team of "Tontons", whose friendship is very important to me.

I would like also thank to my parents, and Joël, Nadia & Fréd, Jérémie, Papi & Mamie and Grand-maman, for their love and support, and because they never doubted my capacities. I also warmly thank my family-in-law, Pierrette & Olivier, Aurélia & Mike, Mél & Nico, Anne-Sylvie, René & Hedwige and Esther, for their permanent kindness and love, and for the example they are to me. To all of them, many thanks for the many good discussions, dinners, aperos, jokes, good moments we have shared together.

Most of all, I wish to thank Franceline without whose constant love, support and patience none of this work would mean anything. Thank you, France, to be simply who you are, and to be everyday by my side. I'm looking forward to our next adventure.

Lausanne, 31 July 2013 R. H.

"Bless the Lord, O my soul: and all that is within me, bless his holy name.
Bless the Lord, O my soul, and forget not all his benefits."
— Psalms, The Bible

Abstract

Risk assessment for extreme natural phenomena has become increasingly important, and over the past few years the scientific community has realized the importance of considering the spatial or spatio-temporal extent of extreme events. Historically, the GEV and GPD distributions have played an important role in the statistical modeling of extremes at individual locations, but for risk assessment it is crucial to assess dependence between locations: if dependence is strong, extreme events might occur simultaneously at different locations, thereby increasing the overall risk.

In this thesis, we construct new dependence models for space-time extremes, based on asymptotically justified arguments, and propose novel inference methods for fitting these models to observations exceeding high thresholds. So far, the modeling of spatial extremes has been limited to fitting max-stable processes to block (usually annual) maxima, regarded as mutually independent. Our threshold-based approach is more efficient and enables more detailed analysis of extremes, but requires a more sophisticated treatment of dependence.

The present work also describes how composite likelihoods can be used for inference, establishes the asymptotic distribution of the corresponding estimators, and assesses statistical efficiency for these methods in various contexts.

The methodology is illustrated by application to hourly rainfall data from western Switzerland, and enables realistic modeling of their extremal properties.

Keywords: Asymptotic independence; Composite likelihood; Extreme event; Max-stable process; Rainfall data; Relative efficiency; Spatio-temporal dependence; Threshold exceedance.

Résumé

L'évaluation des risques liés aux phénomènes naturels extrêmes est de plus en plus d'actualité. Au cours des années précédantes, la communauté scientifique a réalisé l'importance de considérer l'ampleur spatiale, voire spatio-temporelle, de ces événements extrêmes. Historiquement, les distributions GEV et GPD ont joué un rôle majeur dans la modélisation statistique des événements extrêmes à des lieux donnés, mais pour l'évaluation des risques, il est également crucial tenir compte de la dépendance qui les lie. En effet, si celle-ci est forte, les événements extrêmes ont tendance à arriver simultanément, augmentant ainsi le risque global.

Dans cette thèse, nous construisons de nouveaux modèles de dépendance, justifiés par des raisonnements asymptotiques, pour les extrêmes spatio-temporels, et nous proposons des méthodes d'inférence novatrices pour ajuster ces modèles aux observations excédant des seuils élevés. Jusqu'à présent, la modélisation des extrêmes spatiaux se limitait à l'ajustement des processus max-stables aux maxima de blocs (par exemple annuels), considérés comme mutuellement indépendants. Notre approche basée sur des seuils est plus efficace et permet une analyse plus détaillée des extrêmes, mais requiert également un traitement plus sophistiqué de la dépendance.

En outre, le présent travail décrit la manière dont les vraisemblances composites peuvent être utilisées pour l'inférence, établit la distribution asymptotique des estimateurs correspondants, et évalue l'efficacité relative de ces méthodes dans des contextes variés.

La méthodologie est illustrée avec un jeu de données de pluie horaires mesurées en Suisse occidentale, et s'avère adéquate pour une modélisation réaliste de leurs propriétés extrémales.

Mots clés : Dépendance spatio-temporelle ; Données de pluie ; Efficacité relative ; Evénement extrême ; Excès de seuil ; Indépendance asymptotique ; Processus max-stable ; Vraisemblance composite.

Contents

List of Figures

List of Tables

Introduction

Motivation

In recent years there has been a major upsurge of research on the statistics of extreme events for spatial settings. One reason for this is the realization among stakeholders, such as climate scientists, environmental engineers and insurance companies, that in an evolving climate it may be changes in the sizes and frequencies of rare events, rather than changes in the averages, that lead to the most devastating losses of life, damage to infrastructure and so forth. As an illustration, Figure 1 shows a radar snapshot taken over Switzerland on August 22, 2005, when intense flooding affected several countries in Central Europe; in particular, the Swiss capital, Bern, was heavily hit after the Aar river burst its banks, forcing the evacuation of many homes in this region. In addition, the village of Lauterbrunnen in the Bernese Alps was completely isolated, and the only exit possibility was by military helicopter or by crossing one of the high Alpine passes. As a result of this extreme event, 6 people were killed in Switzerland, with more than 3 billion CHF of infrastructure damage.

While it is difficult or even impossible to attribute particular events to the effect of climatic change, the types of events that have long been forecast to increase in frequency by the modeling community —such as widespread heavy summer rainfall, but also heatwaves leading to crop failure and major brush fires— do indeed seem to be appearing more often than in the recorded past; see Figure 2. This motivates attempts to model such events, in order to understand their likely future impacts, and to assess the related risks.

Classical geostatistics is a well-developed field surveyed in numerous textbooks (e.g., Cressie, 1993; Wackernagel, 2003; Banerjee *et al.*, 2003; Diggle & Ribeiro, 2007; Cressie & Wikle, 2011), with much software available and a wide range of user communities corresponding to its many applications. Its basis in Gaussian distributions makes it unsuitable for extremal modeling, however, because the Gaussian density function has exceptionally light tails and therefore can badly underestimate probabilities associated

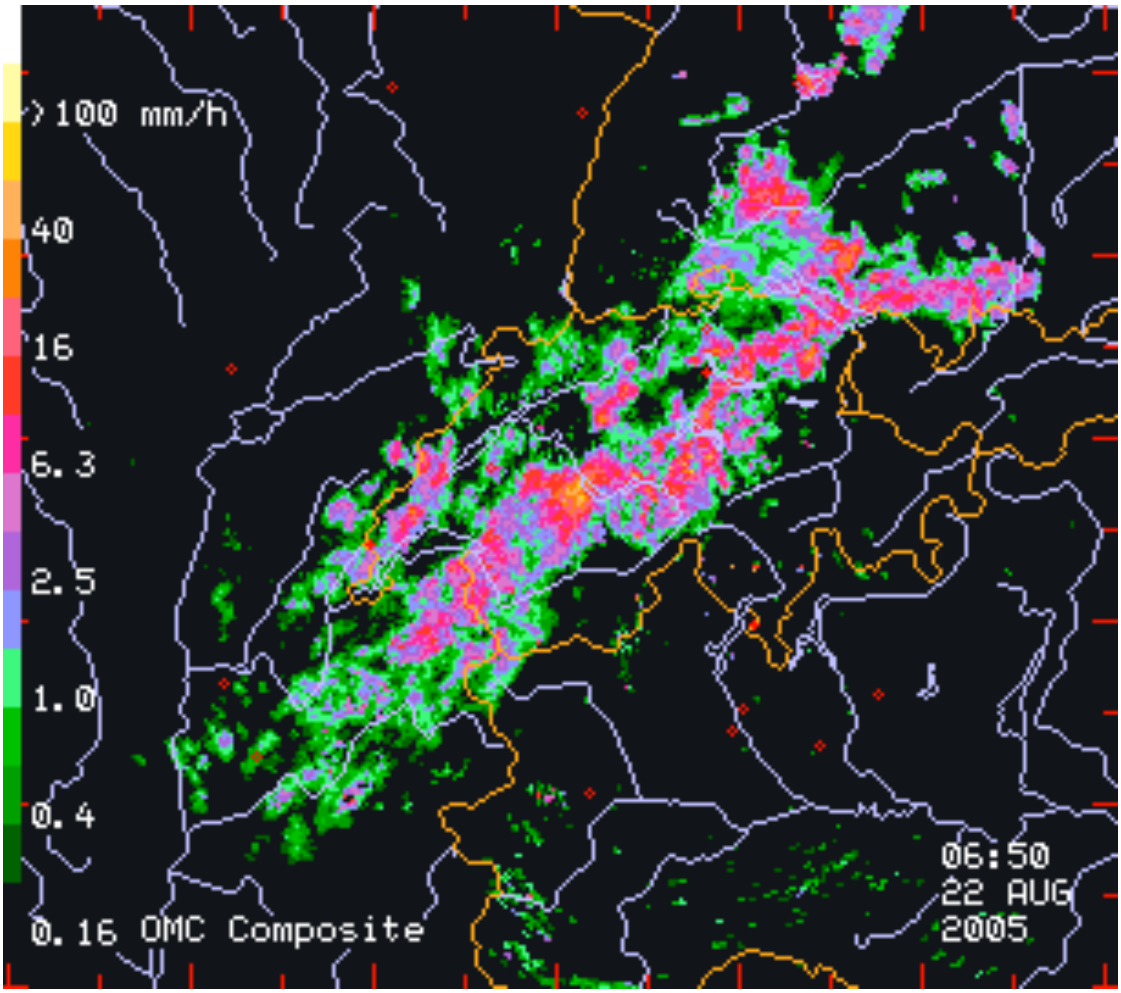

Figure 1: Radar snapshot taken at 06:50 on August 22, 2005, by the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss, illustrating extreme rainfall over Switzerland and neighboring countries. Borders are depicted with orange lines, while rivers are in blue. The color (log-)scale (left side) indicates rainfall amounts [mm/hr] averaged over a 5min time window. Very extreme events appear in yellow or orange.

to extreme events. Moreover, the tails of the multivariate Gaussian distribution lead to independent extremes, for any underlying correlation that is less than unity, resulting in potentially disastrous underestimation of the probabilities of the simultaneous occurrence of two rare events —this is 'the formula that killed Wall Street' which, at least according to *Wired* magazine (Salmon, 2009), has played a key role in the ongoing international financial crisis by providing wildly incorrect assessment of economic risks.

Since Gaussian densities do not provide suitable models for extremes, it is natural to ask what distributions can arise as limits for maxima of independent random

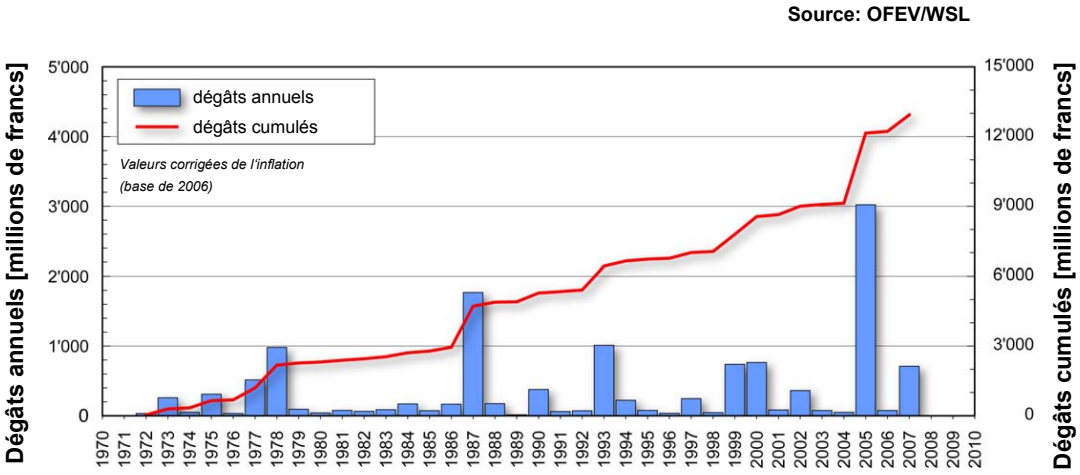

Figure 2: Plot of the overall (flooding, landslide and debris flow) annual damage [million CHF] in Switzerland from 1972 to 2007 (blue histogram), with the corresponding cumulative damage (red curve). This graph is taken from G. R. Bezzola's talk at the press conference on September 18, 2008, in Bern, about the analysis of the water rise in 2005.

variables. Under a suitable rescaling, the latter turn out to be embedded in the class of generalized extreme-value (GEV) distributions, which can have much heavier tails than the Gaussian distribution. Furthermore, this result extends to stationary sequences with short-term dependence, therefore providing even stronger support for the use of the GEV distribution in extremal applications. Alternatively, one can show that the generalized Pareto distribution (GPD) is suitable for the modeling of exceedances above high thresholds. For risk assessment over some spatial region and within some temporal window, space-time dependence of extreme events need to be properly accounted for. By contrast with Gaussian processes, the generalization of the aforementioned extreme-value distributions to multivariate or spatial settings is nonparametric: the classes of so-called *max-stable* distributions and processes cannot be described by a finite number of parameters. However, submodels can be constructed and the challenge is to build flexible but parsimonious models that can capture a large class of dependence structures.

While the justification for the GEV distribution goes back to the late 1920s and the GPD has been applied extensively from the early 1990s, the use of max-stable processes and related extremal dependence models is much more recent. Although the theoretical development of the latter finds its roots in the work of several researchers during the mid-1970s–1980s, useful models, inferential methods, and computer resources to fit these complicated models were lacking until recently. Realistic spatial models for

extremes were proposed in the early 2000s, and the first "true" space-time applications have emerged only in the early 2010s.

This thesis contributes to the extreme-value theory and composite likelihood literatures by developing new models for space-time extremes, as well as novel methods to perform inference, and by assessing the performance of the latter. The following paragraph details the content of the chapters and their specific contributions.

Outline and contributions of the thesis

In Chapter 1, classical extreme-value theory is surveyed in the finite dimensional case, and the theoretical foundations for the use of the GEV distribution, as well as multivariate extreme-value distributions, are established. We also summarize results about alternative point process representation for extremes, which lead to the GPD and the spectral decomposition for multivariate extremes, and form the basis for peaks over thresholds approaches. The novel contribution of this chapter is the comparison of efficiency of several widely-used estimators for bivariate extremes. Using simulations and analytical calculations, we shed some light on the performance of each method, and clarify the connections between them.

In Chapter 2, we tie together classical geostatistics and statistics of extremes, in order to extend the results of Chapter 1 to the spatial setting. In particular, we discuss modeling of extremal dependence based on max-stable processes, and also address the issue of asymptotic independence. The novelty of this chapter is the application of this methodology to two different real datasets, namely daily cumulative rainfall and daily temperature minima, for which we discuss the usefulness of asymptotic dependence, versus asymptotic independence, models.

In Chapter 3, inference based on composite likelihood is addressed, and asymptotic properties of the resulting estimators are described. The main new contributions are to show how this can be applied to the estimation of max-stable processes, and to study the loss in efficiency of weighted pairwise likelihood estimators, compared to maximum likelihood estimation, in a large variety of contexts. We also detail how inference based on the full likelihood can be performed for the extreme-value logistic distribution, and extend this to a max-stable time series model.

In Chapter 4, we consider a special class of max-stable processes, the so-called Brown–Resnick processes, and derive the corresponding full distributions for measurements recorded at an arbitrary set of spatial locations. Using extensive simulations, we then investigate the gain in efficiency of triplewise likelihood estimators compared to

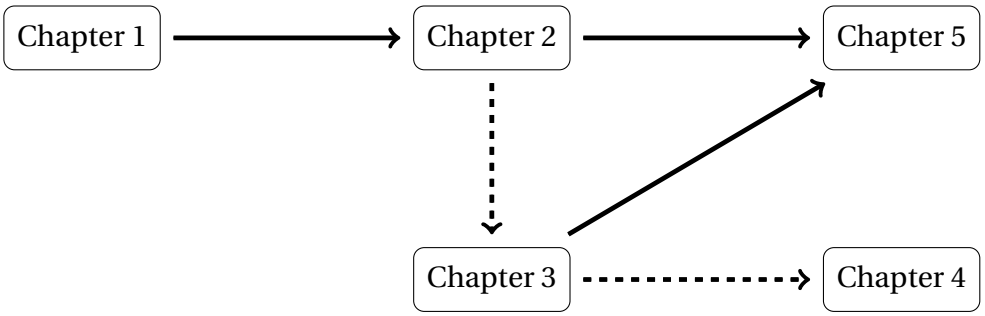

Figure 3: Dependence chart suggesting a possible path through the thesis. A solid arrow means that a chapter is a prerequisite for another chapter. A dashed arrow suggests a natural continuation. Chapter 1: *Classical extreme value theory*; Chapter 2: *Geostatistical modeling of extremes in space and time*; Chapter 3: *Inference based on composite likelihoods*; Chapter 4: *Composite likelihood estimation for the Brown–Resnick process*; Chapter 5: *Real case study: Space-time modeling of extreme rainfall.*

pairwise likelihood estimators in this framework, and explore the potential benefits of using even higher-dimensional composite likelihoods.

In Chapter 5, we propose a censored threshold-based pairwise likelihood estimator for the estimation of extremal dependence models, and prove its asymptotic normality and strong consistency under mild temporal mixing conditions. We illustrate the methodology developed in this thesis with a full space-time application to hourly rainfall extremes recorded in Switzerland. In particular, we develop new models that can capture space-time interactions and are able to mimic the type of process illustrated in Figure 1.

Figure 3 suggests a possible path through the thesis, and summarizes the chapter dependencies.

1 Classical extreme value theory

This chapter is a broad survey of extreme value theory in the finite-dimensional case. The extension to the spatial framework is described in Chapter 2. This chapter is intended to provide a solid background for the rest of the thesis, and most of the content is well-established in the extreme-value literature. An exception, however, is Section 1.2.2.2, which provides a qualitative and quantitative comparison of widely-used estimators for extremal distributions, based on a simulation study and theoretical calculations. The main novelty of this contribution is to highlight the differences and similarities of these approaches, and to show by how much the censored approach adopted in our application discussed in Chapter 5 outperforms its natural competitors.

1.1 Univariate extreme value theory

1.1.1 Asymptotic distribution for linearly renormalized maxima

1.1.1.1 Basic results

We start with $Y, Y_1, Y_2, \ldots$, independent random variables distributed according to a common distribution F with support Supp_F, that is the set of points $y \in \mathbb{R}$ with strictly positive density (or probability mass in the discrete case) with respect to F. For risk assessment purposes, it is natural to be interested in the fluctuations of the maximum of n such variates, which we denote by $M_n = \max(Y_1, \ldots, Y_n)$. The cumulative distribution function of M_n is

$$\Pr(M_n \leq y) = \Pr(Y_1 \leq y, \ldots, Y_n \leq y) = \prod_{i=1}^{n} \Pr(Y_i \leq y) = F^n(y),$$

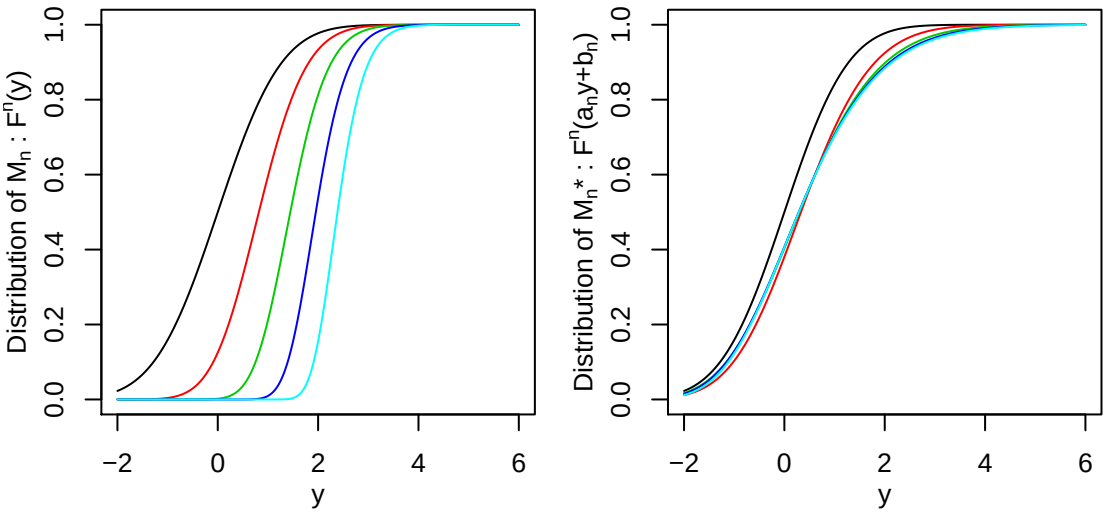

Figure 1.1: Distribution of M_n (left), that is $F^n(y)$, and distribution of M_n^* (right), that is $F^n(a_n y + b_n)$, for suitable normalizing constants $a_n > 0$ and b_n. Here $n = 1$ (black), 3 (red), 9 (green), 27 (dark blue) and 81 (light blue), and $F(y)$ is the Gaussian cumulative distribution function.

which converges to a degenerate distribution function putting mass one at the upper endpoint $y_F = \sup\{y : F(y) < 1\}$ of the underlying distribution F. Indeed, for all $y < y_F$, we have $F^n(y) \to 0$ as $n \to \infty$.

However, in the same way as, under a suitable affine renormalization, sum-stable distributions arise as the only limits of sums of random variables —many of which are attracted to the normal distribution by virtue of the Central Limit Theorem—, the stochastic behavior of M_n may also be stabilized after shift and scale transformations, under some conditions.

Figure 1.1 shows how the distribution of the maximum of n independent standard normal variates evolves as n increases, with (left panel) and without (right panel) renormalization. The left panel reveals that the distribution of M_n moves to the right and that the mass becomes more and more concentrated around a point. The right panel shows the distribution of renormalized maxima $M_n^* = (M_n - b_n)/a_n$ for suitable choices of sequences a_n and b_n. Here, the variable M_n^* converges in distribution to a standard Gumbel variate. In the sequel, conditions are given for the existence of such normalizing constants, and methods will be described to find them explicitly in special cases. Similarly to the Central Limit Theorem for sums of random variables, the hope with this affine renormalization for maxima is to characterize all possible non-degenerate limiting distributions by a single parametric family, and hence to have asymptotically justified models for maxima. As we shall see, this hope can indeed be

fulfilled; see Theorem 4 below. To establish this powerful result, we need to introduce the key notion of max-stability.

Definition 1 (Distributions and random variables of the same type)**.** *Two distributions G_1 and G_2 are said to be of the same type, if there exist norming constants $a > 0$ and $b \in \mathbb{R}$ such that*

$$G_1(ay + b) = G_2(y), \quad \text{for all } y \in \mathbb{R}.$$

Similarly, two random variables Y_1 and Y_2 are said to be of the same type, if their distributions are of the same type.

Definition 2 (Max-stability)**.** *A distribution G is max-stable if for any $k \in \mathbb{N}$, the distribution G^k is of the same type as G. A random variable Z is said to be max-stable if its distribution is max-stable.*

Max-stability is satisfied by distributions for which the operation of taking sample maxima leads to the same distribution, apart from changes in scale and location.

Let us now consider a sequence of independent and identically distributed (i.i.d.) *max-stable* random variates $Z_1, Z_2, \ldots \overset{\text{iid}}{\sim} G$. By definition, the distribution of maxima is of the same type as G, that is, for each $k \in \mathbb{N}$ one can find real constants $a_k > 0$ and b_k such that $G^k(a_k y + b_k) = G(y)$ for all $y \in \mathbb{R}$. Hence, the random variables $Z_n^* = \{\max(Z_1, \ldots, Z_n) - b_n\}/a_n$ and $Z \sim G$ are equal in distribution. In particular, Z_n^* converges in distribution to Z, as $n \to \infty$, which implies that all max-stable distributions are limits of renormalized maxima of i.i.d. random variables. The interesting result resides in the fact that the converse is also true: Max-stable distributions are the *only* possible non-degenerate limit laws of renormalized maxima; see Theorem 3.

Theorem 3 (de Haan, 1970; Embrechts *et al.*, 1997)**.** *The class of all possible non-degenerate limit laws for (properly renormalized) maxima of i.i.d. random variables coincides with the class of max-stable distributions.*

The proof relies on the fact that the maximum of a block of length pq can be written as the maximum of p maxima of little blocks, each of length q: writing $M_{i;j} = \max\{Y_i, \ldots, Y_j\}$, we have

$$M_{1;pq} = \max\{M_{1;q}, \ldots, M_{(p-1)q+1;pq}\}.$$

Hence, the limit law G satisfies the equation $G\left\{(y - b_{pq})/a_{pq}\right\} = G^p\left\{(y - b_q)/a_q\right\}$, and is thus max-stable.

The next theorem, first shown by Fisher & Tippett (1928), is the cornerstone of classical extreme value theory. It states that *if* the maximum of random variables can be shifted

and scaled in such a way that it converges in distribution to a non-degenerate limit, then the latter *has* to be one of three special types; see Theorem 4 and Figure 1.2. Specifically, the class of max-stable distributions is fully described by only three parametric probability laws: the Gumbel, Fréchet and reversed Weibull distributions.

Theorem 4 (Extremal types theorem; Fisher & Tippett, 1928; Gnedenko, 1943; Resnick, 1987)**.** *Let $\{Y_i\}_{i\geq 1}$ be a sequence of i.i.d. random variables and let $M_n = \max(Y_1,\ldots,Y_n)$. If there exist sequences of constants $a_n > 0$ and b_n such that*

$$\Pr\left(\frac{M_n - b_n}{a_n} \leq y\right) \to G(y), \qquad n \to \infty,$$

where $G(y)$ is a non-degenerate distribution function, then $G(y)$ must be of the same type as one of the following distributions:

Type I (Gumbel):

$$\Lambda(y) = \exp\left\{-\exp(-y)\right\}, \quad y \in \mathbb{R}, \tag{1.1}$$

Type II (Fréchet):

$$\Phi_\alpha(y) = \begin{cases} \exp\left(-y^{-\alpha}\right), & y > 0, \\ 0, & y \leq 0, \end{cases} \tag{1.2}$$

Type III (Reversed Weibull):

$$\Psi_\alpha(y) = \begin{cases} \exp\left\{-(-y)^{\alpha}\right\}, & y < 0, \\ 1, & y \geq 0, \end{cases} \tag{1.3}$$

for some $\alpha > 0$.

The three types can be summarized in a single parametric family, the so-called Generalized Extreme Value distribution, GEV(μ,σ,ξ):

$$G(y) = \begin{cases} \exp\left[-\left\{1+\xi(y-\mu)/\sigma\right\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\{-(y-\mu)/\sigma\}\right], & \xi = 0, \end{cases} \tag{1.4}$$

with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$, and where $t_+ = \max(t,0)$. The distribution has support $\text{Supp}_G = \{y \in \mathbb{R} : 1+\xi(y-\mu)/\sigma > 0\}$. The key parameter is the shape parameter ξ. It determines the type of the limit law and thus whether the support of G is bounded from below, from above or unbounded. If $\xi < 0$, the distribution has an upper bound at $y = \mu - \sigma/\xi$ and the reversed Weibull distribution is recovered with $\alpha = -\xi^{-1}$; when $\xi = 0$, $G(y)$ corresponds to the Gumbel distribution whose support is unbounded and whose upper tail decays exponentially;

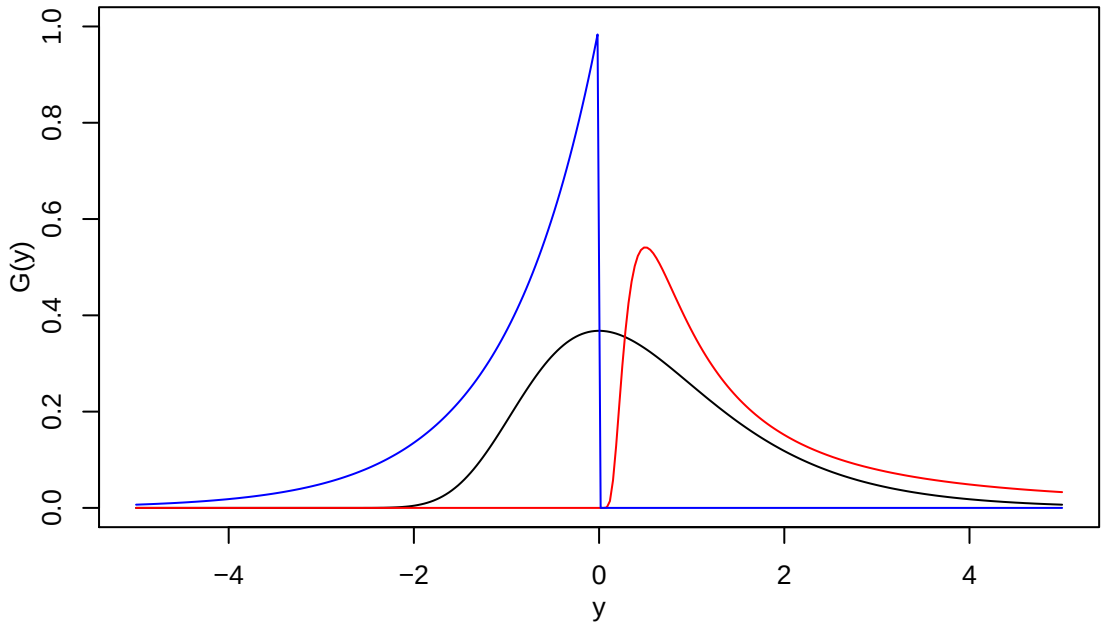

Figure 1.2: Standard Gumbel (black), unit Fréchet (red) and unit reversed Weibull (blue) density functions.

and when $\xi > 0$, the distribution has a lower bound at $y = \mu - \sigma/\xi$ and the Fréchet distribution is recovered with $\alpha = \xi^{-1}$. Furthermore, for the Fréchet distribution, the parameter ξ controls the rate of upper tail decay, and hence the potential severity of future extreme events. The rth moment of the GEV distribution is finite if $r\xi < 1$. Therefore, when $\xi = 1$ (unit Fréchet case), the mean is not well-defined.

The GEV distribution is especially convenient for inference: although ξ is difficult to estimate in practice, one can let the data "decide" its value without previously having to choose a particular type of distribution to fit beforehand. However, at sub-asymptotic levels, ξ is usually estimated with a mis-specification bias.

Very high quantiles are often of particular interest, since they give a quantitative description of the severity of an extreme event that might occur in future time periods. The level that is expected to be exceeded once on average in a specific time period is called a *return level* and is associated to its *return period*; see Definition 6.

Definition 5 (Quantile function)**.** *The generalized inverse, or quantile function, of a distribution function F is*

$$F^{\leftarrow}(p) = \inf\{y \in \mathbb{R} : F(y) \geq p\}, \quad 0 < p < 1.$$

Definition 6 (Return levels and return periods)**.** *Suppose that* annual *maxima of some*

random variable are modeled with the distribution $G \sim \text{GEV}(\mu, \sigma, \xi)$. *The level* y_p *which is exceeded with probability* p, *that is once every* $1/p$ *years on average, is called the* $1/p$-year return level. *We have*

$$y_p = G^{\leftarrow}(1-p) = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{1.5}$$

The return period is $1/p$. *The definition can be extended to other block lengths (e.g., daily blocks, monthly blocks, and so forth).*

1.1.1.2 Maximum domains of attraction

Theorem 4, known as *the extremal types theorem*, identifies the three possible limit laws for renormalized maxima of i.i.d. random variables. Moreover, if there exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that $(M_n - b_n)/a_n$ converges to a non-degenerate distribution function G, then G is uniquely determined up to an affine transformation. That is, if there exist sequences $a'_n > 0$ and $b'_n \in \mathbb{R}$ such that $(M_n - b'_n)/a'_n \xrightarrow{D} G'$, then G' and G must be of the same type. We can therefore define the maximum domain of attraction (MDA) as the class of distributions whose maxima are attracted to a particular limit law, as $n \to \infty$.

Definition 7 (Maximum domain of attraction)**.** *The random variable* Y *belongs to the maximum domain of attraction of the extreme value distribution* G *if there exist constants* $a_n > 0$ *and* $b_n \in \mathbb{R}$ *such that* $(M_n - b_n)/a_n \xrightarrow{D} G$. *We write* $Y \in \text{MDA}(G)$.

The characterization of the MDAs has been extensively studied in the literature. See, e.g., Embrechts *et al.* (1997), von Mises (1964), Resnick (1987), Beirlant *et al.* (2004), Leadbetter *et al.* (1983) and de Haan & Ferreira (2006), where the authors give discussions of necessary and sufficient conditions for different domains of attraction.

The notion of *tail-equivalent distributions* is primordial to fully describe the MDA of each of the three extremal type distributions.

Definition 8 (Tail-equivalence)**.** *The distribution functions* F_1 *and* F_2 *are said to be tail-equivalent if they have the same right endpoint* $y_{F_1} = y_{F_2} = y_F$, *and if*

$$\lim_{y \to y_F} \frac{1 - F_1(y)}{1 - F_2(y)} = c,$$

for some positive constant $0 < c < \infty$.

Each MDA is closed under tail-equivalence, that is if F_1 and F_2 are tail-equivalent,

$F_1 \in \text{MDA}(G)$ if and only if $F_2 \in \text{MDA}(G)$. Loosely speaking, the MDAs are composed of distributions whose right tail decays at the same rate at the right endpoint.

Theorem 9 (Characterization of the MDA)**.** *Let F be a distribution function with upper endpoint y_F, and let $\sim$ denote asymptotic equivalence. The following assertions are true:*

- $F \in \text{MDA}(\Phi_\alpha)$ *if and only if* $1 - F(y) \sim Ky^{-\alpha}$, *as* $y \to y_F = \infty$,
- $F \in \text{MDA}(\Psi_\alpha)$ *if and only if* $1 - F(y) \sim K(y_F - y)^{\alpha}$, *as* $y \to y_F < \infty$,

for some $\alpha > 0$ and $K \in \mathbb{R}$, a constant which does not depend on y, and finally

- $F \in \text{MDA}(\Lambda)$ *if and only if there exists some $z < y_F \leq \infty$ such that $F(y)$ has the representation* $1 - F(y) = c(y) \exp\left[-\int_z^y \{1/a(t)\} dt\right]$, $z < y < y_F$, *where c is a measurable function satisfying $c(y) \to c > 0$, as $y \to y_F$, and $a(y)$ is a positive, absolutely continuous functions (with respect to Lebesgue measure) with density $a'(y)$ having* $\lim_{y \to y_F} a'(y) = 0$.

For distributions that are sufficiently smooth at the right endpoint, von Mises (1964) established sufficient conditions for the convergence $(M_n - b_n)/a_n \to Z \sim \text{GEV}(0, 1, \xi)$, as $n \to \infty$, providing useful tools to determine the type of limit distribution and the choice of normalizing constants $a_n > 0$ and b_n. These conditions can then be employed to classify some known distributions in the different MDAs; see Table 1.1 and Embrechts *et al.* (1997) for more details.

Proposition 10 (von Mises conditions)**.** *Let F be a distribution function with right endpoint y_F and assume there exists some $z < y_F$ such that F is twice differentiable on (z, y_F). Let $f = F'$ be the density of F on (z, y_F). Define the sequences of real numbers $b_n = F^{\leftarrow}(1 - 1/n)$ and $a_n = r(b_n)$, where $r(y) = \{1 - F(y)\}/f(y)$ is the reciprocal hazard function of F. Furthermore, let $\xi = \lim_{y \to y_F} r'(y)$. Then $(M_n - b_n)/a_n \xrightarrow{D} Z \sim \text{GEV}(0, 1, \xi)$. That is, if $\xi > 0$, $F \in \text{MDA}(\Phi_{1/\xi})$; if $\xi < 0$, $F \in \text{MDA}(\Psi_{-1/\xi})$; and if $\xi \to 0$, $F \in \text{MDA}(\Lambda)$.*

The existence of an affine normalization leading to a non-degenerate limiting distribution for maxima is, however, not guaranteed. One can find distributions that do not belong to any maximum domain of attraction. Classical examples are the Poisson, Geometric or Negative Binomial distributions (Embrechts *et al.*, 1997), which are not well-behaved at the right tail. The following results show that convergence of maxima to a non-degenerate distribution can only occur under some continuity condition at the right endpoint, which is not satisfied by the aforementioned discrete distributions.

Table 1.1: Classification of various well-known distributions according to maximum domains of attraction, with the corresponding normalizing constants $a_n > 0$, $b_n \in \mathbb{R}$.

MDA(Λ)

Distribution[†]	$F(y)$	a_n	b_n
$\mathcal{N}(0,1)$	$\Phi(y)$, $y \in \mathbb{R}$	$(2\log n)^{-1/2}$	$(2\log n)^{1/2} - (2\log n)^{-1/2}$ $\times \left(\log 4\pi + \log\log n\right)/2$
Weibull(α)	$1-\exp(-y^{\alpha})$, $y>0$, $\alpha>0$	$(\log n)^{1/\alpha}$	$\frac{1}{\alpha}(\log n)^{1/\alpha-1}$
Exp(λ)	$1-\exp(-\lambda y)$, $y>0$, $\lambda>0$	λ^{-1}	$\lambda^{-1}\log n$

[†]Other examples include the Gamma, the Lognormal and all tail-equivalent distributions.

MDA(Ψ_α)

Distribution[‡]	$F(y)$	a_n	b_n
Uniform	$(y-a)/(b-a)$, $y \in [a,b]$	$(b-a)/n$	$b-(b-a)/n$
Beta	$\int_0^y \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} s^{a-1}(1-s)^{b-1}ds$, $y \in (0,1)$, $a,b>0$	1	$\left(n\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)}\right)^{-1/b}$

[‡]Other examples include distributions with a power law behavior at $y_F < \infty$.

MDA(Φ_α)

Distribution[#]	$F(y)$	a_n	b_n
Pareto	$1-(a/y)^{\alpha}$, $y \geq a$, $\alpha>0$	$a\alpha^{-1}n^{1/\alpha}$	$an^{1/\alpha}$
Cauchy	$1/2+\pi^{-1}\arctan(y)$, $y \in \mathbb{R}$	n/π	0

[#]Other examples include the Burr, the Loggamma, or the Student t distributions.

Proposition 11. *Let u_n be a sequence of real numbers. Then, for all $0 \leq \lambda \leq \infty$, the two assertions are equivalent:*

1. $n\{1-F(u_n)\} \to \lambda$,
2. $\Pr(M_n \leq u_n) \to \exp(-\lambda)$.

Theorem 12 (Continuity at the right endpoint)**.** *Let $0 < \lambda < \infty$. There exists a sequence*

u_n satisfying $n\{1-F(u_n)\} \to \lambda$ if and only if

$$\lim_{y \to y_F} \frac{\{1-F(y)\}}{(1-F_{y_\bullet})} = 1, \tag{1.6}$$

where $F_{y_\bullet} = \lim_{x \uparrow y} F(x)$.

For distributions defined on $\mathbb{Z}$ with infinite right endpoint, condition (1.6) translates into $\{1-F(k)\}/\{1-F(k-1)\} \to 1$, as $k \to \infty$. For the Poisson distribution, for example, this limit equals zero (see Embrechts *et al.*, 1997, p.118), preventing the maxima from converging. Similar results hold for the Geometric and Negative Binomial distributions. Other examples where no non-degenerate limit distribution for maxima exists can be found with super-heavy-tailed distributions. For example, the tail of the distribution $F(y) = 1 - 1/\log(y)$, $y > e$, decays so slowly that no suitable linear normalizing constants $a_n >$ and b_n may be found. Contrasting results may be obtained for Poisson variables with mean m, when we let $n \to \infty$ and $m \to \infty$ simultaneously (Anderson *et al.*, 1997).

1.1.1.3 Extension to stationary series

A strong assumption made in the extremal types theorem is that the sequence $\{Y_i\}_{i \geq 1}$ is independent and identically distributed. However, the data often depart from this assumption in two respects: first, the observations may not be independent. For example, when considering environmental data, short-term temporal dependence may exist (e.g., for hourly rainfall or daily snowfall), and extreme climate conditions might persist for several consecutive observations. Second, the data may not be identically distributed. This is the case, for example, if a seasonal pattern exists (e.g., diurnal and annual cycles for temperature data), if a global trend drives the data (e.g., climate change in environmental applications), or if some sort of volatility clustering is present (e.g., fluctuations in log-returns of financial data are much larger during crises).

In this section, we will only address the first issue, the extension of independence to stationarity. Non-stationarity is often dealt with either by modeling it directly in the marginal parameters, e.g., using linear regression (Katz *et al.*, 2002; Smith, 1989), semi-parametric models based on splines (Chavez-Demoulin & Davison, 2005, 2012), Bayesian hierarchical models (Cooley *et al.*, 2007; Sang & Gelfand, 2009, 2010), or by preprocessing and filtering approaches, e.g., McNeil & Frey (1998) reduce the volatility in a financial dataset by first fitting a AR-GARCH model and then applying extreme value theory techniques to the residuals, or finally by using the most pragmatic

approach: focusing on a stationary subset of the data (for example summer precipitation). For a deep treatment of extreme value theory for independent, non-identically distributed random variables, we refer to Galambos (1987). It turns out that in this framework, the class of limiting distributions is much too wide to be of practical use.

The theory of extreme values for dependent stochastic processes has been extensively developed and summarized in Leadbetter *et al.* (1983). Interest resides here in strictly stationary series, rather than in i.i.d. random variables.

Definition 13 (Stationary time series)**.** *Let* $\{Y_i\}_{i\geq 1}$ *be a time series,* $\mathscr{T} \subset \mathbb{N}$ *be some finite set and let* $\boldsymbol{Y}_{\mathscr{T}}$ *denote the collection of* Y_i*s such that* $i \in \mathscr{T}$*. The time series* $\{Y_i\}_{i\geq 1}$ *is said to be* strictly stationary *if the joint distributions of the vectors* $\boldsymbol{Y}_{\mathscr{T}}$ *and* $\boldsymbol{Y}_{h+\mathscr{T}}$ *are the same, for any time lag* $h \in \mathbb{N}$*. The process is called* weakly stationary *if for any* $i_1, i_2 \geq 1$, $\mathrm{E}(Y_{i_1}) = \mathrm{E}(Y_{i_1+h})$ *and* $\mathrm{cov}(Y_{i_1}, Y_{i_2}) = \mathrm{cov}(Y_{i_1+h}, Y_{i_2+h})$.

Loosely speaking, strict stationarity means that translation does not affect the probabilistic properties of the process. In other words, it corresponds to a series whose variables may be mutually dependent, but whose stochastic properties are homogeneous through time. On the other hand, weak stationarity only assumes temporal homogeneity of the first two moments.

In practice, dependence can take many different forms, and may be felt at short distances only or at longer ones. In fact, long-memory processes can mess up the convergence of renormalized maxima to the GEV distribution. The most obvious counter-example arises for perfectly dependent sequences: If $Y_1, \ldots, Y_n \sim F$, with $Y_i = Y_j$ almost surely, then $\max(Y_1, \ldots, Y_n) \overset{D}{=} Y_1 \sim F$, so the limit distribution of maxima can take essentially any form. In the sequel, we shall see that an analogue of the extremal types theorem can be obtained under strict stationarity and short-term dependence. Leadbetter (1983) formalized the idea of short-term dependence with the so-called $D(u_n)$ condition.

Definition 14 ($D(u_n)$ condition)**.** *Let* $\mathscr{A}, \mathscr{B} \subset \{1, \ldots, n\}$ *denote subsets of indices* $i_1 < \cdots < i_p$ *and* $j_1 < \cdots < j_q$ *respectively, such that* $j_1 > i_p + l$*, and let* $Y_{\mathscr{A}} \leq u$ *denote the event* $\cap_{i\in\mathscr{A}}\{Y_i \leq u\}$*, and similarly for* $Y_{\mathscr{B}} \leq u$*. Then the condition* $D(u_n)$ *is satisfied if*

$$|\Pr(Y_{\mathscr{A}} \leq u_n \cap Y_{\mathscr{B}} \leq u_n) - \Pr(Y_{\mathscr{A}} \leq u_n)\Pr(Y_{\mathscr{B}} \leq u_n)| \leq \alpha(n, l), \tag{1.7}$$

where $\alpha(n, l_n) \to 0$ *for some sequence* $l_n = o(n)$*,* $l_n \to \infty$ *as* $n \to \infty$*.*

This weak condition ensures that rare events sufficiently far apart can be considered to be nearly independent, so that their joint probabilities have no effect on the limit laws

for extremes. For Gaussian sequences with auto-correlation ρ_n at lag n, for example, the $D(u_n)$ condition is satisfied as soon as $\rho_n \log n \to 0$ as $n \to \infty$ (Beirlant *et al.*, 2004; Berman, 1964). In fact, under this condition and stationarity, it holds that for any threshold sequence u_n, $\Pr(M_n \leq u_n) = \left\{\Pr(M_{\lfloor n/k_n \rfloor} \leq u_n)\right\}^{k_n} + o(1)$ for some increasing sequence $k_n \to \infty$ (see, e.g., Beirlant *et al.*, 2004, pp.371–372). That is, for stationary time series satisfying the $D(u_n)$ condition, M_n can be regarded as the maximum of nearly independent smaller blocks. This is the key relation for the proof of Theorem 15.

Theorem 15 (Leadbetter, 1983)**.** *Let $\{Y_i\}_{i\geq 1}$ be a strictly stationary time series. Suppose that $(M_n - b_n)/a_n \xrightarrow{D} Z \sim G$ for some non-degenerate distribution G and suitable sequences $a_n > 0$ and $b_n \in \mathbb{R}$. If $D(u_n)$ holds for $u_n = a_n y + b_n$ for any $y \in \mathbb{R}$, then G is an extreme value distribution.*

Hence, the parametric family of distributions arising as limits for normalized maxima of stationary series with short-term dependence is the same as in the independent case. Let $\{Y_i^\star\}_{i\geq 1}$ denote the independent counterpart of the stationary process $\{Y_i\}_{i\geq 1}$, with maxima $M_n^\star$. Under the conditions of Theorem 15, both $(M_n - b_n)/a_n$ and $(M_n^\star - b_n)/a_n$ converge to a GEV distribution, but the parameters may differ. Theorem 16 and Proposition 18 show how to make the link between the two limit laws.

Theorem 16. *Under suitable conditions, $(M_n^\star - b_n)/a_n \xrightarrow{D} Z^\star \sim G^\star$, as $n \to \infty$, where $G^\star$ is a non-degenerate distribution function and $a_n > 0$ and b_n are normalizing sequences, if and only if $(M_n - b_n)/a_n \xrightarrow{D} Z \sim G$, and $G = G^{\star\theta}$, for some constant* $0 < \theta \leq 1$.

Definition 17 (Extremal index)**.** *The parameter θ arising in Theorem 16 is termed the extremal index. As θ decreases, serial dependence at high levels strengthens, while extremal independence is reached when $\theta = 1$.*

The parameters of the limiting distributions arising in Theorem 16 are linked as follows.

Proposition 18 (Parameters of the GEV under short-term dependence)**.** *Using the same notation as in Theorem 16, and letting $G \sim \text{GEV}(\mu, \sigma, \xi)$ and $G^\star \sim \text{GEV}(\mu^\star, \sigma^\star, \xi^\star)$, we have $\mu = \mu^\star - \sigma^\star(1 - \theta^{\xi^\star})/\xi^\star$, $\sigma = \sigma^\star \theta^{\xi^\star}$ and $\xi = \xi^\star$, where θ is the extremal index.*

According to Proposition 18, the shape parameter of the limiting GEV distribution is not affected by temporal dependence. However, the stronger the dependence, the lower the location and dispersion parameters. In other words, M_n is stochastically smaller than $M_n^\star$, which means that dependence reduces the sizes of the

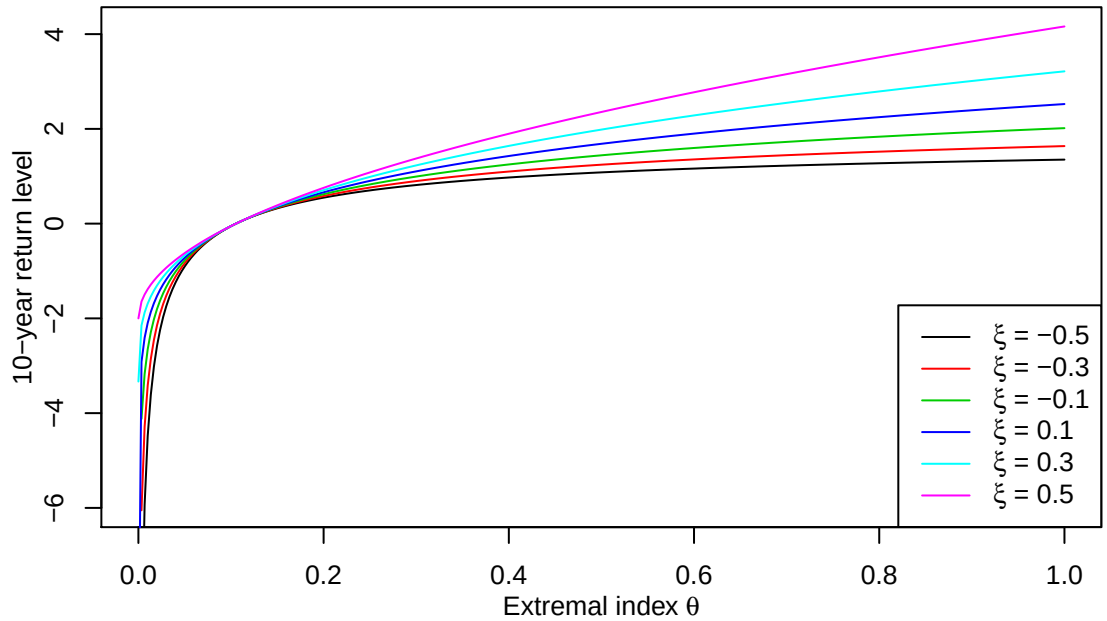

Figure 1.3: 10-year return level curves, that is y_p with $p = 1/10$, plotted against the extremal index θ for $\mu = 0$, $\sigma = 1$ and different values of the shape parameter ξ. The case $\theta = 1$ corresponds to extremal independence, and $\theta \to 0$ corresponds to perfect extremal dependence.

extremes for a series of given length. Consequently, return levels decrease with the strength of the dependence. The $1/p$-year return level in (1.5) becomes $y_p = \mu - \sigma\xi^{-1}\left[1 - \{-\theta^{-1}\log(1-p)\}^{-\xi}\right]$. Thus, y_p decreases as $\theta \to 0$; see Figure 1.3.

As we will see later on, the extremal index controls the degree of clustering at high levels. This parameter has various interpretations, see Section 1.1.2.4, one of these being the reciprocal limiting mean cluster size.

1.1.1.4 Inference

Motivated by the extremal types theorem and Theorem 15, the use of the GEV distribution is asymptotically justified to model the distribution of block maxima. Suppose that a stationary continuous-time process $\{Y_t\}_{t\in\mathcal{T}}$ has been sampled at $n = MN$ regular time points, and that the resulting time series falls into the maximum domain of attraction of some GEV distribution. Let $y_1, \ldots, y_n$ denote the observations.

The classical approach to inference is to form N blocks of M observations, and to fit the distribution GEV(μ, σ, ξ) to the block maxima $m_1 = \max(y_1, \ldots, y_M), \ldots, m_N = \max(y_{M(N-1)+1}, \ldots, y_{MN})$ by maximum likelihood. If the block length M is chosen suffi-

ciently large so that the maxima $m_1, \ldots, m_N$ can be regarded as mutually independent, the log-likelihood is of the form

$$\ell(\psi) = -N\log\sigma - (1+1/\xi)\sum_{i=1}^{N}\log\left\{1+\xi\left(\frac{m_i-\mu}{\sigma}\right)\right\}_+ - \sum_{i=1}^{N}\left\{1+\xi\left(\frac{m_i-\mu}{\sigma}\right)\right\}_+^{-1/\xi}, \quad (1.8)$$

where $\psi = (\mu, \sigma, \xi)^T$ is the vector of parameters. In practice, this expression can be maximized numerically with standard optimization routines available in statistical softwares (e.g., the function `optim` in `R`).

The choice of the block length M is crucial because it corresponds to a trade-off between bias and variance. Indeed, the larger M, the closer the distribution of block maxima to its asymptotic law. Hence, as M increases, the bias becomes smaller, but since the number of sample maxima available for fitting decreases at the same time, the variance of the parameter estimates becomes larger. In practice, different diagnostics can be used to determine a good value for M (parameter stability plot, for example), but usually M is chosen pragmatically so that the blocks correspond to natural periods such as months or years.

In regular situations, the maximum likelihood estimator is asymptotically normal, strongly consistent and has an asymptotic variance equal to the reciprocal Fisher information (see Davison, 2003, p.118). Regularity conditions for maximum likelihood estimation in extreme value theory are discussed in Smith (1985). The latter has established that the limiting behavior of the maximum likelihood estimator depends on the value of the shape parameter ξ:

- when $\xi > -1/2$, the maximum likelihood estimator obeys standard theory;
- when $-1 \leq \xi \leq -1/2$, the maximum likelihood estimator is a solution to the score equation, but it does not have the usual limiting distribution;
- when $\xi \leq -1$, the maximum likelihood estimator is not a solution to the score equation.

The value of ξ is problem-dependent. For rainfall data, ξ is usually found to be rather close to 0.2 (e.g., Tawn, 1988a; Katz *et al.*, 2002), so maximum likelihood estimation is in principle well behaved. In financial or insurance applications, the observations are usually heavy-tailed and ξ often found to be greater than 1, so maximum likelihood estimation may be used in those fields. When dealing with temperature minima data, however, since there exists a physical bound to extreme events, the shape parameter ξ is negative and more attention is needed.

Assessment of the uncertainty of the estimated parameters can be based on asymptotic normality, assuming that the maximum likelihood estimator has reached its limit law and that the reciprocal Fisher information is a suitable estimator for its variance. But in practice, profile likelihood-based methods (Coles, 2001, pp.57–61) are usually preferred since they can better reflect the right-skewness that one expects to see for the likelihood with respect to ξ. The "plug-in" principle, combined with formula (1.5), can be used to estimate return levels, or any other functional of interest of the parameters. Uncertainty for return levels can be assessed either by the delta method or by profile likelihood, provided the likelihood is properly re-parametrized in terms of return levels. In practice, as return levels associated with long return periods usually show a strong asymmetry to the right, profile likelihood is more reliable than the delta method.

Maximum likelihood estimation is not the only possibility for inference. While moment-based fitting techniques are usually inefficient for extremes because moments may not exist, probability-weighted moments have proven to be useful because of their good small-sample properties (Hosking *et al.*, 1985; Katz *et al.*, 2002), though they are relatively difficult to extend to more complex data. Bayesian techniques can also be applied and are especially efficient for high-dimensional problems where maximum likelihood estimates cannot easily be computed: for example in problems involving the computation of a high-dimensional integral, such as hierarchical models (Cooley *et al.*, 2007, 2006a; Sang & Gelfand, 2009; Blanchet & Davison, 2012). The applicability of Bayesian methods in this framework has been made possible thanks to the development of Monte Carlo Markov Chain (MCMC) algorithms (Hastings, 1970). However, though powerful, the Bayesian approach requires prior knowledge for the parameters and the tuning of hyper-parameters is often tricky in practice. Historically, non-parametric methods have also played a prominent role with the development of various non-parametric estimators for ξ, such as the well-known Hill estimator (Hill, 1975), the Pickands estimator (Pickands, 1975) and the moment estimator of Dekkers *et al.* (1989). Lots of effort have been devoted to finding good robust estimators for the key parameter ξ, since it determines the tail weight of the GEV distribution and thus the sizes of future extreme events.

1.1.2 Point process approach

In the point process representation for extremes described below, the notion of "rare events" changes, and extremes are now defined in terms of exceedances over a high threshold, not maxima. However, both approaches are closely linked and are even equivalent as the number of observations n tends to infinity, and as the threshold

converges to the upper endpoint of the underlying distribution. In practice, for finite n, when all original (e.g., hourly or daily) observations are available, threshold-based approaches are usually preferred to the block maxima method of §1.1.1, since more data can be used in the fitting procedure. These methods confer an appreciable reduction in the estimated variance of the parameters, leading to narrower confidence intervals, and thus —hopefully— more reliable conclusions. Difficulties arise when temporal dependence is present, since extremes tend to cluster at high thresholds. However, if dependence is properly modeled, the point process approach can also provide some insight into the structure of clusters of exceedances.

We start the exposition with the mathematical definition of point processes, emphasizing Poisson processes, and then discuss a particular case of interest for extremes: the point process of exceedances. These notions rely on non-trivial arguments from functional analysis and measure theory. The interested reader should refer to Cox & Isham (1980), Daley & Vere-Jones (2002) or Jacobsen (2005).

1.1.2.1 Definitions and basic results

Point processes are mathematical objects that can be thought of as "random distributions" of points in a set. Let $\{P_i\}_{i\geq 1}$ be a sequence of random points in a state space E endowed with a σ-algebra $\mathscr{E}$ of subsets of E. For our purposes, we can think of E as $\overline{\mathbb{R}}^d$, that is, the topological closure of $\mathbb{R}^d$, and of $\mathscr{E}$ as the σ-algebra of Borel sets $\mathscr{B}(E)$. Furthermore, let δ_p denote the Dirac measure at the point $p \in E$, that is, for any $A \in \mathscr{E}$,

$$\delta_p(A) = \begin{cases} 1, & p \in A, \\ 0, & p \notin A. \end{cases} \tag{1.9}$$

Definition 19 (Counting measure and point measure)**.** *For a given sequence of points $\{p_i\}_{i\geq 1} \in E$, the function $m : \mathscr{E} \longrightarrow \overline{\mathbb{R}}$ defined by*

$$m(A) = \sum_{i=1}^{\infty} \delta_{p_i}(A) = \text{card}\{i \,:\, p_i \in A\}$$

is called a counting measure on $\mathscr{E}$. It is a point measure if $m(K) < \infty$ for all compact sets $K \in \mathscr{E}$.

According to Definition 19, a counting measure simply enumerates the number of points p_i within suitable sets A. Intuitively, a point process does a similar job, but with random points $\{P_i\}_{i\geq 1} \in E$. More formally, let $M(E)$ denote the collection of all point measures on E equipped with an appropriate σ-algebra $\mathscr{M}(E)$. A point process on E is a measurable map $N : (\Omega, \mathscr{F}, \Pr) \longrightarrow \{M(E), \mathscr{M}(E)\}$, where $(\Omega, \mathscr{F}, \Pr)$ is a probability

space. It can be written as $N(\cdot) = \sum_{i=1}^{\infty} \delta_{P_i}(\cdot)$. All realizations of a point process are point measures.

Poisson processes play a central role in extreme value theory.

Definition 20 (Poisson process)**.** *Let Λ be a Radon measure on $\mathscr{E}$, that is a locally finite measure: $\Lambda(A) < \infty$ for all compact measurable sets $A \subset E$. A Poisson process —or Poisson random measure— N with mean measure Λ is a point process that satisfies the two following properties:*

1. *For every $k \geq 0$ and measurable set A,*

$$\Pr\{N(A) = k\} = \begin{cases} e^{-\Lambda(A)}\Lambda(A)^k/k!, & \Lambda(A) < \infty, \\ 0, & \Lambda(A) = \infty. \end{cases}$$

2. *For any mutually disjoint sets $A_1, \ldots, A_m \in \mathscr{E}$, $N(A_1), \ldots, N(A_m)$ are independent.*

The first condition ensures that the number of points in any set $A \in \mathscr{E}$ is distributed as a Poisson random variable with mean $\Lambda(A)$. The second condition implies that events happening in some region $A_1 \in \mathscr{E}$ do not influence the stochastic properties of some other disjoint region $A_2 \in \mathscr{E}$. For example, if $E = \mathbb{R}$ represents the time axis, this condition entails that "the past does not influence the future" —and vice versa.

The special class of homogeneous Poisson processes with intensity (or rate) $\lambda > 0$ has mean measure λ times the Lebesgue measure. In $\mathbb{R}_+$, one has $\Lambda([0,t]) = \lambda t$.

The distribution of a point process N is uniquely determined by its finite-dimensional distributions, that is, by the probabilities $\Pr\{N(A_1) = k_1, \ldots, N(A_m) = k_m\}$, for any choice of $A_1, \ldots, A_m \in \mathscr{E}$, any $k_i \geq 0$, $i = 1, \ldots, m$, and any $m \in \mathbb{N}$. Such a complex distribution function is not easily handled and it turns out that a convenient way of "summarizing" the probability law of a point process is through its Laplace functional, the counterpart of the characteristic function for random variables. In the same way as pointwise convergence of characteristic functions implies weak convergence of random variables, pointwise convergence of Laplace functionals implies weak convergence of point processes; see Theorem 23 below. This mathematical tool is useful for the derivation of the limiting point process of exceedances over sequences of increasingly high thresholds; see Theorem 25.

Definition 21 (Laplace functional)**.** *The Laplace functional of the point process N, uniquely defined, is defined as*

$$\mathscr{L}(f) = \mathrm{E}\left\{\exp\left(-\int_E f(p)\mathrm{d}N(p)\right)\right\}, \tag{1.10}$$

where f is a non-negative measurable function on the state space E, and where the integral on the right-hand side of (1.10) is the Lebesgue–Stieltjes integral.

If we set $f(p) = \sum_{i=1}^{m} t_i I(p \in A_i)$ in (1.10), where $t_1, \ldots, t_m \geq 0$ and $I(\cdot)$ is the indicator function, $\mathscr{L}(f)$ reduces to the joint moment generating function of the variables $N(A_1), \ldots, N(A_m)$. Hence, the Laplace functional determines the distribution of a point process completely.

Consider now a sequence of point processes $N_1, N_2, \ldots$, defined on the same state space $E \subset \overline{\mathbb{R}}^d$ endowed with the σ-algebra $\mathscr{E}$ of Borel sets.

Definition 22 (Weak convergence of point processes)**.** *The sequence of point processes $\{N_i\}_{i\geq 1}$ is said to converge weakly to the point process N in $M(E)$ if all its finite-dimensional distributions converge, that is if*

$$\Pr\{N_n(A_1) = k_1, \ldots, N_n(A_m) = k_m\} \to \Pr\{N(A_1) = k_1, \ldots, N(A_m) = k_m\}, \qquad n \to \infty,$$

for all $m \geq 1$, all $k_1, \ldots, k_m \in \mathbb{N}$ and all possible choices of sets $A_1, \ldots, A_m \in \mathscr{E}$ such that $N(\partial A_i) = 0$ almost surely for $i = 1, \ldots, m$, where ∂A denotes the boundary of A.

A criterion which guarantees the weak convergence of a sequence of point processes via Laplace functionals is given by Daley & Vere-Jones (2002) and Embrechts *et al.* (1997, p.234). They show that weak convergence of point processes is equivalent to pointwise convergence of their Laplace functionals.

Theorem 23 (Weak convergence of point processes)**.** *$\{N_i\}_{i\geq 1}$ converges weakly to the point process N in $M(E)$ if and only if the corresponding Laplace functionals converge for all continuous and compactly supported measurable functions $f \geq 0$: $\mathscr{L}_{N_n}(f) \to \mathscr{L}_N(f)$, as $n \to \infty$.*

1.1.2.2 Point process of exceedances

Consider the sequence of i.i.d. random variables $\{Y_i\}_{i\geq 1}$ with marginal distribution F and let $M_n = \max(Y_1, \ldots, Y_n)$. As in §1.1.1, the independence assumption will be extended later on; see §1.1.2.3. Furthermore, assume that there exist normalizing constants $a_n > 0$ and b_n such that $(M_n - b_n)/a_n \xrightarrow{D} Z \sim \text{GEV}(0, 1, \xi)$; this condition implies that there exists a threshold sequence $u_n = a_n y + b_n$ such that $n\{1 - F(u_n)\} \to \lambda = (1 + \xi y)_+^{-1/\xi}$, as $n \to \infty$.

Let P_n denote the set of normalized observations $\{i/(n+1), (Y_i - b_n)/a_n\} \in \mathbb{R}^2\}$, $i = 1, \ldots, n$. The role of the factor $(n+1)^{-1}$ in the first argument is to map the time axis to

the interval $(0,1)$, whereas the affine renormalization in the second argument ensures that the sizes of extreme events are properly "stabilized" to have a non-degenerate limiting distribution. The sequence of points P_n defines a point process on the space $E = [0,1] \times \mathbb{R}$, endowed with the generated σ-algebra of the Borel sets:

$$N_n(\cdot) = \sum_{i=1}^{n} \delta_{P_n}(\cdot).$$

We aim at characterizing the asymptotic behavior of the points P_n on some suitable extremal sets. Notice that evaluating N_n on regions of the form $\mathscr{A}_y = [t_1, t_2] \times [y, \infty)$ corresponds to looking at the the process Y above the level $u_n = a_n y + b_n$ and during some time period $[(n+1)t_1, (n+1)t_2]$. We can write

$$N_n(A \times [y,\infty)) = \sum_{i=1}^{n} \delta_{P_n}(A \times [y,\infty)) = \sum_{i=1}^{n} \delta_{i/(n+1)}(A) I(Y_i > u_n), \tag{1.11}$$

for any $A \in \mathscr{B}([0,1])$. On the right hand-side of (1.11), N_n can also be viewed as a point process on $E^* = [0,1]$, defining the so-called point process of times of exceedances over the threshold u_n.

By stating that the pointwise limit of the Laplace functional of N_n coincides with the Laplace functional of a Poisson process on sets of the form $E = [0,1] \times [u,\infty)$, $u \in \mathbb{R}$, Proposition 24 combined with Theorems 23 and 25 establish that the limiting point process of exceedances is a non-homogeneous Poisson process.

Proposition 24 (Laplace functional of a Poisson process, Embrechts *et al.*, 1997, p.228)**.** *The Laplace functional of a Poisson process on a state space E with mean measure Λ is*

$$\mathscr{L}(f) = \exp\left[-\int_E \left\{ 1 - e^{-f(p)} \right\} \mathrm{d}\Lambda(p) \right],$$

for any measurable function $f \geq 0$.

Theorem 25 (Convergence of the point process of exceedances, Embrechts *et al.*, 1997, p.238)**.** *If $(M_n - b_n)/a_n \xrightarrow{D} Z \sim \mathrm{GEV}(0,1,\xi)$, the Laplace functional of the point process N_n converges to*

$$\mathscr{L}(f) = \exp\left[-\int_E \left\{ 1 - e^{-f(p)} \right\} \mathrm{d}\Lambda(p) \right],$$

on $E = [0,1] \times [u,\infty)$, $u \in \mathbb{R}$, where $\Lambda([t_1,t_2] \times [y,\infty)) = (t_2 - t_1)(1+\xi y)_+^{-1/\xi}$, $y > u$.

According to Theorem 25, the mean measure of the limiting Poisson process of exceedances is Λ; in particular, it turns out that, asymptotically, the point process of exceedance times is homogeneous Poisson with intensity $\lambda = (1+\xi y)_+^{-1/\xi}$.

1.1.2.3 Extension to stationary series

As in §1.1.1, establishing the asymptotic law of block maxima, the hypothesis of temporal independence appears to be far too strong an assumption in most applications. In order to extend the asymptotic results to a more general framework, we shall, similarly to the block maxima approach, restrict our attention to strictly stationary processes with short-range dependence. The $D(u_n)$ condition (1.7), however, needs to be replaced by another criterion, the so-called $\Delta(u_n)$ condition (Hsing *et al.*, 1988), adapted for threshold exceedances. This criterion is stronger than the $D(u_n)$ condition since the focus is not on maxima only, but on the whole point process above some level. But the $\Delta(u_n)$ condition is still weak because it constrains the extremes only, not the body of the distribution of Y. Suppose that $\{Y_i\}_{i\geq 1}$ is a strictly stationary sequence of random variables with common marginal distribution F, in the maximum domain of attraction of the GEV$(0,1,\xi)$ distribution.

Definition 26 ($\Delta(u_n)$ condition; Hsing *et al.*, 1988, Beirlant *et al.*, 2004, p.383)**.** *Let u_n be a sequence of thresholds such that $n\{1-F(u_n)\} \to \lambda$ for some $\lambda > 0$ as $n \to \infty$. Assume that $\mathscr{F}_{\mathscr{A}}(u_n)$ denote the σ-algebra generated by the events $\{Y_i > u_n : i \in \mathscr{A}\}$. Condition $\Delta(u_n)$ is said to be satisfied if for all $\mathscr{A} \in \mathscr{F}_{1,\ldots,m}(u_n)$, all $\mathscr{B} \in \mathscr{F}_{m+l,\ldots,n}(u_n)$, and all $m = 1,\ldots,n-l$,*

$$|\Pr(\mathscr{A}\cap\mathscr{B}) - \Pr(\mathscr{A})\Pr(\mathscr{B})| \leq \alpha(n,l), \tag{1.12}$$

where $\alpha(n,l_n) \to 0$ for some sequence $l_n = o(n)$, $l_n \to \infty$ as $n \to \infty$.

As with the $D(u_n)$ condition, the $\Delta(u_n)$ condition forbids long-range dependence of extremes. In contrast, short-range dependent processes are permitted, and temporal dependence can have a local effect on the behavior of exceedances at high levels: if temporal dependence is strong enough, exceedances above a high threshold tend to occur in clusters. Asymptotically, since the time axis is rescaled to the interval $(0,1)$, extreme values within the same cluster will occur exactly at the same time. Clustering of extreme values is illustrated in Figure 1.4. In this example, we consider a moving maximum process of order 1, that is a process $\{Y_i\}_{i\geq 1}$ such that $Y_0 = Z_0$ and

$$Y_i = (a+1)^{-1}\max(aZ_{i-1}, Z_i), \quad i = 1,2,\ldots, \tag{1.13}$$

where $a \geq 0$ controls the strength of dependence and $\{Z_i\}_{i\geq 0}$ denotes an i.i.d. sequence of unit Fréchet random variables, i.e., $\Pr(Z_i \leq z) = \exp(-1/z)$, $z > 0$. This process has unit Fréchet margins and clearly satisfies the $\Delta(u_n)$ condition. One can show that the extremal index of such a process is $\theta = \max(1,a)/(a+1)$. Hence, extremal independence is reached for $a = 0$ or $a \to \infty$, and perfect dependence is never attained,

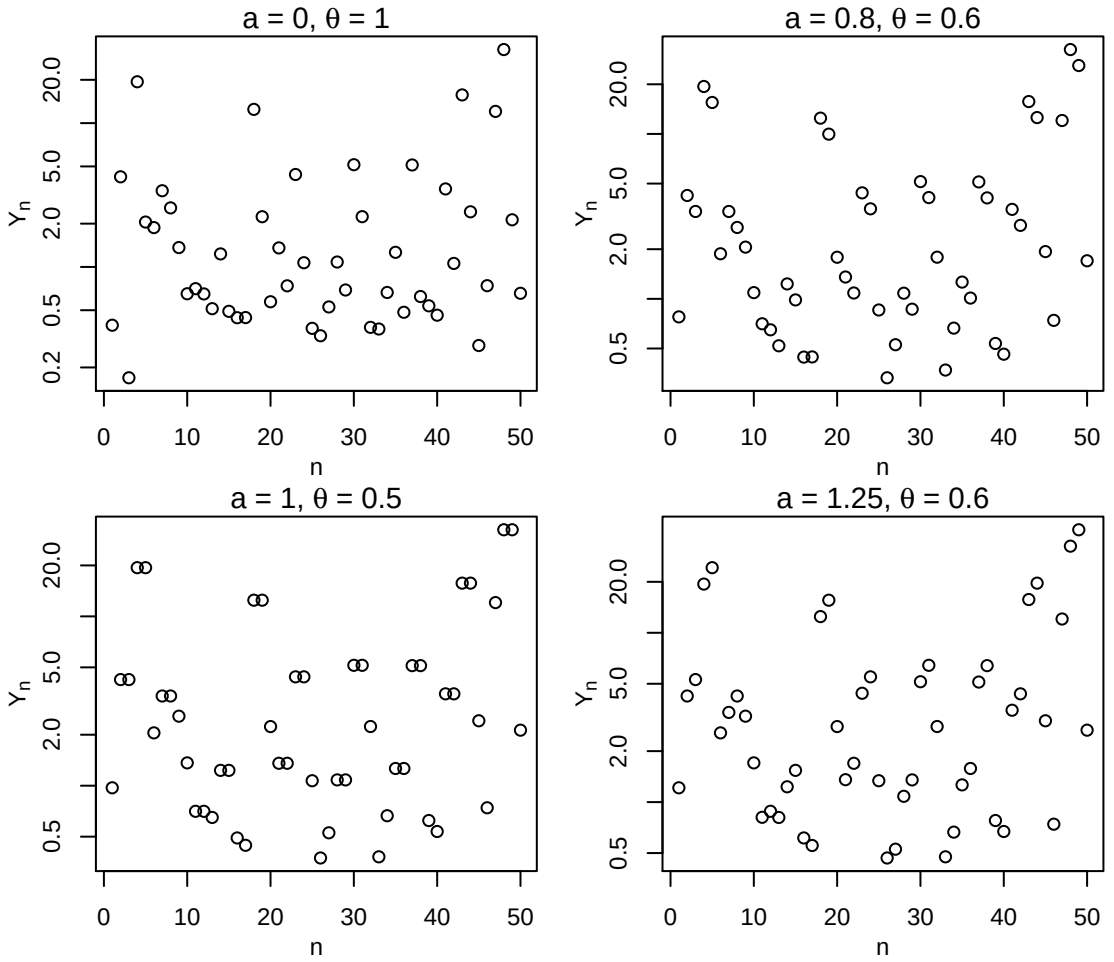

Figure 1.4: Moving maximum process of order 1, see equation (1.13), for different extremal dependence properties; Simulations with the same random seed are shown for $a = 0$ (top left), $a = 0.8$ (top right), $a = 1$ (bottom left) and $a = 1.25$ (bottom right). The extremal index θ is also reported.

whatever the value of a. The case of strongest dependence corresponds to $a = 1$, that is $\theta = 1/2$. Figure 1.4 shows realizations of such a process for $a = 0, 0.8, 1, 1.25$ and highlights the degree of clustering and the structure of the clusters in the four different cases. When $a = 0$, extreme values tend to occur alone, as expected; when $a = 1$, extreme events tend to occur in groups of two; and the intermediate cases $a = 0.8, 1.25$ show other specific cluster patterns.

Formally, the cluster size distribution can be defined as follows: Suppose that u_n is a threshold sequence and split the data into $k_n = \lfloor n/r_n \rfloor$ blocks B_i of length r_n. Moreover, suppose that the number of blocks k_n is such that $k_n \to \infty$ and $r_n \to \infty$, $r_n = o(n)$, as $n \to \infty$. The exceedances over u_n in a block, if any, are said to form a

cluster. In this setting, the times of exceedances within the same cluster will bunch together as $n \to \infty$ since $r_n/n \to 0$. One can now define the cluster size distribution at a specific level u_n and the limiting cluster size distribution.

Definition 27 (Cluster size distribution)**.** *The cluster size distribution at the level u_n is*

$$\pi_n(j) = \Pr\left\{ \sum_{i=1}^{r_n} I(Y_i > u_n) = j \middle| M_{r_n} > u_n \right\}, \quad j \in \mathbb{N}. \tag{1.14}$$

If it exists, $\pi(j) = \lim_{n\to\infty} \pi_n(j)$ is the limiting cluster size distribution.

In the moving maximum process example in (1.13), the cluster size distribution depends on the value of a. One can show that for $0 \le a \le 1$, $\pi(1) = 1-a$, $\pi(2) = a$ and $\pi(j) = 0$, $j \neq 1,2$.

The next result extends Theorem 25 to the stationary case. Under some conditions made precise below, the asymptotic point process turns out to be a compound Poisson process whose multiplicities are distributed according to the limiting cluster size distribution π.

Theorem 28 (Convergence of the point process of exceedances under stationarity; Hsing *et al.*, 1988, Beirlant *et al.*, 2004, p.384)**.** *Suppose that $\Delta(u_n)$ condition holds for the threshold sequence $u_n = a_n y + b_n$. Let there exist positive sequences l_n and r_n and a distribution π such that $l_n = o(r_n)$, $r_n = o(n)$, $n\alpha(n,l_n) = o(r_n)$, where the constants $\alpha(n,l_n)$ are given by the $\Delta(u_n)$ condition, and $\pi_n(j) \to \pi(j)$ for all $j \in \mathbb{N}$ as $n \to \infty$. Then, the point process of cluster maxima $N'_n = \sum_{i=1}^{n} \delta_{P'_n}$, defined by*

$$P'_n = \left\{ \left(\frac{(i-1)r_n + 1}{n}, \frac{M_{B_i} - b_n}{a_n} \right), i = 1, \ldots, k_n \right\},$$

converges to a Poisson process N' on $E = [0,1] \times [u,\infty)$, $u \in \mathbb{R}$, with mean measure

$$\Lambda_\theta\{[t_1,t_2] \times [y,\infty)\} = \theta(t_2 - t_1)(1+\xi y)_+^{-1/\xi},$$

and the point process of exceedances N_n converges to a process N with Laplace functional

$$\mathscr{L}_N(f) = \exp\left[-\int_E \left\{ 1 - \sum_{j=1}^{\infty} \pi(j) \exp\left(-jf\right) \right\} \mathrm{d}\Lambda_\theta \right].$$

Two main conclusions can be drawn from this result. First, the cluster maxima are asymptotically distributed according to a non-homogeneous Poisson process on E

whose intensity function is the same as in the independent case except that it is multiplied by the extremal index θ. That is, clusters arise at a rate θ times less often than in the independent case. Second, the point process of exceedances N_n converges to a compound Poisson process with mean measure $\theta(t_2 - t_1)\lambda$, where $\lambda = (1+\xi y)_+^{-1/\xi}$, and multiplicities following the cluster size distribution π. If π is reduced to the atom at one, then $\sum_{j=1}^{\infty} \pi(j) \exp\left(-jf\right) = \exp(-f)$ and we recover the limiting Poisson process arising in the independent case.

1.1.2.4 Interpretation and estimators for the extremal index

There are various ways of interpreting the extremal index, giving rise to several estimators having their own advantages and drawbacks. A detailed exposition can be found in Embrechts *et al.* (1997, pp.418-429), Ancona-Navarrete & Tawn (2000), or more recently in Beirlant *et al.* (2004, p.390). These methods are usually constructive and provide a way to identify clusters, hence giving simultaneously "declustering" procedures that can be used in practice to remove temporal dependence in the series.

θ as the reciprocal mean cluster size. Using the notation specified above, θ can be expressed in the following fashion (Beirlant *et al.*, 2004, p.377):

$$\theta^{-1} = \lim_{n\to\infty} \frac{r_n\{1-F(u_n)\}}{\Pr(M_{r_n} > u_n)} = \lim_{n\to\infty} \mathrm{E}\left\{ \sum_{i=1}^{r_n} I(Y_i > u_n) \mid M_{r_n} > u_n \right\}, \tag{1.15}$$

Therefore, θ^{-1} can be interpreted as the limiting mean size of a cluster, that is the limiting mean number of exceedances in a block of length r_n, given that at least one observation exceeds the threshold u_n in that block. Equivalently, we can express θ in terms of the cluster size distribution:

$$\theta^{-1} = \lim_{n\to\infty} \sum_{j=1}^{r_n} j\pi_n(j). \tag{1.16}$$

Compared to the independent case, the mean distance between clusters at high thresholds is hence increased by a factor θ^{-1}. This representation of the extremal index motivates the so-called "block-estimator" (Smith & Weissman, 1994), which is simply the empirical counterpart of expression (1.15). In practice, one needs to specify a suitable block length k and compute the average number of exceedances within blocks having at least one extreme observation. A drawback of this estimator is that it is very sensitive to the choice of r.

An improved estimator, proposed by Robert *et al.* (2009), consists in using sliding blocks instead of disjoint, nearly independent, blocks. They prove that their estimator

is more efficient, has a smaller asymptotic bias and is asymptotically normal under mild conditions.

θ as conditional probability. Alternatively, the extremal index can also be regarded as the probability that a high threshold exceedance is the final element in a cluster of exceedances (O'Brien, 1987; Embrechts *et al.*, 1997, p.422):

$$\theta = \lim_{n\to\infty} \Pr\{\max(Y_2,\ldots,Y_{r_n}) \leq u_n \mid Y_1 > u_n\}. \tag{1.17}$$

This definition leads to the "runs-estimator" (Smith & Weissman, 1994), the empirical counterpart of expression (1.17). It consists in counting the number of exceedances over a high threshold u that are followed by a run of r consecutive observations below it, and divide it by the total number of exceedances. Again, this estimator has the drawback of being very sensitive to the choice of the run parameter r.

θ in terms of times between exceedances. A moment-based approach relying on the distribution of inter-exceedance times has been proposed by Ferro & Segers (2003). It deserves particular attention because it was the first method providing an automatic declustering procedure, which, therefore, does not require the preliminary choice of a run or block parameter. Suppose that $\{Y_i\}_{i\geq 1}$ is a strictly stationary sequence of random variables with marginal function F, and let $T(u)$ be the inter-exceedance time, that is $T(u) = \min\{t \geq 1 \,:\, Y_{t+1} > u \mid Y_1 > u\}$. It turns out that under mild regularity conditions and a mixing condition, similar in spirit to the $\Delta(u_n)$ condition, if u_n is a suitable threshold sequence, then $\{1-F(u_n)\}T(u_n)$ converges to a mixture distribution (Ferro & Segers, 2003):

$$\{1-F(u_n)\}T(u_n) \xrightarrow{D} (1-\theta)\delta_0 + \theta E_\theta, \qquad n\to\infty, \tag{1.18}$$

where δ_0 is a point-mass at zero, corresponding to exceedances within the same cluster, and E_θ is the exponential distribution with mean θ^{-1}. Consequently, the extremal index plays a double role: on the one hand, θ is the proportion of non-zero inter-exceedance times; on the other hand, it corresponds to the reciprocal mean of non-zero inter-exceedance times. Letting $T_i = S_{i+1} - S_i$, where $S_1,\ldots,S_{N_u}$ are the N_u times of the exceedances over u, an estimator for θ based on the first two moments of $(1-\theta)\delta_0 + \theta E_\theta$ is (Ferro & Segers, 2003)

$$\widehat{\theta}_u = 2\left(\sum_{i=1}^{N_u-1} T_i\right)^2 \left\{(N_u-1)\sum_{i=1}^{N_u-1} T_i^2\right\}^{-1}. \tag{1.19}$$

An estimator that is shown to correct the first-order bias of $\widehat{\theta}_u$ is

$$\widehat{\theta}_u^* = 2\left\{\sum_{i=1}^{N_u-1}(T_i-1)\right\}^2\left\{(N_u-1)\sum_{i=1}^{N_u-1}(T_i-1)(T_i-2)\right\}^{-1}. \tag{1.20}$$

To avoid problematic effects, one can pool $\widehat{\theta}_u$ and $\widehat{\theta}_u^*$, leading to the so-called "intervals estimator":

$$\tilde{\theta}_u = \begin{cases} \min(1,\widehat{\theta}_u), & \max_{1\le i\le N_u-1} T_i \le 2, \\ \min(1,\widehat{\theta}_u^*), & \max_{1\le i\le N_u-1} T_i > 2. \end{cases} \tag{1.21}$$

This corresponds to the runs-estimator with run parameter the Cth largest inter-exceedance time, that is $T_{(C)}$, where $C = \lfloor \theta N_u \rfloor + 1$.

Ferro & Segers (2003)'s methodology was further developed by Süveges (2007, 2009), who proved that the asymptotic result (1.18) is also valid for the K-gaps quantity $T^K(u_n) = \max\{T(u_n) - K, 0\}$ under the same conditions. A likelihood-based estimator for independent observations was proposed, and is valid under the $D^K(u_n)$ condition (Süveges, 2007).

Other estimators for the extremal index have been proposed; for example the iterative least squares estimator of Süveges (2007), or the two-threshold based estimator of Laurini & Tawn (2003). The latter has better performance than classical methods and allows more realistic modeling of threshold exceedances, but is more complex to use in practice because of the need to choose a second threshold. An alternative automatic approach was advocated by Robert (2013a).

1.1.2.5 Inference

When the observations $y_1, \ldots, y_n$ are independent, one can base inference on the limiting Poisson process for exceedances as follows. Let u be a high threshold and let $\mathscr{A}_u$ be the extreme set $\mathscr{A}_u = [t_1, t_2] \times [u, \infty)$ for $0 \le t_1 < t_2 \le 1$. If N_u exceedances over u are observed at times $t_1, \ldots, t_{N_u}$, the likelihood for the region $\mathscr{A}_u$ is proportional to

$$L(\psi) \propto \exp\left[-n_{\text{year}}(t_2-t_1)\left\{1+\xi\left(\frac{u-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right]\prod_{i=1}^{N_u}\sigma^{-1}\left\{1+\xi\left(\frac{y_{t_i}-\mu}{\sigma}\right)\right\}_+^{-1/\xi-1}, \tag{1.22}$$

where $\psi = (\mu, \sigma, \xi)^T$ is the parameter vector and n_{year} denotes the number of years of observation. Maximization of (1.22) yields the maximum likelihood estimator $\widehat{\psi} = (\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})^T$ based on the point process approach. When the threshold u is high enough, the parameter estimates should not be threshold-dependent, but in practice there is a bias-variance trade-off: the higher the threshold, the lower the bias and the

larger the variance, and vice versa. Compared to the block maximum approach, where only a few observations are available in general, this procedure has the advantage of using more data for inference. Extreme events are scarce by nature, and this approach allows better use of the information available at high levels, thus improving the precision of the parameter and return level estimates.

When the observations come from a temporally dependent process, several strategies for inference exist. The simplest approach is to decompose the data into disjoint independent clusters and to model cluster maxima only, thanks to Theorem 28 (see, e.g., Coles & Tawn, 1994). Declustering schemes were discussed above. Another popular method is to ignore temporal dependence for the estimation of parameters and return levels —hence using a misspecified likelihood—, and to inflate the standard errors appropriately (see, e.g., Fawcett & Walshaw, 2007). A third more difficult approach would be to specify an explicit model for the dependence structure, such as a first-order Markov chain (see, e.g., Smith *et al.*, 1997; Bortot & Tawn, 1998).

Peaks over threshold (POT) approach. An alternative threshold-based representation of extremes has been developed by Davison & Smith (1990), giving rise to extensive discussions and applications. Let $a_n > 0$ and b_n be the normalizing constants arising in Theorem 28. Assuming that the limiting Poisson process is a good approximation for the process of exceedances above the finite threshold u, we have for $y > 0$ that

$$\begin{aligned}
\Pr\{(Y_n - b_n)/a_n > u + y \mid (Y_n - b_n)/a_n > u\} &= \frac{\Pr\{(Y_n - b_n)/a_n > u + y\}}{\Pr\{(Y_n - b_n)/a_n > u\}} \\
&\approx \frac{\Lambda\{[0,1] \times [u + y, \infty)\}}{\Lambda\{[0,1] \times [u, \infty)\}} \\
&= \frac{\left\{1 + \xi(u + y - \mu)/\sigma\right\}_+^{-1/\xi}}{\left\{1 + \xi(u - \mu)/\sigma\right\}_+^{-1/\xi}} \\
&= \left(1 + \xi y/\tau\right)_+^{-1/\xi},
\end{aligned}$$

where $\tau = \sigma + \xi(u - \mu)$. Hence, the exceedances $Y - u$, *conditional on* $Y > u$, can be modeled with the Generalized Pareto Distribution (GPD) with scale parameter τ and shape parameter ξ, although the value of τ depends on u. A similar theoretical result was shown by Balkema & de Haan (1974) and Pickands (1975). It gives rise to the independent conditional likelihood

$$L(\psi) \propto \prod_{i=1}^{N_u} \tau^{-1} \left(1 + \xi y_{t_i}/\tau\right)_+^{-1/\xi - 1}, \tag{1.23}$$

where $\psi = (\tau, \xi)^T$ is the parameter vector. The maximum likelihood estimator for ψ

can be derived by maximizing expression (1.23) numerically. The third parameter is the rate at which the exceedances over u occur, that is $\zeta_u = \Pr(Y > u)$, which can be estimated by its empirical counterpart $\widehat{\zeta}_u = N_u/n$ with approximate variance $N_u(n - N_u)/n^3$. The GPD representation is maybe more convenient than the point process point of view but since the scale parameter τ is threshold-dependent, interpretation needs care. The N-year return level, that is the level expected to be exceeded once in N years on average, is

$$y_N = u + \frac{\sigma}{\xi}\{(N m_{\text{year}} \zeta_u)^{\xi} - 1\}, \tag{1.24}$$

where m_{year} is the number of observations per year. Expression (1.24) can be estimated using the "plug-in" principle, and uncertainty assessment can be based on the delta method or the profile likelihood, as for block maxima.

r-largest order statistics approach. The r-largest order statistics model is very similar to the POT approach and is also a consequence of the point process representation. Basically, the full likelihood is constructed from a product of block contributions, each being an individual point process contribution corresponding to the exceedances over the rth largest order statistics in that block. Mathematically speaking, the likelihood may be written as

$$L(\psi) \propto \prod_{i=1}^{m} \exp\left[-\left\{ 1 + \xi\left(\frac{y_i^{(r)} - \mu}{\sigma} \right) \right\}_+^{-1/\xi} \right] \prod_{j=1}^{r} \sigma^{-1} \left\{ 1 + \xi\left(\frac{y_i^{(j)} - \mu}{\sigma} \right) \right\}_+^{-1/\xi - 1}, \tag{1.25}$$

where $\psi = (\mu, \sigma, \xi)^T$ is the vector of parameters, i is the block index (for a total of m blocks) and $y_i^{(j)}$ is the jth largest order statistic of the ith block. More details can be found in Coles (2001, Sections 3.5 and 7.9). Parameter estimates can be obtained by numerical maximization of (1.25). Their interpretation is the same as for the block maximum and point process approaches.

Choice of the threshold. A natural concern is the choice of the threshold u (or equivalently the value of r in the r-largest order statistics approach), so that the asymptotic approximation provides a reliable model for the tail of the distribution. As mentioned above, the choice of the threshold yields a bias-variance trade-off. To minimize the variance, one is tempted to choose the threshold as low as possible. However, to reduce the bias, one should increase the threshold. In practice, several diagnostics have been developed (see Davison & Smith, 1990) to help decide whether a given threshold is suitable or not. These decision tools rely on the following obvious assertion, stated for the POT approach: If the GPD is a valid model for the exceedances over the threshold u, then for all $v > u$, the GPD also has to be an appropriate distribution for the exceedances over v. Therefore, one might expect some sort of stability when

the GPD is fitted to the exceedances over a sequence of increasing thresholds. This stability translates in at least two ways:

- *In terms of mean excess:* If $Y \sim \text{GPD}(\tau, \xi)$, then $\text{E}(Y \mid Y > u) = (\tau + \xi u)/(1 - \xi)$, which is a linear function of u. The relationship between the empirical mean excess and the threshold u can be viewed in the so-called mean residual life plot. Using this graphical diagnostic, one should therefore choose the lowest threshold u that is sufficiently high to get a linear relationship for all $v > u$ (taking uncertainty into account).
- *In terms of parameter estimates:* If the $\text{GPD}(\tau, \xi)$ is a valid model for $Y - u \mid Y > u$, estimates of ξ and $\tilde{\tau} = \tau - \xi u$ ought to be constant with respect to increasing values of u. Parameter stability plots can help suggest reasonable values for the threshold.

Another pragmatic solution for threshold selection is to choose a high quantile, and to verify its suitability with model diagnostics. In environmental applications, empirical 95%–99% percentiles are often reasonable thresholds, whereas for financial datasets, higher thresholds may be of interest.

1.2 Multivariate extreme value theory

Section 1.1 was concerned with univariate time series. In practice, however, the joint modeling of extremes is often of interest for several reasons. First, one may want to have a qualitative description of the structure and the degree of extremal dependence between two or more series of observations. This is especially important for risk assessment. In hydrology, for example, if extreme rainfall events occur simultaneously over a whole catchment, it would increase the overall risk of floods in that region (Thibaud *et al.*, 2013). In finance, if several stock markets have huge losses on the same day, this would increase the risk of a global financial collapse. Therefore, if extremal dependence is not properly encounted for, one might misestimate the associated measure of risk. Second, the use of a multivariate model permits us to treat the observations in a general and coherent way, and the interpretation of the results is sometimes easier in a multivariate framework. And third, if extremal dependence is well modeled, joint modeling allows us to borrow strength from "neighboring" time series, in order to better estimate marginal parameters. In spatial statistics, this is often referred to as a trade-off between space and time.

Although appealing, the joint modeling of extremes is difficult in several respects. First, there is no obvious way to order multivariate observations, so the definition of

an *extreme* multivariate observation is not as clear as for univariate data. Second, as explained in more details below, the class of multivariate extreme value distributions is nonparametric; unlike the univariate case (with GEV/GPD distributions), they cannot be characterized by a finite number of parameters. Furthermore, bivariate extreme value theory is fairly well understood and developed, but flexible parametric models for extremes in dimension greater than $D = 2$ are still lacking. Finally, the curse of dimensionality cannot be avoided: modeling, fitting, simulation and model checking are much more tricky and computationally intensive for large D.

Another important issue concerns asymptotic independence. Several common multivariate distributions (e.g., the multivariate Gaussian) show decreasing dependence at higher levels. At a high but finite level, extreme events can be nearly independent and the use of classical models for multivariate extremes, which are asymptotically dependent, can yield misleading conclusions. In particular, extrapolation in the tail can be specious and joint return levels badly estimated. Coles *et al.* (1999) have proposed the quantities χ and $\overline{\chi}$ to help discriminate between asymptotic independence and dependence. Moreover, Ledford & Tawn (1996, 1997, 2003) and Ramos & Ledford (2009, 2011) have proposed asymptotically independent models, detecting and measuring the strength of the decay towards independence at high levels. Wadsworth & Tawn (2012) also tackle this problem and provide hybrid spatial models that can handle both asymptotic dependence and asymptotic independence at different spatial distances.

In Sections 1.2.1 and 1.2.2, approaches based on block maxima and point processes for multivariate extremes are presented. Then in §1.2.3, a brief introduction is given to the theory of copulas, which are mathematical objects allowing one to treat the marginal distributions of random vectors separately from their dependence structure. Asymptotic independence is addressed in §1.2.4, and finally, measures for extremal dependence are discussed in §1.2.5. But before going further, some notation needs to be introduced.

Vector notation. To clarify matters, all vectors will be denoted by symbols in bold, and they will usually be assumed to be of dimension D. That is, $\boldsymbol{y} = (y_1, \ldots, y_D)$, $\mathbf{0} \in \mathbb{R}^D$ is a vector of zeros, $\boldsymbol{\infty} \in \mathbb{R}^D$ is a vector of ∞'s, etc. Unless otherwise specified, all operations are done componentwise: for instance, $\boldsymbol{a} \leq \boldsymbol{y}$ means $a_j \leq y_j$ for all $j = 1, \ldots, D$, $\boldsymbol{ay}$ is a vector with jth component $a_j y_j$, etc. Furthermore, $\boldsymbol{a} \not\leq \boldsymbol{y}$ indicates that there exists at least one $j = 1, \ldots, D$ such that $a_j > y_j$. If a comparison or an operation is done between a vector and a scalar, it holds for each component of the vector: $\boldsymbol{y} > a$ means that $y_j > a$, $j = 1, \ldots, D$; similarly, $a\boldsymbol{y}$ is a vector with components ay_j, etc. When sets are involved, $[\boldsymbol{a}, \boldsymbol{b}] \in \mathbb{R}^D$ is the product space $[a_1, b_1] \times \cdots \times [a_D, b_D]$ and $[-\infty, u)^D = [-\infty, u) \times \cdots \times [-\infty, u)$.

1.2.1 Componentwise maximum approach

1.2.1.1 Multivariate extreme value distributions and max-stability

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ be a D-dimensional random vector with marginals $F_1, \ldots, F_D$ and joint distribution F, and let $\{\boldsymbol{Y}_i\}_{i \geq 1}$ be an i.i.d. sequence of random vectors distributed as $\boldsymbol{Y}$. Denote by $Y_{i,j}$ the jth component of the vector $\boldsymbol{Y}_i$, and let $M_{n,j} = \max(Y_{1,j}, \ldots, Y_{n,j})$, $j = 1, \ldots, D$. For essentially arbitrary underlying joint distributions F, one aims at characterizing the family of possible asymptotic distributions for the vector of componentwise maxima $\boldsymbol{M}_n = (M_{n,1}, \ldots, M_{n,D})$, suitably renormalized. Since maxima may not occur at the same time in each margin, $\boldsymbol{M}_n$ does not always correspond to a real observation.

As in the univariate case, we consider an affine renormalization for $\boldsymbol{M}_n$ in order to get a non-trivial limit law as the sample size tends to infinity. Specifically, suppose that there exist sequences $\boldsymbol{a}_n = (a_{n,1}, \ldots, a_{n,D}) \in \mathbb{R}_+^D$ and $\boldsymbol{b}_n = (b_{n,1}, \ldots, b_{n,D}) \in \mathbb{R}^D$ such that the vector of renormalized componentwise maxima $\boldsymbol{M}_n^* = (\boldsymbol{M}_n - \boldsymbol{b}_n)/\boldsymbol{a}_n$ converges to a random variable with joint distribution G and non-degenerate margins $G_1, \ldots, G_D$. If such sequences can be found, the limiting distribution is called a *multivariate extreme-value distribution*. Balkema & Resnick (1977) showed that if convergence occurs, such a limiting distribution G has to be *max-infinitely divisible*.

Definition 29 (Max-infinite divisibility)**.** *A distribution G is max-infinitely divisible if for any $k \in \mathbb{N}$, $G^{1/k}$ is a distribution function. A random variable Z is said to be max-infinitely divisible if its distribution is max-infinitely divisible.*

A max-infinitely divisible distribution G yields a collection of "root" distributions $\{F_k\}_{k \in \mathbb{N}}$ such that $F_k^k = G$, for all $k \in \mathbb{N}$; in other words, G is the distribution function of the maximum of k independent random variates distributed according to F_k. Notice that all one-dimensional distributions are max-infinitely divisible.

Due to univariate extreme value theory and the extremal types theorem, see §1.1, one knows that if the margins G_j are non-degenerate, they have to be GEV, that is $M_{n,j}^* = (M_{n,j} - b_{n,j})/a_{n,j} \to Z_j \sim \mathrm{GEV}(\mu_j, \sigma_j, \xi_j)$, as $n \to \infty$, for any $j = 1, \ldots, D$. Or equivalently, the limiting margins are *max-stable*. Consequently, since the limiting joint distribution G is max-infinitely divisible with max-stable margins, it has to be max-stable itself (de Haan & Ferreira, 2006, §9.2), that is, for every $k \in \mathbb{N}$ and $\boldsymbol{z} = (z_1, \ldots, z_D) \in \mathbb{R}_+^D$, there must exist constants $\boldsymbol{a}_k \in \mathbb{R}_+^D$ and $\boldsymbol{b}_k \in \mathbb{R}^D$ such that

$$G^k(\boldsymbol{a}_k \boldsymbol{z} + \boldsymbol{b}_k) = G(\boldsymbol{z}). \tag{1.26}$$

By definition, a max-stable distribution is also max-infinitely divisible; the converse is

not true, because for some distributions G, one can find root distributions $\{F_k\}_{k\in\mathbb{N}}$ for which $F_k^k = G$ but G is not of the same type as F_k. The characterization of multivariate extreme value distributions therefore reduces to that of (multivariate) max-stable distributions with non-degenerate margins.

In practice, it is common for the study of multivariate extremes to proceed in two stages: The marginal distributions are typically estimated initially using the univariate methodology (by fitting the GEV distribution to maxima or the GPD to threshold exceedances), and then used with the probability integral transform to convert the data to a common scale, in order to handle the dependence structure using multivariate extreme value theory. For reasons of mathematical elegance, the transformation of the data is frequently to the unit Fréchet distribution, involving the maps $t_j(\cdot) = -1/\log\{G_j(\cdot)\}$, $j = 1,\ldots,D$. More precisely, if $\boldsymbol{Z} = (Z_1,\ldots,Z_D)$ has joint distribution G, then the transformed random variates $t_1(Z_1),\ldots,t_D(Z_D)$ are standard Fréchet with the same dependence structure: indeed, it holds that $\tilde{G}_j(z) = \Pr\{t_j(Z_j) \leq z\} = \exp(-1/z)$, for $z > 0$, and $G(z_1,\ldots,z_D) = \tilde{G}\{t_1(z_1),\ldots,t_D(z_D)\}$.

1.2.1.2 The exponent measure

Balkema & Resnick (1977) showed that max-infinitely divisible distributions yield a measure μ on $[-\infty,\infty)^D$ such that for all $\boldsymbol{z} \in \mathbb{R}^D$,

$$G(\boldsymbol{z}) = \exp\{-\mu(A_{\boldsymbol{z}})\} = \exp\{-V(\boldsymbol{z})\}, \tag{1.27}$$

where $A_{\boldsymbol{z}} = [-\infty,\boldsymbol{z}]^c$, with B^c denoting the complement of the set B. The *exponent measure* μ contains all the information about dependence among the variables $Z_1,\ldots,Z_D$. By abuse of language, the function $V = -\log G$ is also frequently referred to as the exponent measure. On the unit Fréchet scale, one can assume that the exponent measure $\tilde{\mu}$ is concentrated on $[0,\infty]^D \setminus \{\boldsymbol{0}\}$, so that $\tilde{G}(\boldsymbol{z}) = \exp\{-\tilde{\mu}(\tilde{A}_{\boldsymbol{z}})\} = \exp\{-\tilde{V}(\boldsymbol{z})\}$, where $\tilde{A}_{\boldsymbol{z}} = [0,\boldsymbol{z}]^c$ for $\boldsymbol{z} > 0$.

For simplicity, but without loss of generality, we shall restrict the discussion to the unit Fréchet case, where $\boldsymbol{a}_k = k^{-1}$ and $\boldsymbol{b}_k = 0$ in (1.26). Therefore, we write $G_j \equiv \tilde{G}_j$, $\mu \equiv \tilde{\mu}$, $A_{\boldsymbol{z}} \equiv \tilde{A}_{\boldsymbol{z}}$ and $V \equiv \tilde{V}$, dropping the tilde for simplicity.

Since the margins of G are assumed to be standard Fréchet, it can be verified that the exponent measure satisfies the constraint $V(z_1,\infty,\ldots,\infty) = 1/z_1$, and similarly for the other margins. Furthermore, owing to the max-stability property (1.26), the function V is homogeneous of order -1, that is $V(t\boldsymbol{z}) = t^{-1}V(\boldsymbol{z})$, for all $\boldsymbol{z} > 0$ and $t > 0$. The homogeneity of the exponent measure justifies, at least theoretically, extrapolation in the joint upper tail; specifically, suppose that estimation of the probability $p_{\mathscr{A}} =$

$\Pr(\boldsymbol{Z} \in \mathscr{A})$ is required for some extreme set $\mathscr{A}$ of the form $[0, \boldsymbol{z}]^c$. If $\boldsymbol{Z}$ follows an extreme value distribution, we can shrink $\mathscr{A}$ towards the origin by a factor $t \in (0,1)$, estimate $p_{t\mathscr{A}}$ instead (using more data points) and back-transform the estimated probability using the homogeneity property, noting that

$$p_{\mathscr{A}} = 1 - \exp\{-V(\boldsymbol{z})\} = 1 - \exp\{-tV(t\boldsymbol{z})\} = 1 - (1 - p_{t\mathscr{A}})^t.$$

1.2.1.3 Spectral representation for multivariate extreme value distributions

Theorem 30 gives a spectral representation of the exponent measure, thus characterizing all the possible limiting distributions for componentwise maxima under affine renormalization.

Theorem 30 (Characterization of multivariate extreme value distributions)**.** *If the renormalized vector $\boldsymbol{M}_n^* \xrightarrow{D} \boldsymbol{Z} \sim G$, where G is a non-degenerate distribution function, then G has the form*

$$G(\boldsymbol{z}) = \exp\{-V(\boldsymbol{z})\}, \quad \boldsymbol{z} > 0, \tag{1.28}$$

where

$$V(\boldsymbol{z}) = D \int_{S_D} \max(\boldsymbol{w}/\boldsymbol{z}) \, \mathrm{d}H(\boldsymbol{w}), \tag{1.29}$$

and $\mathrm{d}H$ is a measure on the $(D-1)$-dimensional simplex $S_D = \left\{\boldsymbol{w} \in \mathbb{R}_+^D : \sum_{i=1}^D w_j = 1\right\}$, satisfying the mean constraints $\int_{S_D} w_j \mathrm{d}H(\boldsymbol{w}) = 1/D$, $j = 1, \ldots, D$.

When $D = 2$, $\mathrm{d}H$ is a distribution on the interval $[0,1]$, subject to the constraint $\int_0^1 w \mathrm{d}H(w) = 1/2$. The measure $\mathrm{d}H$ is often called the *spectral measure* due to its interpretation in terms of the pseudo-radius $r = z_1 + \cdots + z_D$ and pseudo-angles $w_1 = z_1/r, \ldots, w_D = z_D/r$ (see below and Beirlant *et al.*, 2004, p.258). Contrary to the univariate case, where a parametric family of distributions covers all possible limits (recall Theorem 4), expression (1.29) implies that multivariate extreme value distributions cannot be fully described by a finite number of parameters. Indeed, since it is indexed by an essentially arbitrary spectral measure, each such suitable measure $\mathrm{d}H$ provides a valid multivariate extreme value distribution. So when we come to modeling and inference in practice, we need to rely on non-parametric techniques, or to have flexible parametric models at our disposal; see §1.2.1.5–1.2.1.6.

1.2.1.4 Pickands' dependence function

In dimension $D = 2$, an alternative representation of equation (1.29) (Pickands, 1981) leads to the so-called Pickands' dependence function, denoted by $A(w)$. It turns out

that we can rewrite the exponent measure as

$$V(z_1, z_2) = (z_1^{-1} + z_2^{-1}) A\left(\frac{z_1}{z_1 + z_2}\right),$$

where the function $A(w)$ satisfies $A(w) = 2\int_0^1 \max\{(1-w)q, w(1-q)\}\mathrm{d}H(q)$. Here, $A(w)$ is a function defined on the interval $[0,1]$, and is such that: i) $A(0) = A(1) = 1$, ii) $A(w)$ is convex and iii) $A(w)$ is contained in a triangular region defined by $\max(w, 1-w) \leq A(w) \leq 1$ for all $w \in [0,1]$; see Figure 1.5. The function $A(w)$ lies between the two bounding cases of complete dependence when $A(w) = \max(w, 1-w)$, and asymptotic independence when $A(w) \equiv 1$. The scalar $A(w)$ can be interpreted as a measure of the strength of dependence between Z_1 and Z_2 in the "direction" w, where $w = z_1/(z_1 + z_2)$ is the pseudo-angle between z_1 and z_2. The link between Pickands' dependence function $A(w)$ and the spectral measure $\mathrm{d}H$ is explained carefully in Beirlant *et al.* (2004, pp.268–270).

1.2.1.5 Parametric models

The specification of a parametric model for the exponent measure $V(\cdot)$ in (1.29), or equivalently for the spectral measure $\mathrm{d}H$, amounts to restricting the dependence to have a particular structure. It is therefore essential to build flexible, but parsimonious, dependence models that can also be readily interpreted. On the one hand, asymptotic independence arises if $V(\boldsymbol{z}) = z_1^{-1} + \cdots + z_D^{-1}$, or equivalently if $\mathrm{d}H(\boldsymbol{e}_j) = D^{-1}$ for each vertex $\boldsymbol{e}_j$ of the $(D-1)$-dimensional simplex S_D, since the distribution $G(\boldsymbol{z}) = \exp\{-V(\boldsymbol{z})\}$ factorizes. On the other hand, complete dependence is attained when $V(\boldsymbol{z}) = \max\left(\boldsymbol{z}^{-1}\right)$, or $\mathrm{d}H\left(\boldsymbol{D}^{-1}\right) = 1$, where $\boldsymbol{D}^{-1} = \left(D^{-1}, \ldots, D^{-1}\right)$. Many reasonable models lie between these two bounding cases; see below for popular bivariate examples. In high dimensions, however, flexible parametric models for extremes are difficult to build. Only a few have been proposed so far (see, e.g., Tawn, 1990; Coles & Tawn, 1991; Cooley *et al.*, 2010; Ballani & Schlather, 2011; Segers, 2012), but they usually suffer from a lack of flexibility for large D and lead to computational and inferential issues, discussed at the beginning of §1.2.1.6.

The most famous, though somehow rigid and simplistic, parametric model for $D = 2$ is the logistic model, due to Gumbel (1961):

$$V(z_1, z_2) = \left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha}, \quad z_1, z_2 > 0, \tag{1.30}$$

for some dependence parameter $\alpha \in (0,1]$. The limiting case $\alpha = 1$ corresponds to independence, whereas the case $\alpha \to 0$ corresponds to complete dependence. This model can readily be extended to higher dimensions; see §3.3.3.1. However, due to its

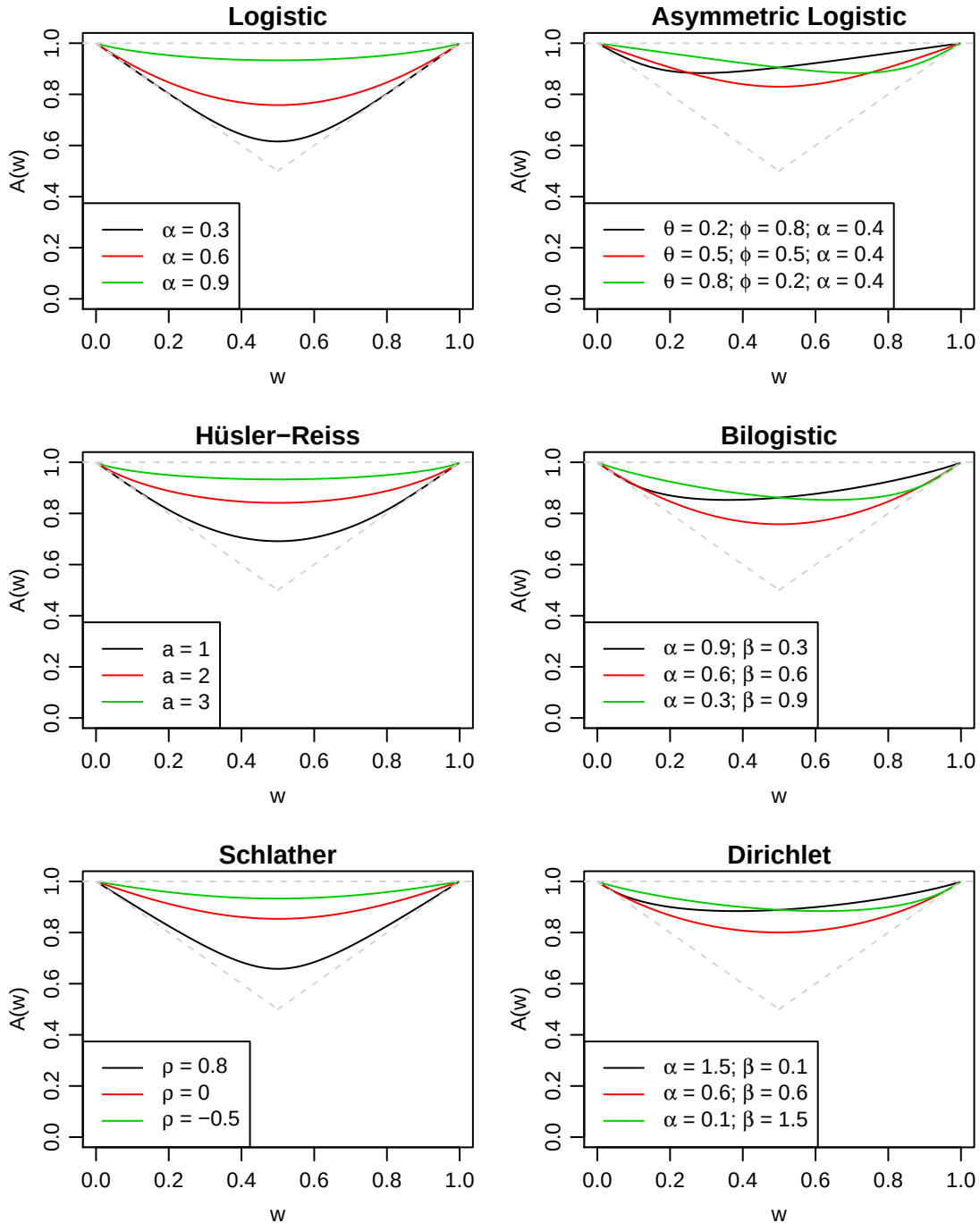

Figure 1.5: Pickands' dependence function for the logistic (top left), Hüsler–Reiss (middle left), Schlather (bottom left), asymmetric logistic (top right), bilogistic (middle right) and Dirichlet (bottom right) models, for different parameter values. Models in the left column are symmetric, whereas models in the right column allow for asymmetry.

symmetry, it often suffers from a lack of flexibility.

A more sophisticated model which can capture asymmetry in the dependence structure is the asymmetric logistic model proposed by Tawn (1988b), available also in dimensions $D \geq 3$ (Coles & Tawn, 1991; Stephenson, 2009):

$$V(z_1, z_2) = (1-\theta)/z_1 + (1-\phi)/z_2 + \left\{(\theta/z_1)^{1/\alpha} + (\phi/z_2)^{1/\alpha}\right\}^{\alpha}, \quad z_1, z_2 > 0, \tag{1.31}$$

for some dependence parameter $\alpha \in (0,1]$ and asymmetry parameters $\theta, \phi \in [0,1]$. When $\theta = \phi = 1$, this model boils down to the logistic model. It yielded the best fit among various bivariate extreme value models in a study by Ferrez *et al.* (2011), analyzing extreme temperatures under forest cover compared to an open field.

Another extension of the logistic model for which the variables are not exchangeable is the bilogistic model (Smith, 1990a):

$$V(z_1, z_2) = z_1^{-1} q^{1-\alpha} + z_2^{-1}(1-q)^{1-\beta}, \quad z_1, z_2 > 0.$$

Here, $q = q(z_1, z_2, \alpha, \beta)$ is the root of the equation $(1-\alpha)z_1^{-1}(1-q)^{\beta} - (1-\beta)z_2^{-1}q^{\alpha} = 0$, and $\alpha, \beta > 0$. The value of $|\alpha - \beta|$ determines the extent of asymmetry in the dependence structure. In particular, when $\alpha = \beta$, the bilogistic model reduces to the logistic model. Complete dependence is attained in the limit as $\alpha = \beta$ approaches zero.

Coles & Tawn (1991) proposed another asymmetric model, the Dirichlet model, for which

$$V(z_1, z_2) = \{1 - B_{\alpha+1;\beta}(t)\}/z_1 + B_{\alpha;\beta+1}(t)/z_2, \quad z_1, z_2 > 0,$$

where $\alpha, \beta > 0$ and $B_{a,b}(t)$ is the beta cumulative distribution function with shape parameters a and b, evaluated at $t = \alpha z_1/(\alpha z_1 + \beta z_2)$. The Dirichlet model is symmetric when $\alpha = \beta$.

Among symmetric models, the Hüsler–Reiss model (Hüsler & Reiss, 1989) based on the standard normal distribution has gained a lot of attention. This model is essentially the only possible limit of rescaled maxima of Gaussian variables, thus providing support for its use in many applications. Hüsler & Reiss (1989) proved that if $\{\boldsymbol{Y}_i\}_{i\geq 1}$ is a sequence of independent bivariate normal random variables with zero mean, unit variance and correlation ρ_n satisfying $4(1-\rho_n)\log n \to a^2$, as $n \to \infty$, then $\Pr(\boldsymbol{M}_n^* \leq \log \boldsymbol{z}) \to G(\boldsymbol{z}) = \exp\{-V(\boldsymbol{z})\}$, where $\boldsymbol{M}_n^* = (\boldsymbol{M}_n - \boldsymbol{b}_n)/\boldsymbol{a}_n$ is the vector of renormalized componentwise maxima, $\boldsymbol{a}_n = (2\log n)^{1/2}$, $\boldsymbol{b}_n = \{2\log n - \log\log n - \log(4\pi)\}^{1/2}$, and

$$V(z_1, z_2) = \frac{1}{z_1}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_1}{z_2}\right)\right\} + \frac{1}{z_2}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_2}{z_1}\right)\right\}, \tag{1.32}$$

where $\Phi(\cdot)$ denotes the normal cumulative distribution function. The parameter $a > 0$ is a dependence parameter. Independence occurs as $a \to \infty$, and complete dependence as $a \to 0$. Extensions to (1.32) can be obtained for the general case in dimension D (Hüsler & Reiss, 1989; Huser & Davison, 2013a).

Another symmetric model is the bivariate Schlather model for extremes (Schlather, 2002; Davison & Gholamrezaee, 2012):

$$V(z_1, z_2) = \frac{1}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)\left\{1 + \frac{1}{z_1 + z_2}\left(z_1^2 - 2z_1 z_2 \rho + z_2^2\right)^{1/2}\right\}. \tag{1.33}$$

The parameter $\rho \in [-1, 1]$ is a dependence parameter, interpretable as the correlation between original zero-truncated Gaussian events. Independence is obtained when $\rho = -1$, and complete dependence is reached when $\rho = 1$.

Extensions of the Hüsler–Reiss and Schlather models to the infinite-dimensional framework play a prominent role in the modeling of spatial extremes; see §2.3.2.

Figure 1.5 depicts the Pickands' dependence function of the aforementioned models for $D = 2$ and typical parameter values. Other relevant models, not presented here, have also been proposed in the literature, and include the negative logistic model (Galambos, 1975), the asymmetric negative logistic model (Joe, 1990), and the mixed and asymmetric mixed models (Tawn, 1988b).

1.2.1.6 Inference

For simplicity, assume that the data can be decomposed into N blocks of M independent D-variate observations, that is $n = MN$, with $M, N \in \mathbb{N}$. One can form N componentwise block maxima $\boldsymbol{z_i} = (z_{1;i}, \ldots, z_{D;i})$, $i = 1, \ldots, N$, and fit an extreme value model to them, either parametrically (for example by maximum likelihood) or non-parametrically.

The main assumption behind the fitting procedures described below is that the *asymptotic* extreme value model is reasonable for finite n. However, investigations in the univariate case show that the slow convergence to the limiting distribution can induce substantial bias if an extreme value distribution is fitted at subasymptotic levels (Smith, 1982; Balkema & de Haan, 1990; Beirlant *et al.*, 2004, p.289); and when $D > 1$, the uniform rate of convergence is bounded by the individual marginal rates of convergence and by the rate of convergence of the dependence function (Omey & Rachev, 1991). When $D \gg 1$, these convergence issues may be even more pronounced, and could entail potential misestimation of return levels and probabilities of simultaneous

extremes. Furthermore, if extreme events are independent in the limit, it is of crucial importance to capture the rate at which extremal dependence vanishes, and so-called asymptotic independent models may be more appropriate; see §1.2.4. In summary, to be comfortable with the use of an (asymptotic) extreme value model for finite n, careful model checking and uncertainty assessment are needed.

Maximum likelihood estimation. The first step is to choose a suitable parametric model for $G = \exp(-V)$ in our "dictionary" of existing extreme value distributions, recall §1.2.1.5. Since the data are scarce by nature, the subfamily of extreme value models chosen needs to be well-balanced between simplicity and flexibility in order to capture a wide variety of dependence structures. In practice, the distribution G depends on a finite number of unknown parameters, summarized by the vector ψ, which must be estimated from the data. The likelihood function is

$$L(\psi) = \prod_{i=1}^{N} g(\boldsymbol{z}_i), \tag{1.34}$$

where $g(\boldsymbol{z}) = \partial^D G(\boldsymbol{z})/(\partial z_1 \cdots \partial z_D)$, and the parameter estimates $\widehat{\psi}$ are obtained by maximizing $L(\psi)$ with respect to ψ. If the chosen model is correct, the variance of $\widehat{\psi}$ can be estimated by the inverse of the Fisher information matrix, that is $\widehat{\mathrm{var}}(\widehat{\psi}) = \left\{-\sum_{i=1}^{N} \partial^2 \log g(\boldsymbol{z}_i)/\left(\partial\psi\partial\psi^T\right)\right\}^{-1}_{\psi=\widehat{\psi}}$. In dimension $D = 2$, the joint density can be expressed as $g(z_1, z_2) = (V_1 V_2 - V_{12})\exp(-V)$, where for simplicity we have written $V_1 = \partial V(z_1, z_2)/\partial z_1$, and so forth. However, the likelihood quickly becomes intractable as D increases. Indeed, the number of terms involved in the joint density $g(\boldsymbol{z})$ equals the Bell number B_D, that is, the number of partitions of the set $\{1, \ldots, D\}$. For example, the number of terms in the likelihood is $115'975$ when $D = 10$, and is around 10^{47} when $D = 50$. Hence, alternative methods need to be found for inference in high dimensions. A possibility is to consider a surrogate for the full likelihood such as, for instance, pairwise likelihood (Lindsay, 1988). Under mild conditions, maximum pairwise likelihood estimators turn out to be asymptotically normal and strongly consistent (Varin, 2008). Provided the parameter ψ is identifiable from the lower-order marginal densities involved, this solution seems reliable and has the advantage that bivariate margins only need to be specified. Furthermore, pairwise likelihood inference is robust against misspecification of higher-order interactions, though with a loss in efficiency compared to maximum likelihood. For more details about composite likelihoods and their application to the spatial framework, see Section 2.6 and Chapters 3–4. Another elegant possibility is to use the information, if available, on the occurrence times of extreme events. Stephenson & Tawn (2005) show that if z_1 and z_2 denote two simultaneous extreme events, their density is $g(z_1, z_2) = -V_{12}\exp(-V)$, and if they did not occur together, $g(z_1, z_2) = V_1 V_2 \exp(-V)$, which yields the Stephenson–Tawn

likelihood

$$L(\psi) = \prod_{i=1}^{N} \left[\left\{ V_1(z_{1;i}, z_{2;i}) V_2(z_{1;i}, z_{2;i}) \right\}^{1-s_i} - \left\{ V_{12}(z_{1;i}, z_{2;i}) \right\}^{s_i} \right] \exp\{-V(z_{1;i}, z_{2;i})\}, \tag{1.35}$$

where $s_i = I(\text{Maxima } z_{1;i} \text{ and } z_{2;i} \text{ occurred together})$ and $I(\cdot)$ is the indicator function. In dimension D, Wadsworth & Tawn (2013) explain how we can greatly benefit from this fact to reduce the number of terms in the likelihood from B_D to one, see (4.10).

Non-parametric approaches. Nonparametric estimation of multivariate extreme value models is natural because of the functional representation in (1.29). Several non-parametric estimators for the Pickands' dependence $A(w)$ function have been proposed in the literature. The estimator proposed by Pickands (1981) is the simplest. Suppose that $\boldsymbol{z}_i = (z_{1;i}, z_{2;i})$, $i = 1, \ldots, N$, follow a bivariate extreme value distribution. Then Pickands' estimator

$$\widehat{A}^P(w) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \min\left(\frac{z_{1;i}}{1-w}, \frac{z_{2;i}}{w} \right) \right\}^{-1}$$

is consistent but does not yield admissible dependence functions for finite N, in the sense that the estimate is not convex, has an end-point bias, and does not lie in its triangular domain in general. Improved estimators have been proposed by Deheuvels & Tiago de Oliveira (1989), Deheuvels (1991), Smith *et al.* (1990), Capéraà *et al.* (1997) and Hall & Tajvidi (2000).

1.2.2 Point process approach

1.2.2.1 Basic results and connection to componentwise maxima

An alternative approach to modeling multivariate extremes is based on a point process representation, similar to §1.1.2 in the univariate case. When original events are available, more information can be used, providing an elegant solution to the recurrent problem of the waste of data of the block maxima approach. The theory is well summarized in Coles & Tawn (1991) and detailed in Resnick (1987), Beirlant *et al.* (2004) and Fougères (2004). The main result is a straightforward extension of the univariate case, stated in Theorem 25. Again, the limiting point process of exceedances turns out to be a Poisson process.

Mathematically speaking, suppose that $\{\boldsymbol{Y}_i\}_{i \geq 1}$ is a sequence of i.i.d. random vectors on $\mathbb{R}^D$, and let $Y_{i,j}$ be the jth component of the vector $\boldsymbol{Y}_i$. Without loss of generality, assume that $\boldsymbol{Y}_i$ has unit Fréchet margins, that is $\Pr(Y_{i,j} \leq y) = \exp(-1/y)$, $y > 0$,

$i \geq 1$, $j = 1, \ldots, D$. Consider the point process $N_n = \sum_{i=1}^{n} \delta_{P_n}(\cdot)$ defined on $\mathbb{R}_+^D$, where $P_n = \{\boldsymbol{Y}_i / n\}_{i=1}^n \in \mathbb{R}^D$ and $\delta_p(\cdot)$ is the Dirac measure defined in (1.9). The scaling factor n corresponds to the normalizing constants $\boldsymbol{a}_n = n$, $\boldsymbol{b}_n = 0$ in (1.26) for unit Fréchet random variables. The following theorem holds.

Theorem 31 (Convergence of the point process of exceedances, Resnick, 1987, p.154)**.** *Suppose that the assumption of Theorem 30 holds, and that μ is the exponent measure of the limiting multivariate extreme value distribution; recall (1.27). Then, N_n converges to a non-homogeneous Poisson process on $\mathbb{R}_+^D \setminus \{\mathbf{0}\}$ with mean measure μ.*

By the Poisson property, we have, for any sets A bounded away from the origin,

$$\Pr\left(P_n \subset A^c\right) = \Pr(P_n \text{ has no point in } A) \to \exp\{-\mu(A)\}, \qquad n \to \infty. \tag{1.36}$$

Therefore, letting $A_{\boldsymbol{z}} = [0, \boldsymbol{z}]^c$, and $\boldsymbol{M}_n^* = \left(M_{n,1}^*, \ldots, M_{n,D}^*\right)$ be the vector of renormalized componentwise maxima defined in §1.2.1.1, we have that $\Pr(\boldsymbol{M}_n^* \leq \boldsymbol{z}) \to \exp\{-\mu(A_{\boldsymbol{z}})\} = \exp\{-V(\boldsymbol{z})\}$, as $n \to \infty$. Hence, Theorem 31 is consistent with the result on maxima. In fact, the correspondance between Theorems 30 and 31 goes much further. One can show (Proposition 5.17 of Resnick, 1987; Beirlant *et al.*, 2004, p.280, equation 8.76) that a vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ is in the max-domain of attraction of the multivariate extreme value distribution G with exponent measure μ if and only if

$$\mu_n(B) = n\Pr(n^{-1}\boldsymbol{Y} \in B) \to \mu(B), \tag{1.37}$$

as $n \to \infty$, where B is any Borel set in $[0, \infty]^D \setminus \{\mathbf{0}\}$ with compact closure and such that $\mu(\partial B) = 0$, where ∂B denotes the topological boundary of B; in other words, the measure μ_n converges vaguely to μ on $[0, \infty]^D \setminus \{\mathbf{0}\}$, or $\mu_n \xrightarrow{v} \mu$.

The spectral decomposition of multivariate extremes through pseudo-radial and pseudo-angular components turns out to be useful. Let us transform the data $\{\boldsymbol{Y}_i\}_{i=1}^n$ into pseudo-polar coordinates with radius $r_i \in \mathbb{R}_+$ and angle $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,D}) \in \mathbb{R}_+^D$, $i = 1, \ldots, n$, as follows:

$$r_i = \sum_{j=1}^{D} \frac{Y_{i,j}}{n}, \qquad w_{i,j} = \frac{Y_{i,j}}{n r_i}, \quad j = 1, \ldots, D. \tag{1.38}$$

Clearly, $\sum_{j=1}^{D} w_{i,j} = 1$ for all $i = 1, \ldots, n$. With this parametrization, the intensity measure $\mathrm{d}\mu$ of the limiting Poisson process, if it exists, factorizes as

$$\mathrm{d}\mu(r, \boldsymbol{w}) = D r^{-2} \mathrm{d}r \, \mathrm{d}H(\boldsymbol{w}), \tag{1.39}$$

where $\mathrm{d}H$ is the spectral measure, introduced in (1.29), which describes the depen-

dence structure of extreme observations. Expression (1.39) implies that the radial and the angular components are asymptotically independent. Loosely speaking, the relative magnitude of extreme events is independent of the magnitude itself. This factorization underpins extrapolation to high levels, since it implies that the angular component can be estimated from observed data, and then used to extrapolate beyond the observations, through the regularity of the radial component. Probabilities of various extreme subspaces can be deduced from property (1.36) combined with (1.39); see the following examples.

Example 32 (Consistency with maxima)**.** *Letting* $A_{\boldsymbol{z}} = [0, \boldsymbol{z}]^c$*, we have*

$$\mu(A_z) = \int_{S_D} \int_{\min\{\boldsymbol{z}/\boldsymbol{w}\}}^{\infty} D \frac{\mathrm{d}r}{r^2} \mathrm{d}H(\boldsymbol{w}) = D \int_{S_D} \max(\boldsymbol{w}/\boldsymbol{z}) \, \mathrm{d}H(\boldsymbol{w}),$$

recovering expression (1.29).

Example 33. *Letting* $\boldsymbol{r} = (r_1, \ldots, r_D) \in \mathbb{R}_+^D$ *and* $A_{\boldsymbol{r}} = \{\boldsymbol{z} \in \mathbb{R}_+^D : \sum_{j=1}^D z_j/r_j > 1\}$*, we have*

$$\mu(A_{\boldsymbol{r}}) = D \int_{s_D} \int_{\left(\sum_{j=1}^D w_j/r_j\right)^{-1}}^{\infty} \frac{\mathrm{d}r}{r^2} \mathrm{d}H(w) = D \int_{s_D} \left(\sum_{j=1}^D w_j/r_j \right) \mathrm{d}H(w) = \sum_{j=1}^D r_j^{-1},$$

which does not depend on the spectral measure $\mathrm{d}H$*. In particular, considering the set* $A_{r_0} = \{\boldsymbol{z} \in \mathbb{R}_+^D : z_1 + \cdots + z_D > r_0\}$ *for some* $r_0 > 0$*, we have*

$$\mu(A_{r_0}) = D/r_0. \tag{1.40}$$

1.2.2.2 Inference methods

We describe four different parametric threshold-based inference methods, some of which rely explicitly on the asymptotic Poisson model stated in Theorem 31. For simplicity, the exposition is for known margins, but it can readily be extended to unknown margins. Suppose the observations $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,D}) \in \mathbb{R}_+^D$, $i = 1, \ldots, n$, are i.i.d. and distributed as $\boldsymbol{Y} = (Y_1, \ldots, Y_D) \sim F$, with unit Fréchet margins. Furthermore, assume that the joint distribution F is in the max-domain of attraction of a multivariate extreme value distribution G, that is $F^n(n\boldsymbol{y}) \to G(\boldsymbol{y})$, as $n \to \infty$, and that $G(\boldsymbol{y})$ depends on the unknown parameter vector ψ. The first approach, developed by Coles & Tawn (1991), consists in choosing high marginal thresholds $\boldsymbol{u} = (u_1, \ldots, u_D)$ and maximizing a likelihood built from the Poisson approximation for extremes on the subspace $[0, \boldsymbol{u}]^c$. The second approach is similar in spirit but relies on another extreme subspace, and uses the spectral decomposition of the exponent measure. This method corresponds in practice to setting a high "diagonal" threshold r_0 and considering as extreme all points whose radial component exceeds r_0. The third approach is essentially a gen-

eralization of the GPD approach in (1.23) to the multivariate case, where a point is considered as extreme if at least one of its component exceeds a high marginal threshold. Technical details can be found in Rootzén & Tajvidi (2006), Buishand *et al.* (2008) and Beirlant *et al.* (2004, pp.277–278). Finally, the fourth approach goes back to Smith (1993) and was applied among others by Ledford & Tawn (1996), Smith *et al.* (1997), Bortot *et al.* (2000), Coles (2001) and Wadsworth & Tawn (2012). It consists in maximizing a censored likelihood, which relies on a penultimate approximation of the limiting mean measure μ of the point process of exceedances. These apparently quite different methods are in fact closely related, and we hope to make the links clearer in the sequel.

Poisson likelihood based on marginal thresholds. Let $\boldsymbol{u} = (u_1, \ldots, u_D) > 0$ be high marginal thresholds and $A_{\boldsymbol{u}}$ be the extreme set $A_{\boldsymbol{u}} = [0, \boldsymbol{u}]^c \subset \mathbb{R}^D$. Furthermore, denote by $N_{\boldsymbol{u}}$ the number of points in $A_{\boldsymbol{u}}$ and let $\boldsymbol{y}^i = (y_1^i, \ldots, y_D^i)$, $i = 1, \ldots, N_{\boldsymbol{u}}$ be these (extreme) points. In virtue of Theorem 31, if the marginal thresholds $u_1, \ldots, u_D$ are large enough, the extreme points scaled by n should be approximately distributed according to a Poisson process with intensity $\mathrm{d}\mu$. Hence, the likelihood over the region $A_{\boldsymbol{u}}$ based on the limiting Poisson process is

$$\begin{aligned} L(\psi) &= \exp\{-\mu(A_{n^{-1}\boldsymbol{u}})\} = i = 1^{N_{\boldsymbol{u}}} \mathrm{d}\mu\left(n^{-1}\boldsymbol{y}^i\right) \propto \exp\left\{-V\left(n^{-1}\boldsymbol{u}\right)\right\} \prod_{i=1}^{N_{\boldsymbol{u}}} -V_{1:D}(\boldsymbol{y}^i) \\ &\propto \exp\{-nV(\boldsymbol{u})\} \prod_{i=1}^{N_{\boldsymbol{u}}} -V_{1:D}(\boldsymbol{w}^i) \propto \exp\{-nV(\boldsymbol{u})\} \prod_{i=1}^{N_{\boldsymbol{u}}} \mathrm{d}H(\boldsymbol{w}^i), \end{aligned} \tag{1.41}$$

where $\boldsymbol{w}^i = \boldsymbol{y}^i \left(\sum_{j=1}^D y_j^i\right)^{-1}$ is the pseudo-angle of the point $\boldsymbol{y}^i$, $\mathrm{d}H$ is the spectral measure and $V_{1:D} = \partial^D V(\boldsymbol{y}) / (\partial y_1 \cdots \partial y_D) = -\mathrm{d}\mu(\boldsymbol{y})$. The proportionality signs above are justified since V is homogenous of order -1, and so $V_{1:D}$ is homogenous of order $-(D+1)$. Using this approach, only the observations for which at least one component is extreme contribute a density component to the likelihood. The observations in the subspace $[0, \boldsymbol{u}] \subset \mathbb{R}^D$ enter in the likelihood through the term $\exp\{-n^{-1}V(\boldsymbol{u})\}$ only. The more general case with unknown margins was developed, discussed and applied successfully to oceanographic data in Coles & Tawn (1991).

Poisson likelihood based on a diagonal threshold. This estimation procedure is similar in spirit to the first method, based on exceedances of high marginal thresholds, but another extreme subspace is used here. Let $A_{r_0} = \{\boldsymbol{y} \in \mathbb{R}_+^D : y_1 + \cdots + y_D > r_0\}$, where $r_0 > 0$ is a high threshold. As before, let the N_{r_0} extreme points lying in A_{r_0} be denoted by $\boldsymbol{y}^i = (y_1^i, \ldots, y_D^i)$, $i = 1, \ldots, N_{r_0}$. Owing to the spectral representation of the exponent measure, the likelihood based on the approximate Poisson process over the

region A_{r_0} is

$$L(\psi) = \exp\{-\mu(A_{n^{-1}r_0})\} \prod_{i=1}^{N_{r_0}} \mathrm{d}\mu\left(n^{-1}\boldsymbol{y}^i\right) \propto \prod_{i=1}^{N_{r_0}} -V_{1:D}(\boldsymbol{w}^i) \propto \prod_{i=1}^{N_{r_0}} \mathrm{d}H(\boldsymbol{w}^i), \tag{1.42}$$

where $\boldsymbol{w}^i$, $V_{1:D}$ and $\mathrm{d}H$ are defined as above. The first proportionality sign in (1.42) is justified by Example 33, using (1.40), and releases us from the need to compute the factor $\exp\{-\mu(A_{n^{-1}r_0})\}$. This approach can also easily be generalized to extreme sets of the form $A_{\boldsymbol{r}} = \{\boldsymbol{y} \in \mathbb{R}_+^D : y_1/r_1 + \cdots + y_D/r_D > 1\}$ for some $\boldsymbol{r} = (r_1, \ldots, r_D) > 0$.

Multivariate GPD approach. This approach find its roots in the PhD thesis of Nader Tajvidi (Tajvidi, 1996), followed by the work of Falk & Reiss (2001, 2002, 2003a,b, 2005), Rootzén & Tajvidi (2006) and Buishand *et al.* (2008). Since the joint distribution F of the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ is assumed to be in the max-domain of attraction of the distribution $G(\boldsymbol{y}) = \exp\{-V(\boldsymbol{y})\}$, one obtains after some calculations that for $\boldsymbol{y}, \boldsymbol{u} > 0$,

$$\Pr\{n^{-1}\boldsymbol{Y} \leq \boldsymbol{y} \mid n^{-1}\boldsymbol{Y} \not\leq \boldsymbol{u}\} \to \frac{1}{-\log G(\boldsymbol{u})} \log\left\{\frac{G(\boldsymbol{y})}{G(\boldsymbol{y} \wedge \boldsymbol{u})}\right\}, \tag{1.43}$$

as $n \to \infty$, where $a \wedge b = \min(a, b)$, and $\boldsymbol{y} \not\leq \boldsymbol{u}$ means that there exists at least one $j = 1, \ldots, D$ such that $y_j > u_j$. This result gives rise to the class of multivariate generalized Pareto distributions (mGPD).

Definition 34 (mGPD class)**.** *A distribution function H is a multivariate generalized Pareto distribution (mGPD) with reference vector $\boldsymbol{u} = (u_1, \ldots, u_D)$ if*

$$H(\boldsymbol{y}) = \frac{1}{-\log G(\boldsymbol{u})} \log\left\{\frac{G(\boldsymbol{y})}{G(\boldsymbol{y} \wedge \boldsymbol{u})}\right\} \tag{1.44}$$

for some multivariate extreme value distribution G with non-degenerate margins and with $0 < G(\boldsymbol{u}) < 1$. If G has unit Fréchet margins, that is $G(\boldsymbol{y}) = \exp\{-V(\boldsymbol{y})\}$ where V is the underlying exponent measure of G, one has

$$H(\boldsymbol{y}) = \frac{1}{V(\boldsymbol{u})} \left\{V(\boldsymbol{y} \wedge \boldsymbol{u}) - V(\boldsymbol{y})\right\}. \tag{1.45}$$

and its density, if it exists, has the form $h(\boldsymbol{y}) = -V_{1:D}(\boldsymbol{y})/V(\boldsymbol{u})$, $\boldsymbol{y} \not\leq \boldsymbol{u}$, where $V_{1:D} = \partial^D V(\boldsymbol{y})/(\partial y_1 \cdots \partial y_D)$,

This definition is given slightly differently in Rootzén & Tajvidi (2006) or Beirlant *et al.* (2004). Contemplating the right-hand side of equation (1.43), one can see that it corresponds to a multivariate generalized Pareto distribution, with reference vector $\boldsymbol{u} = (u_1, \ldots, u_D)$. Hence, this result states that the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$,

conditioned on being large in at least one component, is approximately multivariate generalized Pareto distributed, and this can be easily generalized to other marginal distributions (see, e.g., Beirlant *et al.*, 2004, p.277). The result (1.43) suggests a likelihood constructed from the points that are extreme in at least one component. Let $\boldsymbol{u} = (u_1, \ldots, u_D) > 0$ denote high marginal thresholds and consider the points $\boldsymbol{y}^i = (y_1^i, \ldots, y_D^i) \in [0, \boldsymbol{u}]^c$, $i = 1, \ldots, N_{\boldsymbol{u}}$ as extreme, where $N_{\boldsymbol{u}}$ denotes their total number. The likelihood based on the limit (1.43) is

$$L(\psi) = \prod_{i=1}^{N_{\boldsymbol{u}}} \left\{ -\frac{V_{1:D}(n^{-1}\boldsymbol{y}^i)}{V(n^{-1}\boldsymbol{u})} \right\} \propto \prod_{i=1}^{N_{\boldsymbol{u}}} \left\{ -\frac{V_{1:D}(\boldsymbol{y}^i)}{V(\boldsymbol{u})} \right\}, \tag{1.46}$$

where $V_{1:D}$ is defined as above. This likelihood is closely linked to the likelihood (1.41). By manipulating these two likelihoods, one discovers that the log-likelihood from (1.41), $\ell_1(\psi)$ say, can be decomposed into three terms: $\ell_1(\psi) = K + \ell_{N_{\boldsymbol{u}}}(\psi) + \ell_2(\psi)$. The first term K is a positive constant, which does not have any influence on the fit; the middle term $\ell_{N_{\boldsymbol{u}}}(\psi)$ is the log-likelihood based solely on the Poisson-distributed number of exceedances $N_{\boldsymbol{u}}$; and the last term $\ell_2(\psi)$ is the multivariate GPD log-likelihood from (1.46), which is conditioned on the number of exceedances $N_{\boldsymbol{u}}$. In terms of the corresponding Fisher information quantities, one has $i_1(\psi) = i_{N_{\boldsymbol{u}}}(\psi) + i_2(\psi) > i_2(\psi)$. In other words, the first approach, based on the Poisson approximation for exceedances, is asymptotically more efficient than the multivariate GPD approach. This is due to the fact that the Poisson process approach uses part of the information contained in $[0, \boldsymbol{u}]$ (in fact how many points there are), while the multivariate GPD approach throws it away. Simulations below confirm this, but reveal that the difference in efficiency is slight.

Censored likelihood approach. According to equation (1.37), the joint distribution F is in the max-domain of attraction of the multivariate extreme value distribution G if and only if $\mu_n \xrightarrow{v} \mu$, as $n \to \infty$, where $\mu_n(B) = n\Pr\{n^{-1}\boldsymbol{Y} \in B\}$ and $B \subset [0,\infty]^D \setminus \{\boldsymbol{0}\}$ is a Borel set. In particular, taking $B = [0, \boldsymbol{y}]^c$, we have that

$$n\{1 - F(n\boldsymbol{y})\} \to -\log G(\boldsymbol{y}) = V(\boldsymbol{y}), \qquad n \to \infty.$$

Hence, for large n, we have the approximation $F(n\boldsymbol{y}) \approx 1 - V(\boldsymbol{y})/n$, that is $F(\tilde{\boldsymbol{y}}) \approx 1 - V(\tilde{\boldsymbol{y}})$ for $\tilde{\boldsymbol{y}} = n\boldsymbol{y}$, by the homogeneity of V. Choosing high marginal thresholds $\boldsymbol{u} = (u_1, \ldots, u_D) > 0$, one can therefore approximate the right joint tail of F as

$$F(\boldsymbol{y}) \approx 1 - V(\boldsymbol{y}) \approx \exp\{-V(\boldsymbol{y})\} = G(\boldsymbol{y}), \qquad \boldsymbol{y} > \boldsymbol{u}. \tag{1.47}$$

The second approximation in (1.47) is justified by the first-order Taylor expansion $\exp(-y) \approx 1 - y$, for $y \approx 0$. This means that the joint distribution of large observations

is approximately the same as that of maxima. This approximation is valid for large $\boldsymbol{y}$ (above high marginal thresholds), and thus motivates a censored approach. The idea is to censor observations that are below the threshold. For $i \geq 1$, denote by $\boldsymbol{\delta}_i = (\delta_{i,1}, \ldots, \delta_{i,D}) \in \{0,1\}^D$ the indicator variables reporting whether $\boldsymbol{y}_i > \boldsymbol{u}$, that is $\delta_{i,j} = 1$ if and only if $y_{i,j} > u_j$. Each vector $\boldsymbol{y}_i$ can then be split apart into a vector of exceedances, $\boldsymbol{y}_i^{>}$, and a vector of non-exceedances, $\boldsymbol{y}_i^{\leq}$. Specifically, the vector $\boldsymbol{y}_i^{>}$ (of dimension less than or equal to D) contains the elements of $\boldsymbol{y}_i$ for which $\delta_{i,j} = 1$, and the elements of the vector $\boldsymbol{y}_i^{\leq}$ correspond to $\delta_{i,j} = 0$. The censoring scheme that we consider supposes that the non-exceedances, $\boldsymbol{y}_i^{\leq}$, are not observed, but that we always know when an exceedance occurs. Hence, the censored set of observations is composed of $\left(\boldsymbol{\delta}_i, \boldsymbol{y}_i^{>}\right)$, $i = 1, \ldots, n$.

In order to derive the likelihood of such a dataset, we further define the vectors $\boldsymbol{u}_i^{>}$ and $\boldsymbol{u}_i^{\leq}$ containing the elements of the threshold vector $\boldsymbol{u}$ for which $\delta_{i,j} = 1$ and $\delta_{i,j} = 0$ respectively. Moreover, assume for simplicity that $\boldsymbol{Y}$ has a density f, and denote by $F_{\boldsymbol{\delta}}$ partial differentiation of the distribution F with respect the variables corresponding to $\boldsymbol{\delta} = 1$. Then, for each $i = 1, \ldots, n$, the contribution to the likelihood of a censored observation $\left(\boldsymbol{\delta}_i, \boldsymbol{y}_i^{>}\right)$ is

$$p_{\boldsymbol{u}}(\boldsymbol{y}_i) = \int_{\boldsymbol{0}}^{\boldsymbol{u}_i^{\leq}} f(\boldsymbol{y}_i) \mathrm{d}\boldsymbol{y}_i^{\leq} = F_{\boldsymbol{\delta}_i}(\boldsymbol{c}_i), \tag{1.48}$$

where the vector $\boldsymbol{c}_i = (c_{i,1}, \ldots, c_{i,D})$ has components

$$c_{i,j} = \max(y_{i,j}, u_j) = \begin{cases} y_{i,j}, & y_{i,j} > u_j, \\ u_j, & y_{i,j} \leq u_j. \end{cases} \tag{1.49}$$

Thanks to the approximation (1.47), the distribution $F(\boldsymbol{y})$ can be replaced either by $1 - V(\boldsymbol{y})$ or by the extreme value model $\exp\{-V(\boldsymbol{y})\}$, and a likelihood can be constructed by multiplying all censored contributions:

$$L(\psi) = \prod_{i=1}^{n} p_{\boldsymbol{u}}(\boldsymbol{y}_i). \tag{1.50}$$

This approach has been applied in the bivariate case by several authors (Ledford & Tawn, 1996; Smith *et al.*, 1997; Bortot *et al.*, 2000; Coles, 2001, p.155), and more recently by Wadsworth & Tawn (2012), Huser & Davison (2013b) and Thibaud *et al.* (2013) in the spatial context using a censored pairwise likelihood. For more details about the use of censored pairwise likelihoods in spatial applications, see Chapters 2 and 5. However, when the censored approach is applied in higher dimensions, inference problems may arise in practice in case of low to moderate dependence,

because the probability that an observed D-uplet falls into the subspace $[u,\infty)^D$, so-called "upper right D-dimensional quadrant", decays geometrically with D when the data are independent. A major difference of this approach, with respect to the aforementioned point process-based and mGPD methods, is that the fitting is only based on the full information of points in $[\boldsymbol{u},\infty)$, rather than all those in $[0,\boldsymbol{u}]^c$. In all previous approaches, a point close to the axes also provides full (rather than censored) information, even though it may be extreme in only one component. The censored likelihood approach might therefore suffer from a loss in efficiency, but surely benefits from a gain in robustness against misspecification of the model below the marginal thresholds.

In order to assess the efficiency and robustness properties of the different methods, we conducted a simulation study in dimension 2. To our knowledge, this is the first comparative study of most estimators of reference for bivariate extremes. For $\alpha = 0.1,\ldots,0.9$, we simulated $R = 300$ bivariate datasets of size $n = 10'000$ from an Archimedean copula with generator $\varphi(t) = (t^{-1}-1)^{1/\alpha}$ (see (1.57)), where the margins were transformed to the unit Fréchet scale. In other words, the joint distribution function $F(y_1,y_2)$ of our simulated observations is

$$F(y_1,y_2) = \varphi\left[\varphi^{-1}\left\{\exp\left(-1/y_1\right)\right\} + \varphi^{-1}\left\{\exp\left(-1/y_2\right)\right\}\right], \qquad y_1,y_2 > 0. \tag{1.51}$$

This distribution is known to be in the max-domain of attraction of the logistic model (1.30) with parameter α (Fougères, 2004; Nelsen, 2006), that is $F^n(ny_1,ny_2) \to \exp\{-V(y_1,y_2)\}$ with $V(y_1,y_2) = (y_1^{-1/\alpha}+y_2^{-1/\alpha})^\alpha$, $\alpha \in (0,1]$. Let $\boldsymbol{u}(p) = (u_1(p),u_2(p))$ denote the empirical $p-$quantiles of our observations. For $p = 0.9, 0.95, 0.98$, we then estimated the dependence parameter α of the asymptotic logistic model with the following threshold-based estimators:

- $\widehat{\alpha}_1$: Maximizes the Poisson likelihood with diagonal threshold, see (1.42), based on the extreme set $A(p) = \{(y_1,y_2) \in \mathbb{R}_+^2 \,:\, y_1/u_1(p) + y_2/u_2(p) > K\}$, with $K = 1$.
- $\widehat{\alpha}_2$: Similar to $\widehat{\alpha}_1$, but with $K = 2$.
- $\widehat{\alpha}_3$: Maximizes the Poisson likelihood (1.41), with marginal thresholds $\boldsymbol{u}(p)$.
- $\widehat{\alpha}_4$: Maximizes the multivariate GPD likelihood (1.46), with marginal thresholds $\boldsymbol{u}(p)$. According to the discussion above, this estimator should be slightly less efficient than $\widehat{\alpha}_3$.
- $\widehat{\alpha}_5$: Maximizes the censored likelihood in (1.50), replacing $F(y_1,y_2)$ by $1-V(y_1,y_2)$, with marginal thresholds $\boldsymbol{u}(p)$. Since part of the information is censored, this estimator should be less efficient than $\widehat{\alpha}_3$.

- $\widehat{\alpha}_6$: Maximizes a conditional version of the censored likelihood (1.50) with $F(y_1, y_2)$ replaced by $1 - V(y_1, y_2)$. Conditional on the event "$\boldsymbol{Y}_i \in [0, \boldsymbol{u}(p)]^c$", the censored contributions (1.48) are modified as $p_{\boldsymbol{u}(p)}(\boldsymbol{y}_i) = F_{\boldsymbol{\delta}_i}(\boldsymbol{y}_i)/[1 - F\{\boldsymbol{u}(p)\}]$. $\widehat{\alpha}_6$ should be slightly less efficient than the estimator $\widehat{\alpha}_5$ since the data in the "bottom left quadrant" are dropped, and also less efficient than the estimators $\widehat{\alpha}_3$ and $\widehat{\alpha}_4$ since part of the data is censored.

- $\widehat{\alpha}_7$: Similar to $\widehat{\alpha}_5$, but replacing $F(y_1, y_2)$ by $\exp\{-V(y_1, y_2)\}$.

- $\widehat{\alpha}_8$: Similar to $\widehat{\alpha}_6$, but replacing $F(y_1, y_2)$ by $\exp\{-V(y_1, y_2)\}$.

In order to assess the performance of the threshold-based approaches compared to naive and traditional methods, we also estimated α with

- $\widehat{\alpha}_{\text{naive}}$: Naive estimator maximizing the likelihood formed from logistic contributions for *all* data points, even though this is the *asymptotic* model.

- $\widehat{\alpha}_B$: Classical estimator based on block-maxima, with block sizes $b = 20, 50, 100$.

- $\widehat{\alpha}_{ST}$: Stephenson–Tawn estimator based on block-maxima, with block sizes $b = 20, 50, 100$. Unlike $\widehat{\alpha}_B$, this estimator uses the occurrence times of extremes, see (1.35), so it should be more efficient than $\widehat{\alpha}_B$. Stephenson & Tawn (2005) showed that for the logistic distribution, the gain in efficiency is more pronounced in case of low dependence, and becomes substantial as the dimension of the data increases.

The R replicates were then used to compute the empirical bias, the standard deviation and the root mean squared error (RMSE) of the different estimators. The results are summarized in Figure 1.6 for the estimators $\widehat{\alpha}_B$, $\widehat{\alpha}_{ST}$ with $b = 50$ and $\widehat{\alpha}_1$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\alpha}_5$, $\widehat{\alpha}_7$ with $p = 0.95$. The output of the whole simulation is reported in Tables A.1, A.2 and A.3 in the Appendix. As we can see, all estimators have increasing absolute bias as the data become more independent, that is, when α approaches unity. As expected, the estimators based on block maxima are best in terms of bias, but are much more variable than threshold-based estimators. Non-censored estimators that are based on the limiting Poisson process are not very variable, but suffer from a pronounced bias owing to their sensitivity to misspecification (due to the lack of convergence of the dependence function). Surprisingly, when $\alpha > 0.5$ (low dependence case), block maximum estimators are much better than the non-censored ones in terms of RMSE, and when $\alpha > 0.9$, the Stephenson–Tawn estimator slightly beats the censored estimator $\widehat{\alpha}_5$. Censored methods, and especially $\widehat{\alpha}_7$, seem to offer the best compromise between robustness (small bias) and efficiency (low uncertainty). At the

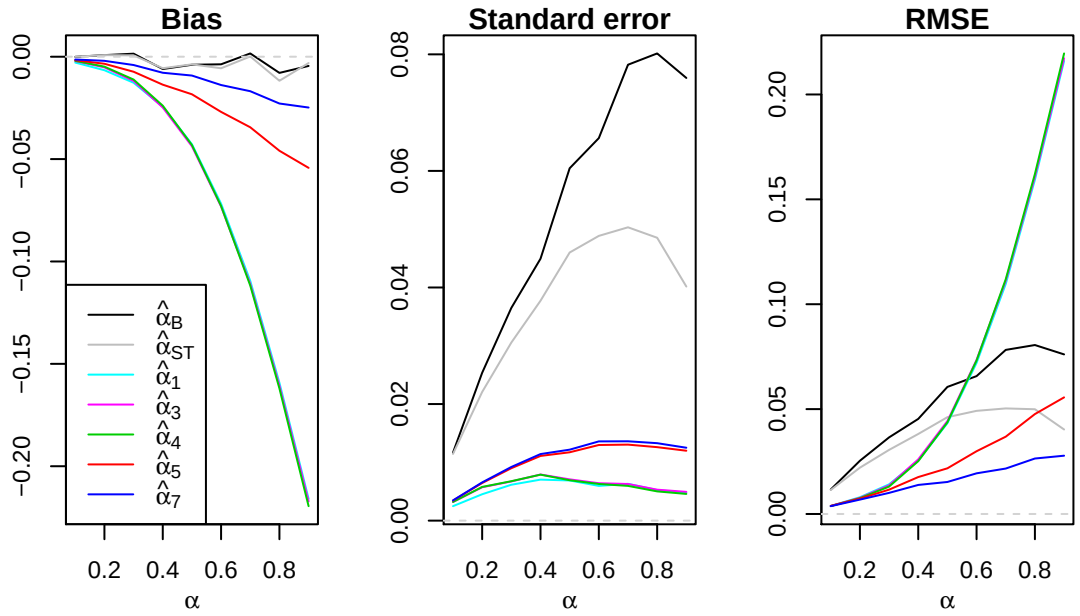

Figure 1.6: Bias (left), Standard error (middle) and RMSE (right) of the estimators $\widehat{\alpha}_B$, $\widehat{\alpha}_{ST}$ with $b = 50$, and $\widehat{\alpha}_1$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\alpha}_5$, $\widehat{\alpha}_7$ with $p = 0.95$, based on 300 bivariate datasets generated independently from the joint density (1.51). Curves for $\widehat{\alpha}_1$, $\widehat{\alpha}_3$ and $\widehat{\alpha}_4$ are almost identical.

95%-threshold, censored approaches outperform their competitors in terms of RMSE over the whole range of possible values for α. When $p = 0.98$, censored methods are still very good when $\alpha > 0.5$, but are slightly beaten by the non-censored ones when $\alpha < 0.5$, which was expected since, at higher thresholds, the Poisson process model is likely to fit better. Overall, at extreme levels often considered in practice and for a large range of dependence strengths, the censored estimators turn out to yield much better results than the other methods. Furthermore, in this study we have not considered the convergence of margins, which may favor censored methods even more.

Since the censored estimator $\widehat{\alpha}_7$ seems to be best overall for the logistic model, compared to most benchmark competitors, we have done further theoretical calculations to derive its exact asymptotic relative efficiency, as well as the Fisher information contributions of the different subspaces $P_{00} = [0, \boldsymbol{u})$, $P_{01} = [0, u_1) \times [u_2, \infty)$, $P_{10} = [u_1, \infty) \times [0, u_2)$ and $P_{11} = [\boldsymbol{u}, \infty)$. Assuming now that $\boldsymbol{Y} = (Y_1, Y_2) \in \mathbb{R}_+^2$ is exactly distributed according to the logistic model (1.30), the Fisher information $i_{\boldsymbol{u}}(\alpha)$ of the estimator $\widehat{\alpha}_7$ may be written, for a single contribution, as

$$i_{\boldsymbol{u}}(\alpha) = \mathrm{E}\left\{-\frac{\partial^2}{\partial \alpha^2} \log p_{\boldsymbol{u}}(\boldsymbol{Y})\right\} = \int_{\boldsymbol{0}}^{\infty} \left\{-\frac{\partial^2}{\partial \alpha^2} \log p_{\boldsymbol{u}}(\boldsymbol{y})\right\} g(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}, \tag{1.52}$$

where $p_{\boldsymbol{u}}$ is defined as in (1.48) and $g(\boldsymbol{y}) = (V_1 V_2 - V_{12})\exp(-V)$, $V_1 = \partial V(\boldsymbol{y})/\partial y_1$, etc., with V denoting the exponent measure of the logistic model (1.30). Expression (1.52) can be split into four parts corresponding respectively to the subspaces P_{00}, P_{01}, P_{10} and P_{11}, namely

$$\begin{aligned}
i_{\boldsymbol{u}}(\alpha) &= i_{00}(\alpha) + i_{01}(\alpha) + i_{10}(\alpha) + i_{11}(\alpha) \\
&= \int_{P_{00}} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log G(u_1, u_2) \right\} g(y_1, y_2) \mathrm{d}\boldsymbol{y} + \int_{P_{01}} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log G_2(u_1, y_2) \right\} g(y_1, y_2) \mathrm{d}\boldsymbol{y} \\
&\quad + \int_{P_{10}} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log G_1(y_1, u_2) \right\} g(y_1, y_2) \mathrm{d}\boldsymbol{y} + \int_{P_{11}} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log g(y_1, y_2) \right\} g(y_1, y_2) \mathrm{d}\boldsymbol{y} \\
&= \left\{ -\frac{\partial^2}{\partial \alpha^2} \log G(u_1, u_2) \right\} G(u_1, u_2) + \int_{u_2}^{\infty} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log G_2(u_1, y_2) \right\} G_2(u_1, y_2) \mathrm{d}y_2 \\
&\quad + \int_{u_1}^{\infty} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log G_1(y_1, u_2) \right\} G_1(y_1, u_2) \mathrm{d}y_1 \\
&\quad + \int_{u_1}^{\infty} \int_{u_2}^{\infty} \left\{ -\frac{\partial^2}{\partial \alpha^2} \log g(y_1, y_2) \right\} g(y_1, y_2) \mathrm{d}y_1 \mathrm{d}y_2,
\end{aligned}$$

where $G(\boldsymbol{y}) = \exp(-V)$ is the logistic distribution, and G_1, G_2 denote partial derivatives of G with respect to the first and second variables respectively. Variants of Bartlett's identities then yield

$$\begin{aligned}
i_{00}(\alpha) &= V_{\alpha^2} \exp(-V)\Big|_{(u_1,u_2)}, \\
i_{01}(\alpha) &= \left(V_\alpha^2 - V_{\alpha^2}\right)\exp(-V)\Big|_{(u_1,u_2)} + \int_{u_2}^{\infty} \left(\frac{V_{1\alpha}}{V_1} - V_\alpha\right)^2 (-V_1)\exp(-V)\Big|_{(u_1,y_2)} \mathrm{d}y_2, \quad (1.53) \\
i_{10}(\alpha) &= \left(V_\alpha^2 - V_{\alpha^2}\right)\exp(-V)\Big|_{(u_1,u_2)} + \int_{u_1}^{\infty} \left(\frac{V_{2\alpha}}{V_2} - V_\alpha\right)^2 (-V_2)\exp(-V)\Big|_{(y_1,u_2)} \mathrm{d}y_1, \quad (1.54) \\
i_{11}(\alpha) &= -\left(V_\alpha^2 - V_{\alpha^2}\right)\exp(-V)\Big|_{(u_1,u_2)} + \\
&\quad + \int_{u_1}^{\infty} \int_{u_2}^{\infty} \left(\frac{V_{1\alpha} V_2 + V_1 V_{2\alpha} - V_{12\alpha}}{V_1 V_2 - V_{12}} - V_\alpha\right)^2 (V_1 V_2 - V_{12})\exp(-V)\Big|_{(y_1,y_2)} \mathrm{d}y_1 \mathrm{d}y_2, \quad (1.55)
\end{aligned}$$

where we have written $V_\alpha = \partial V/\partial\alpha$, $V_{\alpha^2} = \partial^2 V/\partial\alpha^2$, and so forth. The integral in (1.53) can be transformed into a definite integral by considering the change of variables $t = V(u_1, y_2)$, and analogously for the integral in (1.54). After some calculations, one finds that

$$\begin{aligned}
\int_{u_2}^{\infty} \left(\frac{V_{1\alpha}}{V_1} - V_\alpha\right)^2 (-V_1)\exp(-V)\Big|_{(u_1,y_2)} \mathrm{d}y_2 &= \int_{u_1^{-1}}^{(u_1^{-1/\alpha} + u_2^{-1/\alpha})^\alpha} h(t, u_1) \mathrm{d}t, \\
\int_{u_1}^{\infty} \left(\frac{V_{2\alpha}}{V_2} - V_\alpha\right)^2 (-V_2)\exp(-V)\Big|_{(y_1,u_2)} \mathrm{d}y_1 &= \int_{u_2^{-1}}^{(u_1^{-1/\alpha} + u_2^{-1/\alpha})^\alpha} h(t, u_2) \mathrm{d}t,
\end{aligned}$$

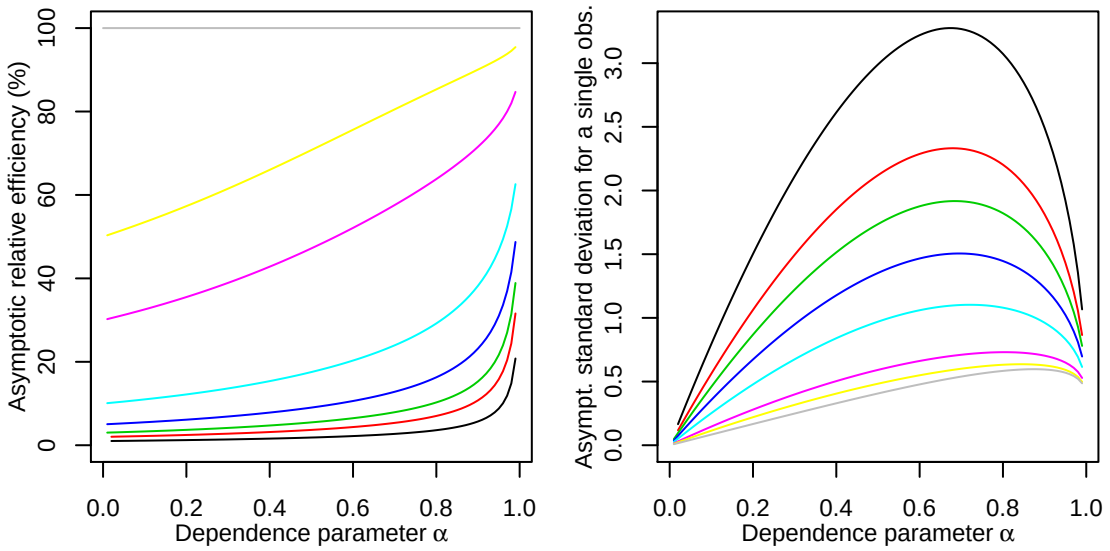

Figure 1.7: *Left*: Asymptotic relative efficiency of the estimator $\widehat{\alpha}_7$, censored at the pth-quantile, with respect to the full likelihood estimator for the logistic model (1.30) with different dependence parameters α. The different curves correspond to censoring at the level $p = 0\%$ (grey), 50% (yellow), 70% (purple), 90% (light blue), 95% (dark blue), 97% (green), 98% (red) and 99% (black). *Right*: Corresponding asymptotic standard deviation for a single observation, that is $i_{\boldsymbol{u}}(\alpha)^{-1/2}$; see (1.52).

where

$$h(t,u) = \frac{e^{-t}}{\alpha^2}\left[(1-t)t^{1/\alpha}(\log u + \log t) - \left\{1+\alpha(1-t)\left(t^{1/\alpha}-u^{-1/\alpha}\right)\right\}\log\left\{-1+(ut)^{1/\alpha}\right\}\right]^2.$$

A finite difference or standard Monte Carlo methods can then be used to compute the above integrals with high accuracy. Following Shi (1995), the integral in (1.55) can be markedly simplified by considering the change of variables $t_1 = V(y_1, y_2)$, $t_2 = \{y_1 V(y_1, y_2)\}^{-1/\alpha}$. The domain of integration P_{11} is hence transformed into the set $\{(t_1, t_2) \in \mathbb{R}^2; 0 < t_1 < B_1, B_2(t_1) < t_2 < B_3(t_1)\}$, with bounds $B_1 = (u_1^{-1/\alpha} + u_2^{-1/\alpha})^{\alpha}$, $B_2(t_1) = \max\{0, 1-(u_2 t_1)^{-1/\alpha}\}$ and $B_3(t_1) = \min\{0, (u_1 t_1)^{-1/\alpha}\}$. The software `Mathematica` can then help in computing this integral analytically with respect to t_2, although it involves the polylogarithmic function which is not expressible in closed form, and a straightforward finite integral with compact support can be used to compute the remaining complicated integral with respect to t_1. If we set $u_1 = u_2 = 0$ in the above formulae, the full Fisher information of the logistic model is recovered, so theoretical asymptotic relative efficiencies can be computed for $\widehat{\alpha}_7$ with respect to the maximum likelihood estimator, at different censoring levels. Figure 1.7 displays the asymptotic relative efficiencies of $\widehat{\alpha}_7$, censored at the pth-quantile with

$p = 0\%, 50\%, 70\%, 90\%, 95\%, 97\%, 98\%, 99\%$, along with the corresponding asymptotic standard deviation for a single observation, that is $i_{\boldsymbol{u}}^{-1/2}$. Comparing the blue lines in the middle panel of Figure 1.6 and the right panel of Figure 1.7, one can see that they look very similar, except that the former is scaled by a factor $\sqrt{n} = 100$, as it should be. Figure 1.7 shows that the use of the 99%-quantile threshold instead of 95% more than doubles the asymptotic standard deviation of $\widehat{\alpha}_7$. This underlies the need to select the threshold carefully in practice, in order to avoid a useless loss of statistical efficiency.

Figure 1.8 shows the contributions of each quadrant, that is P_{00}, P_{01}, P_{10} and P_{11}, to the asymptotic efficiency, and their relative importance at fixed thresholds. At a very low threshold, e.g., at the 5%-quantile, points in P_{11} are almost the only source of information, as expected. At a moderate threshold, e.g., at the 50%-quantile, the importance of points in the quadrant P_{11} decreases dramatically, asymptotically contributing less than 40% of the total Fisher information when $\alpha \approx 0.8$. In this scenario, the points lying in the "off-diagonal" quadrants P_{01} and P_{10} contribute much to the total information, especially in case of low to moderate dependence with $\alpha \in (0.5, 0.9)$. At higher thresholds, e.g., at the 95%-quantile, the relative importance of points in P_{11} is at least 43% whatever the dependence strength, and even larger than 60% when $\alpha < 0.2$ (very dependent case) or $\alpha > 0.8$ (near-independent case). Further computations show that these results also hold for other practical thresholds used in most extreme value applications, that is for thresholds higher than the 90%-quantile. Another comment is that the relative importance of completely censored data points, lying in the subspace P_{00}, is very low at all thresholds and for all dependence strengths, reaching 10% in special cases but at most 1% when the threshold is higher than the 93%-quantile. This explains why the estimator $\widehat{\alpha}_4$, based on mGPD distributions and solely on points lying in $[0, \boldsymbol{u}]^c$, has efficiency properties very similar to those of the estimator $\widehat{\alpha}_3$, which gives censored contributions to the data points in $[0, \boldsymbol{u}]$.

1.2.3 Copula modeling for multivariate extremes

When multivariate data are considered, the theory of copulas is appealing because it separates the modeling of the dependence structure and that of the margins. Sklar's theorem (Sklar, 1959; Nelsen, 2006, p.18) shows that knowledge of the joint distribution is equivalent to the specification of the margins and the dependence function, called the copula, and that the representation is unique if the marginal distributions are continuous.

Definition 35 (Copula)**.** *A copula C is a joint distribution in $[0,1]^D$ with uniform margins in $[0,1]$.*

Theorem 36 (Sklar, 1959)**.** *Let F be a joint distribution function in $\mathbb{R}^D$ with margins*

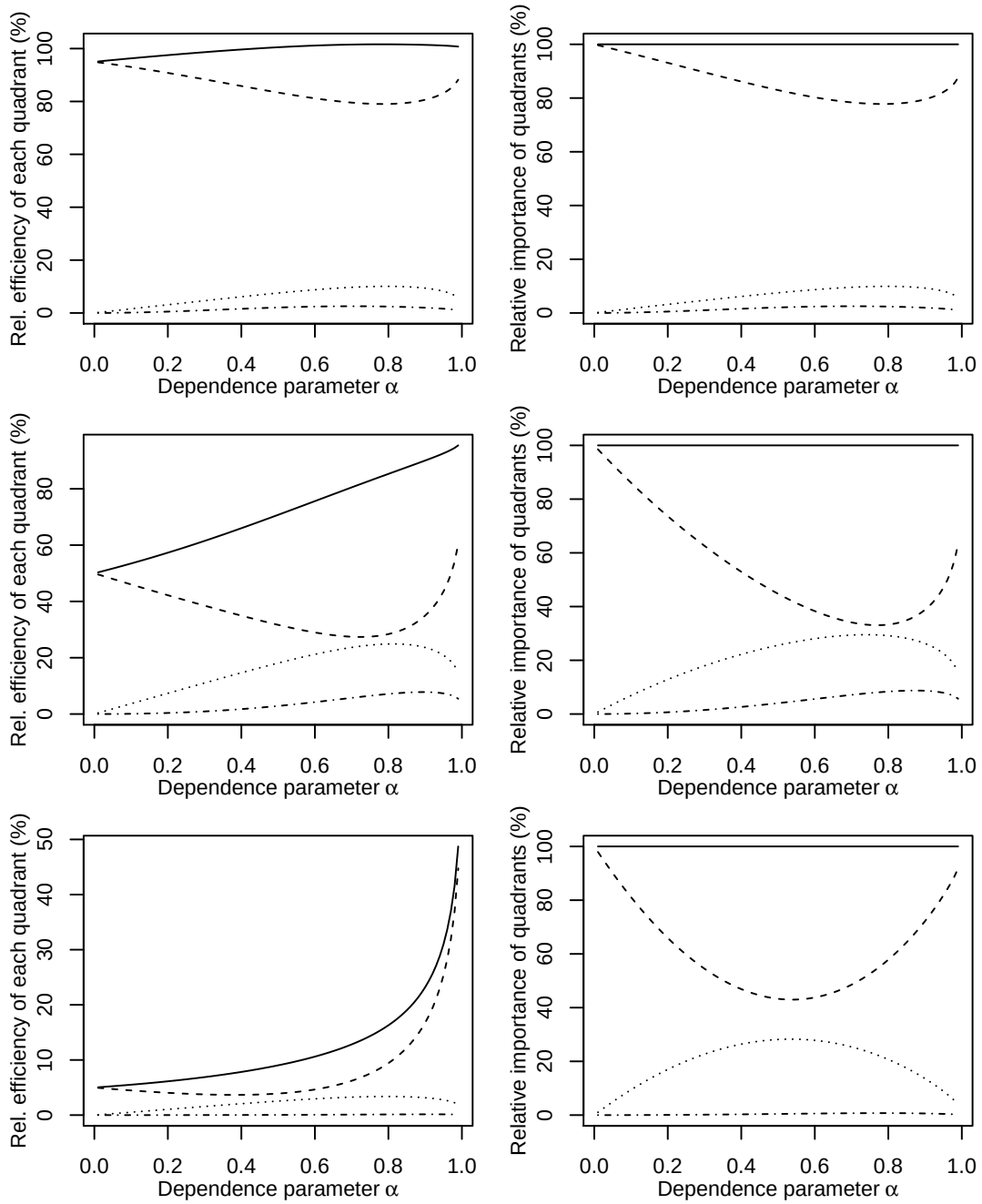

Figure 1.8: *Left column*: Asymptotic relative efficiency of the censored estimator $\widehat{\alpha}_7$ (solid) with respect to the full likelihood estimator for the logistic model (1.30), and the contributions of each quadrant, P_{00} (dash-dotted), P_{01}/P_{10} (dotted) and P_{11} (dashed), for different dependence parameters α. *Right column*: Proportion of Fisher information, in %, explained by each quadrant, i.e., $i_{00}(\alpha)/i_{\boldsymbol{u}}(\alpha)$, $i_{01}(\alpha)/i_{\boldsymbol{u}}(\alpha)$, $i_{10}(\alpha)/i_{\boldsymbol{u}}(\alpha)$ and $i_{11}(\alpha)/i_{\boldsymbol{u}}(\alpha)$. The censoring level of $\widehat{\alpha}_7$ is set at the pth-quantile, with $p = 5\%$ (top row), 50% (middle row) and 95% (bottom row).

$F_1, \ldots, F_D$. Then there exists a copula C such that for all $\boldsymbol{y} = (y_1, \ldots, y_D) \in \mathbb{R}^D$,

$$F(\boldsymbol{y}) = C\{F_1(y_1), \ldots, F_D(y_D)\}. \tag{1.56}$$

If $F_1, \ldots, F_D$ are continuous, then C is unique. Conversely, if C is a copula and $F_1, \ldots, F_D$ are distribution functions, then the function F defined in (1.56) is a joint distribution function with margins $F_1, \ldots, F_D$.

Hence, the copula underlying some joint distribution F may be written as $C(\boldsymbol{u}) = F\{F_1^{-1}(u_1), \ldots, F_D^{-1}(u_D)\} = \Pr\{F_1(Y_1) \leq u_1, \ldots, F_D(Y_D) \leq u_D\}$, where $\boldsymbol{u} = (u_1, \ldots, u_D) \in [0,1]^D$. For instance, the Gaussian copula may be written as $C(\boldsymbol{u}) = \Phi_D\{\Phi(\boldsymbol{u}); \Sigma\}$, where $\Phi_D(\cdot; \Sigma)$ and $\Phi(\cdot)$ are joint and marginal Gaussian distributions, and this generalizes to the t-copula or any other elliptic distributions. Sometimes it is more convenient to manipulate the survival copula defined as $\overline{C}(\boldsymbol{u}) = \Pr\{F_1(Y_1) > u_1, \ldots, F(Y_D) > u_D\}$.

For a good introduction to the theory of copulas, and to have an overview of the existing methods for constructing, simulating and fitting copulas, see the monograph by Nelsen (2006) or the PhD thesis by Hofert (2010). A critical review of such an approach can be found in Mikosch (2006) and in the discussion following the paper. An extensive catalogue of copula models with different theoretical properties has been proposed in the literature; see Joe (1997) and Nelsen (2006). Among this wide variety, Archimedean copulas (Nelsen, 2006, pp.116–119) have played an important role due to their relatively simple form,

$$C(\boldsymbol{u}) = \varphi^{-1}\{\varphi(u_1) + \cdots + \varphi(u_D)\}, \tag{1.57}$$

where the function $\varphi(t)$, called the generator of C, is defined on $(0,1]$ and has a completely monotone inverse.

Definition 37 (Completely monotone function)**.** *A continuous function $\psi(u)$, $u \geq 0$, is said to be* completely monotone *if it possesses derivatives $\psi^{(n)}(u)$ of all orders such that*

$$(-1)^n \psi^{(n)}(u) \geq 0, \qquad u > 0, n = 1, 2, \ldots.$$

Another large class of copulas, the so-called extreme value copulas, has gained a lot of attention in the extreme value community; see e.g., Joe (1990), Capéraà *et al.* (1997), Heffernan (2000), Sang & Gelfand (2010), Padoan (2011) and Davison *et al.* (2012). They may be written as

$$C(\boldsymbol{u}) = \exp\left[-V\left\{-1/\log(\boldsymbol{u})\right\}\right], \tag{1.58}$$

where V is the exponent measure of some multivariate extreme value distribution. The only Archimedean extreme value copula has generator $\varphi(t) = (-\log t)^{1/\alpha}$, for some $0 < \alpha < 1$, and is known as the Gumbel copula. It corresponds to the dependence structure of the logistic extreme value model (1.30). Another useful copula, the extremal t-copula (see Davison *et al.*, 2012, Demarta & McNeil, 2005, and §2.3.2.4), generalizes the Hüsler–Reiss copula defined by (1.32), and is quite flexible.

1.2.4 Asymptotic independence

Suppose that the estimation of the probability $p_A = \Pr(\boldsymbol{Y} \in A)$ is required for some extreme set A. The multivariate models developed in the earlier sections provide a tool to assess probabilities of rare events that have not yet been observed. The homogeneity property of the exponent measure underlying these models, closely linked to the max-stability assumption, enables to extrapolate beyond the range of the existing data, by noting that from (1.37), one has for large n,

$$p_A = \frac{1}{n}\mu_n\left(\frac{A}{n}\right) \approx \frac{1}{n}\mu\left(\frac{A}{n}\right) = \mu(A),$$

which in turn yields the following approximation for suitable constants $t > 0$,

$$p_A \approx \mu(A) = t\mu(tA) \approx t p_{tA}, \tag{1.59}$$

see Figure 1.9. However, it may be true that the level of dependence among the variables in $\boldsymbol{Y}$ varies with the "extremeness" of the event. In practice, extremal dependence is often observed to weaken at high levels, and it can happen that dependence is observed at finite levels, but that the random vector $\boldsymbol{Y}$ is in fact in the max-domain of attraction of independence. For such processes, asymptotic independence models are therefore needed to provide a realistic assessment of probabilities of rare events.

Definition 38 (Asymptotic independence)**.** *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ be a random vector in the max-domain of attraction of some multivariate extreme value distribution with underlying exponent measure V. We say that $\boldsymbol{Y}$ is* asymptotically independent *if $V(y_1, \ldots, y_D) = y_1 + \cdots + y_D$ for any $y_1, \ldots, y_D > 0$, that is, if G factorizes into its marginal components.*

In dimension $D = 2$, asymptotic independence may be more easily understood in terms of threshold exceedances. Let (Y_1, Y_2) denote a bivariate vector with joint distribution F and marginals F_1, F_2, and consider the variates $U_1 = F_1(Y_1)$ and $U_2 = F_2(Y_2)$ which are uniform in $[0,1]$. In order to measure extremal dependence, it is natural to look at the probability that the first variable exceeds a high quantile,

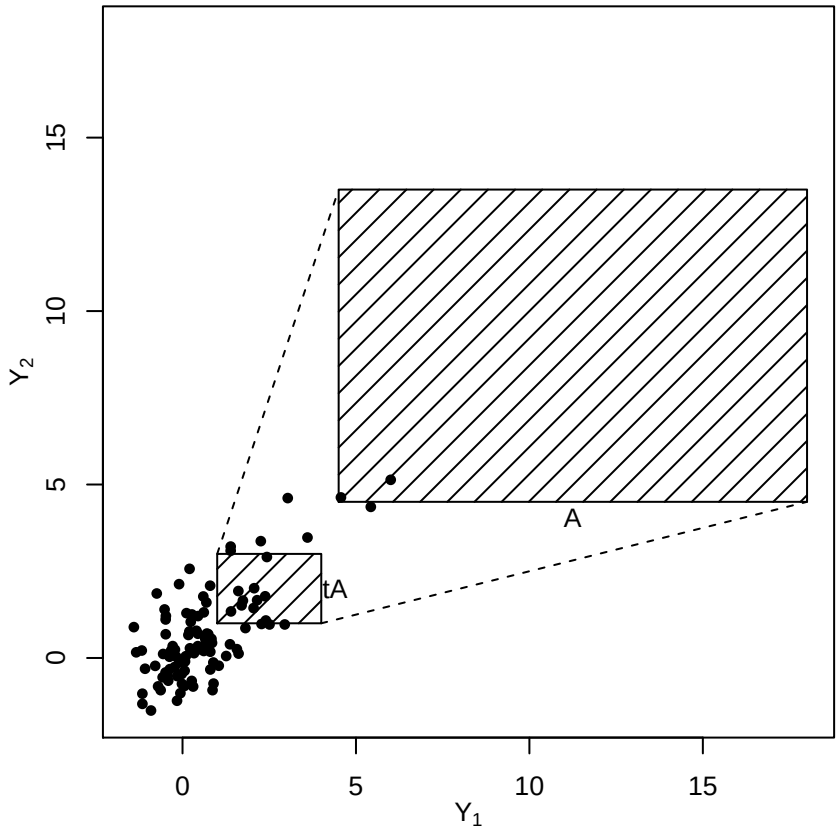

Figure 1.9: Extrapolation in the tail for a distribution in the max-domain of attraction of some multivariate extreme value distribution. In order to estimate the probability $p_A = \Pr\{(Y_1, Y_2) \in A\}$, one can compute instead p_{tA} with $t \in (0,1)$, where more data points are available, and then set $p_A \approx t p_{tA}$.

given that the second is similarly extreme. More precisely, considering the limiting conditional probability

$$\chi = \lim_{u \to 1} \Pr(U_1 > u \mid U_2 > u), \tag{1.60}$$

the variables Y_1 and Y_2 are asymptotically independent if $\chi = 0$, and are asymptotically dependent otherwise. This criterion, which focuses on the dependence on the "diagonal" $y_1 = y_2$, can in fact be shown to be equivalent to Definition 38 (see Ledford & Tawn, 1997). In practice, testing for asymptotic independence is very difficult since a powerful test requires insight very far in the tail, given that such variables might still exhibit quite strong residual dependence at high levels; see Figure 1.10. Ledford & Tawn (1996) provide a score test for asymptotic independence, as well as a joint tail estimation method that is able to discriminate between asymptotic independence and asymptotic dependence when $D = 2$. Moreover, their model provides an estimate of the rate of decay towards independence through the so-called parameter of tail

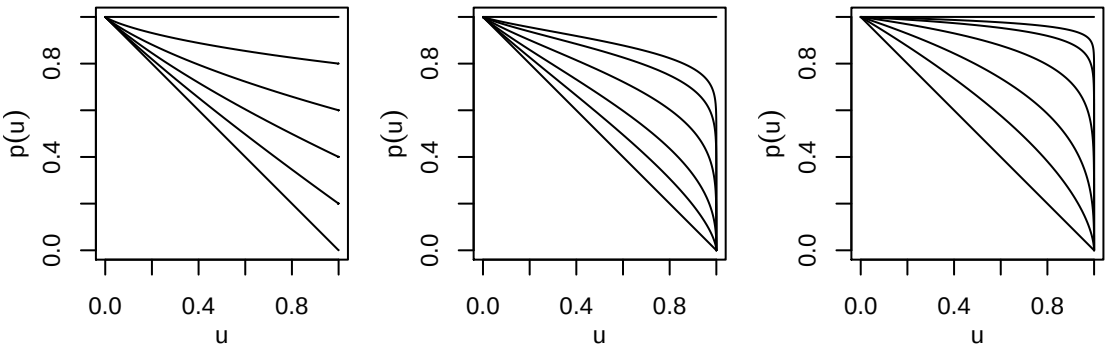

Figure 1.10: Theoretical conditional exceedance probability $p(u) = \Pr\{U_1 > u \mid U_2 > u\}$ for various random vectors $\boldsymbol{U} = (U_1, U_2)$ with uniform margins. *Left*: Max-stable copula with extremal coefficient $\theta_2 = 1, 1.2, 1.4, 1.6, 1.8, 2$ (from top to bottom); *Middle*: Gaussian copula with correlation $\rho = 0, 0.25, 0.75, 0.9, 0.95, 1$ (from bottom to top); *Right*: Inverted max-stable copula with coefficient of tail dependence $\eta = (1+\rho)/2$, using the same values for ρ. The variables in the left panel are asymptotically dependent, whereas those in the two other panels are asymptotically independent.

dependence. Since then, other authors have proposed more complex models for asymptotic independence; see for example Ledford & Tawn (1997), Ledford & Tawn (2003), Heffernan & Tawn (2004), Heffernan *et al.* (2007), Ramos & Ledford (2009) or Ramos & Ledford (2011). de Carvalho & Ramos (2012) provide a good review of the current state of statistical modeling of asymptotically independent data. Davison *et al.* (2013) suggest that, although useful to describe the dependence structure of asymptotically independent processes, these models should be used with care, especially for the derivation of extremely high quantiles, if there is no guarantee that the data considered are indeed asymptotically independent. Max-stable models, which should yield good fits at high levels, provide more conservative bounds for risk.

Theory about asymptotic independence, hidden regular variation and related concepts has been established for example in Resnick (2002), Maulik *et al.* (2002), Maulik & Resnick (2004) and Resnick (2008).

1.2.4.1 Ledford & Tawn model

Suppose that the random variables Y_1 and Y_2 have a common marginal unit Fréchet distribution with a bivariate normal dependence structure, with correlation $0 < \rho < 1$. Then, the joint tail satisfies

$$\Pr(Y_1 > y, Y_2 > y) \sim C_\rho y^{-2/(1+\rho)} (\log y)^{-\rho/(1+\rho)}, \qquad y \to \infty, \tag{1.61}$$

where C_ρ is a positive constant depending on ρ. On the one hand, if $\rho \to 0$, we have $\Pr(Y_1 > y, Y_2 > y) \sim y^{-2}$ (exact independence) and on the other hand, if $\rho \to 1$, we have $\Pr(Y_1 > y, Y_2 > y) \sim y^{-1}$ (complete dependence). Based on this result, Ledford & Tawn (1996) propose a model that smoothly links these two bounding cases, and introduce a new parameter that captures the rate of tail decay towards independence, asymptotic dependence being a particular case. Given two unit Fréchet variables Y_1 and Y_2 that may be associated at high levels, their approach relies on the assumption that the joint tail behaves as

$$\Pr(Y_1 > y, Y_2 > y) \sim \mathscr{L}(y) y^{-1/\eta}, \qquad y \to \infty, \tag{1.62}$$

where $\eta \in (0,1]$ is the coefficient of tail dependence, and where $\mathscr{L}(y)$ is a slowly varying function, that is $\mathscr{L}(ty)/\mathscr{L}(y) \to 1$ as $y \to \infty$ for all fixed $t > 0$. In the case of Gaussian association, $\mathscr{L}(y) = C_\rho (\log y)^{-\rho/(1+\rho)}$ and $\eta = (1+\rho)/2$, so the multivariate Gaussian distribution is asymptotically independent when $\rho < 1$. The approximation (1.62) holds for many known joint distributions. Since the marginal distributions are identical, the parameter η provides a measure of the dependence between the marginal tails. This can be seen more easily by looking at the conditional upper tail,

$$\Pr(Y_1 > y \mid Y_2 > y) \sim \mathscr{L}(y) y^{1-1/\eta}, \qquad y \to \infty, \tag{1.63}$$

meaning that Y_1 and Y_2 are asymptotically independent if $\eta < 1$ or $\eta = 1$ and $\mathscr{L}(y) \to 0$. Asymptotic dependence arises otherwise. Based on these modeling assumptions, equation (1.59) becomes

$$p_A \approx t^{1/\eta} p_{tA} \leq t p_{tA}, \qquad t \in (0,1),$$

for suitable extreme sets of the form $A = [y, \infty]^D$. The approximation (1.59) for max-stable models is recovered by setting $\eta = 1$. Therefore, the use of an extreme value model to extrapolate in the tail of an asymptotically independent distribution results in overestimating probabilities of very large events, and the extent of the misestimation is determined by the parameter of tail dependence.

Several dependence scenarios can be handled within this modeling framework. If $\eta = 1$ and $\mathscr{L}(y) > 0$, the variables are asymptotically dependent; if $1/2 < \eta < 1$, the variables are asymptotically independent, but positively associated ; if $\eta = 1/2$ and $\mathscr{L}(y) \geq 1$, the variables are near-independent; and if $0 < \eta < 1/2$, the variables are negatively associated. Ledford & Tawn (1996, 1997) and Heffernan (2000) provide a classification of some copulas according to these categories.

Ledford & Tawn (1997) extend this model by proposing a more flexible second-order

joint tail approximation for simultaneously large excesses of y_1 and y_2, $\Pr(Y_1 > y_1, Y_2 > y_2) \sim \mathscr{L}_1(y_1, y_2) y_1^{-c_1} y_2^{-c_2} + \mathscr{L}_2(y_1, y_2) y_1^{-(c_1+d_1)} y_2^{-(c_2+d_2)} + \cdots$, where $\mathscr{L}_1, \mathscr{L}_2$ are bivariate slowly varying functions. In this case, the parameter of tail dependence equals $\eta = (c_1 + c_2)^{-1}$. In a recent closely related paper, Ramos & Ledford (2009) provide new flexible parametric models for the so-called "ray dependence function" underlying these models, and show how likelihood-based inference and model selection can be worked out.

1.2.4.2 Inverted multivariate extreme value distributions

The lower tail of a multivariate extreme value distribution with exponent measure V may be described by its joint survivor copula,

$$\overline{C}(\boldsymbol{u}) = \exp\left[-V\left\{-1/\log(1-\boldsymbol{u})\right\}\right], \qquad \boldsymbol{u} = (u_1, \ldots, u_D) \in [0,1]^D. \tag{1.64}$$

The corresponding distribution is called an *inverted multivariate extreme value distribution* or *inverted max-stable distribution* (Ledford & Tawn, 1996; Heffernan & Tawn, 2004). It is asymptotically independent with parameter of tail dependence $\eta = 1/\theta_D$, where $\theta_D = V(1, \ldots, 1)$ is the extremal coefficient; see §1.2.5.1. This result provides a straightforward recipe for the construction of new asymptotically independent models with positive association from known extreme value distributions. Generalizing this to the infinite dimensional framework, Wadsworth & Tawn (2012) propose new classes of spatial models for extremes that can capture asymptotic independence based on known max-stable processes, and Thibaud *et al.* (2013) and Davison *et al.* (2013) show that such models provide a reasonable fit for the daily rainfall data analyzed in those papers.

1.2.5 Measures of extremal dependence

Different measures of extremal dependence, with their own estimation methods, have been developed in the literature, and a concise review can be found in Davison *et al.* (2013). The extremal coefficient is better suited for asymptotically dependent variables, that is, in the max-domain of attraction of some non-trivial max-stable distribution, whereas the coefficient of tail dependence is adequate for asymptotically independent variables. As mentioned above, discriminating between asymptotic dependence and asymptotic independence is not an easy task since the dependence between variables may vanish very slowly as the level increases. However, Coles *et al.* (1999) have developed a pair of diagnostic coefficients $(\chi, \overline{\chi})$ to aid this. For time series dependence measures, see Davis & Mikosch (2009) and Ledford & Tawn (2003).

In the sequel, we shall assume that $\boldsymbol{Y} = (Y_1, \ldots, Y_D)$ has unit Fréchet margins and underlying copula C.

1.2.5.1 Extremal coefficient θ_D

Suppose that $\boldsymbol{Y}$ is in the max-domain of attraction of some extreme value distribution G with exponent measure V and unit Fréchet margins, that is $G(\boldsymbol{y}) = \exp\{-V(\boldsymbol{y})\}$, $\boldsymbol{y} > 0$. The structure of extremal dependence is embedded in the exponent measure. Since V is homogeneous of order -1, a useful summary of extremal dependence among the variables $Y_1, \ldots, Y_D$ is the extremal coefficient

$$\theta_D = V(1, \ldots, 1) \in [1, D]. \tag{1.65}$$

It can easily be seen that

$$\Pr(\boldsymbol{Y} \leq y) = \left\{\exp(-1/y)\right\}^{\theta_D},$$

so θ_D can be interpreted as the effective number of independent variables. For perfectly dependent data, we have $\theta_D = 1$, and for asymptotically independent data, $\theta_D = D$. In dimension $D = 2$, one has $\theta_2 = V(1,1) = 2A(1/2)$, where A is the Pickands' dependence function; recall §1.2.1.4. In reality, this coefficient measures extremal dependence "on the diagonal $y_1 = \cdots = y_D = y$". In this sense, it does not give an exhaustive description of extremal dependence, especially for large D, but can still be useful for explanatory purposes or as a tool for model checking, by comparing empirical estimates (Schlather & Tawn, 2003; Naveau *et al.*, 2009) and fitted extremal coefficients.

The naive Schlather–Tawn estimator of the extremal coefficient (Schlather & Tawn, 2003) is based on the asymptotic distribution for $Z = \max(Y_1, \ldots, Y_D)$. Let $z_1, \ldots, z_n$ denote the observed values of the statistics Z, and let N_u be the number of z_i's exceeding a large threshold u. Since $\Pr(Z \leq z) \approx \exp(-\theta_D/z)$, $z > u$, a censored maximum likelihood estimator $\widehat{\theta}_D$ may be obtained in closed form as

$$\widehat{\theta}_D = \left[\frac{1}{N_u} \sum_{i=1}^{n} \left\{ \frac{1}{z_i} I(z_i > u) + \frac{1}{u} I(z_i \leq u) \right\} \right]^{-1}, \tag{1.66}$$

where $I(\cdot)$ is the indicator function. Another estimator for the extremal coefficient based on the concept of madograms, analogues of variograms, is proposed by Cooley *et al.* (2006b) and Naveau *et al.* (2009).

1.2.5.2 Coefficient of tail dependence η

The coefficient of tail dependence was introduced for dimension $D = 2$ by Ledford & Tawn (1996). It measures the strength of extremal dependence within the class of asymptotically independent models; see (1.62). For its estimation, Ledford & Tawn (1996) consider the structure variable $T = \min(Y_1, Y_2)$ which, conditional on exceeding a large threshold u, is approximately Generalized Pareto distributed with scale parameter $u\eta$ and shape parameter η. An estimate may thus be obtained by fitting the GPD to the observations $T_i = \min(Y_{i,1}, Y_{i,2})$, such that $T_i > u$, $i = 1, \ldots, N_u$. This yields a maximum likelihood estimator available in closed form,

$$\widehat{\eta} = \frac{1}{N_u} \sum_{i=1}^{N_u} \log\left(\frac{T_i}{u}\right), \tag{1.67}$$

which turns out to be the Hill estimator (Hill, 1975; Beirlant *et al.*, 2004, p.101).

1.2.5.3 Coefficients χ and $\overline{\chi}$

Another natural summary of extremal dependence for $D = 2$, introduced by Coles *et al.* (1999), is the quantity χ in equation (1.60). Getting rid of the marginal effects, this coefficient can also be seen as a function of the underlying copula C,

$$\chi = \lim_{u \to 1} \frac{\overline{C}(u,u)}{1-u} = \lim_{u \to 1} \frac{1 - 2u + C(u,u)}{1-u} = \lim_{u \to 1} 2 - \frac{1 - C(u,u)}{1-u}.$$

Letting

$$\chi(u) = 2 - \frac{\log C(u,u)}{\log u}, \tag{1.68}$$

one has $\chi = \lim_{u \to 1} \chi(u) \in [0,1]$. This alternative function can thus be interpreted as a quantile-dependent measure of dependence (Coles *et al.*, 1999). In particular, since $\chi(u) > 0$ if and only if $C(u,u) > u^2$, the sign of $\chi(u)$ determines whether the variables are positively or negatively associated at the quantile level u (Nelsen, 2006, p.187). If the variables are asymptotically dependent, then $\chi(u) \to \chi > 0$ as $u \to 1$, and $\chi(u) = 1$, $u \in [0,1]$, if they are perfectly dependent. When the variables are completely independent, we have $\chi(u) = 0$, $u \in [0,1]$. However, since the only situation corresponding to asymptotic independence is the bounding case $\chi(u) \to 0$, as $u \to 1$, a complementary dependence measure is proposed by Coles *et al.* (1999). By analogy to (1.68), they define

$$\overline{\chi}(u) = \frac{2\log(1-u)}{\log \overline{C}(u,u)} - 1, \tag{1.69}$$

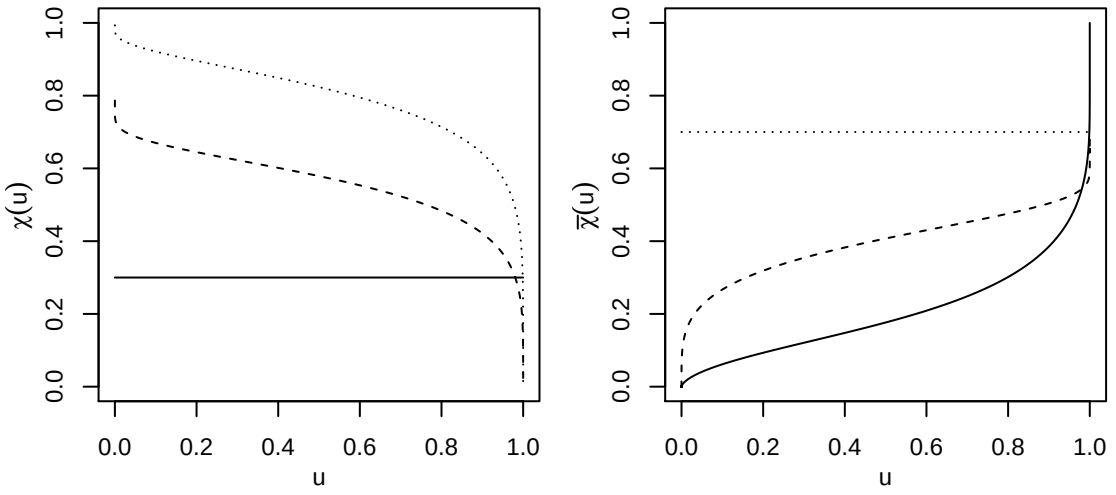

Figure 1.11: Coefficients $\chi(u)$ (left) and $\overline{\chi}(u)$ (right) for a max-stable copula with extremal coefficient $\theta_2 = 1.7$ (solid), a Gaussian copula with correlation $\rho = 0.7$ (dashed), and an inverted max-stable copula with coefficient of tail dependence $\eta = (1+\rho)/2 = 0.85$ (dotted).

and $\overline{\chi} = \lim_{u\to 1} \overline{\chi}(u) \in [-1,1]$. Similarly to the coefficient $\chi(u)$, the sign of $\overline{\chi}(u)$ is informative about the positive (or negative) association between the random variables. Moreover, under the Ledford–Tawn model (1.62), we have that $\overline{\chi} = 2\eta - 1$, where η is the parameter of tail dependence, and the only case corresponding to asymptotic dependence is $\overline{\chi} = 1$. Hence, this coefficient can be interpreted as a measure of dependence within the class of asymptotic independence models.

Coles *et al.* (1999) suggest estimating $\chi(u)$ and $\overline{\chi}(u)$ by their empirical counterparts and considering the complete pair of diagnostics $(\chi, \overline{\chi})$ to discriminate between asymptotic independence and asymptotic dependence; on the one hand, $(\chi = 0, \overline{\chi} < 1)$ signifies asymptotic independence and on the other hand $(\chi > 0, \overline{\chi} = 1)$ corresponds to asymptotic dependence. In practice, the shape of the curves described by $\chi(u)$ and $\overline{\chi}(u)$, for u close to 1, may be indicative about the type of asymptotic dependence of the data. Figure 1.11 displays $\chi(u)$ and $\overline{\chi}(u)$ for max-stable, Gaussian and inverted max-stable models. For max-stable models, $\chi(u)$ takes the constant value $2-\theta_2$, where θ_2 is the extremal coefficient and $\overline{\chi}(u)$ is increasing to unity; for Gaussian models, $\chi(u)$ is decreasing to zero and $\overline{\chi}(u)$ increasing to $\rho = 2\eta - 1$, where ρ is the underlying correlation; and for inverted max-stable processes, $\chi(u)$ is decreasing to zero and $\overline{\chi}(u)$ equals the constant value $2\eta - 1$. Hence, in this sense, the multivariate Gaussian distribution has even lighter tails than the left tail of a multivariate max-stable distribution.

1.3 Summary

In this chapter, we have given an overview of classical extreme value theory. In particular, in §1.1 we have shown that the generalized extreme value (GEV) distribution, that is, the only univariate max-stable distribution, is asymptotically justified for modeling maxima of a wide variety of processes, which may not be independent. Alternatively, the generalized Pareto distribution (GPD) is suitable for the modeling of high threshold exceedances, and permits to use more data for inference. We have also shown that the point process of extreme events can be approximated by a non-homogeneous Poisson process above high thresholds, and that threshold-based and block maximum approaches, both of which may be viewed as stemming from this characterization, are actually asymptotically equivalent.

In §1.2, we have seen that the family of distributions that generalize the GEV distribution to the multivariate framework is nonparametric. Hence, these limiting distributions cannot be characterized by a finite number of parameters, as in the univariate case, although they remain max-stable. Some well-known multivariate models have been described, and inference methods based on componentwise maxima or threshold exceedances have been discussed. In particular, Section 1.2.2.2 is a novel contribution, in which we compare different estimators for bivariate extreme value distributions. Focusing on the logistic model, our results reveal that the threshold approach based on marginal censoring, which we use in Chapter 5 to fit space-time models for rainfall extremes, is best in terms of root mean squared error. We have also clarified the links between these different estimators.

When the data are asymptotically independent, we have seen that extrapolation in the upper tail may be biased, and we have discussed some other related models that may be more appropriate in this context. Finally, we have also described several measures of extremal dependence, including the extremal coefficient, the coefficient of tail dependence and the diagnostics χ and $\overline{\chi}$, which may be used for model checking, for example, or to discriminate between asymptotic dependence and asymptotic independence.

In Chapter 2, we extend these ideas to provide models for spatial (or spatio-temporal) extremes, and generalize some of the methods developed earlier in this chapter to the infinite-dimensional case.

2 Geostatistical modeling of extremes in space and time

Many extreme value problems are spatial, or spatio-temporal, in nature. For instance, Westra & Sisson (2011) use spatial extreme-value processes to investigate whether significant trends drive extreme precipitation at sub-daily and daily timescales, and simulate spatial fields comprising observations from multiple point locations. Aryal *et al.* (2009) use extreme value theory along with classical geostatistics tools to estimate extreme rainfall and associated return levels for ungauged locations. Blanchet & Davison (2011) fit spatial models to extreme snowfall data, which may be used for risk management in Alpine regions of Switzerland. Davison & Gholamrezaee (2012) model annual temperature maxima, a relatively large-scale phenomenon compared to rainfall, and we have extended this to space-time modeling of extreme rainfall (see Chapter 5 and Huser & Davison, 2013b). Sang & Gelfand (2009) fit a space-time hierarchical model to extreme precipitation data recorded over the Cape Floristic Region in South Africa, assuming conditional independence of the data, given spatially correlated model parameters, and provide spatio-temporal forecasts with uncertainty assessment based on Markov Chain Monte Carlo (MCMC) methods.

Let us suppose that $Y(s,t)$ denotes some random quantity of interest recorded at location $s \in \mathscr{S}$ and time $t \in \mathscr{T}$ in some subspace $\mathscr{X} = \mathscr{S} \times \mathscr{T}$, and whose extremes must be understood for appropriate risk assessment. To illustrate the idea, $Y(s,t)$ could be some pollution index, precipitation or temperature measurement. In practice, it might be required to make inference on some risk measure of the type

$$R_{\mathscr{X}} = \iint_{\mathscr{X}} p(s,t)\{Y(s,t) - y_{\text{danger}}\}_+ \mathrm{d}s\mathrm{d}t,$$

where $p(s,t)$ denotes the population at risk at location s and time t, if $Y(s,t)$ exceeds some high critical threshold y_{danger}, and $a_+ = \max(0,a)$; see Davison & Gholamrezaee (2012). In order to derive probabilities for $R_{\mathscr{X}}$, space-time modeling of the extremes

of the phenomenon $Y(s,t)$ is needed within the space-time window $\mathscr{X}$. In particular, the dependence structure must be properly modeled to assess the risk correctly. A huge literature addresses the question of spatial extrapolation, or kriging, at points where data have not been collected (see, e.g., the textbooks Cressie, 1993; Stein, 1999; Wackernagel, 2003; Banerjee *et al.*, 2003; Diggle & Ribeiro, 2007; Cressie & Wikle, 2011), and Section §2.1 reviews some basics about random processes and related properties, along with covariance or variogram models proposed in the literature. However, since classical geostatistics is usually based on the Gaussian distribution, which has an exceptionally light tail, these standard methods tend to badly underestimate probabilities associated to extreme events. Moreover the upper joint tail of the multivariate Gaussian distribution leads to independent extremes for any underlying correlation that is less than unity, resulting in even more drastic underestimation of the probabilities of the simultaneous occurrence of two rare events.

In §2.2, we discuss the approach advocated by Cooley *et al.* (2007) among others. They propose hierarchical models that can suitably reflect the marginal distributions, but again, since these models are constructed from Gaussian processes, they are not adapted to describe the joint behavior of extremes.

More recently, max-stable processes have been proposed to model extreme events because under suitable conditions, they turn out to be asymptotically justified for the modeling of maxima of independent replications of random fields; see §2.3.1. Since they extend the multivariate extreme-value distributions to the spatial setting, they thus appear to be natural models for spatial extremes. Reviews about max-stable processes include Davison *et al.* (2012), Cooley *et al.* (2012), Davison *et al.* (2013) and Ribatet (2013). In Section 2.3, we first describe de Haan's (1984) spectral representation of such processes and then the main parametric models that have been proposed in the literature, including the Smith, Schlather, Brown–Resnick and extremal-t processes. Max-stable models based on latent α-stable random effects (Reich & Shaby, 2012; Shaby & Reich, 2012) are mentioned in §2.3.3. Then in §2.4, we address the question of asymptotic independence; the models by Wadsworth & Tawn (2012), which extend (1.64) and generalize max-stable models to hybrid spatial dependence models, able to capture and handle both asymptotic dependence and asymptotic independence, are discussed.

In Section 2.5, we present measures of extremal dependence that may be used in the spatial context, and in Section 2.6, we give an overview of different methods that may be used to make inference for max-stable and asymptotic independence models; the classical approach based on composite likelihoods is further discussed in Chapter 3. Finally, we close in Section 2.7 with a short spatial application.

2.1 Fundamentals of spatial random processes

Some basics of spatial random processes are summarized in this section. More details can be found in Cressie (1993), Stein (1999), Banerjee *et al.* (2003), Diggle & Ribeiro (2007), Cressie & Wikle (2011) and reviews are provided by Abrahamsen (1997) and Schlather (1999).

2.1.1 Definitions and notation

A random process may be defined as follows.

Definition 39 (Random process)**.** *A random process is a collection of random variables* $\{Y(x)\}_{x\in\mathscr{X}}$ *defined on a common probability space* $(\Omega, \mathscr{F}, \Pr)$ *and indexed by a parameter* $x \in \mathscr{X} \subset \mathbb{R}^d$*. Various notations and designations are adopted in specific frameworks:*

- *A random process in dimension* $d \geq 2$ *is also called a random field.*
- *When* $Y(x)$*,* $x \in \mathscr{X}$*, is a purely temporal process, it is common to replace* $x \in \mathscr{X}$ *by* $t \in \mathscr{T} \subset \mathbb{R}_+$ *to indicate time, in which case the points* $t \in \mathscr{T}$ *are called "times". Furthermore, if* $\mathscr{T} \subset \mathbb{N}$*, the process is usually referred to as a time series.*
- *When* $Y(x)$*,* $x \in \mathscr{X}$*, is a purely spatial process, it is common to replace* $x \in \mathscr{X}$ *by* $s \in \mathscr{S} \subset \mathbb{R}^d$ *to indicate space, in which case the points* $s \in \mathscr{S}$ *are called "stations".*
- *When* $Y(x)$*,* $x \in \mathscr{X}$*, is a spatio-temporal process, one usually writes* $Y(s,t)$*,* $(s,t) \in \mathscr{X} = \mathscr{S} \times \mathscr{T} \subset \mathbb{R}^d \times \mathbb{R}_+$*, where* $s \in \mathscr{S}$ *denotes the spatial coordinate and* $t \in \mathscr{T}$ *is the temporal coordinate.The points* $x \in \mathscr{X}$ *are called "sites" or "locations" and the space* $\mathscr{X}$ *is the spatio-temporal domain.*

Throughout this chapter, we shall always assume that the state-space $\mathscr{X}$ is a metric space endowed with the σ-algebra of Borel sets $\mathscr{B}(\mathscr{X})$ and that random processes are measurable with respect to the product σ-algebra $\mathscr{F} \times \mathscr{B}(\mathscr{X})$. Moreover, if not otherwise explicitly specified, we shall use the notation $x \in \mathscr{X}$ to denote a generic location in some abstract space $\mathscr{X}$, which can be time, or space, or a space-time domain. Lowercase d will be used to denote the dimension of the spatial field, while uppercase D shall refer to the dimension of the problem. Furthermore, we shall use n to denote the number of replicates of the process, and N for the total number of observations sampled in $\mathscr{X}$. In the space-time framework, the letters S and T will be used for the numbers of sampled stations in $\mathscr{S}$ and sampled times in $\mathscr{T}$, respectively, so that $N = ST$.

Definition 40 (Identically distributed random processes)**.** *Two processes $Y_1(x)$ and $Y_2(x)$, $x \in \mathscr{X}$, which need not be defined on the same probability space, are said to have the same finite-dimensional distributions, or to be identically distributed, if the vectors $\{Y_1(x_1), \ldots, Y_1(x_k)\}$ and $\{Y_2(x_1), \ldots, Y_2(x_k)\}$ have the same distributions for any k and any choices of $x_1, \ldots, x_k \in \mathscr{X}$.*

A random process $Y(x)$, $x \in \mathscr{X}$, is usually described by its first two moments, that is, the expectation $m(x) = \mathrm{E}\{Y(x)\}$, $x \in \mathscr{X}$, and the covariance function $C(x_1, x_2) = \mathrm{E}[\{Y(x_1) - m(x_1)\}\{Y(x_2) - m(x_2)\}]$, $x_1, x_2 \in \mathscr{X}$. The variance is defined as $v(x) = C(x, x)$ and the correlation function is $\rho(x_1, x_2) = C(x_1, x_2)\{v(x_1)\, v(x_2)\}^{-1/2}$, $x_1, x_2 \in \mathscr{X}$. An important category of random processes is the class of Gaussian processes, which are fully determined by $m(x)$ and $C(x_1, x_2)$.

Definition 41 (Gaussian process)**.** *A random process $Y(x)$, $x \in \mathscr{X}$, is called Gaussian if all its finite-dimensional distributions are multivariate Gaussian, that is if for any $k \in \mathbb{N}$ and any $x_1, \ldots, x_k \in \mathscr{X}$, the joint distribution of $Y(x_1), \ldots, Y(x_k)$ is Gaussian. A Gaussian process with mean $m(x)$ and covariance function $C(x_1, x_2)$ is denoted* $\mathrm{GP}(m, C)$.

Due to their sum-stability properties, conditional densities available in closed form and ease of simulation, Gaussian processes form the foundation of classical geostatistics, and are also building blocks of most hierarchical models, max-stable models and asymptotic independence models for spatial extremes. Brownian motions $B(t)$, $t \in \mathscr{T} \subset \mathbb{R}_+$, are particular cases of Gaussian processes in time; they have almost surely continuous sample paths, independent increments and are such that $B(t_2) - B(t_1)$ has mean zero and variance $|t_2 - t_1|$, for all $t_1, t_2 \in \mathscr{T}$.

2.1.2 Important properties

2.1.2.1 Stationarity

Natural phenomena can often be modeled in terms of an underlying homogeneous spatial process. In practice, different sorts of homogeneity may be assumed; in particular, extending Definition 13 to the functional setting, various definitions of a *stationary* process can be considered.

Definition 42 (Strict stationarity)**.** *A random process $Y(x)$, $x \in \mathscr{X}$, is called* strictly stationary *if the distributions of the vectors $\{Y(x_1), \ldots, Y(x_k)\}$ and $\{Y(x_1 + h), \ldots, Y(x_k + h)\}$ are the same for any choice of $k \in \mathbb{N}$, any sites $x_1, \ldots, x_k \in \mathscr{X}$ and any lag $h \in \mathbb{R}^d$, provided that $x_1 + h, \ldots, x_k + h \in \mathscr{X}$.*

Definition 43 (Weak stationarity)**.** *A random process $Y(x)$, $x \in \mathcal{X}$, is called* weakly stationary*, or* wide-sense stationary*, if its mean is constant, that is $m(x) = m$, $x \in \mathcal{X}$, and if its covariance function depends only on the separation vector, that is $C(x_1, x_2) = C(h)$, where $h = x_2 - x_1$ and $x_1, x_2 \in \mathcal{X}$. In particular, the variance takes the constant value $C(0)$ over $\mathcal{X}$, and the correlation function equals $\rho(h) = C(h)/C(0)$.*

Definition 44 (Intrinsic stationarity and variogram)**.** *A random process $Y(x)$, $x \in \mathcal{X}$, is called* intrinsically stationary *if its increments are weakly stationary; in particular, there exists a function $\gamma(h)$, called the* semi-variogram*, for which $2\gamma(h) = \mathrm{var}\{Y(x) - Y(x+h)\}$ for any $x \in \mathcal{X}$ such that $x + h \in \mathcal{X}$. The function $2\gamma(h)$ is called the* variogram.

In the space-time framework, one speaks of spatial stationarity if the above conditions are satisfied for the projection of the space-time process $Y(s,t)$, $(s,t) \in \mathcal{S} \times \mathcal{T}$, onto its spatial subdomain, $\mathcal{S}$. Temporal stationarity is defined analogously.

Although all these definitions correspond to a property of translation invariance, the three "levels" of stationarity are not equivalent. It is easy to see that strict stationary implies weak stationarity, which in turn entails intrinsic stationarity. But the converse relations do not hold in general. However, for Gaussian processes, whose finite dimensional densities only depend on the first two moments, strict stationarity and weak stationarity are equivalent.

2.1.2.2 Isotropy

The notion of isotropy is closely related to stationarity in the sense that both are geometric invariance properties of random processes. In the same way as stationarity is a shift invariance property, isotropy is a rotation invariance property.

Definition 45 (Isotropic random field)**.** *A random process $Y(x)$, $x \in \mathcal{X} \subset \mathbb{R}^d$, is called* isotropic *if for any $k \in \mathbb{N}$, $x_1, \ldots, x_k \in \mathcal{X}$ and any rotation, that is, orthogonal matrix $O \in \mathrm{Isom}(\mathbb{R}^d)$ such that $Ox_1, \ldots, Ox_k \in \mathcal{X}$, the joint distributions of the vectors $\{Y(x_1), \ldots, Y(x_k)\}$ and $\{Y(Ox_1), \ldots, Y(Ox_k)\}$ are the same.*

If the random process $Y(x)$, $x \in \mathcal{X}$, is weakly stationary and isotropic, its covariance function depends only on the length of the separation vector, not on its orientation, that is $C(x_1, x_2) = C(\|h\|)$, where $h = x_2 - x_1$ and $\|\cdot\|$ is some norm in $\mathcal{X}$ (usually the Euclidean norm). Similarly to any sort of stationarity, one can speak of spatial or temporal isotropy if a space-time process is considered.

2.1.2.3 Ergodicity

Loosely speaking, an ergodic random process $Y(x)$ defined on a set $\mathscr{X}$ has the property that distant events are nearly independent. For purely spatial applications, with $\mathscr{X} \subset \mathbb{R}^d$, the words "distant events" can be interpreted as usual with respect to the Euclidean distance. For purely temporal applications, with $\mathscr{X} \subset \mathbb{R}_+$, ergodicity means that two events barely influence each other if they occur at sufficiently different times. And generally, when $Y(x)$ is a random process in space and time, where $\mathscr{X} \subset \mathbb{R}^d \times \mathbb{R}_+$ denotes the space-time domain, one can distinguish spatial ergodicity and temporal ergodicity. Mathematically, this may be formulated as follows. Let $Y(x)$, $x \in \mathscr{X} \subset \mathbb{R}^d$, be a strictly stationary random field and denote by A, B and B_h the events

$$A = \bigcap_{x \in \mathscr{D}_1} \{Y(x) \leq y_1(x)\}, \quad B = \bigcap_{x \in \mathscr{D}_2} \{Y(x) \leq y_2(x)\}, \quad B_h = \bigcap_{x \in \mathscr{D}_2} \{Y(x+h) \leq y_2(x)\}, \quad (2.1)$$

where $\mathscr{D}_1, \mathscr{D}_2 \subset \mathscr{X}$ are finite subsets of sites, $h \in \mathbb{R}^d$ is a lag vector, $y_1(x), y_2(x)$ are real functions, and where it is implicitly assumed that the points $x + h \in \mathscr{X}$ for the definition of B_h. By strict stationarity, we have $\Pr(B) = \Pr(B_h)$. Furthermore, in order for such a notion to be well defined, $\mathscr{X}$ must be of "infinite size", meaning that there must exist a non-null vector $x \in \mathscr{X}$ such that $tx \in \mathscr{X}$ for all $t > 0$.

Definition 46 (Mixing and ergodic random field)**.** *The process $Y(x)$, $x \in \mathscr{X}$, is called* mixing *if for any events A, B and B_h defined in (2.1),*

$$\lim_{\|h\| \to \infty} \Pr(A \cap B_h) = \Pr(A)\Pr(B),$$

that is, if dependence vanishes at infinity, and the process is called ergodic *if*

$$\lim_{h \to \infty} (2h)^{-d} \int_{[-h,h]^d} \Pr(A \cap B_u) \mathrm{d}u = \Pr(A)\Pr(B).$$

One can show that a mixing process is also ergodic, but that the two properties are not equivalent. Assuming that a process is mixing is therefore a stronger assumption than being ergodic. Furthermore, a weakly stationary Gaussian process is mixing if and only if its covariance function satisfies $C(h) \to 0$, as $\|h\| \to \infty$ and it is ergodic if the Fourier transform of its correlation function, see (2.4) below, is atomless (Samorodnitsky, 2006).

For a detailed review of other technical mixing conditions, we refer to the three volumes by Bradley (2007a,b,c). For more details about ergodicity theory, see Aaronson (1997), Cornfeld *et al.* (1982) and Krengel (1985).

2.1.3 Covariance functions and variograms

2.1.3.1 Basic notions and classical models

Let $Y(x)$, $x \in \mathscr{X}$, denote a weakly stationary random process indexed by some continuous subspace $\mathscr{X} \subset \mathbb{R}^d$, and let $C(h)$, $\rho(h)$ and $\gamma(h)$ be its covariance function, correlation function and semi-variogram, respectively. The following relations hold:

$$\begin{aligned} 2\gamma(h) &= \operatorname{var}\{Y(x) - Y(x+h)\} \\ &= \operatorname{var}\{Y(x)\} + \operatorname{var}\{Y(x+h)\} - 2\operatorname{cov}\{Y(x) - Y(x+h)\} \\ &= 2\{C(0) - C(h)\}. \end{aligned}$$

Thus,

$$\gamma(h) = C(0) - C(h) = C(0)\{1 - \rho(h)\}. \tag{2.2}$$

Hence, $\gamma(h)$ can be readily recovered from $C(h)$. Moreover, if the spatial process is mixing, one has

$$C(h) = C(0) - \gamma(h) = \lim_{\|u\| \to \infty} \gamma(u) - \gamma(h). \tag{2.3}$$

In general, the limit on the right-hand side of (2.3) need not exist, but if it does, then the process is weakly stationary with covariance $C(h)$ as defined in (2.3). If a variogram $\gamma(h)$ is unbounded, a relationship of the form (2.2) with some covariance function $C(h)$ cannot hold. However, a covariance function might exist such that the correspondence holds locally, that is for $h \leq R$, in which case it is called a locally equivalent weakly stationary covariance (Gneiting *et al.*, 2001; Schlather & Gneiting, 2006). This enables the approximation of some classes of intrinsically stationary processes with locally weakly stationary representations.

Hence, variograms and covariance functions are closely related, and both need to satisfy specific constraints, ensuring that variances of linear combinations of the process at different sites are positive. More precisely, covariance and correlation functions are nonnegative definite, that is

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j C(x_i - x_j) \geq 0, \qquad \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j \rho(x_i - x_j) \geq 0,$$

for any positive integer k, any set of locations $x_1, \ldots, x_k \in \mathscr{X}$ and any set of real numbers $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. Likewise, (semi-)variograms are conditionally nonpositive definite in the sense that

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \beta_i \beta_j \gamma(x_i - x_j) \leq 0,$$

for any positive integer k, any set of locations $x_1, \ldots, x_k \in \mathscr{X}$ and any set of real numbers $\beta_1, \ldots, \beta_k \in \mathbb{R}$ satisfying $\sum_{i=1}^{k} \beta_i = 0$. Due to these constraints, valid covariance or variogram models are difficult to invent. However, Bochner's Theorem provides a characterization of nonnegative definite functions which has been used extensively to construct covariance functions.

Theorem 47 (Bochner, 1955)**.** *The function $C(h)$ is nonnegative definite if and only if it may be written as*

$$C(h) = \int \cos(w^T h) G(\mathrm{d}w), \tag{2.4}$$

where the measure G in $\mathbb{R}^d$ is bounded, positive and symmetric about 0.

Therefore, stationary correlation functions correspond to the collection of all probability measures $\tilde{G}(\mathrm{d}w) = G(\mathrm{d}w)/C(0)$ that are symmetric about 0. The latter are called *spectral measures*. If the density associated to $\tilde{G}$ exists, i.e., if we can write $\tilde{G}(\mathrm{d}w) = \tilde{g}(w)\mathrm{d}w$, then $\tilde{g}$ is referred to as the *spectral density*. Since G is symmetric about 0, expression (2.4) may be rewritten as $C(h) = \int e^{iw^T h} G(\mathrm{d}w)$. Hence, $g(w)$ is a rescaled version of the Fourier transform $\widehat{C}(w)$ of $C(h)$, namely $g(w) = (2\pi)^{-d}\widehat{C}(w)/C(0)$, implying that each valid covariance function must have a positive and integrable Fourier transform. This can be used in practice (using the fast Fourier transform algorithm) to check the permissibility of a given covariance function. As far as isotropic covariance functions in $\mathbb{R}^d$ are concerned, Matérn (1986) has shown that they can be characterized as

$$C(\|h\|) = \int_0^\infty \frac{J_\nu(t\|h\|)}{(t\|h\|)^\nu} H(\mathrm{d}t), \tag{2.5}$$

where H is a nondecreasing and integrable measure on $\mathbb{R}_+$ and $J_\nu(\cdot)$ is the Bessel function of the first kind of order $\nu = d/2 - 1$. According to Abrahamsen (1997), the measure H is linked to the spectral measure by the relation $H(t) = G(\{w \in \mathbb{R}^d : \|w\| < t\})$. Obviously, a covariance function defined in $\mathbb{R}^d$ is also valid in $\mathbb{R}^p$ for $p = 1, \ldots, d$, since the nonnegative definiteness property remains true on restrictions of the space. However, some isotropic covariance functions valid in $\mathbb{R}^d$ may not be valid in $\mathbb{R}^p$ for $p > d$. This is the case for example with the spherical model, which in valid up to 3 dimensions, but not for $d \geq 4$. Necessary conditions for the validity of covariances are discussed, e.g., in Cressie (1993), Stein (1999) and Banerjee *et al.* (2003).

The most basic covariance model, called *nugget effect*, is that of a white noise process, namely

$$C_{\text{nugget}}(h) = I(h = 0)\tau^2, \tag{2.6}$$

where $\tau > 0$ and $I(\cdot)$ is the indicator function. It can be shown (see Abrahamsen, 1997; Gneiting & Sasvári, 1999, and the references therein) that any isotropic covariance

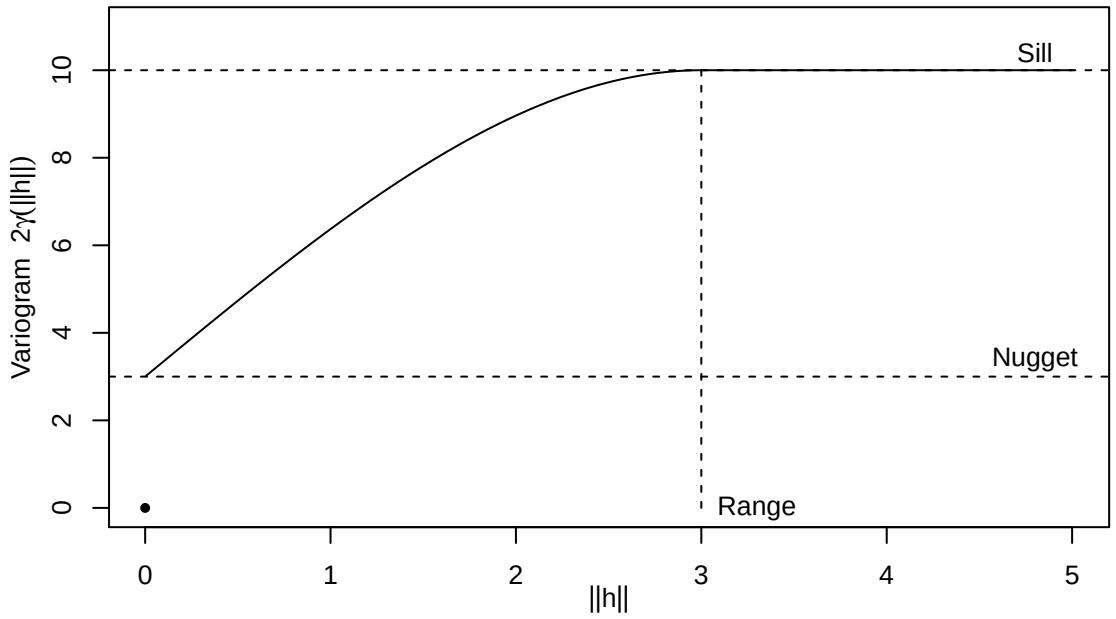

Figure 2.1: Generic isotropic variogram $2\gamma(\|h\|)$ and illustration of the nugget, the sill and the range. The discontinuity at the origin, represented by the dot separated from the main curve, is the nugget effect.

function in $\mathbb{R}^d$, with $d \geq 2$, admits a decomposition as the sum of a nugget effect (purely chaotic part), and a continuous covariance function. In practice, covariance models fitted to observed data frequently involve a nugget effect to account for small-scale variations, such as measurement errors, since the latter cannot easily be captured by a random process whose covariance function is continuous everywhere. In terms of variograms, this result implies that any isotropic semi-variogram $\gamma(\|h\|)$ can be represented as

$$\gamma(\|h\|) = I(\|h\| > 0)\left\{\tau^2 + \tilde{\gamma}(\|h\|)\right\},$$

where $\tilde{\gamma}(\|h\|)$ is a semi-variogram having no discontinuity at the origin (Gneiting *et al.*, 2001). Furthermore, if the corresponding process is weakly stationary and mixing, this representation can be rewritten as $\gamma(\|h\|) = I(\|h\| > 0)\left[\tau^2 + \sigma^2\left\{1 - \rho(\|h\|)\right\}\right]$, where the function $\rho(\|h\|)$ is an isotropic correlation function and $\sigma^2 > 0$ is the global effect variance. The limit $\lim_{\|u\|\to\infty} 2\gamma(u)$ is called the sill and the range of a variogram (resp. a correlation function) is defined as the minimum distance for which the sill, is attained (resp. the correlation reaches zero); see Figure 2.1. The effective range of a process is the minimum distance $R > 0$, possibly infinite, such that the correlation drops below 0.05, that is $R = \inf\{r > 0 : \rho(r) < 0.05\}$.

Table 2.1 lists several variograms that may be used in spatial applications. In particular,

Table 2.1: Common parametric isotropic (semi-)variogram models in $\mathbb{R}^d$. Since some of these models are not valid for all $d \in \mathbb{N}$, the maximum dimension d^M is also given. For all models, the parameter $\lambda > 0$ denotes a range parameter, $\nu > 0$ and $\alpha \in (0,2]$ are smoothness parameters, $\sigma^2 > 0$ is the spatial effect variance and $\tau^2 > 0$ is the nugget effect. The modified Bessel function of order ν is denoted by $K_\nu(\cdot)$.

Model	Semi-variogram, $\gamma(\|h\|)$	d^M
Nugget	$I(\|h\| > 0)\tau^2$	∞
Spherical	$I(0 < \|h\| \leq \lambda)\left[\tau^2 + \sigma^2\left\{\frac{3}{2}\left(\frac{\|h\|}{\lambda}\right) - \frac{1}{2}\left(\frac{\|h\|}{\lambda}\right)^3\right\}\right] + I(\|h\| > \lambda)(\tau^2 + \sigma^2)$	3
Powered exponential	$I(\|h\| > 0)\left(\tau^2 + \sigma^2\left[1 - \exp\left\{-\left(\frac{\|h\|}{\lambda}\right)^\alpha\right\}\right]\right)$	∞
Exponential	Powered exponential model with $\alpha = 1$	∞
Gaussian	Powered exponential model with $\alpha = 2$	∞
Cauchy	$I(\|h\| > 0)\left(\tau^2 + \sigma^2\left[1 - \left\{1 + \left(\frac{\|h\|}{\lambda}\right)^\alpha\right\}^{-\nu}\right]\right)$	∞
Power law	$I(\|h\| > 0)\left\{\tau^2 + \left(\frac{\|h\|}{\lambda}\right)^\alpha\right\}$	∞
Linear	Power law model with $\alpha = 1$	∞
Matérn	$I(\|h\| > 0)\left[\tau^2 + \sigma^2\left\{1 - \frac{(2\sqrt{\nu}\|h\|\lambda^{-1})^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu\left(2\sqrt{\nu}\|h\|\lambda^{-1}\right)\right\}\right]$	∞
Damped sine	$I(\|h\| > 0)\left[\tau^2 + \sigma^2\left\{1 - \frac{\lambda}{\|h\|}\sin\left(\frac{\|h\|}{\lambda}\right)\right\}\right]$	3

Brownian motions have linear semi-variograms, $\gamma(h) = \|h\|/\lambda$, $h \in \mathbb{R}^2$, and more generally, fractional Brownian motions have power law semi-variograms, $\gamma(h) = (\|h\|/\lambda)^\alpha$, $h \in \mathbb{R}^d$. These processes are examples of intrinsically stationary random processes that are not weakly stationary, but which admit a locally equivalent weakly stationary representation, provided the smoothness parameter satisfies $\alpha \neq 2$ (Gneiting *et al.*, 2001). The Matérn model has received a lot of attention and has become quite popular since its smoothness parameter $\nu > 0$ has an appealing interpretation in terms of the differentiability of the realized random field. A Gaussian process with Matérn covariance (without nugget effect) has sample paths that are $\lfloor \nu - 1 \rfloor$ times mean square differentiable (Cressie, 1993). Special cases of the Matérn model include the exponential ($\nu = 1/2$) and Gaussian ($\nu \to \infty$) models.

Other covariance functions may be constructed from known models by mixing, products or convolutions, since nonnegative definiteness is closed under these operations. Moreover, given a semi-variogram $\gamma(h)$ in $\mathbb{R}^d$, the function $\exp\{-\gamma(h)\}$ is a valid correlation function in $\mathbb{R}^d$ (Gneiting *et al.*, 2001), so new correlation functions can be

constructed from known semi-variograms in this manner.

2.1.3.2 Space-time correlation functions and related properties

Let now consider a weakly stationary spatio-temporal process $Y(s,t)$, $(s,t) \in \mathscr{X} = \mathscr{S} \times \mathscr{T}$, where $\mathscr{S} \subset \mathbb{R}^d$ is space and $\mathscr{T} \subset \mathbb{R}_+$ is the time axis. We denote by $C(h_s, h_t)$, $\rho(h_s, h_t)$ and $\gamma(h_s, h_t)$ its covariance function, correlation function and variogram, respectively, where $h_s \in \mathbb{R}^d$ and $h_t \in \mathbb{R}$ are spatial and temporal lags. Moreover, let $h = (h_s, h_t) \in \mathbb{R}^d \times \mathbb{R}$ be the space-time lag vector. Since there is no reason to assume that the marginal process in $\mathscr{T}$ behaves similarly to the marginal process in $\mathscr{S}$ (in particular, isotropy may not hold in $\mathscr{X}$), other more complex models than those in Table 2.1 are needed.

In practice, for modeling and estimation purposes, it is common to assume some structure for the covariance function. Gneiting *et al.* (2007) provide a good review of properties of space-time covariances, such as stationarity, separability and full-symmetry.

Definition 48 (Full-symmetry)**.** *A weakly stationary space-time covariance function $C(h_s, h_t)$ is said to be* fully-symmetric *if*

$$C(h_s, h_t) = C(h_s, -h_t) = C(-h_s, h_t) = C(-h_s, -h_t),$$

for any spatial lag $h_s \in \mathbb{R}^d$ and temporal lag $h_t \in \mathbb{R}$.

In the spatial context, this property is also referred to as axial, directional or reflection symmetry (Lu & Zimmerman, 2005). If a weakly stationary space-time covariance function is isotropic in space and in time, that is, if it only depends on the absolute distance $\|h_s\|$ and the time difference $|h_t|$, then it is also fully-symmetric.

Definition 49 (Separability)**.** *A weakly stationary space-time covariance function $C(h_s, h_t)$ is said to be* separable *if it can be expressed as the product of a purely spatial covariance function $C_{\mathscr{S}}(h_s)$ and a purely temporal covariance function $C_{\mathscr{T}}(h_t)$, i.e.,*

$$C(h_s, h_t) = C_{\mathscr{S}}(h_s) C_{\mathscr{T}}(h_t) = \frac{1}{C(0,0)} C(h_s, 0) C(0, h_t),$$

for any lags $h_s \in \mathbb{R}^d$ and $h_t \in \mathbb{R}$.

Definitions 48 and 49 can be extended more generally to nonstationary covariance and correlation functions; see Gneiting *et al.* (2007). Separable covariance functions

are easy to construct from known spatial and temporal covariances, but a major drawback is that they cannot capture space-time interactions in the dependence structure, making them physically unrealistic in many applications (Brown *et al.*, 2000, 2001). However, as Gneiting *et al.* (2007) and Genton (2007) point out, the simplified structure of separable covariances entails a dramatic drop in the number of parameters and facilitates computational procedures for large space-time datasets. Indeed, if Σ denotes the covariance matrix associated to a separable covariance function $C(h_s, h_t) = C_{\mathscr{S}}(h_s)C_{\mathscr{T}}(h_t)$, i.e., $\Sigma_{ij} = C(s_j - s_i, t_j - t_i)$, for some space-time locations $(s_1, t_1), \ldots, (s_N, t_N) \in \mathscr{S} \times \mathscr{T}$, then there exist two matrices ${}^S M \in \mathbb{R}^{S\times S}$ and ${}^T M \in \mathbb{R}^{T\times T}$, uniquely defined up to positive multiplicative constants, such that

$$\Sigma = {}^S M \otimes {}^T M = \begin{pmatrix} {}^S M_{11}\, {}^T M & \ldots & {}^S M_{1S}\, {}^T M \\ \vdots & \ddots & \vdots \\ {}^S M_{S1}\, {}^T M & \ldots & {}^S M_{SS}\, {}^T M \end{pmatrix}, \tag{2.7}$$

where $N = ST$ and $\otimes$ denotes the Kronecker product between two matrices. In particular, if the data are collected at regular times $t^{(1)}, \ldots, t^{(T)} \in \mathscr{T}$ at the stations $s^{(1)}, \ldots, s^{(S)} \in \mathscr{S}$, then the matrices ${}^S M$ and ${}^T M$ in (2.7) have entries ${}^S M_{ij} = aC_{\mathscr{S}}\left\{s^{(j)} - s^{(i)}\right\}$ and ${}^T M_{ij} = a^{-1}C_{\mathscr{T}}\left\{t^{(j)} - t^{(i)}\right\}$, where $a > 0$ can be any positive constant. In many spatial statistics applications, for example for kriging (Cressie, 1993; Cressie & Wikle, 2011; Anderes *et al.*, 2012), it is required to compute the inverse of such a covariance matrix Σ at a cost of $O\left(N^3\right)$ operations. But if (2.7) holds, $\Sigma^{-1} = \left({}^T M\right)^{-1} \otimes \left({}^S M\right)^{-1}$, so that the computations are much less intensive provided $S, T \ll N$. Genton (2007) discusses optimal separable approximation of nonseparable covariance matrices, with respect to the Frobenius matrix norm. This ingenious idea permits one to model the data using a realistic nonseparable covariance structure, and to do the computations using a separable approximation.

Testing for separability of space-time covariance functions has also been investigated; see Mitchell *et al.* (2005, 2006), Fuentes (2006), Li *et al.* (2007) and others.

Cressie & Huang (1999) introduced classes of nonseparable stationary covariance functions. Their approach is based on Bochner's theorem for space-time covariances, recall Theorem 47, and depends on Fourier transform pairs in $\mathbb{R}^d$. Gneiting (2002) extended their work by providing very general classes of valid nonseparable covariance functions, which admit Cressie & Huang's models as special cases. These classes are characterized by completely monotone functions, recall Definition 37. Table 2.2 lists some known completely monotone functions and Table 2.3 provides three examples of positive functions with completely monotone derivatives (see also Gneiting, 2002). Each combination of a function $\psi(s)$ from Table 2.2 and a function $\varphi(t)$ from Table 2.3

Table 2.2: Some completely monotone functions $\psi(s)$, $s > 0$. The modified Bessel function of order ν is denoted by $K_\nu(\cdot)$.

Function $\psi(s)$	Parameters
$\exp\left\{-(\sqrt{s}/\lambda)^\alpha\right\}$	$\lambda > 0$, $\alpha \in (0,2]$
$\{2^{\nu-1}\Gamma(\nu)\}^{-1}\left(2\sqrt{\nu s}\lambda^{-1}\right)^\nu K_\nu\left(2\sqrt{\nu s}\lambda^{-1}\right)$	$\lambda > 0$, $\nu > 0$
$\left\{1+\left(\sqrt{s}/\lambda\right)^\alpha\right\}^{-\nu}$	$\lambda > 0$, $\nu > 0$, $\alpha \in (0,2]$
$2^\nu\left\{\exp\left(\sqrt{s}/\lambda\right)+\exp\left(-\sqrt{s}/\lambda\right)\right\}^{-\nu}$	$\lambda > 0$, $\nu > 0$

Table 2.3: Some positive functions $\varphi(t)$, $t > 0$, with completely monotone derivative.

Function $\varphi(t)$	Parameters
$\{1+(\sqrt{t}/\lambda)^\alpha\}^\varrho$	$\lambda > 0$, $\alpha \in (0,2]$, $\varrho \in [0,1]$
$\log\{(\sqrt{t}/\lambda)^\alpha+\delta\}/\log(\delta)$	$\lambda > 0$, $\delta > 1$, $\alpha \in (0,2]$
$\{(\sqrt{t}/\lambda)^\alpha+\delta\}/[\delta\{(\sqrt{t}/\lambda)^\alpha+1\}]$	$\lambda > 0$, $\delta \in (0,1]$, $\alpha \in (0,2]$

may be chosen to construct a valid space-time fully-symmetric correlation function $\rho(h_s, h_t)$ as follows (Gneiting, 2002):

$$\rho(h_s,h_t) = \frac{1}{\varphi(|h_t|^2)^{d/2}}\psi\left(\frac{\|h_s\|^2}{\varphi(|h_t|^2)}\right), \tag{2.8}$$

where h_s and h_t are lags in space and time, d is the spatial dimension and $\|\cdot\|$ denotes the Euclidean norm. For example, by combining the first examples in Tables 2.2–2.3 with expression (2.8), we obtain

$$\rho(h_s,h_t) = \frac{1}{\left\{1+\left(\frac{|h_t|}{\lambda_t}\right)^{\alpha_t}\right\}^{d\varrho/2}}\exp\left[-\frac{\left(\frac{\|h_s\|}{\lambda_s}\right)^{\alpha_s}}{\left\{1+\left(\frac{|h_t|}{\lambda_t}\right)^{\alpha_t}\right\}^{\varrho\alpha_s/2}}\right], \tag{2.9}$$

where $\lambda_t, \lambda_s > 0$ determine spatial and temporal scale parameters, $\alpha_s, \alpha_t \in (0,2]$ are spatial and temporal smoothness (or shape) parameters and $\varrho \in [0,1]$ is a separability parameter quantifying the space-time interactions. However, this correlation function does not admit a separable model as a restriction. To remedy this, one can multiply it

by the purely temporal correlation function $\left\{1 + (|h_t|/\lambda_t)^{\alpha_t}\right\}^{-1}$, yielding

$$\rho(h_s, h_t) = \frac{1}{\left\{1 + \left(\frac{|h_t|}{\lambda_t}\right)^{\alpha_t}\right\}^{1+d\varrho/2}} \exp\left[-\frac{\left(\frac{\|h_s\|}{\lambda_s}\right)^{\alpha_s}}{\left\{1 + \left(\frac{|h_t|}{\lambda_t}\right)^{\alpha_t}\right\}^{\varrho\alpha_s/2}} \right]. \tag{2.10}$$

Hence, when $\varrho = 0$, the correlation function (2.10) is separable, i.e., it reduces to the product of a purely temporal correlation and a purely spatial correlation, whereas as ϱ approaches 1, the spatial and temporal components become increasingly entwined. This class of flexible correlation functions is used in our analysis of extreme rainfall in Chapter 5. Contours of model (2.10) are shown in Figure 2.2 for typical scale and smoothness parameters and for increasing values of the separability parameter. The bottom right panel with $\varrho = 1$ corresponds to the fitted model in §5. Corresponding covariance functions $C(h_s, h_t)$ can be obtained by multiplying the correlations by the spatial variance σ^2, and space-time variograms $\gamma(h_s, h_t)$ are derived using equation (2.2). Further extensions of the Gneiting class are provided by Schlather (2010).

Alternative space-time covariance models include the so-called product-sum models and their extensions proposed by de Cesare *et al.* (2001) and de Iaco *et al.* (2002), and covariance functions generated from mixtures of separable models (Ma, 2002, 2003a,b). In the context of modeling rainfall, "physically motivated" covariance functions have been advocated by Cox & Isham (1988) and Schlather (2010), an example of which may be constructed as

$$C(h_s, h_t) = \mathrm{E}_V \left\{ C_{\mathscr{S}}(\|h_s - V h_t\|) \right\}, \tag{2.11}$$

where V is a d-dimensional random velocity and $C_{\mathscr{S}}(\cdot)$ is a "motion invariant" spatial covariance function. If the speed vector V has relatively small variability and can be approximated by its expectation $\mathrm{E}(V)$, then model (2.11) satisfies the Taylor hypothesis, i.e., $C(0, h_t) \doteq C\{\mathrm{E}(V)h_t, 0\}$, which has found widespread interest in fluid dynamics, meteorology and hydrology (Cox & Isham, 1988; Cenedese *et al.*, 1991; Li *et al.*, 2009; Schlather, 2010; Davoust & Jacquin, 2011). Furthermore, the model (2.11) is closely related to the random set-based max-stable models for spatial extremes developed later in §2.3.2.2 and applied to rainfall data in Chapter 5. A different, though related, approach is based on stochastic differential equations; see Brown *et al.* (2000, 2001). The latter promote the use of so-called blur-generated spatio-temporal covariance models, which are motivated by physically plausible processes but cannot be expressed in closed form in full generality and must be computed using integral approximations.

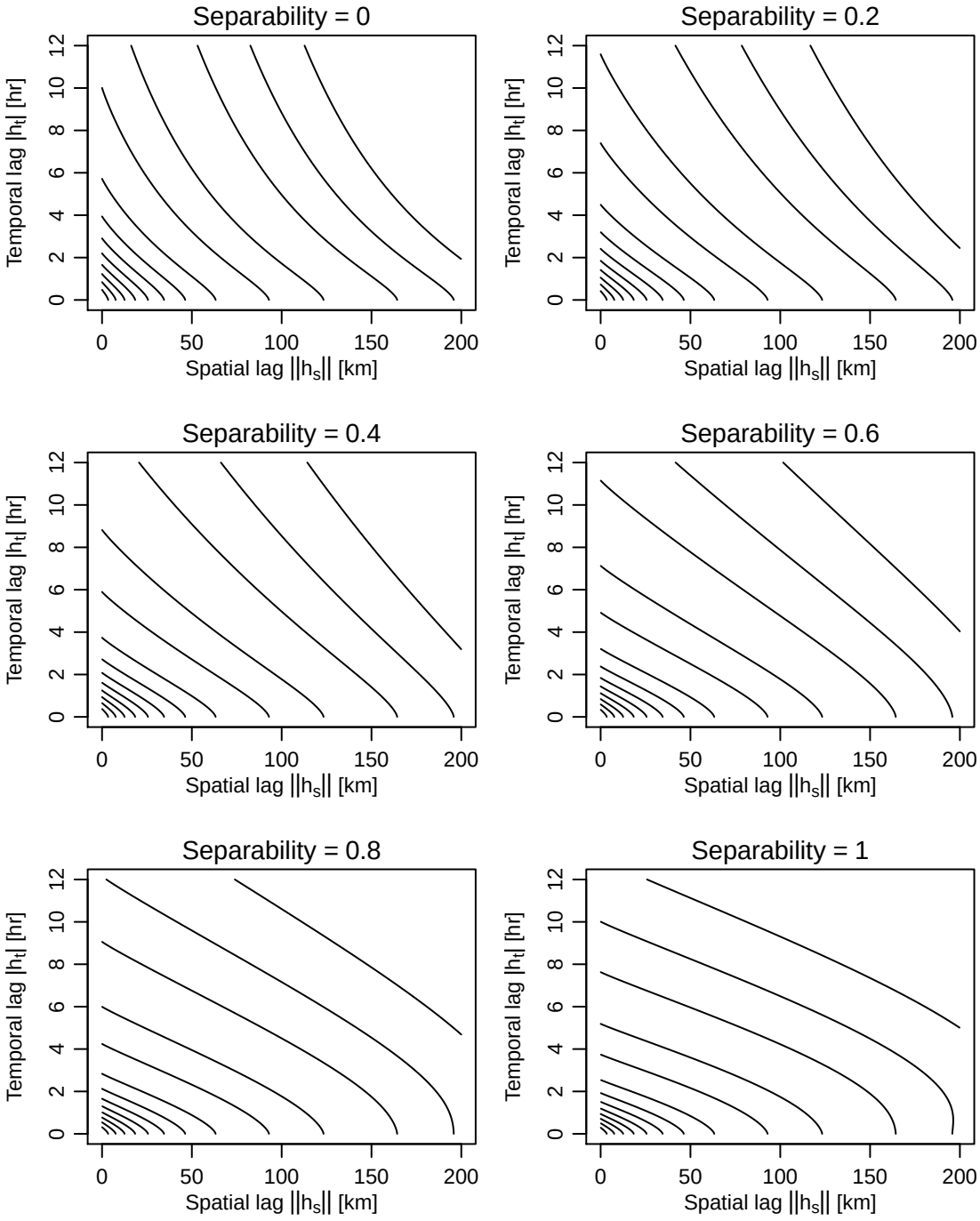

Figure 2.2: Contours of the Gneiting correlation function (2.10) for $\rho(h_s, h_t) = 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005$ (from the origin on). The scale and shape parameters are fixed to $\lambda_s = 37.9$, $\lambda_t = 2.2$, $\alpha_s = 0.93$ and $\alpha_t = 1.45$, as estimated in §5. The separability parameter varies from $\varrho = 0$ (top left), for which the model is separable, to $\varrho = 1$ (bottom right), with increments of 0.2.

2.2 Hierarchical models for extremes

An approach for modeling spatial extremes, typically used in Bayesian modeling using Markov chain Monte Carlo simulation, is to represent variation in extremal parameters through spatial Gaussian processes (see, e.g., Coles & Casson, 1998; Casson & Coles, 1999; Cooley *et al.*, 2007; Sang & Gelfand, 2009, 2010; Cooley & Sain, 2010). One major difference between these hierarchical models and the max-stable models presented below in §2.3 is that the former fit univariate extremal distributions to data at particular spatial locations, but usually treat the margins as independent conditional on the covariates. An exception to this is the model proposed by Sang & Gelfand (2010) which is based on a Gaussian copula, so risk estimates for spatial quantities may be poor, though those at individual locations can be expected to improve because of borrowing of strength across the spatial domain. By contrast the max-stable models described in §2.3 aim to capture joint spatial properties of extremes in addition to their marginal variation. Bayesian hierarchical modeling based on max-stable models has also been investigated, though fitting such models is much more complicated than for conditional independence models based on Gaussian processes. For example, Ribatet *et al.* (2012) fit such a model, using a pseudo-MCMC sampler, and Reich & Shaby (2012) and Shaby & Reich (2012) consider max-stable models constructed from latent α-stable random effects; see §2.3.3 for more details. Davison *et al.* (2012) provide a good comparative study of some of these models.

Cooley *et al.* (2007) is nowadays the main reference for Bayesian hierarchical conditional independence models fitted to spatial extremes, while Sang & Gelfand (2009) are almost the only researchers to fit such a model to space-time data. Related work includes Huerta & Sansó (2007). Hence for completeness, we include below a brief description of the Cooley *et al.* and Sang & Gelfand models, although the approach proposed in our application in §5 is based on the max-stable models of Section 2.3.

2.2.1 Cooley et al.'s model

Cooley *et al.* (2007) fit a Bayesian hierarchical model to rainfall exceedances using a GPD model (1.23) in which the scale parameter follows a log-normal process but the shape parameter has two values. More precisely, at the top of the hierarchy, the cluster peaks recorded at the sth station, denoted by $Y(s,t)$, $t = 1,\ldots,T$, are assumed to be GPD$\{\tau(s),\xi(s)\}$ above some predetermined high threshold u, that is

$$\Pr\{Y(s,t) - u > y \mid Y(s,t) > u\} = \left\{1 + \frac{\xi(s)}{\tau(s)} y\right\}_+^{-1/\xi(s)}.$$

Furthermore, the random variates $Y(s,t)$, $s = 1,\ldots,S$, $t = 1,\ldots,T$, are assumed to be mutually independent given the parameters $\tau(s)$ and $\xi(s)$. In the second layer of the hierarchy, a Gaussian spatial model is assumed for $\log\tau(s)$ while $\xi(s)$ can take only two values depending on the altitude of the station s,

$$\log\tau(s) \sim \mathrm{GF}(m_\tau, C_\tau) \qquad \xi(s) = \begin{cases} \xi_{\text{plains}}, & s \text{ is in the plains;} \\ \xi_{\text{mtn}}, & s \text{ is in the mountains,} \end{cases} \tag{2.12}$$

where the mean m_τ follows a linear regression in terms of covariates, such as elevation or mean annual precipitation, in order to capture the latent process that drives the climatological extreme precipitation for the region, and where the covariance function C_τ is assumed to be exponential. Chavez-Demoulin *et al.* (2011) point out that inclusion of covariates in the usual parametrization of the GPD model leads to a lack of invariance to the choice of threshold that can be resolved using the GEV parametrization, or by considering the modified scale parameter $\tilde{\tau} = \tau - \xi u$ (Thibaud *et al.*, 2013).

The third layer of the hierarchy is composed by prior distributions for the regression and covariance parameters, where the choice of hyperparameters was driven by subject-matter knowledge.

The posterior distribution for the parameter vector is obtained by multiplying the distributions in each layer, and fitting, inference and model selection can be made using the standard MCMC techniques. More complex models than the one presented above were also considered by Cooley *et al.* (2007), but did not yield significant improvements. A separate Bayesian hierarchical model was also fitted to the exceedance rates over the threshold u, so that they could produce return level maps for the entire front range region in Colorado.

2.2.2 Sang–Gelfand model

Sang & Gelfand (2010) propose a hierarchical model for spatially-referenced time series of extreme values, and consider a gridded interpolated precipitation dataset recorded over the Cape Floristic Region in South Africa. They assume that annual maxima follow the GEV distribution (1.4) whose location and scale parameters are jointly spatially dependent, and where the dependence is captured using conditional autoregressive (CAR) models.

More precisely, let $Z(s,t)$, $s = 1,\ldots,S$, $t = 1,\ldots,T$ denote the precipitation annual maxima. The first layer of the hierarchy assumes a conditional independence model

based on the GEV distribution, that is

$$Z(s,t) \mid \mu(s,t), \sigma(s), \xi \overset{\text{ind}}{\sim} \text{GEV}\{\mu(s,t), \sigma(s), \xi\}.$$

In the second layer, different space-time models for the location $\mu(s,t)$ may be used. A possibility proposed by Sang & Gelfand (2010) is to take

$$\mu(s,t) \mid \beta, W(s,t), \kappa^2 \overset{\text{ind}}{\sim} \mathcal{N}\{X(s)^T\beta + W(s,t), \kappa^2\},$$

where $X(s)$ is a station-specific vector of covariates, β is a vector of regression parameters, κ^2 is a nugget effect and $W(s,t)$ is a spatio-temporal random effect defined through

$$W(s,t) = U(s) + V(t), \tag{2.13}$$

where the temporal component $V(t)$ follows an AR(1) model, that is $V(t) = \lambda V(t-1) + \omega(t)$, with $\omega(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, W_0^2)$. A coregionalization CAR model is then assumed for the joint modeling of the purely spatial components $U(s)$ and $\log\sigma(s)$. Mathematically speaking, one has

$$\{U(s), \log\sigma(s)\} = \begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} E(s) \\ F(s) \end{pmatrix},$$

where $E(s)$ and $F(s)$ are two independent univariate CAR models (see the Chapter 3 of Banerjee *et al.*, 2003, for more details about CAR models). More complex models allowing for space-time interactions in (2.13) are also possible.

The full hierarchical Bayesian model, completed with a third layer specifying prior distributions for hyperparameters, can then be fitted using MCMC methods.

2.3 Max-stable processes

2.3.1 Generalities

In short, max-stable processes are spatial extensions of the max-stable distributions satisfying (1.26). A formal definition is given below.

Definition 50 (Max-stable process)**.** *Let $Z(x)$, $x \in \mathcal{X}$, be a random process indexed by a compact subspace $\mathcal{X} \subset \mathbb{R}^d$. The process $Z(x)$ is called* max-stable *if for each $k = 1, 2, \ldots$, there exist continuous functions $a_k(x) > 0$ and $b_k(x)$ such that for any function $z(x)$,*

$$\Pr\{Z(x) \leq a_k(x)z(x) + b_k(x), x \in \mathcal{X}\}^k = \Pr\{Z(x) \leq z(x), x \in \mathcal{X}\}, \tag{2.14}$$

or equivalently $Z(x)$ and the maximum of k independent copies of $\{Z(x) - b_k(x)\}/a_k(x)$

have the same distribution.

By analogy with the finite dimensional case, max-stable processes arise as the only possible class of limits for rescaled componentwise maxima of spatial processes. Specifically, consider i.i.d. stochastic processes $\{Y_i(x) : x \in \mathscr{X} \subset \mathbb{R}^d\}$, $i = 1, 2, \ldots$, with continuous sample paths on a compact set $\mathscr{X}$ equipped with the infinity norm, and suppose that there exist sequences of continuous functions $\{a_n(x)\} > 0$ and $\{b_n(x)\}$ such that as $n \to \infty$, the rescaled process of maxima,

$$Z^n(x) = \frac{\max\{Y_1(x), \ldots, Y_n(x)\} - b_n(x)}{a_n(x)}, \tag{2.15}$$

converges in distribution to a continuous random process $Z(x)$, $x \in \mathscr{X}$, all of whose univariate margins are non-degenerate. Then it can be shown that the class of possible limiting processes coincides with the class of max-stable processes with non-degenerate margins (de Haan & Ferreira, 2006, §9.2). For each $D = 1, 2, \ldots$ and any finite collection of sites $\mathscr{D} = \{x_1, \ldots, x_D\} \subset \mathscr{X}$, the corresponding variates $Z(x_1), \ldots, Z(x_D)$ have a multivariate extreme-value distribution, recall Section 1.2.1.1. In particular, the marginal distributions of $Z(x)$ are GEV$\{\mu(x), \sigma(x), \xi(x)\}$, where the real surfaces $\mu(x)$, $\sigma(x) > 0$ and $\xi(x)$ denote site-specific location, scale and shape parameters, respectively. By considering the transformation

$$\tilde{Z}(x) = t_x\{Z(x)\} = \left[1 + \frac{\xi(x)}{\sigma(x)}\left\{Z(x) - \mu(x)\right\}\right]_+^{-1/\xi(x)}, \tag{2.16}$$

one has that for each site $x \in \mathscr{X}$, $\Pr\{\tilde{Z}(x) \leq z\} = \exp(-1/z)$, $z > 0$, so that the marginal distributions of $\tilde{Z}(x)$ are unit Fréchet; such a process is called *simple max-stable*.

Definition 51 (Simple max-stable process)**.** *A random process $Z(x)$, $x \in \mathscr{X}$, is simple max-stable if it is max-stable with unit Fréchet marginals, or equivalently if the processes $Z(x)$ and $k^{-1}\max\{Z_1(x), \ldots, Z_k(x)\}$ have the same finite-dimensional distributions for any $k \in \mathbb{N}$, where $Z_1(x), \ldots, Z_k(x)$ denote i.i.d. copies of $Z(x)$.*

As in the multivariate case, the transformation (2.16) enables the modeling spatial extremes in two distinct steps: after first transforming the margins using fitted GEV distributions, an extremal model for spatial dependence can be fitted. Hence, without loss of generality, we shall restrict the discussion to simple max-stable random fields.

When a max-stable process $Z(x)$ is simple, the renormalizing constants in (2.15) are $a_n(x) \equiv n$, $b_n(x) \equiv 0$ and the joint distribution of $Z(x_1), \ldots, Z(x_D)$ can be written as

$$\Pr\{Z(x_1) \leq z_1, \ldots, Z(x_D) \leq z_D\} = \exp\{-V_{\mathscr{D}}(z_1, \ldots, z_D)\}, \qquad z_1, \ldots, z_D > 0, \tag{2.17}$$

where $V_{\mathscr{D}}(\cdot)$ is the exponent measure arising in (1.29), which summarizes the extremal dependence structure. The exponent measure is homogeneous of order -1, i.e. $V_{\mathscr{D}}(t\boldsymbol{z}) = t^{-1}V_{\mathscr{D}}(\boldsymbol{z})$ for any $\boldsymbol{z} = (z_1, \ldots, z_D) \in \mathbb{R}_+^D$, and satisfies $V_{\mathscr{D}}(\infty, \ldots, z, \ldots, \infty) = 1/z$ for any permutation of the D arguments. By analogy with the multivariate case in (1.65), the extremal coefficient, defined for $\mathscr{D} = \{x_1, \ldots, x_D\}$ as

$$\theta_D(x_1, \ldots, x_D) = V_{\mathscr{D}}(1, \ldots, 1) \in [1, D], \tag{2.18}$$

can be seen as a summary of extremal dependence. In the stationary case, it may be rewritten as

$$\theta_D(x_1, \ldots, x_D) = \theta_D(\boldsymbol{h}), \tag{2.19}$$

where $\boldsymbol{h} \in \mathbb{R}^{D-1}$ is a $(D-1)$-dimensional lag vector. When $\theta_D(\boldsymbol{h}) = 1$, the variables $Z(x_1), \ldots, Z(x_D)$ are perfectly dependent, and when $\theta_D(\boldsymbol{h}) = D$, they are independent, meaning that the original observations $Y_i(x_1), \ldots, Y_i(x_D)$, $i \geq 1$, are *asymptotically independent*, recall Definition 38.

Theorem 52 provides a useful representation of simple max-stable processes, which is valid under mild technical conditions.

Theorem 52 (de Haan & Ferreira, 2006, §9.4)**.** *Let $Z(x)$ be a simple max-stable process with continuous sample paths and defined on a compact subspace $\mathscr{X} \subset \mathbb{R}^d$ equipped with the infinity norm. Then, there exist i.i.d. continuous positive stochastic processes $W, W_1, W_2, \ldots$, with $\mathrm{E}\{W(x)\} = 1$ for all $x \in \mathscr{X}$ and $\mathrm{E}\{\sup_{x\in\mathscr{X}} W(x)\} < \infty$, such that*

$$Z(x) = \sup_{i\geq 1} W_i(x)/P_i, \tag{2.20}$$

where the P_is are the points of a unit rate Poisson process on $\mathbb{R}_+$. Conversely, each process with this representation is simple max-stable.

A process constructed from (2.20) is still max-stable when $W(x)$ is a stationary but not necessarily continuous random field, and when the condition $\mathrm{E}\{\sup_{x\in\mathscr{X}} W(x)\} < \infty$ is relaxed to $\mathrm{E}[\max\{0, W(x)\}] < \infty$ (see Schlather, 2002, Theorem 2).

Since the random process $W(x)$ is essentially arbitrary, this representation implies that no finite parametrization exists for max-stable processes. A common interpretation of expression (2.20) is to think of $Z(x)$ as a pointwise maximum of random storms $W_i(x)$ with corresponding intensities P_i^{-1}. Moreover, the geometric properties of the "storms" $W_i(x)$ are usually transferrable to the resulting max-stable process $Z(x)$; for example, if $W(x)$ is stationary over $\mathscr{X}$, then $Z(x)$ is stationary as well.

Assuming that we can simulate from the random process $W(x)$, the representation

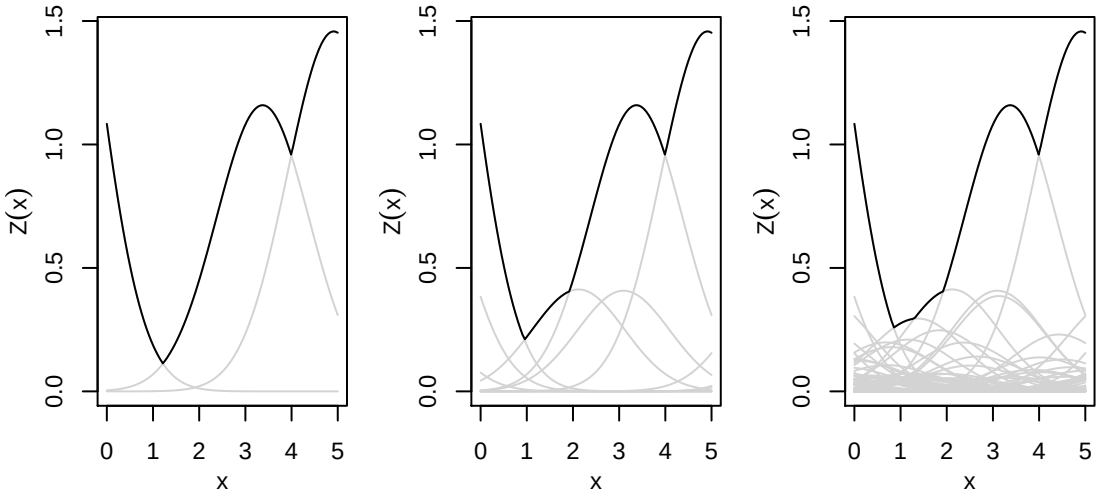

Figure 2.3: Illustration of the simulation of max-stable processes, with cut-off parameter $K = 3$ (left), $K = 10$ (middle) and $K = 100$ (right). The light grey lines are the simulated storms $W_1(x)/P_1, \ldots, W_K(x)/P_K$, and the solid line is the approximate max-stable random process.

(2.20) can be used to generate approximate realizations of the simple max-stable process $Z(x)$ on $\mathscr{X}$ as follows (see Figure 2.3 for an illustration):

- Choose the cut-off parameter $K \in \mathbb{N}$, controlling the quality of the finite approximation. Large K means better approximation.
- Generate $E_1, \ldots, E_K \overset{\text{iid}}{\sim} \text{Exp}(1)$, and set $P_l = \sum_{k=1}^{l} E_k$, $l = 1, \ldots, K$. It can easily be seen that the scalars $P_1, \ldots, P_K$ form a set of *increasing* points from a unit rate Poisson process on $\mathbb{R}_+$.
- Generate $W_1(x) \ldots, W_K(x) \overset{\text{iid}}{\sim} W(x)$ on a fine grid approximating the region of interest $\mathscr{X}$.
- Set $Z^\star(x) = \max_{i=1,\ldots,K} W_i(x)/P_i$.

Since the points P_i are generated in increasing order, the remaining part dropped in the finite approximation $Z^\star(x)$, i.e., $W_{K+1}/P_{K+1}, W_{K+2}/P_{K+2}, \ldots$, has a negligible contribution to the max-stable process $Z(x)$ if K is large enough, though the choice of K is not always straightforward. The simulation of the processes $W_i(x)$ can be intensive in practice, and a stopping rule has been proposed by Schlather (2002). For more details about simulation of max-stable random processes; see Schlather (2002), Oesting *et al.* (2012) and Ribatet (2013).

Let $Q(w_1,\ldots,w_D)$ denote the distribution of the random vector $\{W(x_1),\ldots,W(x_D)\}$. It follows from (2.17) that the exponent measure of a max-stable process $Z(x)$ defined through (2.20) is

$$\begin{aligned}
V_{\mathscr{D}}(z_1,\ldots,z_D) &= -\log\left[\Pr\{Z(x_1)\le z_1,\ldots,Z(x_D)\le z_D\}\right] \\
&= -\log\left[\Pr\left\{\sup_{i\ge 1} W_i(x_1)/P_i \le z_1,\ldots,\sup_{i\ge 1} W_i(x_D)/P_i \le z_D\right\}\right] \\
&= -\log\left[\Pr\{P_i \ge W_i(x_1)/z_1,\ldots,P_i \ge W_i(x_D)/z_D, i\ge 1\}\right] \\
&= -\log\left[\Pr\left\{P_i \ge \max_{j=1,\ldots,D} W_i(x_j)/z_j, i\ge 1\right\}\right] \\
&= \iint_{p<\max_{j=1,\ldots,D} w_j/z_j} \mathrm{d}p\mathrm{d}Q(w_1,\ldots,w_D) \qquad (2.21)\\
&= E\left[\max_{j=1,\ldots,D}\left\{\frac{W(x_j)}{z_j}\right\}\right], \qquad (2.22)
\end{aligned}$$

where equation (2.21) is justified by the fact that the points P_i stem from a unit rate Poisson process. More generally, if $\mathscr{D}\subset\mathscr{X}$ denotes some (possibly infinite) compact subspace of $\mathscr{X}$, similar calculations yield

$$\Pr\{Z(x)\le z(x), x\in\mathscr{D}\} = \exp\left(-E\left[\sup_{x\in\mathscr{D}}\left\{\frac{W(x)}{z(x)}\right\}\right]\right), \qquad (2.23)$$

where $z(x)$ can be any suitable function defined on $\mathscr{D}$. Expression (2.22) may be used to compute $V_{\mathscr{D}}(z_1,\ldots,z_D)$ for certain choices of $W(x)$, see §2.3.2, but typically it is only explicitly available for $D=2$ (but see Genton *et al.*, 2011, Wadsworth & Tawn, 2013, Huser & Davison, 2013a, and Chapter 4), so that full likelihood inference based on (2.17) or (2.23) seems unattainable in general; see §2.6.

An alternative slightly different spectral representation for simple max-stable random processes is formulated in the next theorem; see below and de Haan & Ferreira (2006, §9.6).

Theorem 53 (de Haan, 1984; de Haan & Ferreira, 2006)**.** *Let $\{P_i,U_i\}_{i\ge1}$, be a realization of a Poisson process on $\mathbb{R}_+\times[0,1]$ with mean measure $\mathrm{d}p\times d\lambda$, where λ denotes Lebesgue measure. If $Z(x)$ is a simple max-stable process with continuous sample paths in compact $\mathscr{X}$, there exists a family of functions $f(x,u)$, $x\in\mathscr{X}$, $u\in[0,1]$ with the following properties:*

1. *for each $u\in[0,1]$, we have a non-negative continuous function $f(x,u):\mathscr{X}\to[0,\infty)$,*
2. *for each $x\in\mathscr{X}$, $\int_0^1 f(x,u)\mathrm{d}u=1$,*

3. *for each compact subspace* $\mathscr{D} \subset \mathscr{X}$, $\int_0^1 \sup_{x\in\mathscr{D}} f(x,u)\mathrm{d}u < \infty$, *and* $f(x,u)$ *is such that*

$$Z(x) = \sup_{i\geq 1} f(x, U_i)/P_i. \tag{2.24}$$

Conversely, every process of the form on the right-hand side of (2.24), along with the conditions stated, is a simple max-stable process with continuous sample paths in $\mathscr{X}$.

The functions $f(x,\cdot) : [0,1] \to [0,\infty)$, which are not uniquely defined, are called the *spectral functions* or *storm profiles* of the simple max-stable process, and (2.22) becomes

$$V_{\mathscr{D}}(z_1,\ldots,z_D) = E\left[\max_{j=1,\ldots,D}\left\{\frac{f(x_j,U)}{z_j}\right\}\right], \tag{2.25}$$

where $U \sim \text{Unif}(0,1)$. The representation (2.24) remains valid if the spectral functions are defined on an abstract compact set $\mathscr{C} \subset \mathbb{R}^p$, instead of the interval $[0,1]$, and if they satisfy $\int_{\mathscr{C}} f(x,u) = |\mathscr{C}|$, for each $x \in \mathscr{X}$; in this case, U is uniformly distributed on $\mathscr{C}$ in (2.25).

Kabluchko & Schlather (2010) discuss mixing and ergodic properties of stationary max-stable processes defined on $\mathbb{Z}$. Their result is summarized in the next theorem.

Theorem 54 (Mixing and ergodicity for a stationary max-stable process)**.** *Let* $Z(x)$, $x \in \mathscr{X} \subset \mathbb{Z}$, *be a stationary simple max-stable process on the integers, and let* $\theta_2(h)$ *denote its bivariate extremal coefficient; recall (2.19). Then,*

1. $Z(x)$, $x \in \mathscr{X}$, *is mixing if and only if* $\theta_2(h) \to 2$, *as* $\|h\| \to \infty$.
2. $Z(x)$, $x \in \mathscr{X}$, *is ergodic if and only if* $(2h)^{-1}\int_{[-h,h]} \theta_2(u)\mathrm{d}u \to 2$, *as* $\|h\| \to \infty$.

2.3.2 Stationary parametric models

The representations (2.20) and (2.24) imply that there are infinitely many max-stable processes, and in practice the challenge is to build flexible but parsimonious models that can capture a wide range of extremal dependencies. Parsimony is important since extremal data are often scarce, but flexibility is also crucial since a poor fit might lead to misestimation of the risk. Several stationary parametric models for simple max-stable processes have been suggested; for example, Smith (1990b) proposes a max-stable model with deterministic storm shapes, Schlather (2002) proposes one based on a truncated Gaussian process, and the so-called Brown–Resnick process (Brown & Resnick, 1977; Kabluchko & Schlather, 2010) is contructed from log-normal

processes. Other models include the extremal-t process (Demarta & McNeil, 2005; Davison *et al.*, 2012; Opitz, 2013; Ribatet & Sedki, 2013), so-called max-max-stable processes (Robert, 2013b), a Brownian motion model proposed by Buishand *et al.* (2008), which has the drawback of not being invariant with respect to coordinate axes, or also Voronoï max-stable processes based on indicator functions of Poisson polytopes (Lantuéjoul *et al.*, 2011; Robert, 2013b). Bayesian non-parametric max-stable models, based on Dirichlet processes, have been advocated by Fuentes *et al.* (2011).

2.3.2.1 Smith model

Smith (1990b) proposes to take storm profiles of the type $f(x,u) = g(x-u)$ in (2.24), where $g(\cdot)$ is a probability density function defined on $\mathbb{R}^d$. For example, if

$$f(x,u) = \phi_d(x-u;\Sigma), \qquad x,u \in \mathscr{X}, \tag{2.26}$$

$\mathscr{X}$ compact, where $\phi_d(\cdot;\Sigma)$ denotes the d-variate Gaussian density with mean zero and covariance matrix Σ, the exponent measure of the resulting stationary max-stable process can be calculated for $D = 2$ using (2.25), and we get

$$V_{\mathscr{D}}(z_1,z_2) = \frac{1}{z_1}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_1}{z_2}\right)\right\} + \frac{1}{z_2}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_2}{z_1}\right)\right\}, \tag{2.27}$$

where $\mathscr{D} = \{x, x+h\} \subset \mathscr{X}$, $a = \{h^T\Sigma^{-1}h\}^{1/2}$, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. In this case, the bivariate extremal coefficient equals $\theta_2(h) = 2\Phi(a/2)$, and from Theorem 54, we see that the Smith model is mixing.

The covariance matrix Σ controls the dependence range and the degree of anisotropy of the realized random field. If $\Sigma = \mathrm{diag}(\lambda^2,\ldots,\lambda^2)$, the contours of the storm profiles are spherical (that is, the random process is isotropic) and λ determines the spatial extent of the storms. In the general case, the contours are elliptic. For example in $\mathbb{R}^2$, if Σ has ones on the diagonal and $\rho \in (0,1)$ in the off-diagonal entries, the degree of anisotropy can be measured by ρ; see Figure 2.4. Different types of storms may be obtained by varying these parameters. But because the Smith model has deterministic storm shapes, it is not very flexible and seems too smooth to be useful in environmental applications. In this respect it is unfortunate that it was the first model to be fitted (Padoan *et al.*, 2010) and thus has in some sense become a standard.

Extension of this model, involving the Student t or Laplace densities, have been proposed by Smith (1990b) and de Haan & Pereira (2006), but they do not resolve the problem mentioned above.

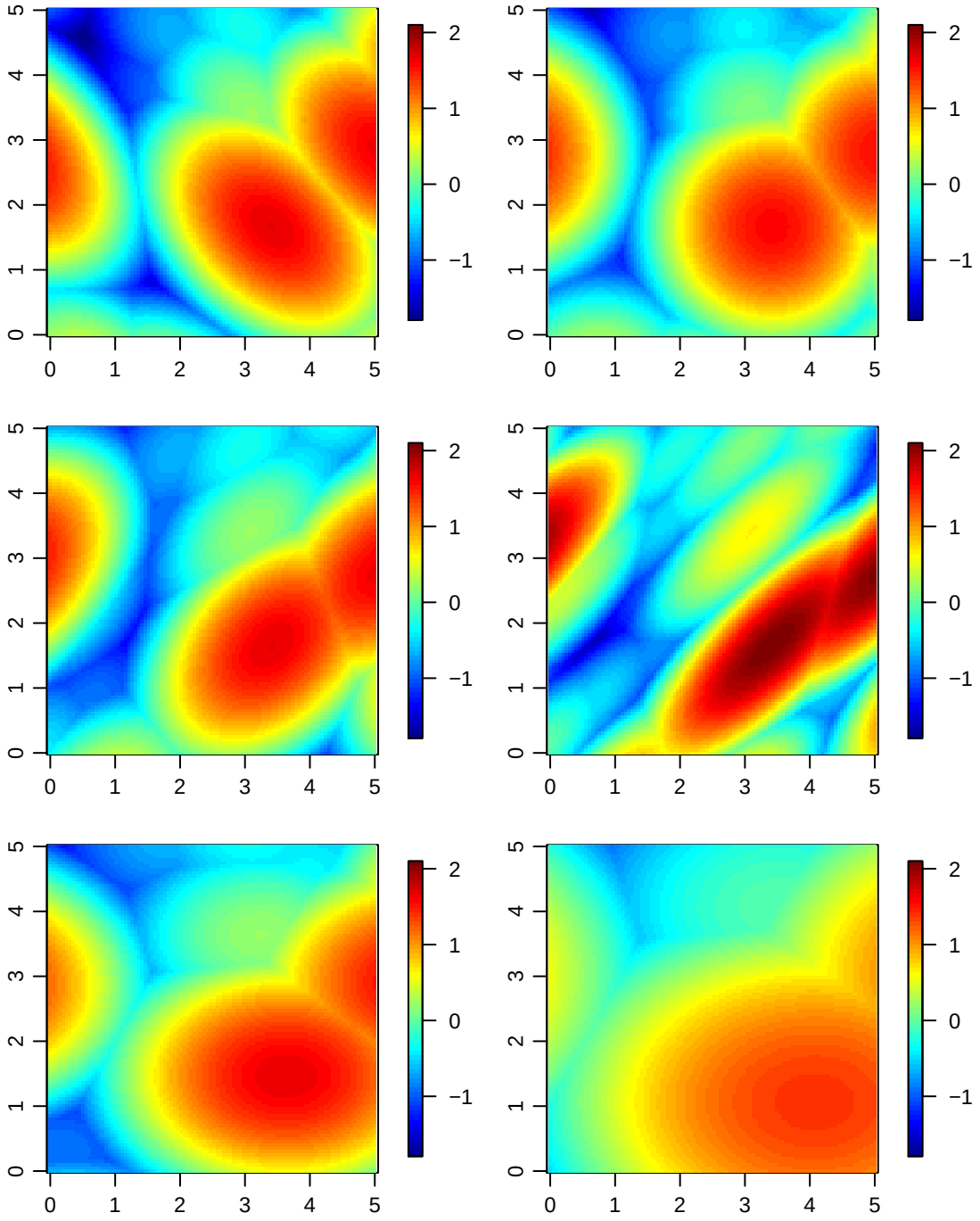

Figure 2.4: Realizations from the Smith model in $\mathbb{R}^2$ using Gaussian storm profiles with various covariance matrices $\Sigma = (\Sigma_{ij})$, $i, j = 1, 2$. *Top left*: $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = -0.4$; *Top right*: $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = 0$ (isotropic case); *Middle left*: $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = 0.4$; *Middle right*: $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = 0.8$; *Bottom left*: $\Sigma_{11} = 2$, $\Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = 0$. *Bottom right*: $\Sigma_{11} = 5$, $\Sigma_{22} = 2.5$, $\Sigma_{12} = \Sigma_{21} = 0$.

2.3.2.2 Schlather model

Non-smooth processes are better suited to modeling natural phenomena. An approach proposed by Schlather (2002) is take stochastic storm shapes of the form

$$W(x) = \sqrt{2\pi}\max\{\varepsilon(x), 0\}, \qquad x \in \mathscr{X}, \tag{2.28}$$

where $\varepsilon(x) \sim \mathrm{GF}\{0, \rho(h)\}$, so that $W(x)$ is the positive part of a Gaussian random field with zero mean, unit variance and correlation function $\rho(h)$, suitably rescaled to satisfy $\mathrm{E}\{W(x)\} = 1$. The resulting stationary max-stable process (2.20) has bivariate exponent measure

$$V_{\mathscr{D}}(z_1, z_2) = \frac{1}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)\left[1 + \frac{1}{z_1 + z_2}\left\{z_1^2 - 2z_1 z_2 \rho(h) + z_2^2\right\}^{1/2}\right], \tag{2.29}$$

where $\mathscr{D} = \{x, x+h\} \subset \mathscr{X}$, and is therefore an extension of the bivariate extreme-value distribution (1.33) to the spatial case. As for Gaussian processes, the smoothness of the max-stable process constructed from (2.28) can be controlled by the choice of correlation function $\rho(h)$ (see typical realizations in Figure 2.5) and the bivariate extremal coefficient equals $\theta_2(h) = 1 + \sqrt{\{1-\rho(h)\}/2}$. However, since in practice, one usually has $\rho(h) \to 0$, as $\|h\| \to \infty$, $\theta_2(h)$ is bounded above by 1.707 as $\|h\| \to \infty$, meaning that this model is not mixing, and complete independence cannot be captured, even at very large distances.

To circumvent this, Schlather (2002) proposed to introduce a random set element that ensures that sites that are distant enough cannot be covered by the same random function $W_i(x)$ in (2.20) and thus yields exact independence between maxima at such sites. Specifically, let

$$W(x) = \sqrt{2\pi}\frac{|\mathscr{X}|}{\mathrm{E}(|\mathscr{A}|)}\max\{\varepsilon(x), 0\} I_{\mathscr{A}}(x - X), \qquad x \in \mathscr{X}, \tag{2.30}$$

where $\mathscr{A} \subset \mathscr{X}$ is a compact random set lying in the compact state space $\mathscr{X}$, $\varepsilon(x)$ is defined as above, $I_{\mathscr{A}}(\cdot)$ is the indicator function over $\mathscr{A}$, X is a point from a unit rate Poisson process on $\mathscr{X}$, and $|\cdot|$ denotes the volume of a set. The resulting max-stable process (2.20) is stationary and built from random sets each with a truncated Gaussian process inside. Hence, the short-range dependence is largely determined by the correlation function $\rho(h)$, while the longer-range dependence is regulated by the geometry of the random set $\mathscr{A}$. The exponent measure can be written as

$$V_{\mathscr{D}}(z_1, z_2) = \left(\frac{1}{z_1} + \frac{1}{z_2}\right)\left(1 - \frac{\delta(h)}{2}\left[1 - \frac{1}{z_1 + z_2}\left\{z_1^2 - 2z_1 z_2 \rho(h) + z_2^2\right\}^{1/2}\right]\right), \tag{2.31}$$

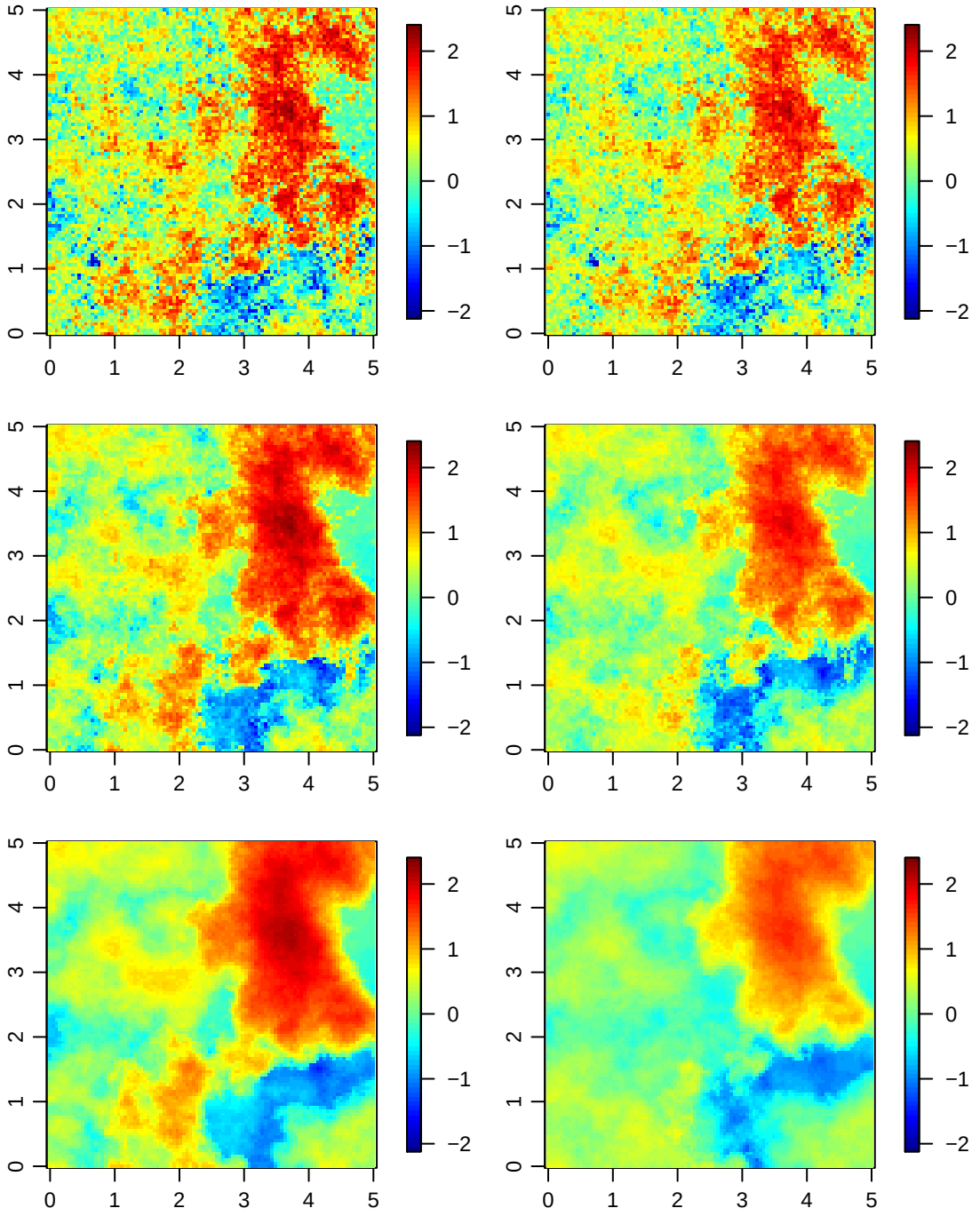

Figure 2.5: Realizations from the Schlather model with Gumbel margins in $\mathbb{R}^2$, using the powered exponential correlation function $\rho(h) = \exp\{-(\|h\|/\lambda)^\alpha\}$ and typical parameter values. *Top left*: $\lambda = 2$, $\alpha = 0.5$; *Top right*: $\lambda = 4$, $\alpha = 0.5$; *Middle left*: $\lambda = 2$, $\alpha = 1$; *Middle right*: $\lambda = 4$, $\alpha = 1$; *Bottom left*: $\lambda = 2$, $\alpha = 1.5$. *Bottom right*: $\lambda = 4$, $\alpha = 1.5$. The same random seed was used in all simulations.

where $\delta(h) = \mathrm{E}\{|\mathscr{A} \cap (h + \mathscr{A})|\}/\mathrm{E}(|\mathscr{A}|)$ lies in the unit interval. This model is clearly an extension of the Schlather model without random set (2.29), since the latter is recovered with $\delta(h) \equiv 1$. The extremal coefficient can be expressed as $\theta_2(h) = 2 - \delta(h)[1 - \sqrt{\{1-\rho(h)\}/2}]$, and since the random set $\mathscr{A}$ is chosen to be compact, we can choose $\mathscr{A}$ so that $\delta(h) \to 0$ and thus $\theta_2(h) \to 2$ as $\|h\| \to \infty$ for any correlation function $\rho(h)$. Consequently, this model is mixing for suitable choices of $\mathscr{A}$ and independence can be captured; see the illustration in Figure 2.6.

A drawback of this model, intrinsically related to the random set element, is that realizations are not continuous. However, it can be adapted to model processes with very local effects and abrupt changes, or phenomena which reflect complete independence after some fixed lag, as is usually observed in space-time applications. The Schlather model with random set was first fitted to annual temperature maxima in Switzerland by Davison & Gholamrezaee (2012) and a space-time version of it is fitted satisfactorily to hourly rainfall extremes in Chapter 5 (see also Huser & Davison, 2013b).

2.3.2.3 Brown–Resnick model

Another possibility, which can be viewed as extending the Hüsler–Reiss distribution (1.32) to the spatial framework, and the Smith model (2.28) to non-smooth processes (see Chapter 4 and Huser & Davison, 2013a), is the Brown–Resnick process (Brown & Resnick, 1977; Kabluchko *et al.*, 2009), sometimes called the geometric Gaussian process. This is constructed by using (2.20) with

$$W(x) = \exp\{\varepsilon(x) - \gamma(x)\}, \qquad x \in \mathscr{X}, \tag{2.32}$$

where $\varepsilon(x)$ is an intrinsically stationary Gaussian process with mean zero, semi-variogram $\gamma(h)$, and $\varepsilon(0) = 0$ almost surely. Although Brown–Resnick processes are based on nonstationary Gaussian models (in the strict sense), the construction in equation (2.20) ensures that their distributions are strictly stationary and highly non-Gaussian. The bivariate exponent measure of the resulting stationary max-stable process may be written as

$$V_{\mathscr{D}}(z_1, z_2) = \frac{1}{z_1}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_1}{z_2}\right)\right\} + \frac{1}{z_2}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_2}{z_1}\right)\right\}, \tag{2.33}$$

where $\mathscr{D} = \{x, x+h\} \subset \mathscr{X}$, $a = \{2\gamma(h)\}^{1/2}$, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Notice the correspondence with (2.27); as for the Smith model, the bivariate extremal coefficient is $\theta_2(h) = 2\Phi(a/2)$, but a takes a different value. Therefore, if the semi-variogram $\gamma(h)$ is unbounded, $\theta_2(h) \to 2$, as $\|h\| \to \infty$,

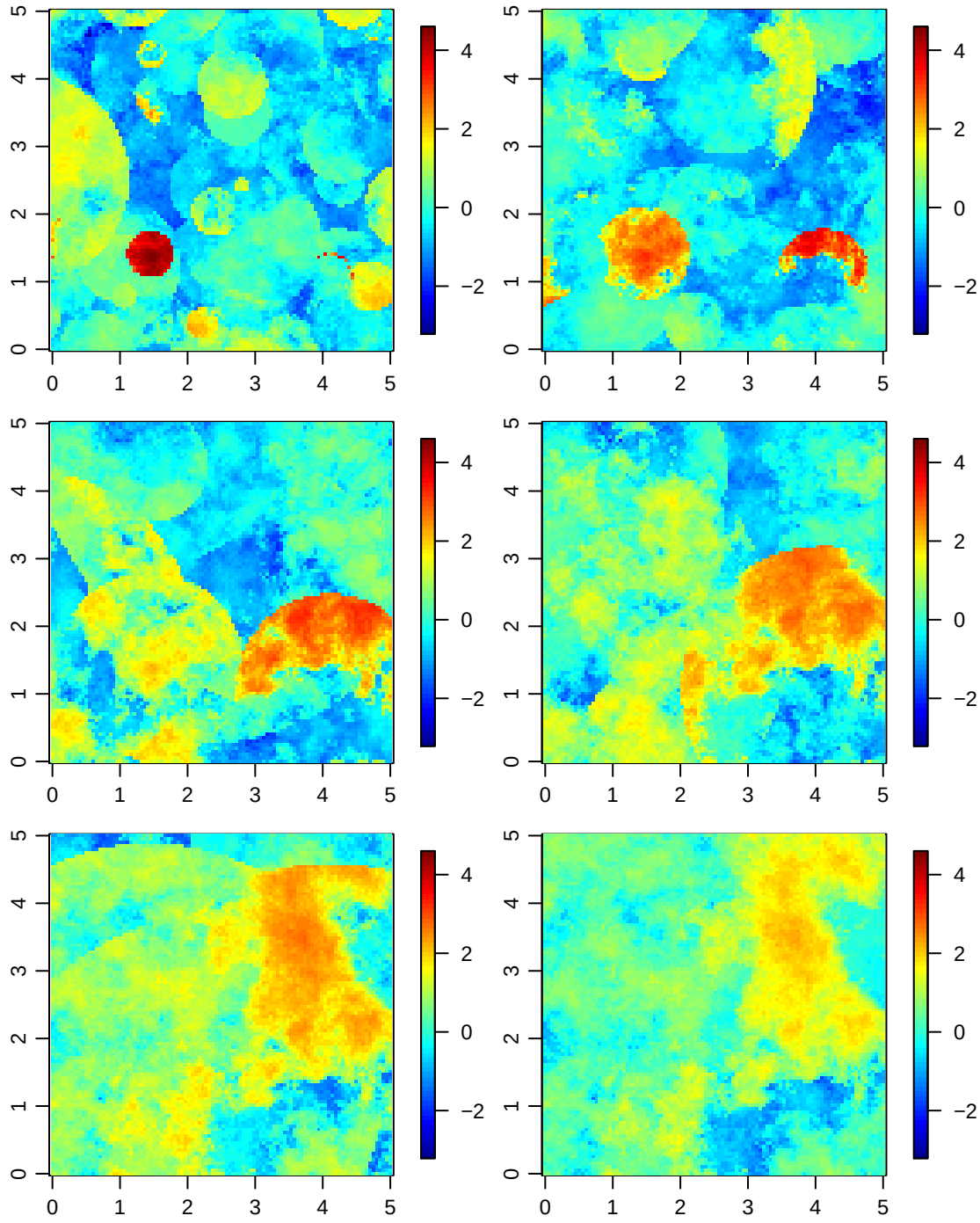

Figure 2.6: Realizations from the Schlather model with random set in $\mathbb{R}^2$, using Gumbel margins. The correlation function is $\rho(h) = \exp\{-\|h\|/2\}$ and the random sets are disks with radius distributed as a Gamma random variable with shape parameter 2 and mean $0.5, 1, 2, 3, 5$ (from left to right and top to bottom), the last having an infinite radius. The random seed used in these simulations was the same as for Figure 2.5.

and the process is mixing.

Gaussian processes are known to be asymptotically independent, recall (1.61). However, owing to the next result, Brown–Resnick processes turn out to be essentially the only limit of properly renormalized and stretched maxima of Gaussian processes.

Theorem 55 (Kabluchko *et al.*, 2009)**.** *Let $Y(x)$, $x \in \mathcal{X}$, be a zero mean stationary Gaussian process with a covariance function satisfying some mild conditions detailed in Kabluchko* et al. *(2009), and let $Y_i(x)$, $i = 1,2,\ldots$, be i.i.d. continuous replicates of $Y(x)$. Then there exist sequences b_n and $s_n > 0$ such that the process*

$$Z^n(x) = \max_{i=1,\ldots,n} b_n\{Y_i(s_n x) - b_n\}$$

converges in distribution, as $n \to \infty$, to a Brown–Resnick process $Z(x)$ with Gumbel margins.

In the result above, the sequence s_n acts as a stretching factor, so that the correlation of the process $Y(s_n x)$ is increasingly strong, as $n \to \infty$. A further result shown by Kabluchko *et al.* (2009) establishes that among all Brown–Resnick processes, only those corresponding to variograms of fractional Brownian motions, that is, of the form

$$\gamma(h) = \left(\frac{\|h\|}{\lambda}\right)^{\alpha},$$

where $\lambda > 0$ and $\alpha \in (0,2]$, arise as limits of suitably renormalized and stretched maxima of stationary and isotropic Gaussian processes. A more general statement is provided for non-isotropic random fields.

Brown–Resnick processes have been found to yield good fits in applications (Davison *et al.*, 2012; Jeon & Smith, 2012). Their flexibility, and the recent development of efficient inference procedures (Engelke *et al.*, 2012; Wadsworth & Tawn, 2013), make them particularly attractive.

Methods for simulating Brown–Resnick processes have been developed (Oesting *et al.*, 2012; Schlather, 2002; Ribatet, 2011, 2013), though they are more computationally intensive than for Schlather or Smith models. In Chapter 4, we show such simulations in $\mathbb{R}$ for different underlying semi-variograms.

More details about inference for the Brown–Resnick process can be found in Chapter 4. In particular, the question of efficiency of pairwise and triplewise likelihoods used in this context is addressed.

2.3.2.4 Extremal-t model

Recently, another interesting max-stable model, known as the extremal-t process and generalizing the Schlather model (2.28), has been proposed (Demarta & McNeil, 2005; Nikoloulopoulos *et al.*, 2009; Davison *et al.*, 2012; Opitz, 2013; Ribatet & Sedki, 2013; Ribatet, 2013). It assumes in the representation (2.20) that

$$W(x) = \sqrt{\pi} 2^{-\nu/2+1} \Gamma\left(\frac{\nu+1}{2}\right)^{-1} \max\{\varepsilon(x), 0\}^{\nu}, \qquad x \in \mathscr{X}, \tag{2.34}$$

where $\nu \geq 1$, $\Gamma(\cdot)$ is the Gamma function and $\varepsilon(x) \sim \mathrm{GF}\{0, \rho(h)\}$. The Schlather process is recovered when $\nu = 1$, while the limit when the correlation may be expressed as $\rho(h) \sim \exp\{-2\gamma(h)/\nu\}$, for some function $\gamma(h)$ and $\nu \to \infty$, is the Brown–Resnick process with semi-variogram $\gamma(h)$; see Nikoloulopoulos *et al.* (2009) and Davison *et al.* (2012). Furthermore, the extremal-t process can be viewed as the limiting process for rescaled maxima of independent Student t processes with correlation function $\rho(h)$ and ν degrees of freedom (Demarta & McNeil, 2005). In particular, it appears that Cauchy processes are in the max-domain of attraction of the Schlather model. The bivariate exponent measure of the extremal-t process is

$$V_{\mathscr{D}}(z_1, z_2) = \frac{1}{z_1} T_{\nu+1}\left[b\left\{(z_2/z_1)^{1/\nu} - \rho(h)\right\}\right] + \frac{1}{z_2} T_{\nu+1}\left[b\left\{(z_1/z_2)^{1/\nu} - \rho(h)\right\}\right], \tag{2.35}$$

where $\mathscr{D} = \{x, x+h\} \subset \mathscr{X}$, $T_{\nu}(\cdot)$ denotes the cumulative distribution function of a Student t random variable with ν degrees of freedom, and $b = \sqrt{(\nu+1)/\{1-\rho(h)^2\}}$. The bivariate extremal coefficients equals $\theta_2(h) = 2T_{\nu+1}[\sqrt{(\nu+1)\{1-\rho(h)\}/\{1+\rho(h)\}}]$. Hence, if $\rho(h) \to 0$, as $\|h\| \to \infty$, one has $\theta_2(h) \to 2T_{\nu+1}(\sqrt{\nu+1})$ and the process is not mixing, unless $\nu \to \infty$. In fact, $\theta_2(h) \to 2$, as $\nu \to \infty$, for any correlation $\rho(h) > -1$, so the parameter ν controls the degree of long-range dependence.

Simulated extremal-t processes with increasing degrees of freedom are displayed in Figure 2.7. Owing to the representation (2.34), naive simulation of the extremal-t process may be tricky for large ν.

2.3.2.5 Other models

In the same vein as the extremal-t model, so-called max-max-stable random fields (Robert, 2013b) are specified with

$$W(x) = \Gamma\left(1-\beta\right)^{-1} Z'(x)^{\beta}, \qquad x \in \mathscr{X}, \tag{2.36}$$

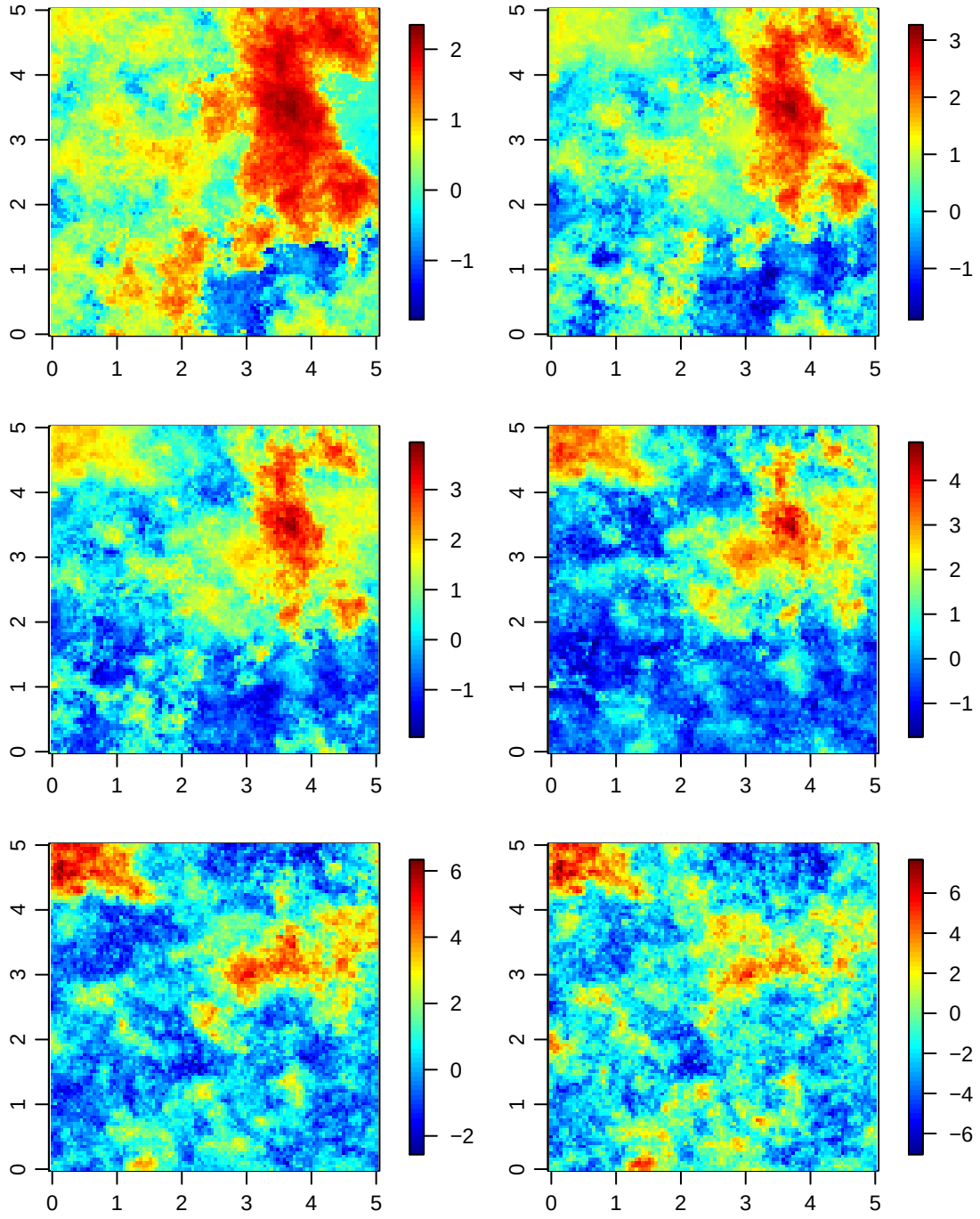

Figure 2.7: Realizations from the extremal-t model with correlation function $\rho(h) = \exp(-\|h\|/2)$ and degrees of freedom $\nu = 1$ (top left), which corresponds to the Schlather model, $\nu = 2$ (top right), $\nu = 3$ (middle left), $\nu = 5$ (middle right), $\nu = 10$ (bottom left) and $\nu = 20$ (bottom right). The random seed used in these simulations was the same as for Figure 2.5.

where $\beta \in (0,1]$ and $Z'(x)$ is a stationary max-stable random field itself with exponent measure $V'_{\mathscr{D}}(x_1,\ldots,x_D)$. The resulting stationary max-stable random field $Z(x)$, constructed from plugging (2.36) in (2.20), has exponent measure

$$V_{\mathscr{D}}(z_1,\ldots,z_D) = V'_{\mathscr{D}}(z_1^{1/\beta},\ldots,z_D^{1/\beta})^{\beta}, \tag{2.37}$$

where $\mathscr{D} = \{x_1,\ldots,x_D\} \subset \mathscr{X}$, and its bivariate extremal coefficient can be expressed as $\theta_2(h) = \theta'_2(h)^{\beta}$, where $\theta'_2(h)$ is the bivariate extremal coefficient of the underlying max-stable random field $Z'(x)$. Hence, the resulting process $Z(x)$ is increasingly dependent as $\beta \to 0$, whereas $Z(x) \equiv Z'(x)$ when $\beta = 1$. The simulation of max-max-stable processes can be computationally very intensive since it requires the simulation of replicates of the max-stable process $Z'(x)$, which may already be burdensome.

Another max-stable model in $\mathbb{R}^2$, proposed by Buishand *et al.* (2008) for the study of extreme precipitation in North Holland, but which is not very realistic due to its lack of invariance with respect to the coordinate axes, assumes in (2.20) that

$$W(x_1,x_2) = \exp\{B_1(\lambda x_1) + B_2(\lambda x_2) - \lambda(|x_1|+|x_2|)/2\}, \qquad (x_1,x_2) \in \mathscr{X} \subset \mathbb{R}^2, \tag{2.38}$$

where $B_1(t)$ and $B_2(t)$, $t \in \mathbb{R}$, are double-sided Brownian motions. In fact, the model (2.38) is a special Brown–Resnick model, using the non-isotropic semi-variogram $\gamma(x_1,x_2) = \lambda(|x_1|+|x_2|)/2$.

A last model, the so-called Voronoï max-stable random field (Lantuéjoul *et al.*, 2011; Robert, 2013b), consists in letting $W(x) = U(x)^{\kappa}/\mathrm{E}\{U(x)^{\kappa}\}$, $x \in \mathscr{X}$, provided the expectation exists, where $\kappa > 0$ and $U(x)$ is a random process defined as

$$U(x) = \begin{cases} \sum_{j\geq 1} c_j I(x \in R_j), & x \in \mathbb{R}^d \setminus \Delta, \\ 0, & x \in \Delta, \end{cases}$$

where the c_j's are i.i.d. positive random variables, and the R_j's and Δ are sets specified in terms of a Poisson–Voronoï tessellation. Although the exponent measure is known, realizations from this model are piecewise constant, so not realistic for the modeling of natural phenomena such as rainfall or temperature, and it is not useful to fit it to such data.

2.3.3 Models based on α-stable random effects

Positive stable, or α-stable, random variates, denoted PS(α), are known to have a Laplace transform of the form

$$\mathrm{E}\{\exp(-tS)\} = \exp(-t^{\alpha}), \qquad t \geq 0, \tag{2.39}$$

where $\alpha \in (0,1]$ (Fougères *et al.*, 2009), and although their density cannot be expressed in closed form, they can be easily simulated using the representation

$$S = \left\{ \frac{h(\pi U)}{-\log(W)} \right\}^{(1-\alpha)/\alpha}, \tag{2.40}$$

where U and W are independent uniform random variables on the interval $[0,1]$, and

$$h(\omega) = \left\{ \frac{\sin(\alpha\omega)}{\sin(\omega)} \right\}^{1/(1-\alpha)} \frac{\sin\{(1-\alpha)\omega\}}{\sin(\alpha\omega)};$$

see Stephenson (2009). Taking advantage of (2.39), Stephenson (2009) showed that the multivariate extension of the asymmetric logistic distribution (1.31) admits a conditional representation in terms of α-stable variates, and Fougères *et al.* (2009), Reich & Shaby (2012) and Shaby & Reich (2012) used this fact to construct hierarchical max-stable processes based on latent α-stable spatial random effects. Specifically, let $Z(x) = R(x)S(x)$, $x \in \mathscr{X}$, where $R(x)$ is a multiplicative "nugget effect" and $S(x)$ is a spatial random field defined as

$$\begin{aligned} R(x) &\overset{\text{iid}}{\sim} \mathrm{GEV}(1,\alpha,\alpha) \qquad \Longleftrightarrow \qquad Z(x) \mid S(x) \overset{\text{iid}}{\sim} \mathrm{GEV}\{S(x), \alpha S(x), \alpha\}, \\ S(x) &= \left\{ \sum_{l=1}^{L} S_l w_l(x)^{1/\alpha} \right\}^{\alpha}, \\ S_l &\overset{\text{iid}}{\sim} \mathrm{PS}(\alpha), \end{aligned} \tag{2.41}$$

and $w_l(x) \geq 0$ are deterministic spatial kernels satisfying $\sum_{l=1}^{L} w_l(x) = 1$ for all $x \in \mathscr{X}$. Then, the joint distribution of the process $Z(x)$ at D sites $x_1, \ldots, x_D \in \mathscr{X}$ can be expressed as

$$\Pr\{Z(x_1) \leq z_1, \ldots, Z(x_D) \leq z_D\} = \exp\left(-\sum_{l=1}^{L} \left[\sum_{i=1}^{D} \left\{ \frac{z_i}{w_l(x_i)} \right\}^{-1/\alpha} \right]^{\alpha} \right). \tag{2.42}$$

Hence, the process $Z(x)$, $x \in \mathscr{X}$, is simple max-stable with asymmetric logistic dependence structure (Coles & Tawn, 1991; Stephenson, 2009). This model is neither

stationary nor isotropic, since the bivariate extremal coefficient

$$\theta_2(x_1, x_2) = \sum_{l=1}^{L} \left\{ w_l(x_1)^{1/\alpha} + w_l(x_2)^{1/\alpha} \right\}^{\alpha},$$

is not homogeneous over space. However, using Gaussian kernels centered at regularly spaced knots, this model can be shown to converge to the Smith model (2.26), as the number of knots $L \to \infty$ (Reich & Shaby, 2012). Direct computation of the density $f(z_1, \ldots, z_D)$ from expression (2.42) is awkward, but the conditional representation (2.41) permits one to write down the likelihood as

$$L(\alpha) = \mathrm{E}\left\{ f(z_1, \ldots, z_D) \mid S_1, \ldots, S_L \right\} = \mathrm{E}\left\{ \prod_{i=1}^{D} g_i(z_i) \right\}, \tag{2.43}$$

where $g_i(\cdot)$ is the density of the GEV distribution with location parameter $S(x_i)$, scale parameter $\alpha S(x_i)$ and shape parameter α. Maximum likelihood inference for the estimation of α is difficult (but see §3.3.3.1), since the approximation of the L-fold integration in the right-hand side of (2.43) may be too variable to be reliable using Monte Carlo techniques, but Bayesian inference can be performed with MCMC methods, combining (2.43) and the efficient simulation scheme for the auxiliary α-stable variates (2.40).

In Chapter 3, we construct a similar model for time series, in order to assess the loss in efficiency of pairwise likelihood estimators for extremal models in this context, compared to maximum likelihood.

2.3.4 Max-stable processes for threshold exceedances

Most of the max-stable processes described above are defined in terms of an underlying correlation function or a variogram, and the model (2.30) also relies on a random set element. By carefully choosing suitable space-time correlation functions $\rho(h_s, h_t)$, variogram $\gamma(h_s, h_t)$ and random sets $\mathscr{A}$, one may also fit these max-stable models to space-time data; see Chapter 5 for an application to rainfall extremes.

Suppose that a stationary space-time process $Y(s,t)$, $(s,t) \in \mathscr{S} \times \mathscr{T}$, has been sampled at S stations and T times, and let $Y_{s,t}$ denote the observation at the sth station and tth time. Without loss of generality, assume that $Y(s,t)$ has unit Fréchet margins. Under mild conditions, maxima of independent replications of $Y(s,t)$ may be well approximated by a simple max-stable process in space and time, $Z(s,t)$ say. Usually in the space-time framework, we only have one replicate of the process at our disposal, but treating different parts of the data as independent (for example summers in

distinct years), we can "create" artificial space-time replications of the process $Y(s,t)$, and fit the max-stable model $Z(s,t)$ to the resulting process of maxima. However, since this approach requires a lot of data to be reliable, it has not been used in practice.

When the original observations $Y_{s,t}$, $s = 1,\dots,S$, $t = 1,\dots,T$, are available, approaches based on threshold exceedances are preferred. Loosely speaking, like the multivariate case with (1.47), it can be shown that under mild technical conditions, the dependence structure of very high spatial events, not necessarily maxima, converges to that of a max-stable process (see de Haan & Ferreira, 2006, p.297), so max-stable models can also be applied for excesses of very large thresholds. That is, letting $u > 0$ be a high threshold, one has

$$\Pr\{Y(s,t) \le y(s,t), (s,t) \in \mathscr{D}\} \approx \Pr\{Z(s,t) \le y(s,t), (s,t) \in \mathscr{D}\}, \tag{2.44}$$

where $y(s,t)$ is a function defined on the compact set $\mathscr{D} \subset \mathscr{S} \times \mathscr{T}$ such that $y(s,t) > u$. In Section 3.2.2, we shall see how the approximation (2.44) may be used for inference, using a censored pairwise likelihood approach.

2.4 Asymptotic independence and related models for spatial extremes

For a broad class of stationary processes with unit Fréchet margins, we have (recall §1.2.4.1)

$$\Pr\{Z(x_1) > z \mid Z(x_2) > z\} \sim \mathscr{L}_h(z) z^{1-1/\eta(h)}, \quad x_1, x_2 \in \mathscr{X}, \tag{2.45}$$

where $h = x_1 - x_2$ is the lag vector, $\eta(h)$ is the coefficient of tail dependence, and $\mathscr{L}_h(z)$ is a slowly varying function for any h (Ledford & Tawn, 1996). In particular, simple max-stable processes have $\eta(h) = 1$ for all h. Thus, they are asymptotically dependent in the sense that

$$\chi_h = \lim_{z\to\infty} \Pr\{Z(x_1) > z \mid Z(x_2) > z\} = 2 - \theta_2(h), \quad x_1, x_2 \in \mathscr{X}, \tag{2.46}$$

where $\theta_2(h)$ is the bivariate extremal coefficient (2.19), and χ_h may be strictly positive. Recall the definition of asymptotic independence in (1.60) for the bivariate case. In practice it may be difficult to identify independence of extremes based on finite samples, since the data may display residual dependence for any finite threshold, however high, recall Figure 1.10. Asymptotic independence models, for which $\eta(h) < 1$ and the limit in equation (2.46) equals zero, but which can also model the dependence present before the limit is reached, may therefore be preferred for modeling at finite thresholds. The Gaussian model is asymptotically independent for all correlations

$\rho(h) \neq 1$, but Gaussian processes are too restrictive in the bulk of extremal applications, so broader classes of models are needed to allow flexible modeling.

In this section, we present the classical Gaussian copula model, the spatial extension of multivariate inverted max-stable distributions (1.64), and finally hybrid models able to capture both asymptotic independence and asymptotic dependence, which were proposed by Wadsworth & Tawn (2012).

2.4.1 Gaussian copula model

Gaussian processes may be transformed to the unit Fréchet scale as follows:

$$Z(x) = -\frac{1}{\log[\Phi\{S(x)\}]}, \qquad x \in \mathscr{X}, \tag{2.47}$$

where $S(x) \sim \mathrm{GF}\{0, \rho(h)\}$ is a Gaussian process, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Although the process $Z(x)$, $x \in \mathscr{X}$, has unit Fréchet margins, all finite-dimensional distributions are based on the Gaussian copula. Hence, as for Gaussian processes, $Z(x)$ is asymptotically independent, with coefficient of tail dependence $\eta(h) = \{1 + \rho(h)\}/2$.

2.4.2 Inverted max-stable processes

The class of inverted max-stable processes, which extends the multivariate models (1.64) to the spatial framework, has been proposed by Wadsworth & Tawn (2012). They are defined in terms of a max-stable process $Z'(x)$ as

$$Z(x) = -1/\log[1 - \exp\{-1/Z'(x)\}], \quad x \in \mathscr{X}, \tag{2.48}$$

and provide spatial models for asymptotic independence. For these processes, $\eta(h) = 1/\theta_2(h)$, where $\theta_2(h)$ is the bivariate extremal coefficient of $Z'(x)$. With this construction, each max-stable model $Z'(x)$ may be transformed to provide an asymptotically independent counterpart $Z(x)$. The bivariate margins of $Z(x)$ have the distribution

$$\Pr\{Z(x_1) \leq z_1, Z(x_2) \leq z_2\} = -1 + \exp(-1/z_1) + \exp(-1/z_2) + \exp[-V\{s(z_1), s(z_2)\}], \tag{2.49}$$

where V is the exponent measure of the bivariate extreme-value distribution corresponding to $\{Z'(x_1), Z'(x_2)\}$, and $s(z) = -1/\log\{1 - \exp(-1/z)\}$. Hence, the partial derivatives of the right-hand side of (2.49) with respect to z_1 and z_2 involve the first

derivative of the transformation $s(z)$; in particular, the bivariate density is

$$\frac{\partial^2}{\partial z_1 \partial z_2} \Pr\{Z(x_1) \leq z_1, Z(x_2) \leq z_2\} = s'(z_1) s'(z_2) g\{s(z_1), s(z_2)\}, \tag{2.50}$$

where $g(z_1, z_2) = \{V_1(z_1, z_2) V_2(z_1, z_2) - V_{12}(z_1, z_2)\} \exp\{-V(z_1, z_2)\}$, $V_1 = \partial V(z_1, z_2)/\partial z_1$, etc., and $s'(z)$ is the first derivative of the function $s(z)$; see Appendix B.2.2.

2.4.3 Hybrid models

Although max-stable models should be suitable for the modeling of extremely high threshold exceedances (recall §2.3.4), asymptotic independence models may provide better fits at finite thresholds. Since in practice it is difficult, if not impossible, to determine whether a dataset should be modeled using an asymptotically dependent or asymptotically independent process, Wadsworth & Tawn (2012) introduced a novel class of models, so-called hybrid models, which may be dependent in the limit but more flexible than max-stable models at subasymptotic levels. Loosely, the basic idea is to mix max-stable and asymptotic independence processes.

Let $Z_1(x)$, $x \in \mathscr{X}$, be a stationary simple max-stable process with bivariate extremal coefficient $\theta_2(h)$, and $Z_2(x)$, $x \in \mathscr{X}$, be a stationary asymptotic independence model with Fréchet margins and coefficient of tail dependence $\eta(h)$, independent of $Z_1(x)$, and satisfying (2.45). Then for $a \in [0, 1]$, the spatial process defined by the max-mixture

$$Z(x) = \max\{a Z_1(x), (1-a) Z_2(x)\}, \qquad x \in \mathscr{X}, \tag{2.51}$$

has unit Fréchet margins and bivariate conditional exceedance probability of the form

$$\Pr\{Z(x_1) > z \mid Z(x_2) > z\} \sim a\{2 - \theta_2(h)\} + (1-a)^{1/\eta(h)} \frac{\mathscr{L}_h\{z/(1-a)\}}{z^{1/\eta(h)-1}}, \tag{2.52}$$

as $z \to \infty$. Max-stable models, with $a = 1$, may be too restrictive in the sense that they have only the first-order term in (2.52), while asymptotic independence models, with $a = 0$, may be unreliable since they are left with the second term only; recall (2.45) and (2.46). Hence, hybrid models seem to provide a good balance between the two classes.

Moreover, the first term on the right-hand side of (2.52), which corresponds to the max-stable part of $Z(x)$, vanishes if $a = 0$ or $\theta_2(h) = 2$. Hence, in the case where $a \neq 0$ and $\theta_2(h) = 2$ for $h > h_0$, asymptotic dependence is present at short distances while asymptotic independence prevails at larger ones. This can be especially interesting for the modeling of space-time data, where independence is usually observed at moderate time lags.

Although hybrid models are appealing in all these respects, they may be difficult to fit owing to the large number of parameters, and the likely poor identifiability of the proportion parameter a. Inference can be performed using the pairwise margins of $Z(x)$, which may be expressed in terms of the pairwise margins of $Z_1(x)$ and $Z_2(x)$ as

$$\begin{aligned}\Pr\{Z(x_1) \leq z_1, Z(x_2) \leq z_2\} \; = \\ \Pr\left\{Z_1(x_1) \leq \frac{z_1}{a}, Z_1(x_2) \leq \frac{z_2}{a}\right\} \Pr\left\{Z_2(x_1) \leq \frac{z_1}{1-a}, Z_2(x_2) \leq \frac{z_2}{1-a}\right\}. \quad (2.53)\end{aligned}$$

Although the task is tedious, the pairwise density can be derived by differentiating the right-hand side of (2.53) with respect to z_1 and z_2; see Appendix B.2.3.

2.5 Measures of extremal dependence

Several measures of extremal dependence have been proposed for stationary processes, most of which are simple extensions of the multivariate case described in §1.2.5. The extremal coefficient $\theta_D(\boldsymbol{h})$, introduced in (2.19), is suitable for asymptotically dependent processes, whose renormalized maxima converge to a non-trivial max-stable process. Alternatively, the concept of madogram, an analogue of the variogram for extremal processes, finds its roots in the PhD thesis by Cooley (2005) and the pioneering works by Cooley *et al.* (2006b) and Naveau *et al.* (2009). By contrast with the variogram, which depends on second-order moments, the madogram is well-defined for any max-stable process $Z(x)$, $x \in \mathscr{X}$, with arbitrary GEV margins $G(x)$, and may be defined as

$$v(h) = \frac{1}{2}\mathrm{E}\left[|G\{Z(x+h)\} - G\{Z(x)\}|\right]. \quad (2.54)$$

It turns out that this quantity can be expressed in terms of the extremal coefficient as $\theta_2(h) = \{1+2v(h)\}/\{1-2v(h)\}$, and is therefore a measure of dependence between similarly extreme events observed at the sites x and $x+h$. More generally, the λ-madogram, defined as

$$v_\lambda(h) = \frac{1}{2}\mathrm{E}\left[\left|G^\lambda\{Z(x+h)\} - G^{1-\lambda}\{Z(x)\}\right|\right], \quad (2.55)$$

where $\lambda \in (0,1)$, satisfies $V_{\mathscr{D}}(\lambda, 1-\lambda) = \{c(\lambda) + v_\lambda(h)\}/\{1 - c(\lambda) - v_\lambda(h)\}$, and $c(\lambda) = 3/\{2(1+\lambda)(2-\lambda)\}$, where $V_{\mathscr{D}}$ is the underyling exponent measure of $Z(x)$ for the set of sites $\mathscr{D} = \{x, x+h\} \subset \mathscr{X}$, and therefore characterizes completely the joint distribution of $Z(x)$ and $Z(x+h)$. Empirical estimators for the bivariate extremal coefficient, which rely on the assumption of max-stability, have been suggested by Schlather & Tawn (2003), recall (1.66), and Naveau *et al.* (2009), the latter based on the madogram. These

estimators appear to yield satisfactory results within the max-stable framework, but cannot distinguish different degrees of asymptotic independence.

The coefficient of tail dependence $\eta(h)$, introduced in (2.45) and §1.2.4.1, is suited for asymptotically independent variables, and can be estimated by maximum likelihood, recall (1.67). However, since the case $\eta(h) = 1$ corresponds to the entire class of max-stable processes, it provides no information about the strength of dependence of a given max-stable process. Moreover, discriminating between asymptotic dependence and asymptotic independence is difficult, because the dependence between variables may vanish very slowly as the level increases; recall Figure 1.10.

Coles *et al.* (1999) suggested model-free diagnostics $\chi_h(u)$ and $\overline{\chi}_h(u)$ for distinguishing among these different types of tail dependence. Denoting by $Y(x)$, $x \in \mathscr{X}$, the process of interest, these coefficients can be expressed as

$$\chi_h(u) = 2 - \frac{\log C(u,u)}{\log u}, \qquad \overline{\chi}_h(u) = \frac{2\log(1-u)}{\log \overline{C}(u,u)} - 1, \quad 0 \leq u \leq 1, \tag{2.56}$$

where C and $\overline{C}$ are, respectively, the copula and survival copula of the random vector $\{Y(x), Y(x+h)\}$; recall the bivariate analogues in (1.68) and (1.69). The limits $\chi_h = \lim_{u\to 1} \chi_h(u)$ and $\overline{\chi}_h = \lim_{u\to 1} \overline{\chi}_h(u)$ determine the type of tail dependence. Hence, for fixed lag h, the pair of diagnostics $\{\chi_h(u), \overline{\chi}_h(u)\}$ can be used as a tool to detect asymptotic independence when u approaches one; or, alternatively, for a fixed extreme level u, it can serve as a "correlogram" for extreme events, when considered as a function of h. These dependence measures are often estimated by their empirical rank-based counterparts, though the model-based estimators mentioned above will be more efficient, at least when the underlying model is reasonable. Further details and discussion can be found in §1.2.5.3.

Another large family of "correlograms" for extremal time series, so-called extremograms, has been proposed by Davis & Mikosch (2009). Let $\boldsymbol{Y}_t$ be a strictly $\mathbb{R}^D$-valued time series. Under suitable conditions, the sequence of conditional probabilities

$$\Pr\left(\boldsymbol{a}_n^{-1}\boldsymbol{Y}_{t+h} \in B \mid \boldsymbol{a}_n^{-1}\boldsymbol{Y}_t \in A\right) \tag{2.57}$$

converges, as $n \to \infty$, to some real function $\varrho_{AB}(h)$, for some increasing sequence of real vectors $\boldsymbol{a}_n$, and any Borel sets A and B of $\mathbb{R}^d$ bounded away from zero. Such limits are called *extremograms*. When $D = 1$ and $A = B = (y, \infty)$, $y > 0$, the resulting extremogram is just χ_h, seen as a function of h, but in general many different forms are possible. Davis & Mikosch (2009) discuss the large-sample properties of the empirical extremogram under α-mixing conditions, and address estimation in depth.

2.6 Inference for extremal models

Owing to the complicated form of the density stemming from differentiation of (2.23), classical likelihood inference for the parameters of max-stable models is not possible in general, and recourse has been made to composite likelihoods (Varin *et al.*, 2011) based on lower-order marginal densities, such as bivariate margins of all pairs of maxima. These pseudo-likelihoods are very general and are applicable to a large variety of models. Furthermore, under mild conditions, maximum composite likelihood estimators are strongly consistent and have asymptotic normal distributions, though they may be much more variable than ordinary maximum likelihood estimators (see Chapter §3 and Davis & Yau, 2011). The composite likelihood information criterion, CLIC, the analogue of the Akaike information criterion (AIC) for composite likelihoods, allows model comparison (Varin & Vidoni, 2005; Varin, 2008), and CLIC*, a scaled version of the CLIC, has been advocated by Davison & Gholamrezaee (2012). More details about composite likelihoods may be found in Chapter 3.

More recently, inference methods based on a full likelihood have been proposed for a special class of max-stable processes that includes the Brown–Resnick processes (see §4.4.1 and Wadsworth & Tawn, 2013), and Engelke *et al.* (2012) proposed to fit Brown–Resnick processes based on the full likelihood of "extremal increments" of the process. From the Bayesian perspective, Ribatet *et al.* (2012) have developed a Monte Carlo Markov chain algorithm for fitting such models by sampling the pseudo composite posterior distribution using a modified acceptance rate. The Bayesian hierarchical models of Reich & Shaby (2012) and Shaby & Reich (2012), based on α-stable random effects and introduced in §2.3.3, can be fitted using standard MCMC methods, thanks to their nice conditional representation.

When individual events are recorded, more efficient inference is feasible. Following Stephenson & Tawn (2005), Davison & Gholamrezaee (2012) and Wadsworth & Tawn (2013) show how to incorporate the occurrence times of extreme events, use of which both simplifies the likelihood and allows much more efficient inference in cases of moderate to low spatial dependence, recall the simulation study in §1.2.2.2. Alternatively, since the max-stable models are suitable only above some predetermined high threshold, recall §2.3.4, inference can be made using a censored threshold-based approach; see §3.2.2, Jeon & Smith (2012), Huser & Davison (2013b) and Thibaud *et al.* (2013). However, when the data are temporally correlated, the standard results on composite likelihoods do not directly apply; this issue is addressed in Chapter 5.

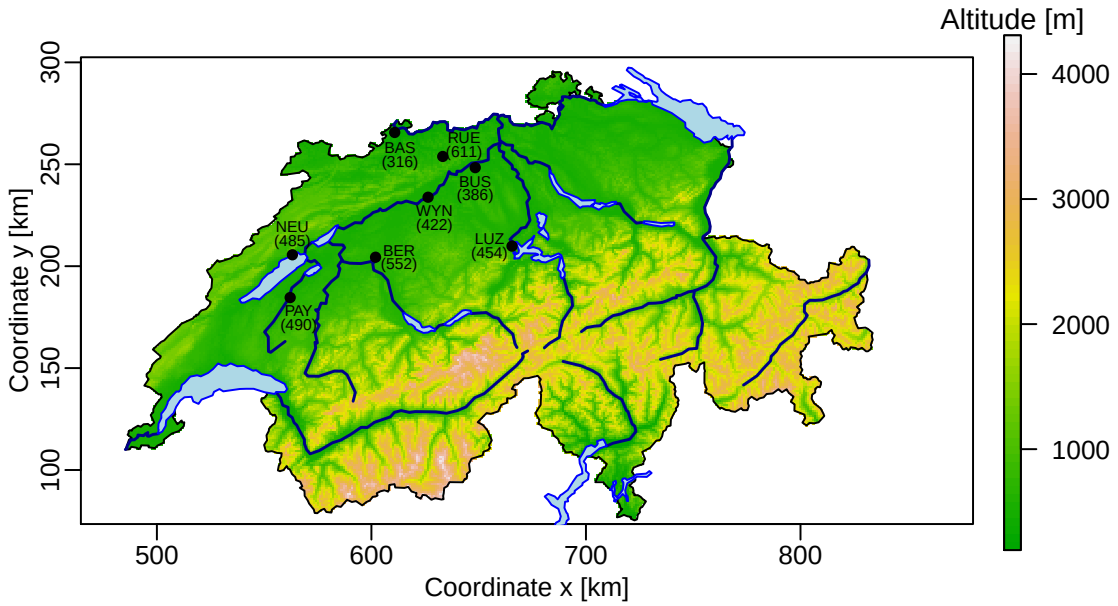

Figure 2.8: Topographic map of monitoring sites (with their altitudes) from *MeteoSwiss*, where temperature and precipitation data were recorded. The most distant sites are 107km apart, while the closest are 16km apart.

2.7 Application

In this section, we discuss a small spatial application of the theory presented in this chapter (also published in Davison *et al.*, 2013), using data on air temperature and precipitation, and find that asymptotic dependence models may be preferred for the first dataset but that asymptotic independence models seem to fit better for the second. Gaussian models perform less well in both cases. We consider winter daily temperature minima and summer daily cumulative rainfall, recorded at eight monitoring sites with similar altitudes and located in the so-called plateau region of Switzerland; see Figure 2.8. The data were available from 1981 to 2012, giving a total of about 2900 observations per site. For simplicity, we treat these daily data as independent over time, although this is false at least for the temperature data. A more complex spatio-temporal study is performed in Chapter 5 (see also Huser & Davison, 2013b), where further details about part of the rainfall data may be found.

We first transformed the temperature data by multiplication by -1, and then fitted the generalized Pareto distribution (1.23) to model events above the 98% quantile, u_{98}, of the time series at each site separately, and used this fitted model to transform the data to have the unit Fréchet distribution. Since Bortot *et al.* (2000) and Coles & Pauli (2002) have shown that the choice between asymptotic dependence and

asymptotic independence models can influence extrapolation and association of extremes far more than the particular model used in each of these model classes, we fitted only a limited selection of spatial correlations for this illustrative analysis. For each dataset, we used the censored pairwise threshold-based (3.17) to fit three models to the exceedances over u_{98}:

(a) the max-stable Brown–Resnick model defined in (2.32), with variogram $2\gamma(h) = \tau^2 I(\|h\| > 0) + (\|h\|/\lambda)^\alpha$, where τ^2 is a nugget effect, $I(\cdot)$ denotes the indicator function, $\lambda > 0$ is a range parameter and $\alpha \in (0,2]$ is a smoothness parameter;

(b) the corresponding inverted max-stable model; recall (2.48);

(c) the Gaussian copula model (2.47) with correlation function $\rho(h) = \exp\{-2\gamma(h)\}$.

Larger values of α give smoother processes in each case. For the temperature data, α was always estimated very close to its upper boundary, so we set $\alpha = 2$ and estimated only the remaining parameters. Table 2.4 reports the estimated parameters and confidence intervals, and the corresponding scaled values of the log pairwise likelihood and CLIC* (see §3.1.4), for each model and dataset. There is a strong difference in the scales and in the smoothness of the processes. The temperature data have $\widehat{\lambda} = 210$km or more, and $\alpha = 2$, corresponding to a relatively smooth process with large-scale dependence, though local variation is accommodated by the nugget τ^2, which is significantly larger than zero. The rainfall data have $\widehat{\lambda} > 3$km and $0 < \widehat{\alpha} < 1$, corresponding to a much rougher process, now with a nugget whose confidence interval almost includes zero. There is a clear trade-off between including a nugget to allow local variation in a max-stable process, and using an asymptotically independent process, in which local variation can be stronger; the availability of data at neighboring locations might allow better discrimination between these. In each case the value of $\widehat{\lambda}$ is smallest for the max-stable process and seems unrealistically large for the Gaussian process. The rather wide confidence intervals for λ for the temperature data probably stem from the difficulty of estimating continental-scale events from data over a limited region of a small country.

Figure 2.9 displays binned empirical estimates of the extremal coefficient $\theta_2(h)$ for the temperature data, for which the best model is max-stable, and of the coefficient of tail dependence $\eta(h)$ for the rainfall data, which seem asymptotically independent. Comparing these estimates to their fitted counterparts, it seems that the models capture spatial dependence quite flexibly. The graphs confirm that the natural processes considered here have decreasing extremal dependence with increasing distance, although dependence remains strong at long distances for temperatures, perhaps explaining

Table 2.4: Estimated dependence parameters $\widehat{\tau}^2$, $\widehat{\lambda}$ and $\widehat{\alpha}$ (with 95% confidence intervals based on block bootstrap using seasonal blocks) from the models (a), (b) and (c), along with the value of the log-pairwise likelihood, $\widehat{\ell}$, evaluated at the parameter estimates and scaled in such a way that its value is comparable to the log likelihood under independence, as suggested by Davison & Gholamrezaee (2012). The values of the CLIC*, a scaled version of the CLIC, are also reported.

Dataset	Model	$\tau^2 \times 10^2$	λ [km] $\times 10^{-1}$	α	$\widehat{\ell}$	CLIC*
Temp.	(a) MS	58 (50,71)	21 (15,579)	2 (–)	−4289.6	8599.2[†]
	(b) IMS	4 (3,6)	77 (58,668)	2 (–)	−4295.6	8604.5
	(c) Gauss	4 (3,6)	80 (60,3062)	2 (–)	−4291.8	8599.9
Rainfall	(a) MS	11 (0,272)	0.3 (0.1,3.7)	0.55 (0.36,1.56)	−4836.9	9711.9
	(b) IMS	6 (0,28)	11 (7,23)	0.92 (0.52,1.95)	−4829.4	9673.0[‡]
	(c) Gauss	4 (0,21)	31 (16,111)	0.65 (0.38,1.76)	−4829.9	9689.9

MS: max-stable model; IMS: inverted max-stable model; Gauss: Gaussian model.
[†] Our preferred model for the temperature data.
[‡] Our preferred model for the rainfall data.

the difficulty in estimating the range parameter. The huge uncertainty in these plots would be reduced by taking more sites. Moreover, given that the monitoring sites are at least 16km apart, small local variation represented by the nugget is difficult to estimate.

While the best model for the temperature data, according to the CLIC* and the value of the log-pairwise likelihood, is max-stable, the rainfall data appear to be asymptotically independent. Since max-stable models provide the only possible limiting dependence structure, the fit of these models should be better for higher thresholds, but as finite thresholds must be considered in practice, asymptotic independence models might provide better fits to the available data. If interest resides in extremely high joint return levels, it could be misleading to base inference on asymptotic independence models. To assess this, Figure 2.10 shows the empirical values for $\chi_h(u)$ and $\overline{\chi}_h(u)$ for the sites NEU (Neuchâtel) and PAY (Payerne), and their model-based counterparts. Although the fits at finite thresholds are similar for the different models, they differ at very extreme thresholds, and the probabilities of very extreme events might be strongly underestimated if an asymptotic independence model is used. The right panels of Figure 2.10 display joint return levels for some spatial functionals of the daily temperature and rainfall data recorded at NEU and PAY. At observed timescales, the curves for the different models give more or less the same predictions, but at the 1000 year-return period, a discrepancy of around 7.6mm appears for the rainfall data, underlining the ability of the max-stable model to capture dependence at high levels.

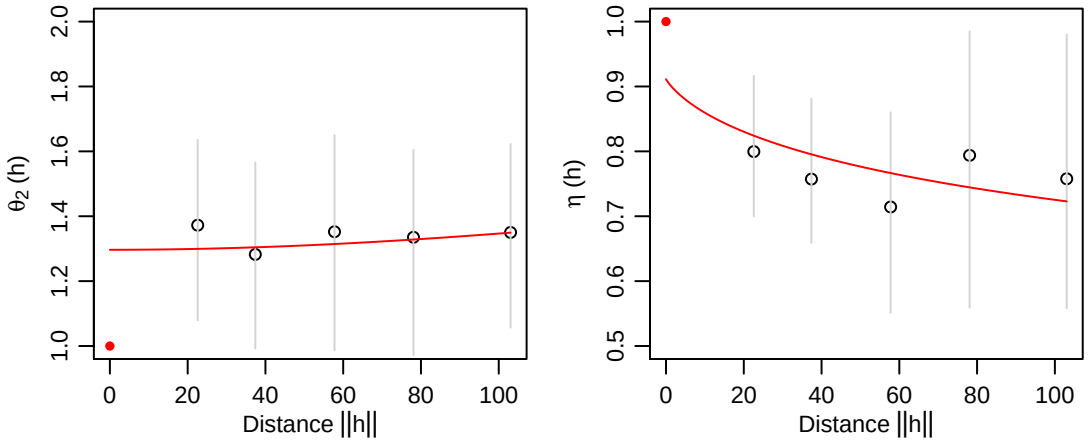

Figure 2.9: Bivariate extremal coefficient $\theta_2(h)$ (left) and coefficient of tail dependence $\eta(h)$ (right), corresponding to the best fitted models for the temperatures and rainfall data, respectively. According to Table 2.4, these models are respectively max-stable and inverted max-stable. Circles are binned empirical estimates with 95% confidence intervals in grey, while red solid lines are model-based estimates. The red points represent theoretical values of $\theta_2(0)$ and $\eta(0)$.

This effect would probably increase for longer return periods and if more sites were to be considered simultaneously.

In summary, if there were any guarantees that the data we consider are indeed asymptotically independent, the Gaussian or inverted max-stable models would potentially be suitable, but since it is very difficult, if not impossible, to have insight so far into the tail, it is safer to base risk assessments on max-stable models, which provide upper bounds for joint probabilities of extreme events. In fact, extrapolation to estimate probabilities of events never yet seen rests entirely on assumptions about the behavior of distribution tails, but these can only be verified with respect to events that have already occurred. Thus the likely validity of the model underlying the extrapolation needs exceptionally careful consideration. This task at first appears impossible, but, if used with attention and supplemented with subject-matter knowledge, the ideas sketched in this chapter provide initial steps towards a quantitative understanding of spatial rare events.

2.8 Summary

In this chapter, we have given a broad overview of models for spatial extremes, and have in particular underlined the usefulness of max-stable processes. We have de-

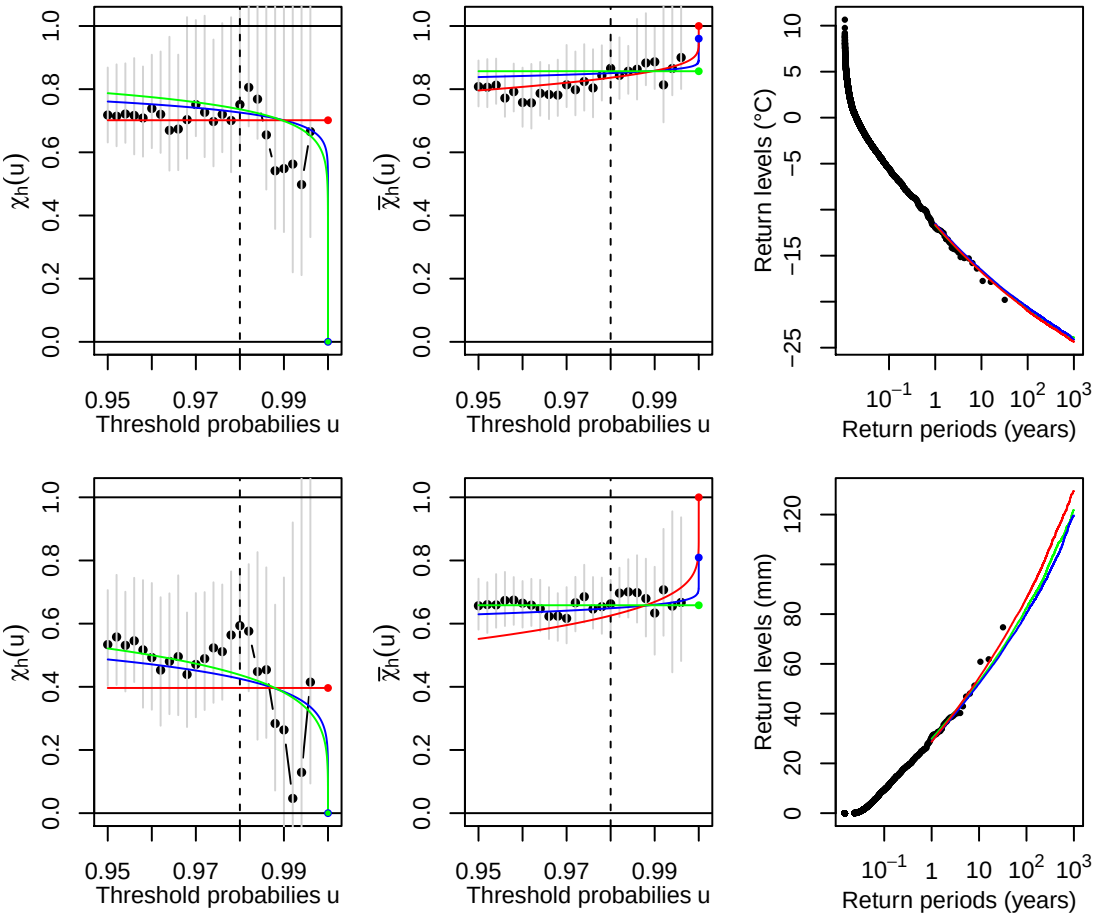

Figure 2.10: Extremal diagnostics computed for the temperature (top row) and rainfall (bottom row) data for the pair of monitoring sites NEU and PAY. *Left–Middle columns*: Empirical estimates (black dots) of $\chi_h(u)$ (left) and $\overline{\chi}_h(u)$ (middle) with their model-based counterparts: The solid lines show the fitted curves for (a) the max-stable model (red), (b) the inverted max-stable model (green) and (c) the Gaussian model (blue). The grey vertical lines are 95% confidence intervals from a block bootstrap using seasonal blocks, and the vertical dashed line is the threshold used when fitting the dependence models. *Right column*: Empirical and model-based return levels for the spatial average (over both sites) of daily temperature minima (top) and cumulative rainfall (bottom), on the original scale.

scribed the main max-stable models that have appeared in the literature, showing their similarities and dissimilarities. We have also seen that other types of model, some of which are asymptotically independent and can capture the rate of decay towards the limit, may be more appropriate when dependence vanishes at high levels. Although this chapter was essentially based on the existing literature, the application in §2.7 is a novel contribution, in which the distinction between max-stable and

asymptotically independent models is clearly made, and studied in depth with two different datasets (rainfall and temperature) recorded in Switzerland. This application has appeared as a review paper about spatial extremes in Davison *et al.* (2013).

With these tools in hands, the next chapter tackles a major issue that arise when one wants to fit such extremal models, namely how to perform inference. As has already been mentioned in this chapter, inference based on the full likelihood is usually unattainable for max-stable processes or related models, and pairwise likelihood approaches seem to be reasonable alternatives. In Chapter 3, we study composite likelihoods in depth, in particular their efficiency properties.

3 Inference based on composite likelihoods

When the full likelihood cannot be obtained analytically or is difficult to evaluate, as it is the case for example for max-stable processes, composite likelihoods turn out to be convenient surrogates for classical frequentist or Bayesian inference. See, e.g., Varin (2008) and Varin *et al.* (2011) for general overviews of composite likelihoods. In the spatial extremes context, the bivariate marginal densities can usually be derived for max-stable or asymptotic independence models (recall Chapter 2), so inference can be based on a pairwise likelihood (see, e.g., Padoan *et al.*, 2010; Blanchet & Davison, 2011; Davison & Gholamrezaee, 2012). Such pseudo-likelihoods are robust against misspecification of higher-order marginal distributions and have nice theoretical properties (see §3.1.3), but so far have been applied only to componentwise maxima. An important extension, which improves inference by incorporating more information, is to perform pairwise threshold-based inference for max-stable processes (or related models for spatial extremes), analogous to the use of the generalized Pareto distribution in the univariate case. In Section 3.1, we give an overview of composite likelihood methods, and in §3.2 we discuss their application to the modeling of spatial extremes. In particular, we show how to perform inference for extremal models based on spatial componentwise maxima, and discuss the extension to models defined in terms of independent threshold exceedances. The question of temporal dependence is postponed to Chapter 5. Then, in §3.3, we address the question of the statistical efficiency of maximum composite likelihood estimators, and discuss how the latter might be improved by carefully weighting the composite likelihood components.

3.1 Composite likelihoods

3.1.1 Definitions

Consider a N-dimensional random vector $\boldsymbol{Y}$ with density function $g(\boldsymbol{y};\psi)$, for some q-dimensional parameter vector $\psi \in \Psi$. Assuming a set of marginal or conditional events $\{\mathscr{I}_1,\ldots,\mathscr{I}_K\}$ with associated likelihoods $g(\boldsymbol{y} \in \mathscr{I}_k;\psi)$, the composite likelihood for n independent replicates of $\boldsymbol{Y}$ is defined as

$$L_C(\psi) = \prod_{i=1}^{n}\prod_{k=1}^{K} g(\boldsymbol{y}_i \in \mathscr{I}_k;\psi)^{w_k}, \tag{3.1}$$

where $w_k > 0$, $k = 1,\ldots,K$, are suitable positive weights (Lindsay, 1988; Padoan *et al.*, 2010; Varin, 2008; Varin *et al.*, 2011). The composite log-likelihood is

$$\ell_C(\psi) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_k \log g(\boldsymbol{y}_i \in \mathscr{I}_k;\psi). \tag{3.2}$$

If the weights are all equal, they can be ignored, but a judicious choice may improve statistical efficiency; see §3.3. The maximum composite likelihood estimator $\widehat{\psi}_C$, which may not be unique, is obtained by maximizing (3.1) or equivalently (3.2), that is

$$\widehat{\psi}_C = \arg\max_{\psi\in\Psi} L_C(\psi) = \arg\max_{\psi\in\Psi} \ell_C(\psi). \tag{3.3}$$

The composite score $U_C(\psi)$ is the vector of first-order partial derivatives of $\ell_C(\psi)$ with respect to the components of ψ, while the matrix of negated second-order partial derivatives yields the composite observed information matrix $H_C(\psi)$, that is

$$U_C(\psi) = \nabla_\psi \ell_C(\psi), \qquad H_C(\psi) = -\nabla_\psi U_C(\psi). \tag{3.4}$$

The score equation $U_C(\psi) = 0$ may be viewed as an unbiased estimating equation for $\widehat{\psi}_C$. One may further define the variability $K_C(\psi)$ and sensitivity $J_C(\psi)$ matrices as

$$K_C(\psi) = \mathrm{var}\left\{U_C(\psi)\right\}, \qquad J_C(\psi) = \mathrm{E}\left\{H_C(\psi)\right\}. \tag{3.5}$$

By analogy, we define $k_C(\psi)$ and $j_C(\psi)$ the matrices in (3.5), normalized by the number of replicates,

$$k_C(\psi) = \frac{K_C(\psi)}{n}, \qquad j_C(\psi) = \frac{J_C(\psi)}{n}. \tag{3.6}$$

3.1.2 Marginal likelihoods

Marginal likelihoods form a subclass of composite likelihoods, that are constructed from lower-order marginal densities. A good overview may be found in Varin (2008). The simplest type of log-marginal likelihood is constructed under working independence assumptions,

$$\ell_{\text{Ind}}(\psi) = \sum_{i=1}^{n} \sum_{k=1}^{N} w_k \log g(y_{i;k}; \psi), \tag{3.7}$$

where $y_{i;k}$ is the kth component of the ith observation $\boldsymbol{y}_i$. This independence likelihood has been used by Fawcett & Walshaw (2007) for the estimation of return levels, but is useless for the estimation of dependence parameters. The straightforward extension to pairwise margins gives rise to the log-pairwise likelihood

$$\ell_{\text{Pair}}(\psi) = \sum_{i=1}^{n} \sum_{k_1=1}^{N-1} \sum_{k_2=k_1+1}^{N} w_{k_1;k_2} \log g(y_{i;k_1}, y_{i;k_2}; \psi), \tag{3.8}$$

and the triplewise version may be written as

$$\ell_{\text{Triple}}(\psi) = \sum_{i=1}^{n} \sum_{k_1=1}^{N-2} \sum_{k_2=k_1+1}^{N-1} \sum_{k_3=k_2+1}^{K} w_{k_1;k_2;k_3} \log g(y_{i;k_1}, y_{i;k_2}, y_{i;k_3}; \psi). \tag{3.9}$$

In Chapter 4, efficiencies of pairwise and triplewise likelihood estimators are compared in the context of dependence estimation for the Brown–Resnick process. The choice of the weights is non-trivial and some discussion about it is given in Section 3.3.

3.1.3 Asymptotics

Standard results for misspecified likelihoods show that, under regularity conditions and provided ψ is identifiable from the component densities, the maximum composite likelihood estimator $\widehat{\psi}_C$ is strongly consistent and asymptotically normal (Kent, 1982; Cox & Reid, 2004; Padoan *et al.*, 2010; Varin *et al.*, 2011; Davison, 2003, p.147):

$$\sqrt{n}(\widehat{\psi}_C - \psi_0) \xrightarrow{D} \mathcal{N}_q\{0, v_C(\psi_0)\}, \tag{3.10}$$

where $\xrightarrow{D}$ denotes convergence in distribution, and where the asymptotic variance has the "sandwich" form $v_C(\psi_0) = j_C(\psi_0)^{-1} k_C(\psi_0) j_C(\psi_0)^{-1}$, and ψ_0 is the "true" underlying parameter. Hence, the asymptotic distribution for $\widehat{\psi}_C$ may be approximated by

$$\widehat{\psi}_C \dot{\sim} \mathcal{N}_q\{\psi_0, V_C(\widehat{\psi}_C)\}, \tag{3.11}$$

as $n \to \infty$, where

$$V_C(\psi) = n v_C(\psi) = J_C(\psi)^{-1} K_C(\psi) J_C(\psi)^{-1}. \tag{3.12}$$

The inverse of the asymptotic sandwich variance matrix, $I_C(\psi) = V_C(\psi)^{-1}$, is called the Godambe information matrix and is the equivalent of the classical Fisher information for composite likelihoods. When the composite likelihood is actually the full likelihood, Bartlett's identities yield $I_C(\psi) = K_C(\psi) = J_C(\psi)$, meaning that the composite expected information is identical to the Fisher information.

3.1.4 Model comparison

To compare the performance of nested models, a composite likelihood ratio test can be performed. Suppose that the parameter vector ψ is partitioned into $(\psi^1, \psi^2) \in \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$, with $q_1 + q_2 = q$, and that we want to test whether the null hypothesis $\psi^1 = \psi^1_\star$ holds. In this testing framework, the parameter $\psi^1 \in \mathbb{R}^{q_1}$ is the parameter of interest, while $\psi^2 \in \mathbb{R}^{q_2}$ acts as a nuisance parameter. Let $\widehat{\psi}_C = (\widehat{\psi}^1, \widehat{\psi}^2)$ denote the unrestricted composite maximum likelihood estimator, and let $\widehat{\psi}_{C;\star} = (\psi^1_\star, \widehat{\psi}^2_{C,\star})$ denote the maximum likelihood composite estimator under the null hypothesis, i.e., $\widehat{\psi}^2_{C,\star}$ is the maximum composite likelihood estimator of ψ^2, when ψ^1 is held fixed to the value $\psi^1_\star$. A two-sided composite likelihood ratio test may be based on the statistic

$$W(\psi^1_\star) = 2[\ell_C(\widehat{\psi}_C) - \ell_C(\widehat{\psi}_{C,\star})]. \tag{3.13}$$

Under the null hypothesis, one can show (Kent, 1982) that the distribution of $W(\psi^1_\star)$ can be approximated by $\sum_{j=1}^{q_1} c_j Z_j$ for large n, where the Z_j's are independent χ^2_1 random variables, and the c_j's are the eigenvalues of the matrix $\{\tilde{J}^1_C(\widehat{\psi}_C)\}^{-1} I^1_C(\widehat{\psi}_C)$, where $\tilde{J}^1_C(\psi)$ and $I^1_C(\psi)$ are the $q_1 \times q_1$ submatrices of $\{J_C(\psi)\}^{-1}$ and $I_C(\psi)$, respectively, with elements corresponding to ψ^1. Since the theoretical quantiles of linear combinations of χ^2_1 random variables are unknown in general, one must resort to simulation to compute approximate p-values for the test statistic (3.13).

For the comparison of non-nested models, the composite likelihood information criterion (CLIC), an analogue of the Akaike information criterion (AIC), may be useful (Padoan *et al.*, 2010; Davison *et al.*, 2013). It is defined as

$$\text{CLIC} = -2\left[\ell_C(\widehat{\psi}_C) - \text{tr}\left\{K_C(\widehat{\psi}_C) J_C^{-1}(\widehat{\psi}_C)\right\}\right], \tag{3.14}$$

where $\text{tr}(\cdot)$ denotes the trace operator. Model selection can be based on minimizing (3.14). Davison & Gholamrezaee (2012) suggested use of a variant, the CLIC*, which is scaled to be comparable with AIC for independent data. For example, when pairwise

likelihood is considered, CLIC* = CLIC/$(N-1)$, and for triplewise likelihood, CLIC* = CLIC/$\{(N-1)(N-2)/3\}$. In general, CLIC* = CLIC$\left\{2N\binom{N}{D}\right\}^{-1}$ for a D-variate full marginal likelihood. When there are missing data, the scaling factor changes, and must be calculated on a case by case basis.

3.1.5 Estimation of the asymptotic variance

The asymptotic variance of the maximum composite likelihood estimator can be approximated by

$$V_C(\psi) \approx V_C(\widehat{\psi}_C) = J_C(\widehat{\psi}_C)^{-1} K_C(\widehat{\psi}_C) J_C(\widehat{\psi}_C)^{-1}.$$

The sensitivity matrix, $J_C(\widehat{\psi}_C)$, is easily approximated by finite differences, using the hessian matrix usually returned by the optimization routine. More precisely, $J_C(\widehat{\psi}_C)$ can be estimated by the matrix $H_C(\widehat{\psi}_C)$, where the latter is computed using numerical approximations. Unless some parameters are close to their boundaries, this estimate of $J_C(\widehat{\psi}_C)$ is often reasonable. Difficulties arise with the estimation of the variability matrix. Since $K_C(\psi) = \text{var}\{U_C(\psi)\} = \text{E}\{U_C(\psi)U_C(\psi)^T\}$, the naive approach would consist in approximating $K_C(\widehat{\psi}_C)$ by $U_C(\widehat{\psi}_C)U_C(\widehat{\psi}_C)^T$. But by definition, $U_C(\widehat{\psi}_C) = 0$, so the resulting estimated matrix contains only zeros. To circumvent this issue, several approaches have been suggested. Assuming independence of the observations, a first approach (Varin *et al.*, 2011) is to compute the empirical variance of the n score contributions, multiplied by n, that is,

$$K_C(\psi) \approx n\widehat{\text{var}}_{i=1,\ldots,n}\left\{\sum_{k=1}^{K} w_k \nabla_\psi \log g(\boldsymbol{y}_i \in \mathscr{I}_k; \psi)\right\}, \tag{3.15}$$

and to plug $\widehat{\psi}_C$ in (3.15). One drawback of this formula is that it involves all partial derivatives of the log-density, explicit calculations of which may be painful in practice. More importantly, (3.15) may be imprecise when n is not sufficiently large compared to the dimension of ψ. Since slight misestimation of $K_C(\psi)$ may have large repercussions on $V_C(\psi)$, and on model selection using the CLIC, it is crucial to have accurate estimators.

If computational resources allow, other estimators for $V_C(\psi)$ based on subsampling, jackknife or bootstrap methods may be used (see, e.g., Shao & Tu, 1995; Davison & Hinkley, 1997; Politis *et al.*, 1999; Varin *et al.*, 2011; Thibaud *et al.*, 2013), and the variability matrix can be estimated by left and right multiplication with an estimate of $J_C(\psi)$.

3.2 Pairwise likelihood for spatial extremes

Let us assume that the spatio-temporal process of interest $Y(s,t)$, $(s,t) \in \mathscr{S} \times \mathscr{T}$, has been sampled at S stations and T time points, resulting in $N = ST$ data points in the space-time domain $\mathscr{S} \times \mathscr{T}$. The study of the extremes of $Y(s,t)$ is usually performed by assuming that a max-stable, or asymptotic independence, process is a reasonable model for block maxima or large threshold-exceedances. However, classical inference is not possible since the full likelihood (here in dimension N) is intractable for most max-stable and asymptotically independent models. By contrast, the pairwise margins are known (see Appendix B), and inference can be made using pairwise likelihood.

The classical approach based on block maxima is first described in §3.2.1, and our novel, more efficient, procedure based on threshold exceedances is then discussed in §3.2.2. The case of temporal dependence is postponed to Chapter 5.

3.2.1 Componentwise maxima

Suppose that the process $Y(s,t)$, $(s,t) \in \mathscr{S} \times \mathscr{T}$, can be split into n independent site-referenced block maxima $M_i(s)$, $i = 1, \ldots, n$. Let $m_{s;i}$ denote the ith block maxima recorded at the sth monitoring station. Assuming that a max-stable, or asymptotic independence, process with bivariate density $g(z_1, z_2; \psi)$ yields a reasonable fit for the spatial process of maxima, one can make inference based on the log-pairwise likelihood

$$\ell(\psi) = \sum_{i=1}^{n} \sum_{s_1=1}^{S-1} \sum_{s_2=s_1+1}^{S} w_{s_1;s_2} \log g(m_{s_1;i}, m_{s_2;i}; \psi). \tag{3.16}$$

Using weights that include only neighboring sites, this methodology has been used satisfactorily by Padoan *et al.* (2010) in a study of US precipitation extremes. Blanchet & Davison (2011) used an equally weighted pairwise likelihood in an analysis of annual maxima of snow depth in the Alpine region, and Davison & Gholamrezaee (2012) adopted the same approach to fit a max-stable process to annual temperature maxima in Switzerland.

3.2.2 Threshold exceedances

When the original data (not only maxima) are available, more efficient inference is feasible. For simplicity, let assume that the observations are independent in time, and let $y_{s;t}$ denote the measurement at the sth station at tth time. Furthermore, suppose that a max-stable, or asymptotic independence, model with bivariate distribution $G(y_1, y_2; \psi)$ is suitable to capture the dependence features of the process at high levels.

The results of the simulation study in §1.2.2.2 established in dimension $D = 2$, along with the considerations in §2.3.4, suggest to make inference based on a censored threshold-based log-pairwise likelihood, defined as

$$\ell(\psi) = \sum_{t=1}^{T} \sum_{s_1=1}^{S-1} \sum_{s_2=s_1+1}^{S} w_{s_1;s_2} \log p_u(y_{s_1;t}, y_{s_2;t}; \psi), \tag{3.17}$$

where the censored pairwise contributions, illustrated in Figure 3.1, are

$$p_u(y_1, y_2; \psi) = \begin{cases} G(u, u; \psi), & y_1, y_2 \leq u, \\ \frac{\partial}{\partial y_1} G(y_1, u; \psi), & y_1 > u, y_2 \leq u, \\ \frac{\partial}{\partial y_2} G(u, y_2; \psi), & y_1 \leq u, y_2 > u, \\ \frac{\partial^2}{\partial y_1 \partial y_2} G(y_1, y_2; \psi), & y_1, y_2 > u, \end{cases} \tag{3.18}$$

and $u \in \mathbb{R}$ is a high threshold. Different marginal thresholds can be used (Bortot *et al.*, 2000) and the approach generalizes to higher dimensions, though the probability that an observed D-uplet falls into the "upper right quadrant" decays rapidly with D, leading to potential inference problems in practice. This approach has been used by several researchers (Ledford & Tawn, 1996; Smith *et al.*, 1997; Bortot *et al.*, 2000; Wadsworth & Tawn, 2012 and Coles, 2001, p.155).

Since the data are assumed to be temporally independent, the results of §3.1.3 apply and the maximum pairwise likelihood estimator based on (3.18) is strongly consistent and asymptotically normal. When the observations exhibit temporal dependence, we can consider the log-pairwise likelihood

$$\ell(\psi) = \sum_{t_1=1}^{T-1} \sum_{t_2=t_1}^{T} \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} w_{t_1;t_2;s_1;s_2} \{1 - I(s_1 \geq s_2 \text{ and } t_2 - t_1 = 0)\} \log p_u(y_{s_1;t_1}, y_{s_2;t_2}; \psi), \tag{3.19}$$

where $I(\cdot)$ denotes the indicator function, and the results continue to hold under some additional mixing condition in time, along with a condition on the rate at which independence is reached for increasing time lags; see Section 5.2.1.

3.3 Efficiency of pairwise likelihoods

Maximum composite likelihood estimators inherit appealing properties from the classical maximum (full) likelihood estimator. Under regularity conditions, they are strongly consistent and asymptotically normal, but the question of their relative efficiency remains to be addressed.

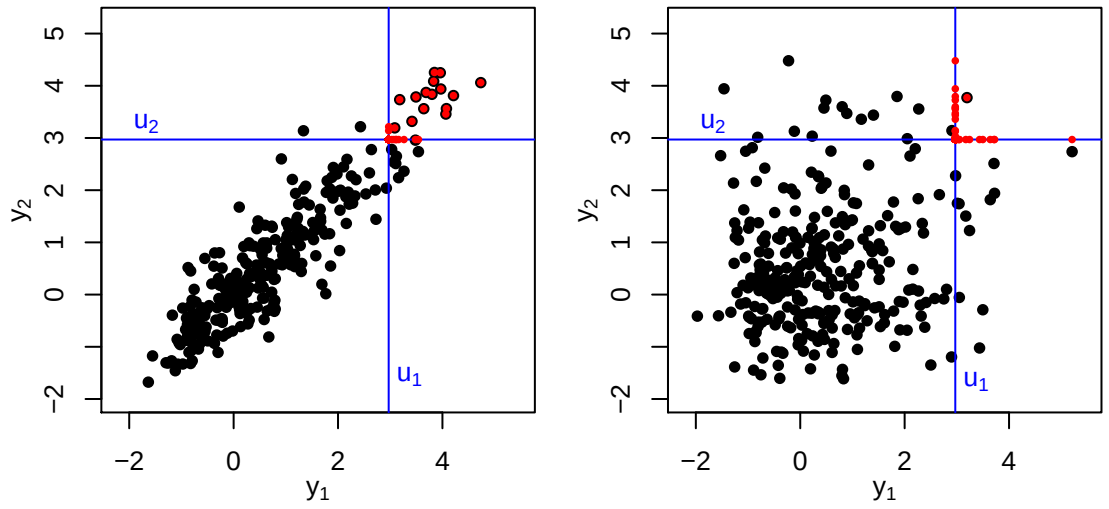

Figure 3.1: Illustration of the censored pairwise likelihood approach, in case of strong (left panel) and weak (right panel) dependence. The black points are the original data points, and the red dots show their censored counterparts, used for fitting. Different likelihood contributions are associated to the four different subspaces delimited by the marginal thresholds u_1, u_2 (blue lines); see (3.18). When the data are strongly dependent, many points lie in the upper right quadrant, and the effect of the censoring is overall weaker than for the low dependence case.

3.3.1 Previous work and weighting strategy

Much has been written on the efficiency of composite likelihoods in a variety of contexts; see Varin *et al.* (2011) and the references therein for a global overview. For example, Renard *et al.* (2004) study pairwise likelihoods for multilevel probit models. Hjort & Varin (2008) work out explicit expressions for the efficiency of pairwise and conditional likelihood estimators for finite-state Markov chains. Cox & Reid (2004) focus on the use of full (i.e., equally weighted) pairwise likelihoods for inference on symmetric normal and dichotomized symmetric normal models, and conclude that the asymptotic relative efficiency depends nonlinearly on the underlying correlation and on the dimension of the data; the higher the dimension, the lower the efficiency. Varin & Vidoni (2008) consider general state-space models and give some efficiency results for the Gaussian AR(1) model with additive observation noise. Bevilacqua *et al.* (2012) use a similar approach to make inference for large data sets, and discuss an application to space-time ozone levels data. They compare pairwise likelihood estimators and pairwise score estimators with respect to restricted maximum likelihood (REML) estimators, and show that the former provide a good compromise between statistical and computational efficiency. Davis & Yau (2011) explore pairwise likelihoods for stationary linear time series models, including ARMA and fractionally

integrated ARMA models. They show that the asymptotic normality of maximum pairwise likelihood estimators holds under a mild condition on the autocorrelation, and compare their performance for AR(1), MA(1) and fractionally integrated ARMA models of order less than 0.5. They underline that the loss of efficiency is moderate for autoregressive models, while it can be substantial for MA-type models. In the spatial extremes framework, Padoan *et al.* (2010) use pairwise likelihood to fit max-stable processes to extreme rainfall, and suggest discarding distant pairs to improve the efficiency.

As Varin & Vidoni (2008), Padoan *et al.* (2010), Davis & Yau (2011) and Bevilacqua *et al.* (2012) highlight, the full pairwise likelihood, for which $w_{k_1;k_2} = 1$ in (3.8) for any pairs of sites (k_1, k_2), may be suboptimal and can usually be largely improved by taking different weights. So far, people have proposed several weighting approaches. One strategy for spatial applications, which acts as a *selection* scheme, consists of including in the pairwise likelihood all the pairs of observations with a lag distance $\|h\|$ not greater than a given predetermined value $h_{\max}$, and discarding the other pairs (Heagerty & Lele, 1998), i.e.,

$$w_{k_1;k_2} = I(\|h\| \leq h_{\max}),$$

where $I(\cdot)$ denotes the indicator function; we call the resulting pairwise likelihood a *consecutive* or *locally weighted* pairwise likelihood, depending on the context. The intuition behind this approach is that one expects neighboring pairs to be strongly dependent, thus providing valuable information for the estimation of dependence parameters. The results of Varin & Vidoni (2008), Davis & Yau (2011) and Bevilacqua *et al.* (2012) show that locally weighted pairwise likelihood can be much more efficient, both statistically and computationally, than full pairwise likelihood, but that a careful choice of weights is essential. The selection of the threshold $h_{\max}$ is critical: when $h_{\max} \to \infty$, the full pairwise likelihood is recovered, and many pairwise contributions are involved, whereas for $h_{\max}$ small, only a few pairs are included. In practice, this choice may be guided by plotting an empirical dependence measure (e.g., the correlation for Gaussian data or the extremal coefficient for extremal data) for all pairs of observations, against their distance. Alternatively, Padoan *et al.* (2010) suggest choosing the threshold $h_{\max}$ that minimizes the total variation of the asymptotic variance, i.e., $h_{\max} = \arg\min \mathrm{tr}\{V_C(\widehat{\psi}_C)\}$, where $\mathrm{tr}(\cdot)$ is the trace operator and $V_C(\widehat{\psi}_C)$ is the sandwich matrix defined in (3.12).

Another possibility would be to consider weights that are inversely proportional to the lag distance, i.e.,

$$w_{k_1;k_2} = 1/\|h\|,$$

so that the distant pairs, that are usually less dependent than close ones, are down-weighted. More generally, one could take weights of the form $w_{k_1;k_2} = \|h\|^{-\alpha}$, with $\alpha > 0$, and choose the exponent α as above, that is, minimizing the total variation of the asymptotic variance: $\alpha = \arg\min \mathrm{tr}\{V_C(\widehat{\psi}_C)\}$. However, with this approach, one needs to computed the contribution of all pairs, so that no computational gains can be expected with respect to the full pairwise likelihood.

A further possibility is to include a mixture of close and distant pairs, in order to be able to capture the independence range while accurately estimating the dependence parameters. This approach is investigated in §3.3.3.2 for max-stable time series constructed from random sets, and our results suggest that it can dramatically improve the fit, especially when the random set parameters (therefore, the independence range) have to be estimated.

In the next two sections, we investigate how the weights should be chosen under different scenarios. Since theoretical calculations seem to be out of reach for max-stable processes, we start in §3.3.2.1 with simple classical time series models, namely AR(1) and MA(1) models, and calculate the exact asymptotic relative efficiency of maximum pairwise likelihood estimators under different schemes of time lag inclusion. We then extend this framework to more general Gaussian ARMA models in §3.3.2.2, and conduct a simulation study to find the best combination of pairs. In §3.3.2.3, we consider the estimation of a single dependence parameter for Gaussian processes defined on $\mathbb{R}^2$ and discuss optimal weighting in a certain sense. Since these theoretical results do not apply directly to max-stable processes, we then conduct two further simulations in dimension one; in the first simulation, in §3.3.3.1, we consider the logistic extreme-value model for which the maximum likelihood estimator can be well approximated, and investigate the loss in efficiency of pairwise likelihoods with respect to the latter; in the second simulation, in §3.3.3.2, we try to understand how to choose the pairs to include in the pairwise likelihood under different fitting procedures, when the Schlather model with random set is considered. The results of this last part are used by analogy in our extreme rainfall application in Chapter 5.

3.3.2 Gaussian models

3.3.2.1 Theoretical results for AR(1) and MA(1) models

AR(1) and MA(1) time series models may be defined as follows. Let $\mu \in \mathbb{R}$ be a location parameter, $\lambda \in (-1,1)$ a dependence parameter and $\sigma > 0$ a dispersion parameter, and let $\psi = (\mu, \lambda, \sigma^2)$ denote the parameter vector. The Gaussian autoregressive model of

order one —or AR(1), Z_t, $t \in \mathbb{Z}$, satisfies

$$Z_t - \mu = \lambda(Z_{t-1} - \mu) + \varepsilon_t, \tag{3.20}$$

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is a Gaussian white noise. It is easy to prove that the process Z_t is Gaussian, weakly stationary, has mean μ, variance $\gamma_0 = \sigma^2/(1-\lambda^2)$ and correlation $\rho_h = \lambda^{|h|}$ at lag $h \in \mathbb{Z}$. As for the moving average model of order one —or MA(1), it can be defined as

$$Z_t - \mu = \varepsilon_t + \lambda\varepsilon_{t-1}, \quad t \in \mathbb{Z}, \tag{3.21}$$

with similar innovations ε_t, and thus shares the properties of the AR(1) model, except that the variance is $\gamma_0 = \sigma^2\left(1+\lambda^2\right)$, and the correlation at lag h may be expressed as $\rho_h = I(h=0) + \lambda(1+\lambda^2)^{-1} I(|h|=1)$, where $I(\cdot)$ is the indicator function.

Let $z_1, \ldots, z_T$ denote a segment of observations from a Gaussian AR(1) or MA(1) model. In order to study the efficiency of pairwise likelihoods under different schemes of lag inclusion, we consider the maximum pairwise likelihood estimator,

$$\widehat{\psi}_{\mathcal{K}} = \arg\min_{\psi} \underbrace{\sum_{t=1}^{T-\mathcal{K}_{\max}} \sum_{h \in \mathcal{K}} \log f(z_t, z_{t+h}; \psi)}_{\ell_P(\psi)}, \tag{3.22}$$

where $\mathcal{K} \subset \mathbb{N}$ is a finite collection of time lags, $\mathcal{K}_{\max}$ is the maximum lag in $\mathcal{K}$, and $f(z_1, z_2; \psi)$ denotes the bivariate density of either time series model. We consider three different sets of time lags: $\mathcal{K}_a^K = \{1, \ldots, K\}$, $K < \infty$, which corresponds to the maximum consecutive pairwise likelihood estimator; $\mathcal{K}_b^K = \{b_k; k = 1, \ldots, K\}$, where b_k is based on the Fibonacci sequence: $1, 2, 3, 5, 8, 13, 21, \ldots$; and $\mathcal{K}_c^K = \{2^{k-1}; k = 1, \ldots, K\}$, where the lags increase geometrically.

In order to compute the asymptotic relative efficiency of $\widehat{\psi}_{\mathcal{K}}$, we need to compute the sensitivity matrix $J_{\mathcal{K}}(\psi)$ and variability matrix $K_{\mathcal{K}}(\psi)$ defined in (3.5). Since AR(1) and MA(1) processes are Gaussian, the log density in the right-hand side of (3.22) is, up to an additive constant,

$$\begin{aligned} \log f(z_t, z_{t+h}; \psi) \quad \equiv \quad & -\log(\gamma_0) - \frac{1}{2}\log\left(1-\rho_h^2\right) \\ & - \frac{1}{2\gamma_0\left(1-\rho_h^2\right)} \left\{ (z_t-\mu)^2 - 2\rho_h (z_t-\mu)(z_{t+h}-\mu) + (z_{t+h}-\mu)^2 \right\}. \end{aligned} \tag{3.23}$$

To derive the asymptotic variance of $\widehat{\psi}_{\mathcal{K}}$, that is $V_{\mathcal{K}}(\psi) = J_{\mathcal{K}}(\psi)^{-1} K_{\mathcal{K}}(\psi) J_{\mathcal{K}}(\psi)^{-1}$, explicitly, we need to compute the first and second derivatives of (3.23) with respect to the model parameters, and then use the moment properties of the process of interest

to deduce the variance of the composite score, and the expectation of the composite information matrix. By definition, the composite score is the vector

$$\nabla_\psi \ell_P(\psi) = \left\{ \sum_{t=1}^{T-\mathcal{K}_{\max}} \sum_{h\in\mathcal{K}} u_\mu(z_t, z_{t+h}), \sum_{t=1}^{T-\mathcal{K}_{\max}} \sum_{h\in\mathcal{K}} u_\lambda(z_t, z_{t+h}), \sum_{t=1}^{T-\mathcal{K}_{\max}} \sum_{h\in\mathcal{K}} u_{\sigma^2}(z_t, z_{t+h}) \right\}, \tag{3.24}$$

where the inner score components may be written as

$$\begin{aligned}
u_\mu(z_{t_1}, z_{t_2}) &= A(h)\left\{(z_{t_1}-\mu)+(z_{t_2}-\mu)\right\}, && (3.25)\\
u_\lambda(z_{t_1}, z_{t_2}) &= B_1(h) + B_2(h)(z_{t_1}-\mu)^2 + B_3(h)(z_{t_1}-\mu)(z_{t_2}-\mu) + B_2(h)(z_{t_2}-\mu)^2, && (3.26)\\
u_{\sigma^2}(z_{t_1}, z_{t_2}) &= C_1(h) + C_2(h)(z_{t_1}-\mu)^2 + C_3(h)(z_{t_1}-\mu)(z_{t_2}-\mu) + C_2(h)(z_{t_2}-\mu)^2, && (3.27)
\end{aligned}$$

and whose coefficients are deterministic and given by

$$\begin{aligned}
A(h) &= \frac{1}{\gamma_0(1+\rho_h)}, && (3.28)\\
B_1(h) &= -\frac{1}{\gamma_0}\frac{\partial\gamma_0}{\partial\lambda} + \frac{\rho_h}{1-\rho_h^2}\frac{\partial\rho_h}{\partial\lambda}, && (3.29)\\
B_2(h) &= \frac{1}{2\gamma_0^2(1-\rho_h^2)}\frac{\partial\gamma_0}{\partial\lambda} - \frac{\rho_h}{\gamma_0(1-\rho_h^2)^2}\frac{\partial\rho_h}{\partial\lambda}, && (3.30)\\
B_3(h) &= -\frac{\rho_h}{\gamma_0^2(1-\rho_h^2)}\frac{\partial\gamma_0}{\partial\lambda} + \frac{1+\rho_h^2}{\gamma_0(1-\rho_h^2)^2}\frac{\partial\rho_h}{\partial\lambda}, && (3.31)\\
C_1(h) &= -\frac{1}{\gamma_0}\frac{\partial\gamma_0}{\partial\sigma^2}, && (3.32)\\
C_2(h) &= \frac{1}{2\gamma_0^2(1-\rho_h^2)}\frac{\partial\gamma_0}{\partial\sigma^2}, && (3.33)\\
C_3(h) &= -\frac{\rho_h}{\gamma_0^2(1-\rho_h^2)}\frac{\partial\gamma_0}{\partial\sigma^2}. && (3.34)
\end{aligned}$$

Now, denoting the ith element of the parameter vector $\psi = (\mu, \lambda, \sigma^2)$ by ψ_i, the variability matrix $K_{\mathcal{K}}(\psi)$ is found by computing for $i, j = 1, 2, 3$,

$$\begin{aligned}
\operatorname{cov}\left\{\frac{\partial}{\partial\psi_i}\ell_P(\psi), \frac{\partial}{\partial\psi_j}\ell_P(\psi)\right\} &= \operatorname{cov}\left\{\sum_{t_1=1}^{T-\mathcal{K}_{\max}}\sum_{h_1\in\mathcal{K}} u_{\psi_i}(z_t, z_{t+h}), \sum_{t_2=1}^{T-\mathcal{K}_{\max}}\sum_{h_2\in\mathcal{K}} u_{\psi_j}(z_t, z_{t+h})\right\}\\
&= (T-\mathcal{K}_{\max})\sum_{t=1}^{T-\mathcal{K}_{\max}}\sum_{h_1\in\mathcal{K}}\sum_{h_2\in\mathcal{K}}\operatorname{cov}\left\{u_{\psi_i}(z_1, z_{1+h_1}), u_{\psi_j}(z_1, z_{1+h_2})\right\}\\
&\quad + \sum_{t=2}^{T-\mathcal{K}_{\max}}(T-\mathcal{K}_{\max}-t+1)\sum_{h_1\in\mathcal{K}}\sum_{h_2\in\mathcal{K}}\operatorname{cov}\left\{u_{\psi_i}(z_1, z_{1+h_1}), u_{\psi_j}(z_t, z_{t+h_2})\right\}
\end{aligned}$$

$$+\sum_{t=2}^{T-\mathscr{K}_{\max}}(T-\mathscr{K}_{\max}-t+1)\sum_{h_1\in\mathscr{K}}\sum_{h_2\in\mathscr{K}}\operatorname{cov}\Big\{u_{\psi_j}(z_1,z_{1+h_1}),u_{\psi_i}(z_t,z_{t+h_2})\Big\}, \quad (3.35)$$

which is justified because the process is stationary. Moreover, owing to the Gaussianity, one has that

$$\begin{aligned}\operatorname{cov}\big\{(z_{t_1}-\mu)(z_{t_2}-\mu),(z_{t_3}-\mu)\big\} &= 0, &(3.36)\\ \operatorname{cov}\big\{(z_{t_1}-\mu)(z_{t_2}-\mu),(z_{t_3}-\mu)(z_{t_4}-\mu)\big\} &= \gamma_0^2\big(\rho_{|t_3-t_1|}\rho_{|t_4-t_2|}+\rho_{|t_4-t_1|}\rho_{|t_3-t_2|}\big). &(3.37)\end{aligned}$$

Hence, combining the expressions (3.24–3.37), we can easily compute the variability matrix $K_{\mathscr{K}}(\psi)$ for various stationary Gaussian time series models using computer softwares. It turns out that the matrix $K_{\mathscr{K}}(\psi)$ is block diagonal because the covariance between the component of the score corresponding to μ and the other components vanishes, but for the remaining entries, it is difficult to give a closed formula. As far as the sensitivity matrix $J_{\mathscr{K}}(\psi)$ is concerned, we must compute the derivative of the negated score components with respect to the model parameters, and then take the expectation; we get again a block diagonal matrix, whose non-zero elements are

$$\begin{aligned}\mathrm{E}\left\{-\frac{\partial^2}{\partial\mu^2}\ell_P(\psi)\right\} &= \frac{2(T-\mathscr{K}_{\max})}{\gamma_0}\sum_{h\in\mathscr{K}}(1+\rho_h)^{-1}, &(3.38)\\ \mathrm{E}\left\{-\frac{\partial^2}{\partial\lambda^2}\ell_P(\psi)\right\} &= \frac{T-\mathscr{K}_{\max}}{\gamma_0^2}\left\{\left(\frac{\partial\gamma_0}{\partial\lambda}\right)^2K-\gamma_0\sum_{h\in\mathscr{K}}D(h)\right\}, &(3.39)\\ \mathrm{E}\left\{-\frac{\partial^2}{(\partial\sigma^2)^2}\ell_P(\psi)\right\} &= \frac{T-\mathscr{K}_{\max}}{\gamma_0^2}\left(\frac{\partial\gamma_0}{\partial\sigma^2}\right)^2K, &(3.40)\\ \mathrm{E}\left\{-\frac{\partial^2}{\partial\lambda\partial\sigma^2}\ell_P(\psi)\right\} &= \frac{T-\mathscr{K}_{\max}}{\gamma_0^2}\frac{\partial\gamma_0}{\partial\sigma^2}\left(\frac{\partial\gamma_0}{\partial\lambda}K-\gamma_0\sum_{h\in\mathscr{K}}\frac{\rho_h}{1-\rho_h^2}\frac{\partial\rho_h}{\partial\lambda}\right), &(3.41)\end{aligned}$$

where $K=|\mathscr{K}|$ is the number of time lags used, and where we have

$$D(h)=\frac{2\rho_h}{1-\rho_h^2}\frac{\partial\rho_h}{\partial\lambda}\frac{\partial\gamma_0}{\partial\lambda}-\frac{\gamma_0(1+\rho_h^2)}{(1-\rho_h^2)^2}\left(\frac{\partial\rho_h}{\partial\lambda}\right)^2. \quad (3.42)$$

Calculations for the AR(1) model. By definition of the AR(1) model, $\gamma_0=\sigma^2/(1-\lambda^2)$, and $\rho_h=\lambda^{|h|}$. Hence, equation (3.23) becomes for $h=1,2,\ldots$,

$$\begin{aligned}\log f(z_t,z_{t+h};\psi) \equiv\ & -\log(\sigma^2)+\log\big(1-\lambda^2\big)-\frac{1}{2}\log\Big(1-\lambda^{2h}\Big) &(3.43)\\ & -\frac{1-\lambda^2}{2\sigma^2\big(1-\lambda^{2h}\big)}\Big\{(z_t-\mu)^2-2\lambda^h(z_t-\mu)(z_{t+h}-\mu)+(z_{t+h}-\mu)^2\Big\}.\end{aligned}$$

But since, for $h > 0$, the partial derivatives of γ_0 and ρ_h are

$$\frac{\partial \gamma_0}{\partial \lambda} = \frac{2\sigma^2\lambda}{(1-\lambda^2)^2}, \qquad \frac{\partial \gamma_0}{\partial \sigma^2} = \frac{1}{1-\lambda^2}, \qquad \frac{\partial \rho_h}{\partial \lambda} = h\lambda^{h-1}, \tag{3.44}$$

the coefficients in equations (3.28–3.34) can be worked out:

$$\begin{aligned}
A(h) &= \frac{1-\lambda^2}{\sigma^2(1+\lambda^h)},\\
B_1(h) &= -\frac{2\lambda}{1-\lambda^2} + \frac{h\lambda^{2h-1}}{1-\lambda^{2h}},\\
B_2(h) &= \frac{1}{\sigma^2(1-\lambda^{2h})}\left\{\lambda - \frac{h\lambda^{2h-1}(1-\lambda^2)}{1-\lambda^{2h}}\right\},\\
B_3(h) &= \frac{1}{\sigma^2(1-\lambda^{2h})}\left\{-2\lambda^{h+1} + \frac{h\lambda^{h-1}(1-\lambda^2)(1+\lambda^{2h})}{1-\lambda^{2h}}\right\},\\
C_1(h) &= -\frac{1}{\sigma^2},\\
C_2(h) &= \frac{1-\lambda^2}{2\sigma^4\left(1-\lambda^{2h}\right)},\\
C_3(h) &= -\frac{(1-\lambda^2)\lambda^h}{\sigma^4\left(1-\lambda^{2h}\right)}.
\end{aligned}$$

The variability matrix $K_{\mathscr{K}}(\psi)$ can hence be calculated by plugging these formulae into the previous expressions. Concerning the sensitivity matrix $J_{\mathscr{K}}(\psi)$, it can be computed using (3.44) and the expressions (3.38–3.41), leading to

$$\begin{aligned}
\mathrm{E}\left\{-\frac{\partial^2}{\partial\mu^2}\ell_P(\psi)\right\} &= \frac{2(T-\mathscr{K}_{\max})(1-\lambda^2)}{\sigma^2}\sum_{h\in\mathscr{K}}\left(1+\lambda^h\right)^{-1},\\
\mathrm{E}\left\{-\frac{\partial^2}{\partial\lambda^2}\ell_P(\psi)\right\} &= (T-\mathscr{K}_{\max})\left[\frac{4\lambda^2 K}{(1-\lambda^2)^2} - \sum_{h\in\mathscr{K}}\left\{\frac{4h\lambda^{2h}}{(1-\lambda^2)(1-\lambda^{2h})} - \frac{h^2\lambda^{2h-2}(1+\lambda^{2h})}{(1-\lambda^{2h})^2}\right\}\right],\\
\mathrm{E}\left\{-\frac{\partial^2}{(\partial\sigma^2)^2}\ell_P(\psi)\right\} &= \frac{(T-\mathscr{K}_{\max})K}{\sigma^4},\\
\mathrm{E}\left\{-\frac{\partial^2}{\partial\lambda\partial\sigma^2}\ell_P(\psi)\right\} &= \frac{T-\mathscr{K}_{\max}}{\sigma^2}\left(\frac{2\lambda K}{1-\lambda^2} - \sum_{h\in\mathscr{K}}\frac{h\lambda^{2h-1}}{1-\lambda^{2h}}\right).
\end{aligned}$$

As far as the full log-likelihood is concerned, it can be written as a sum of conditional log-densities because of the Markovian property of the AR(1) process, that is

$$\ell(\psi) \equiv \log\{f(z_1)\} + \sum_{t=2}^{T}\log\{f(z_t \mid z_{t-1})\}$$

$$-\frac{T}{2}\log(\sigma^2)+\frac{1}{2}\log(1-\lambda^2)-\frac{1-\lambda^2}{2\sigma^2}(z_1-\mu)^2-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\{z_t-\mu-\lambda(z_{t-1}-\mu)\}^2,$$

up to an additive constant. It follows easily that the classical maximum likelihood estimator has an asymptotic variance-covariance matrix of the form

$$V(\psi)=\begin{pmatrix} \frac{1-\lambda^2}{\sigma^2}+\frac{(T-1)(1-\lambda)^2}{\sigma^2} & 0 & 0 \\ 0 & \frac{T-2}{1-\lambda^2}+\frac{1+\lambda^2}{(1-\lambda^2)^2} & \frac{\lambda}{\sigma^2(1-\lambda^2)} \\ 0 & \frac{\lambda}{\sigma^2(1-\lambda^2)} & \frac{T}{2\sigma^4} \end{pmatrix}^{-1}. \tag{3.45}$$

Hence, theoretical marginal asymptotic relative efficiencies ARE_μ, ARE_λ and ARE_{σ^2} can be deduced by calculating the ratio of the diagonal elements of the variance matrices $V(\psi)$ in (3.45) and $V_{\mathcal{K}}(\psi)=J_{\mathcal{K}}(\psi)^{-1}K_{\mathcal{K}}(\psi)J_{\mathcal{K}}(\psi)^{-1}$ using the expressions above. The theoretical global asymptotic relative efficiency for ψ is defined as

$$\mathrm{ARE}_\psi=\left[\frac{\det\{V(\psi)\}}{\det\{V_{\mathcal{K}}(\psi)\}}\right]^{1/3},$$

where $\det(\cdot)$ denotes the determinant. Notice that if the time horizon T is large enough, these quantities are close to the true asymptotic relative efficiencies. For the plots in Figures 3.2 and 3.3, we take $T=5000$, so that the discrepancy with the truth is minor.

Our results, summarized in Figure 3.2, confirm those obtained by Davis & Yau (2011), in that the pairwise likelihood estimator is fully efficient when only pairs at lag 1 are included. However, unlike Davis & Yau (2011), our objective was to understand how the asymptotic relative efficiency depends on the set of time lags $\mathcal{K}$ used to select pairs in the likelihood. In particular, our results shed some light on the question of the choice of lags for autoregressive-type models, when a fixed number K is predetermined (for example to ensure parameter identifiability). When $K>1$, the efficiency for μ remains maximal (whatever its value), while those for λ and σ^2 are affected by the choice of lags. Since, the efficiencies are constant with the value of μ and σ^2, Figure 3.2 reports only the results with respect to the choice of $\mathcal{K}$ (left panels) and the dependence parameter λ (right panels).

As far as the inclusion scheme $\mathcal{K}_a^K$ is concerned, we can see in the top left panel of Figure 3.2 that the efficiencies are 100% when $\mathcal{K}=\{1\}$ and then decrease sharply, before stabilizing at about lag 12. This shape is reproduced qualitatively in the middle and bottom left panels, where $\mathcal{K}_b$ and $\mathcal{K}_c$ are used, but the efficiency curves stabilize at a higher level from about lag 8. Interestingly, when the number of lags K is fixed in advance, the best option is not necessarily to include all strongest dependent pairs: for example, for $K=6$ and $\lambda=0.7$, when the pairs at lags $1,2,3,4,5,6$ are included, the

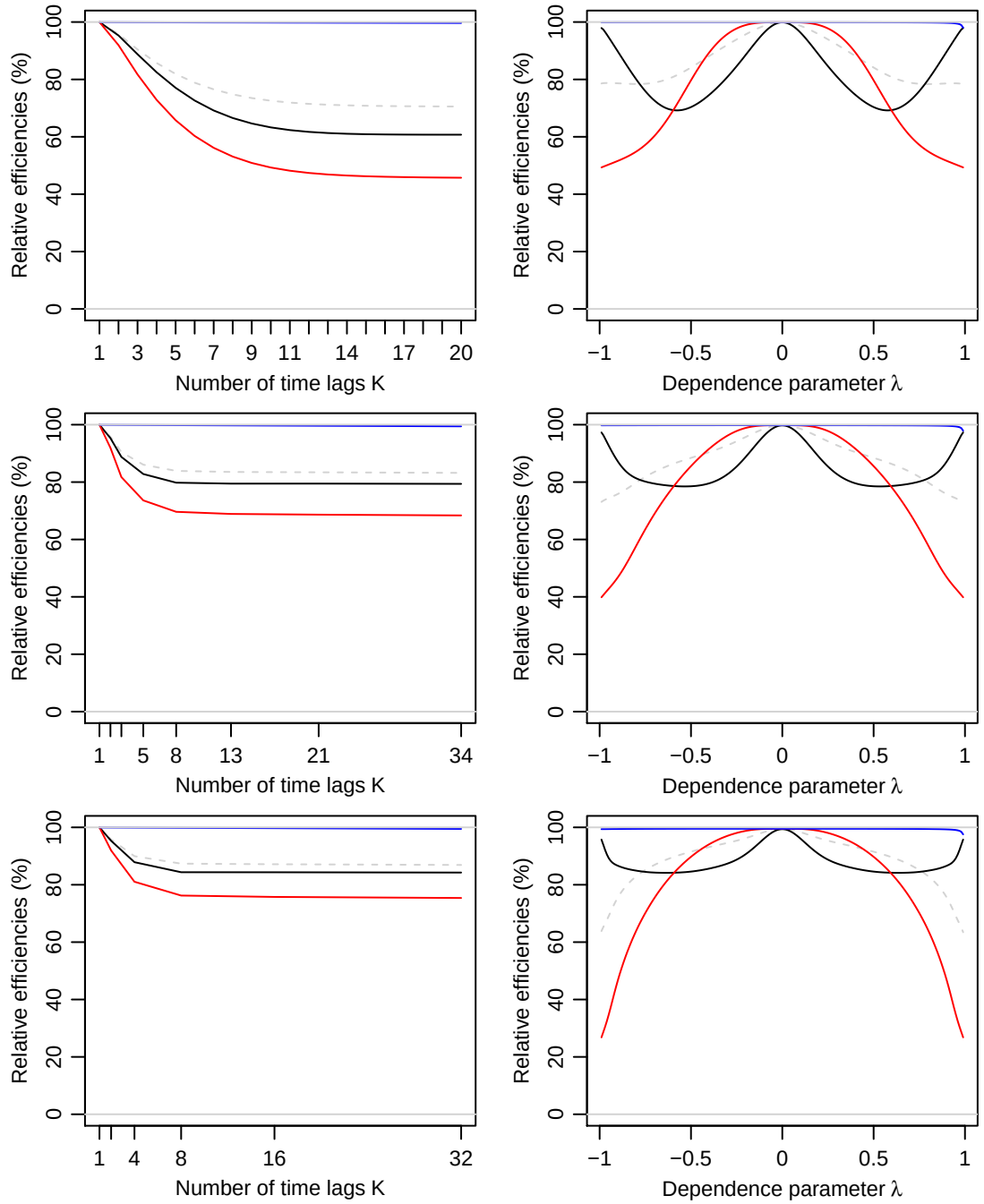

Figure 3.2: Asymptotic relative efficiencies of the pairwise likelihood estimator (3.22) for the AR(1) model (3.20), with respect to (left) the number of time lags included K, fixing $\lambda = 0.7$, and (right) the dependence parameter λ, fixing $K = 6$. The set of time lags is $\mathcal{K}_a^K$ (top), $\mathcal{K}_b^K$ (middle) and $\mathcal{K}_c^K$ (bottom). The solid lines represent the marginal efficiencies ARE_μ (blue), ARE_λ (black) and ARE_{σ^2} (red), while the dashed light grey line is the global efficiency ARE_ψ. We used $\mu = 0$, $\sigma^2 = 1$ and $T = 5000$.

global efficiency of the estimator, around 80%, is significantly lower than when the pairs at lags $1,2,4,8,16,32$, or at lags $1,2,3,5,8,13$, are included. Therefore, for a fixed number of pairs, here 6, it is best to include some distant pairs as well. Although this is not true when $|\lambda|$ is very close to unity (see right panels of Figure 3.2), in general this is the case.

Hence, it seems that two main conclusions can be drawn for AR(1)-type models: including many pairs in the pairwise likelihood can spoil the estimator, suggesting that we should retain as few pairs as possible, provided the parameters remain identifiable; and if further pairs are to be used in addition to adjacent ones, estimation for autoregressive-type processes is usually least damaged by including temporally distant (or weakly correlated) pairs.

Calculations for the MA(1) model. Owing to the second-order properties of the MA(1) process, that is

$$\gamma_0 = \sigma^2(1+\lambda^2), \qquad \rho_h = \begin{cases} 1, & h=0, \\ \lambda(1+\lambda^2)^{-1}, & h=1, \\ 0, & h\geq 2, \end{cases}$$

where $I(\cdot)$ is the indicator function, equation (3.23) becomes, for $h \geq 1$,

$$\log f(z_t, z_{t+h}; \psi) \equiv \begin{cases} -\log(\sigma^2) - \frac{1}{2}\log(1+\lambda^2+\lambda^4) - \frac{1+\lambda^2}{2\sigma^2(1+\lambda^2+\lambda^4)} \\ \qquad \times \left\{(z_t-\mu)^2 - \frac{2\lambda}{1+\lambda^2}(z_t-\mu)(z_{t+1}-\mu) + (z_{t+1}-\mu)^2\right\}, & h=1, \\ -\log(\sigma^2) - \log(1+\lambda^2) - \frac{1}{2\sigma^2(1+\lambda^2)}\left\{(z_t-\mu)^2 + (z_{t+h}-\mu)^2\right\}, & h\geq 2. \end{cases}$$

The composite score has the form of (3.24), with inner components (3.25–3.27) and since

$$\frac{\partial\gamma_0}{\partial\lambda} = 2\lambda\sigma^2, \qquad \frac{\partial\gamma_0}{\partial\sigma^2} = 1+\lambda^2, \qquad \frac{\partial\rho_h}{\partial\lambda} = \begin{cases} (1-\lambda^2)(1+\lambda^2)^{-2}, & h=1, \\ 0, & h\geq 2, \end{cases} \tag{3.46}$$

the coefficients (3.28–3.34) are now

$$\begin{aligned} A(h) &= \begin{cases} \sigma^{-2}(1+\lambda+\lambda^2)^{-1}, & h=1, \\ \sigma^{-2}(1+\lambda^2)^{-1}, & h\geq 2, \end{cases} \\ B_1(h) &= \begin{cases} -\lambda(1+2\lambda^2)(1+\lambda^2+\lambda^4)^{-1}, & h=1, \\ -2\lambda(1+\lambda^2)^{-1}, & h\geq 2, \end{cases} \\ B_2(h) &= \begin{cases} \sigma^{-2}\lambda^3(2+\lambda^2)(1+\lambda^2+\lambda^4)^{-2}, & h=1, \\ \sigma^{-2}\lambda(1+\lambda^2)^{-2}, & h\geq 2, \end{cases} \end{aligned}$$

$$
\begin{aligned}
B_3(h) &= \begin{cases} \sigma^{-2}(1-4\lambda^4-3\lambda^6)(1+\lambda^2)^{-1}(1+\lambda^2+\lambda^4)^{-2}, & h=1, \\ 0, & h\geq 2, \end{cases} \\
C_1(h) &= -\sigma^{-2}, \\
C_2(h) &= \begin{cases} \sigma^{-4}(1+\lambda^2)(1+\lambda^2+\lambda^4)^{-1}/2, & h=1, \\ \sigma^{-4}(1+\lambda^2)^{-1}/2, & h\geq 2, \end{cases} \\
C_3(h) &= \begin{cases} -\sigma^{-4}\lambda(1+\lambda^2+\lambda^4)^{-1}, & h=1, \\ 0, & h\geq 2. \end{cases}
\end{aligned}
$$

These formulae can be used along with expressions (3.25–3.27) and (3.35–3.37) to compute the variability matrix $K_{\mathscr{K}}(\psi)$ for the MA(1) model. Then, assuming that the time lag $h=1$ belongs to the set $\mathscr{K}$ (otherwise λ is not estimable from the bivariate densities), we obtain, using equations (3.38–3.41) and (3.46), that the sensitivity matrix has the following non-zero entries:

$$
\begin{aligned}
\mathrm{E}\left\{-\frac{\partial^2}{\partial\mu^2}\ell_P(\psi)\right\} &= \frac{2(T-\mathscr{K}_{\max})}{\sigma^2}\left(\frac{1}{1+\lambda+\lambda^2}+K-1\right), \\
\mathrm{E}\left\{-\frac{\partial^2}{\partial\lambda^2}\ell_P(\psi)\right\} &= \frac{T-\mathscr{K}_{\max}}{(1+\lambda^2)^2}\left(4\lambda^2 K+\frac{1-3\lambda^2+\lambda^6+\lambda^8}{1+\lambda^2+\lambda^4}\right), \\
\mathrm{E}\left\{-\frac{\partial^2}{(\partial\sigma^2)^2}\ell_P(\psi)\right\} &= \frac{(T-\mathscr{K}_{\max})K}{\sigma^4}, \\
\mathrm{E}\left\{-\frac{\partial^2}{\partial\lambda\partial\sigma^2}\ell_P(\psi)\right\} &= \frac{(T-\mathscr{K}_{\max})\lambda}{\sigma^2(1+\lambda^2)}\left(2K-\frac{1-\lambda^2}{1+\lambda^2+\lambda^4}\right).
\end{aligned}
$$

The asymptotic variance of the maximum pairwise likelihood estimator $\widehat{\psi}_{\mathscr{K}}$ can be calculated using the sandwich formula $V_{\mathscr{K}}(\psi)=J_{\mathscr{K}}(\psi)^{-1}K_{\mathscr{K}}(\psi)J_{\mathscr{K}}(\psi)^{-1}$ as for the AR(1) model.

The maximum full likelihood estimator maximizes the function

$$
\ell(\psi)\equiv -\frac{1}{2}\log(|\Sigma|)-\frac{T}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{z}-\mu\mathbf{1})^T\Sigma^{-1}(\boldsymbol{z}-\mu\mathbf{1}),
$$

where $\boldsymbol{z}=(z_1,\ldots,z_T)^T$, $\mathbf{1}$ is a T-dimensional vector of ones, and $\Sigma\in\mathbb{R}^{T\times T}$ is a Toeplitz matrix with first row $\{(1+\lambda^2),\lambda,0,\ldots,0\}$. Hence, following the development by Klein & Mélard (1995), it can be shown that the variance of the maximum likelihood estimator is, as $T\to\infty$, the diagonal matrix

$$
V(\psi)=\begin{pmatrix} \frac{\sigma^2}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}} & 0 & 0 \\ 0 & \frac{1-\lambda^2}{T} & 0 \\ 0 & 0 & \frac{2\sigma^4}{T} \end{pmatrix},
$$

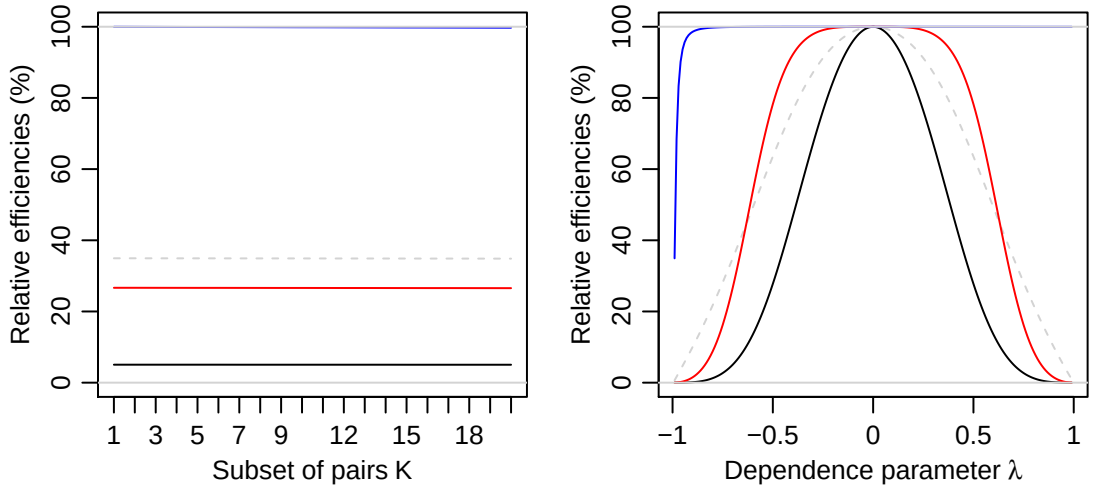

Figure 3.3: Asymptotic relative efficiencies of the pairwise likelihood estimator (3.22) for the MA(1) model (3.21), with respect to (left) the number of time lags included K, fixing $\lambda = 0.7$, and (right) the dependence parameter λ, fixing $K = 1$. The choice of set of time lags, here $\mathcal{K}_a^K$, does not change the results as $T \to \infty$. The solid lines represent the marginal efficiencies ARE_μ (blue), ARE_λ (black) and ARE_{σ^2} (red), while the dashed light grey line is the global efficiency ARE_ψ. We used $\mu = 0$, $\sigma^2 = 1$ and $T = 5000$.

and the asymptotic relative efficiencies ARE_μ, ARE_λ, ARE_{σ^2} and ARE_ψ can be deduced as before for the AR(1) model, by comparing the matrices $V_{\mathcal{K}}(\psi)$ and $V(\psi)$. The left panel of Figure 3.3 illustrates that, assuming the set of time lags $\mathcal{K}$ contains 1 (otherwise λ is not estimable from the bivariate densities), the efficiency of the pairwise maximum likelihood estimator (3.22) does not vary with the number and choice of lags included. Furthermore, unlike the dependence parameter λ, the value of μ and σ^2 do not influence the results, as for the AR(1) model. Concerning λ, the right panel of Figure 3.3 reveals that the asymptotic relative efficiency drops dramatically to zero as λ approaches ± 1, which confirms the results of Davis & Yau (2011). This suggests that for moving average-types models, pairwise likelihood estimators may be much more variable than the maximum full likelihood estimator, and the difference is striking when the process strongly departs from white noise.

3.3.2.2 Simulation study for ARMA models

In order to see how the estimator (3.22) behaves in more complex settings, we now consider zero-mean Gaussian ARMA(p, q) models, which may be defined in the recursive form as

$$Z_t = \lambda_1 Z_{t-1} + \cdots + \lambda_p Z_{t-p} + \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \cdots + \vartheta_q \varepsilon_{t-q}, \tag{3.47}$$

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is a Gaussian white noise and the parameters $\lambda_1, \ldots, \lambda_p$ correspond to the AR part of the process, while $\vartheta_1, \ldots, \vartheta_q$ are the MA coefficients. The case $q = 0$ boils down to the autoregressive model of order p, AR(p), and when $p = 1$, the AR(1) model (3.20) is recovered. Likewise, when $p = 0$, (3.47) corresponds to the moving average process of order q, MA(q), and when $q = 1$, the MA(1) model (3.21) is recovered. In the general setting where $p + q > 1$, it seems impossible to carry out the exact calculations of §3.3.2.1, so we use simulation to assess the loss of efficiency of pairwise likelihood estimators.

For different parameter values, we generate 1000 independent times series of length $T = 1000$ from model (3.47), and then estimate the dependence and variance parameters $\psi = (\lambda_1, \ldots, \lambda_q, \vartheta_1, \ldots, \vartheta_q, \sigma^2)$, using the estimator $\widehat{\psi}_{\mathcal{K}}$ defined in (3.22). Since our goal is to study how the selection of a predetermined number of time lags influences the efficiency of $\widehat{\psi}_{\mathcal{K}}$, we consider all combinations of 4 time lags that can be chosen in the set $\{1, \ldots, 8\}$, therefore resulting in a total of $\binom{8}{4} = 70$ different pairwise likelihood estimators.

The empirical variance of $\widehat{\psi}_{\mathcal{K}} = (\widehat{\lambda}_{1;\mathcal{K}}, \ldots, \widehat{\lambda}_{p;\mathcal{K}}, \widehat{\vartheta}_{1;\mathcal{K}}, \ldots, \widehat{\vartheta}_{q;\mathcal{K}}, \widehat{\sigma}^2_{\mathcal{K}})$ is calculated using the 1000 replicates, and the relative efficiencies are defined as

$$\text{RE}_{\lambda_i} = \frac{\widehat{\text{var}}\widehat{\lambda}_i}{\widehat{\text{var}}\widehat{\lambda}_{i;\mathcal{K}}}, \qquad \text{RE}_{\vartheta_j} = \frac{\widehat{\text{var}}\widehat{\vartheta}_j}{\widehat{\text{var}}\widehat{\vartheta}_{j;\mathcal{K}}}, \qquad \text{RE}_{\sigma^2} = \frac{\widehat{\text{var}}\widehat{\sigma}^2}{\widehat{\text{var}}\widehat{\sigma}^2_{\mathcal{K}}}, \tag{3.48}$$

for $i = 1, \ldots, p$, $j = 1, \ldots, q$, where $\widehat{\psi} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p, \widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_q, \widehat{\sigma}^2)$ is the maximum likelihood estimator, calculated quickly using a Kalman filter (see Shumway & Stoffer, 2004, Chapter 6). In order to reduce the variability of these relative efficiencies, the empirical variance of the maximum likelihood estimator is computed based on 10000 independent time series of length $T = 10000$.

The left panels of Figure 3.4 display the true asymptotic relative efficiencies for the AR(1) model with $\lambda = 0.3, 0.6, 0.9$ and $\sigma^2 = 1$, where the efficiency values are computed using the explicit expressions given in §3.3.2.1. When a selection of four different lags has to be made, we can see that the best strategy for weakly dependent AR(1) models (with almost 100% efficiency) seems to be to include the pairs at lags in $\mathcal{K} = \{1, 6, 7, 8\}$, that is, the strongest dependent pairs at lag 1, and some nearly independent ones. This agrees with the findings of §3.3.2.1. When the pairs at lag 1 are not included, the efficiency drops dramatically. For moderately dependent AR(1) processes, with $\lambda = 0.6$, the best choice is a mixture of very informative, but highly variable, pairs (at lags 1 and 2), and little informative, but less variable, pairs (at lags 7 and 8). For strongly dependent processes, the best option is to include all strongest dependent pairs, because the correlation at lag 8 is far too strong to have little impact on the

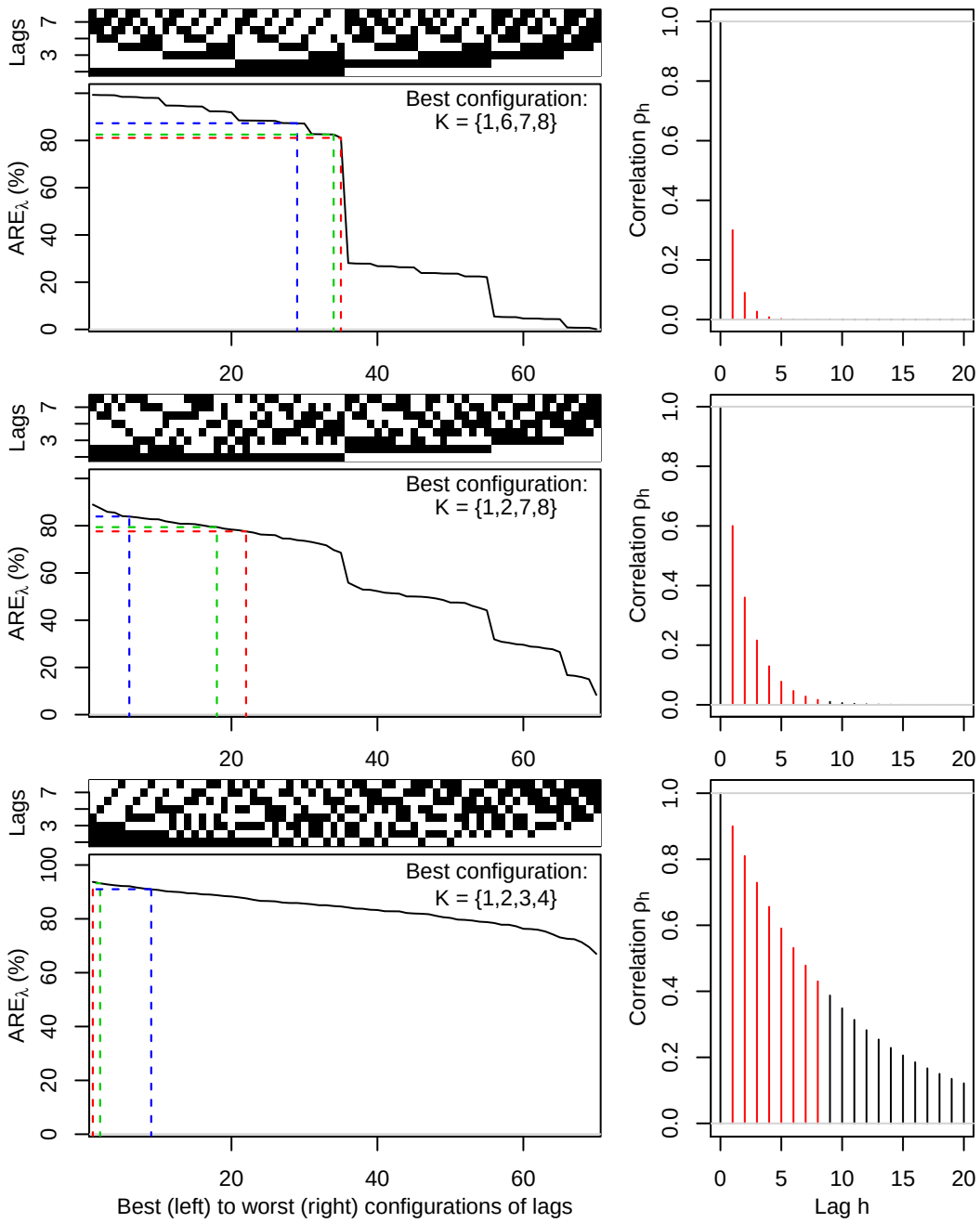

Figure 3.4: *Left*: Asymptotic relative efficiencies ARE_λ (solid line) of the estimator $\widehat{\psi}_{\mathscr{K}}$ for the AR(1) model, with $\sigma^2 = 1$ and $\lambda = 0.3$ (top), 0.6 (middle), 0.9 (bottom), with respect to the choice of $\mathscr{K}$. All configurations $\mathscr{K}$ of 4 time lags among $\{1,\ldots,8\}$ are compared (the black squares mean "lag included" while the blanks mean "not included"). The results are ordered from the best to the worst configuration. The dashed lines correspond to $\mathscr{K}_a^4$ (red), $\mathscr{K}_b^4$ (green) and $\mathscr{K}_c^4$ (blue). *Right*: Corresponding ACFs, with lags $1,\ldots,8$ in red.

variability of $\widehat{\psi}_{\mathscr{K}}$, and the corresponding pairs have less information than those at lower lags. Furthermore, the estimators based on $\mathscr{K}_a^4 = \{1,2,3,4\}$, $\mathscr{K}_b^4 = \{1,2,3,5\}$ and $\mathscr{K}_c^4 = \{1,2,4,8\}$ perform overall very well, though the latter is usually the best among the three options.

Figure 3.5 reports the results for the AR(2) model with $\sigma^2 = 1$ and $\lambda_1 = 0.8$, $\lambda_2 = -0.16$ (top row), $\lambda_1 = 0.15$, $\lambda_2 = 0.55$ (middle row), $\lambda_1 = 1$, $\lambda_2 = -0.9$ (bottom row). These parameter values were chosen to cover a large variety of behaviors: in the first case, the autocorrelation function (ACF) declines monotonically with the time lag h, and at an exponential rate as $h \to \infty$; in the second case, the ACF is a mixture of two exponentially decaying ACFs; and in the last case, the ACF oscillates around zero, and the absolute value of the correlation peaks has an exponentially decreasing behaviour, as h increases. Furthermore, in the first scenario, the dependence is relatively short-range, while in the second it is moderate, and in the last it can be strong at quite large time lags. It seems to be difficult to draw general conclusions about the optimal choice of $\mathscr{K}$ for the AR(2) model, and in fact it is maybe safer to decide the lags to be included on a case by case basis. Nevertheless, it seems that the two strongest dependent lags are often included in the best configuration, perhaps because the AR(2) model has two parameters. However, like for the AR(1) model, the best solution is usually not to include *all* strongest dependent pairs: we should instead retain, whenever possible, a mixture of very correlated pairs, and very uncorrelated ones. Again, the choices of pairs based on $\mathscr{K}_a^4$, $\mathscr{K}_b^4$ and $\mathscr{K}_c^4$ are reasonable.

Figure 3.6 shows the results for the ARMA(1,1) model with innovation variance $\sigma^2 = 1$, MA parameter $\vartheta_1 = 0.7$ and AR parameter $\lambda_1 = 0.3$ (top row), 0.6 (middle row) and 0.9 (bottom row). These three cases correspond to strong, moderate and weak dependence, respectively, as can be seen in the ACFs in the right panels of Figure 3.6. We have also tried other values for ϑ_1, but since it does not change qualitatively the results, we report only the relative efficiencies when $\vartheta_1 = 0.7$. By comparing Figures 3.4 and 3.6, we can see that the efficiencies for the AR coefficient are similar, though for the ARMA model, it is slightly lower because of the MA parameter to estimate. However, concerning the efficiency for the MA coefficient ϑ_1, it appears to be very close to zero, suggesting that unfortunately the ARMA(1,1) model inherits the bad efficiency behaviour of the pairwise likelihood estimator $\widehat{\psi}_{\mathscr{K}}$, observed for the MA(1) model; recall Figure 3.3. In fact, the efficiency of $\widehat{\psi}_{\mathscr{K}}$ is acceptable when ϑ_1 is rather small (that is, when the process resembles an autoregressive process), but when $|\vartheta_1| \approx 1$ the MA part of the process may be very badly estimated by $\widehat{\psi}_{\mathscr{K}}$ (though the AR part remains quite well estimated). Concerning the optimal configuration of time lags, there is a clear trade-off between including many pairs, which increases the information about the dependence parameters, and discarding some pairs in order to control the global

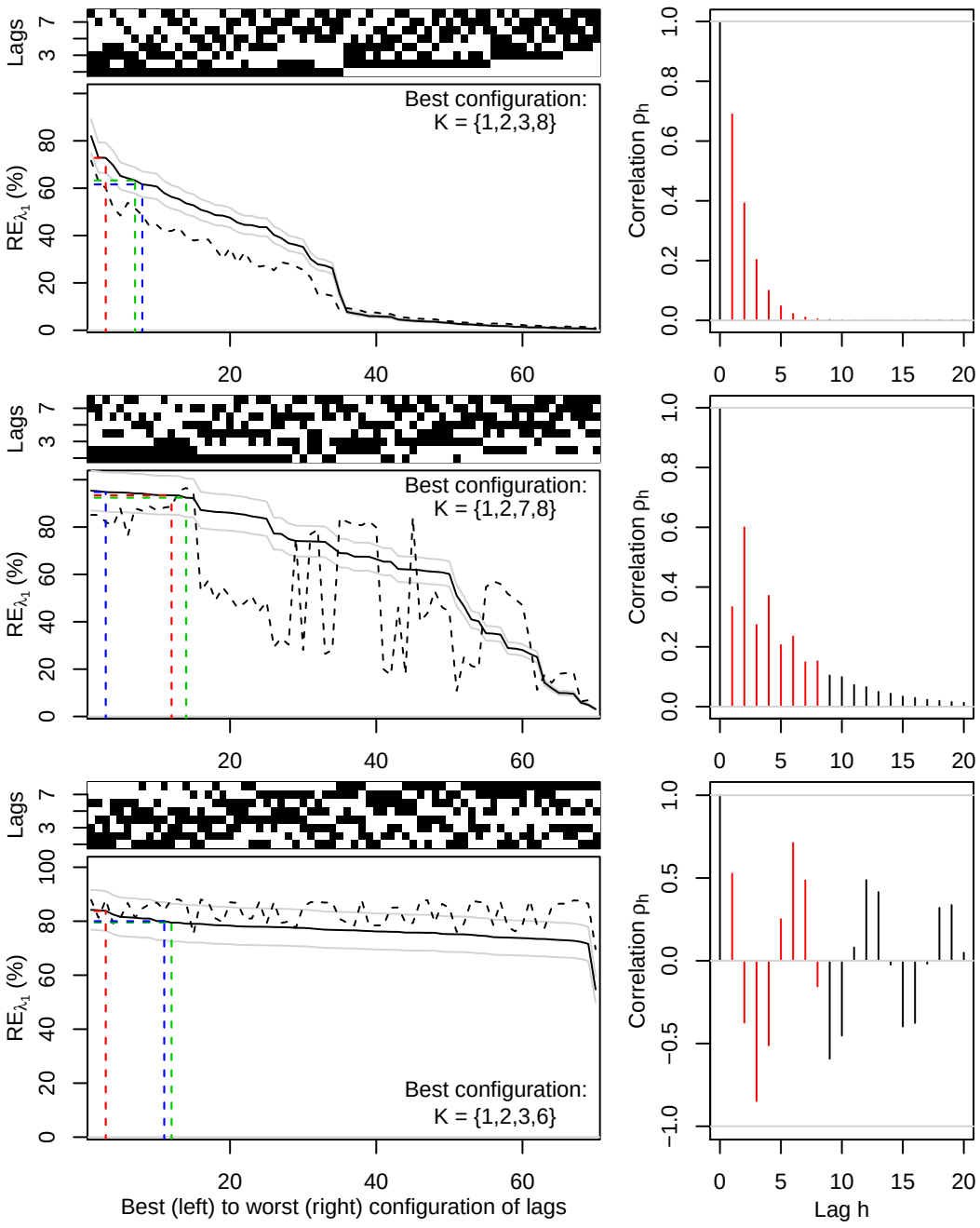

Figure 3.5: *Left*: Relative efficiencies RE_{λ_1} (solid black line) of the estimator $\widehat{\psi}_{\mathscr{K}}$, with 95% confidence bands (solid grey lines), for the AR(2) model, with $\sigma^2 = 1$ and $\lambda_1 = 0.8$, $\lambda_2 = -0.16$ (top), $\lambda_1 = 0.15$, $\lambda_2 = 0.55$ (middle), $\lambda_1 = 1$, $\lambda_2 = -0.9$ (bottom), with respect to the choice of $\mathscr{K}$. All configurations $\mathscr{K}$ of 4 time lags among $\{1,\dots,8\}$ are compared (the black squares mean "lag included"). The results are ordered from the best to the worst configuration. The dashed lines correspond to RE_{λ_2} (black), $\mathscr{K}_a^4$ (red), $\mathscr{K}_b^4$ (green) and $\mathscr{K}_c^4$ (blue). *Right*: Corresponding ACFs, with lags $1,\dots,8$ in red.

variability of $\widehat{\psi}_{\mathcal{K}}$. In the ARMA case, it seems essential to include the pairs at lags 1 and 2, probably because there are two parameters to estimate. And again, since these lags are sufficient to identify the parameters, inclusion of additional pairs damages less the estimator $\widehat{\psi}_{\mathcal{K}}$ when they are uncorrelated. As can be seen in the top panel of Figure 3.6, when $\lambda_1 = 0.3$, the best strategy is to consider $\mathcal{K} = \{1,2,7,8\}$, a mixture of strongly dependent and weakly dependent pairs. When lag 2 is not considered, the efficiency drops, and the configurations which do no include lags 1 and 2 are systematically worst. Furthermore, the choices of pairs based on $\mathcal{K}_a^4$, $\mathcal{K}_b^4$ and $\mathcal{K}_c^4$ are reasonable, though $\mathcal{K}_c^4$ is usually slightly better.

To summarize the findings of §3.3.2.1–3.3.2.2, the efficiency of pairwise likelihood estimators is reasonably high for autoregressive-type models, but may suffer from large losses compared to the maximum likelihood estimator, when moving average-type models, including ARMA models with positive MA coefficients, are considered. Therefore in practice, it seems to be essential to investigate whether the data mimic either type of model. For rainfall data, for example, "physical" considerations could suggest that an autoregressive model is likely to yield a reasonable fit, but since we cannot be sure of this, fitting needs care. Overall, estimators based on $\mathcal{K}_b^K$ or $\mathcal{K}_c^K$ seem to be recommendable, and the number of time lags to consider should be as small as possible, as far as the parameters remain identifiable.

3.3.2.3 Optimal weights for Gaussian processes

We now consider stationary isotropic Gaussian processes defined on $\mathbb{R}^2$, with zero mean, unit variance and for simplicity, we assume that the correlation function is exponential, $\rho(h) = \exp(-\|h\|/\lambda)$, $\lambda > 0$; recall Definition 41 and §2.1.

We want to understand how to best choose the weights in (3.8) using a piecewise linear specification, i.e.,

$$w_{k_1;k_2} = \omega(h) = \sum_{m=1}^{M} \omega_m I(a_m \leq \|h\| < a_{m+1}), \tag{3.49}$$

where $I(\cdot)$ is the indicator function, $0 \leq a_1 < \cdots < a_{M+1} \leq \infty$ define distinct distance classes, and $\omega_m \geq 0$, $m = 1,\ldots,M$, are positive class weights, such that $\sum_{m=1}^{M} \omega_m = 1$. In order to search the optimal weight function $\omega(h)$ for a given set of locations, we need to minimize the asymptotic sandwich variance $V_C(\lambda) = J_C(\lambda)^{-2} K_C(\lambda)$ in (3.12), where $J_C(\lambda) = \mathrm{E}\{-\partial^2 \ell(\lambda)/\partial\lambda^2\}$, $K_C(\lambda) = \mathrm{var}\{\partial\ell(\lambda)/\partial\lambda\}$, and $\ell(\lambda)$ is the log likelihood,

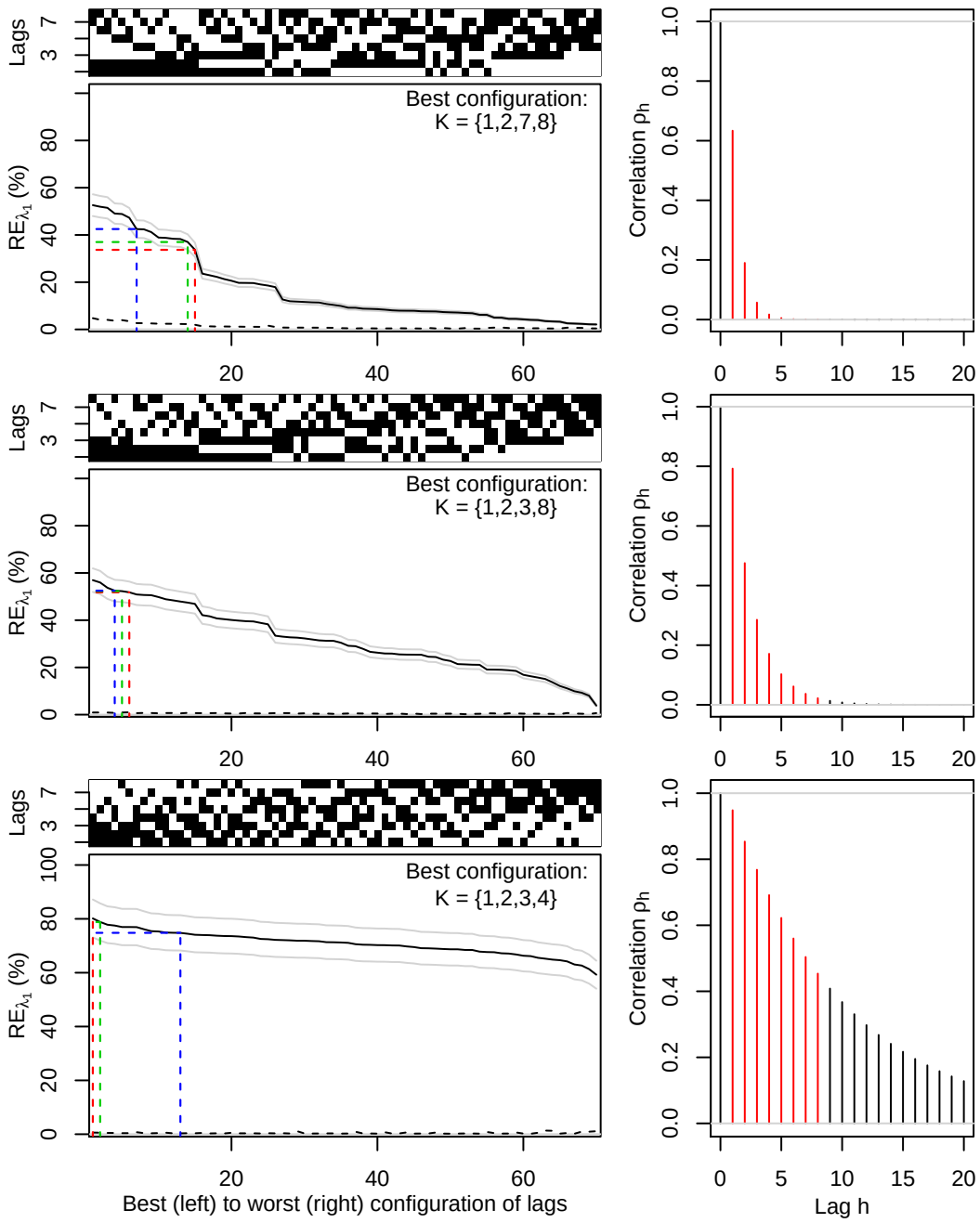

Figure 3.6: *Left*: Relative efficiencies RE_{λ_1} (solid black line) of the estimator $\widehat{\psi}_{\mathscr{K}}$, with 95% confidence bands (solid grey lines), for the ARMA(1,1) model, with $\sigma^2 = 1$ and $\lambda_1 = 0.3$, $\vartheta_1 = 0.7$ (top), $\lambda_1 = 0.6$, $\vartheta_1 = 0.7$ (middle), $\lambda_1 = 0.9$, $\vartheta_1 = 0.7$ (bottom), with respect to the choice of $\mathscr{K}$. All configurations $\mathscr{K}$ of 4 time lags among $\{1,\ldots,8\}$ are compared (the black squares mean "lag included"). The results are ordered from the best to the worst configuration. The dashed lines correspond to $\mathrm{RE}_{\vartheta_1}$ (black), $\mathscr{K}_a^4$ (red), $\mathscr{K}_b^4$ (green) and $\mathscr{K}_c^4$ (blue). *Right*: Corresponding ACFs, with lags $1,\ldots,8$ in red.

i.e., we look for the optimum

$$(\omega_1^\star, \ldots, \omega_M^\star) = \arg \min_{\omega_1,\ldots,\omega_M \geq 0} V_C(\lambda), \tag{3.50}$$

under the linear constraint

$$\sum_{m=1}^{M} \omega_m^\star = 1.$$

The weighted pairwise likelihood for a single replicate is, up to an additive constant,

$$\ell_P(\lambda) \equiv \sum_{s_1=1}^{S-1} \sum_{s_2=s_1+1}^{S} \omega(h_{12}) \left[-\frac{1}{2} \log\left\{1-\rho(h_{12})^2\right\} - \frac{1}{2\left\{1-\rho(h_{12})^2\right\}} \left\{z_1^2 - 2\rho(h_{12}) z_1 z_2 + z_2^2\right\} \right], \tag{3.51}$$

where $h_{12} = s_2 - s_1$ is the spatial lag, z_i denotes the value of the spatial process at the location s_i, and $\rho(h)$ is the exponential correlation with range parameter λ. Following the computations in §3.3.2.1, the composite score equals

$$\frac{\partial \ell_P(\lambda)}{\partial \lambda} = \sum_{s_1=1}^{S-1} \sum_{s_2=s_1+1}^{S} \omega(h_{12}) \frac{e^{-2\|h_{12}\|/\lambda}}{1-e^{-2\|h_{12}\|/\lambda}} \frac{\|h_{12}\|}{\lambda^2} \left\{ 1 - \frac{z_1^2 + z_2^2 - z_1 z_2 \left(e^{-\|h_{12}\|/\lambda} + e^{\|h_{12}\|/\lambda}\right)}{1-e^{-2\|h_{12}\|/\lambda}} \right\}.$$

Therefore, its variance is

$$\begin{aligned} \operatorname{var}\left\{\frac{\partial \ell_P(\lambda)}{\partial \lambda}\right\} &= \sum_{s_1<s_2} \sum_{s_3<s_4} \omega(h_{12})\omega(h_{34}) \Big[E(h_{12})E(h_{34}) \operatorname{cov}(z_1 z_2, z_3 z_4) \\ &\quad + F(h_{12})F(h_{34}) \left\{ \operatorname{cov}\left(z_1^2, z_3^2\right) + \operatorname{cov}\left(z_1^2, z_4^2\right) + \operatorname{cov}\left(z_2^2, z_3^2\right) + \operatorname{cov}\left(z_2^2, z_4^2\right) \right\} \\ &\quad + E(h_{12})F(h_{34}) \left\{ \operatorname{cov}\left(z_1 z_2, z_3^2\right) + \operatorname{cov}\left(z_1 z_2, z_4^2\right) \right\} \\ &\quad + F(h_{12})E(h_{34}) \left\{ \operatorname{cov}\left(z_1^2, z_3 z_4\right) + \operatorname{cov}\left(z_2^2, z_3 z_4\right) \right\} \Big], \end{aligned} \tag{3.52}$$

with

$$E(h) = \frac{\|h\| e^{\|h\|/\lambda} \left(1 + e^{2\|h\|/\lambda}\right)}{\lambda^2 \left(1 - e^{2\|h\|/\lambda}\right)^2}, \qquad F(h) = -\frac{\|h\| e^{2\|h\|/\lambda}}{\lambda^2 \left(1 - e^{2\|h\|/\lambda}\right)^2},$$

where

$$\operatorname{cov}(z_i z_j, z_k z_l) = e^{-(\|h_{ik} + h_{jl}\|)/\lambda} + e^{-(\|h_{il} + h_{jk}\|)/\lambda},$$

and where $h_{ij} = s_i - s_j$, $i, j = 1, 2, 3, 4$. The sensitivity component of the asymptotic sandwich variance reduces to

$$\begin{aligned} E\left\{\frac{\partial^2 \ell_P(\lambda)}{\partial \lambda^2}\right\} &= \sum_{s_1<s_2} \omega(h_{12}) \frac{\|h_{12}\|}{\lambda^4 \left(1 - e^{2\|h_{12}\|/\lambda}\right)^3} \\ &\quad \times \left\{ \|h_{12}\| \left(1 + e^{2\|h_{12}\|/\lambda}\right)^2 - 2e^{2\|h_{12}\|/\lambda} - \lambda \left(1 - e^{2\|h_{12}\|/\lambda}\right) \right\}, \end{aligned} \tag{3.53}$$

and the asymptotic variance $V_C(\lambda)$ of the maximum pairwise likelihood estimator $\widehat{\psi}_C(\lambda)$ is found by dividing (3.52) by the square of (3.53).

For a given configuration of sites, the optimal weighting function $\omega^\star(h)$ can be computed by resolving the constrained nonlinear optimization problem (3.50), using Lagrange multipliers. Since the constraints are linear, we can use the `R` function `constrOptim` for this.

In order to study the optimal weighting strategy in this framework, we first generated 300 sets of 20 locations independently in the unit square $[0,1]^2$, and considered eight distance classes defined by $a_m = (m-1)\sqrt{2}/8$, $m = 1,\ldots,9$. For each of these configurations of points, we then computed the optimal class weights $\omega_1^\star,\ldots,\omega_8^\star$ assuming $\lambda = 0.1$ (short-range dependence), 0.25 (mid-range dependence) and 0.9 (long-range dependence). The results, reported in Figure 3.7, show that when $\lambda = 0.1$, about 58% of the mass is attributed to the closest pairs, distant from 0.35 at most, and about 33% to those whose distance is between 0.71 and 1.24. The remaining pairs, however, are largely downweighted, even though the correlation is sometimes higher than for other distances. It seems to be the case that the estimation of a correlation function within a certain family, may be performed with small variability when one has information on the behaviour at its origin and its tail. When λ is larger, similar conclusions hold, except that some mass attributed to the tail is transferred to the origin. These results corroborate the findings obtained for simple time series in §3.3.2.2, in the sense that most of the mass is given to the strongest dependent pairs, but some nearly independent ones also have some significant weight. This correspondence is not very surprising, because the correlation function used here is the spatial counterpart of the ACF of an AR(1) model. It would be interesting to see whether these considerations hold for other types of correlation families, such as the powered exponential or Matérn ones.

Since these results rely on the distribution of locations over space, we repeated the same simulation study, but generating the locations uniformly in the rectangle $[0,0.5]\times[0,\sqrt{1.75}]$, so that the distribution of distances has changed, but the maximum distance is the same as before. However, the results do not change much qualitatively; see Figure 3.8.

This study could be extended in many respects. First, other correlation functions should be tested to validate the results in more general settings. Second, it could be interesting to see how this generalizes to a higher number of sites, or to other distributions of locations over space. Third, the weight function considered here is piecewise linear, and it would be better to consider other more flexible nonparametric forms, e.g., functions based on cubic splines. It would also be particularly interest-

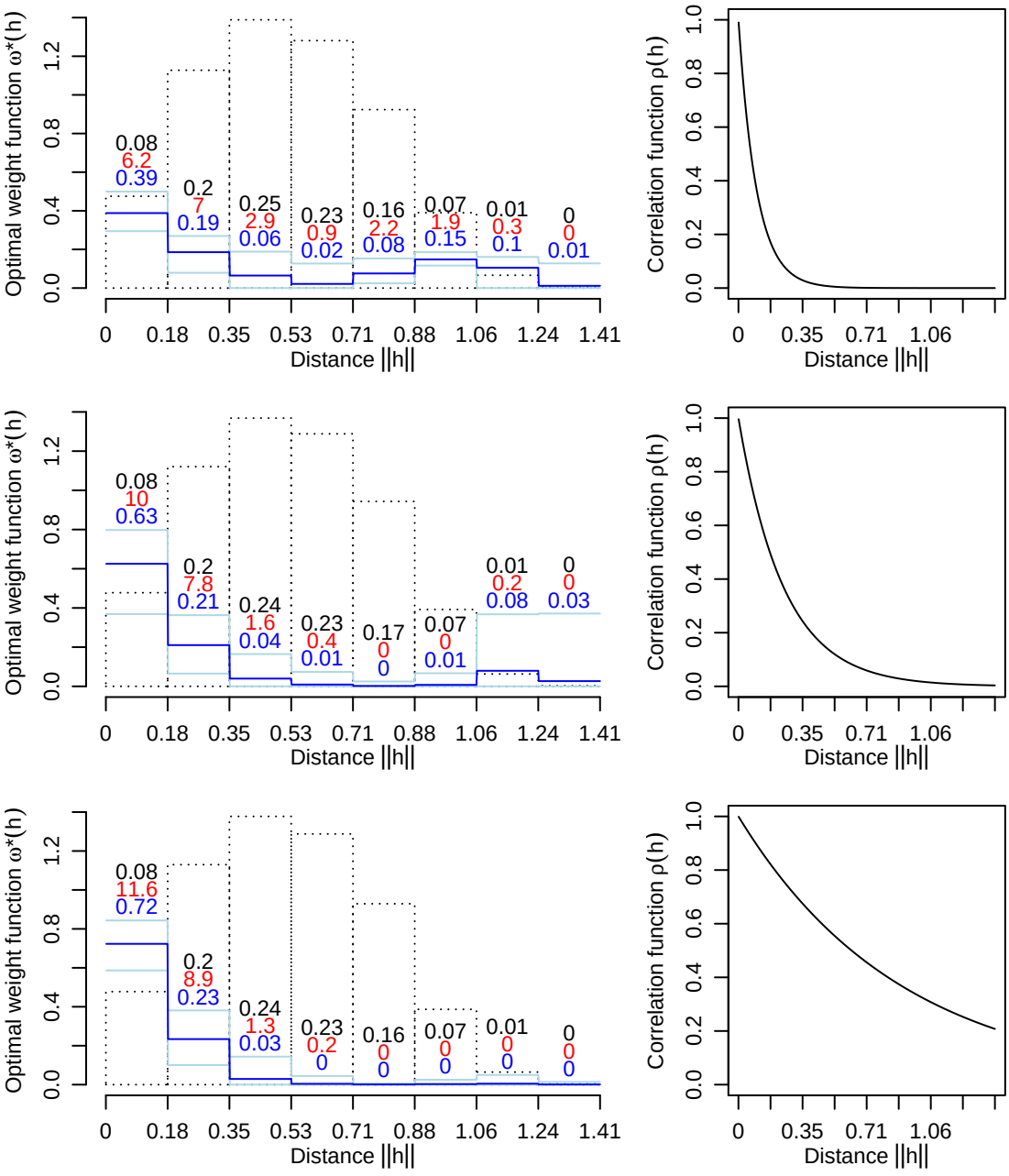

Figure 3.7: *Left*: Mean (blue) and 95%-confidence bands (light blue) of optimal weight functions $\omega^\star(h)$ resulting from 300 constrained minimizations (3.50), each based on 20 sites uniformly generated in $[0,1]^2$. The (dotted) histogram of distances is superimposed. The rows correspond to $\lambda = 0.1, 0.25, 0.9$ (from top to bottom). The numbers reported are the mean weight of a single pair (blue), the mean cumulated weight of all pairs in the distance class (red) and the expected proportion of pairs within the distance class (black). *Right*: Corresponding correlation functions $\rho(h)$.

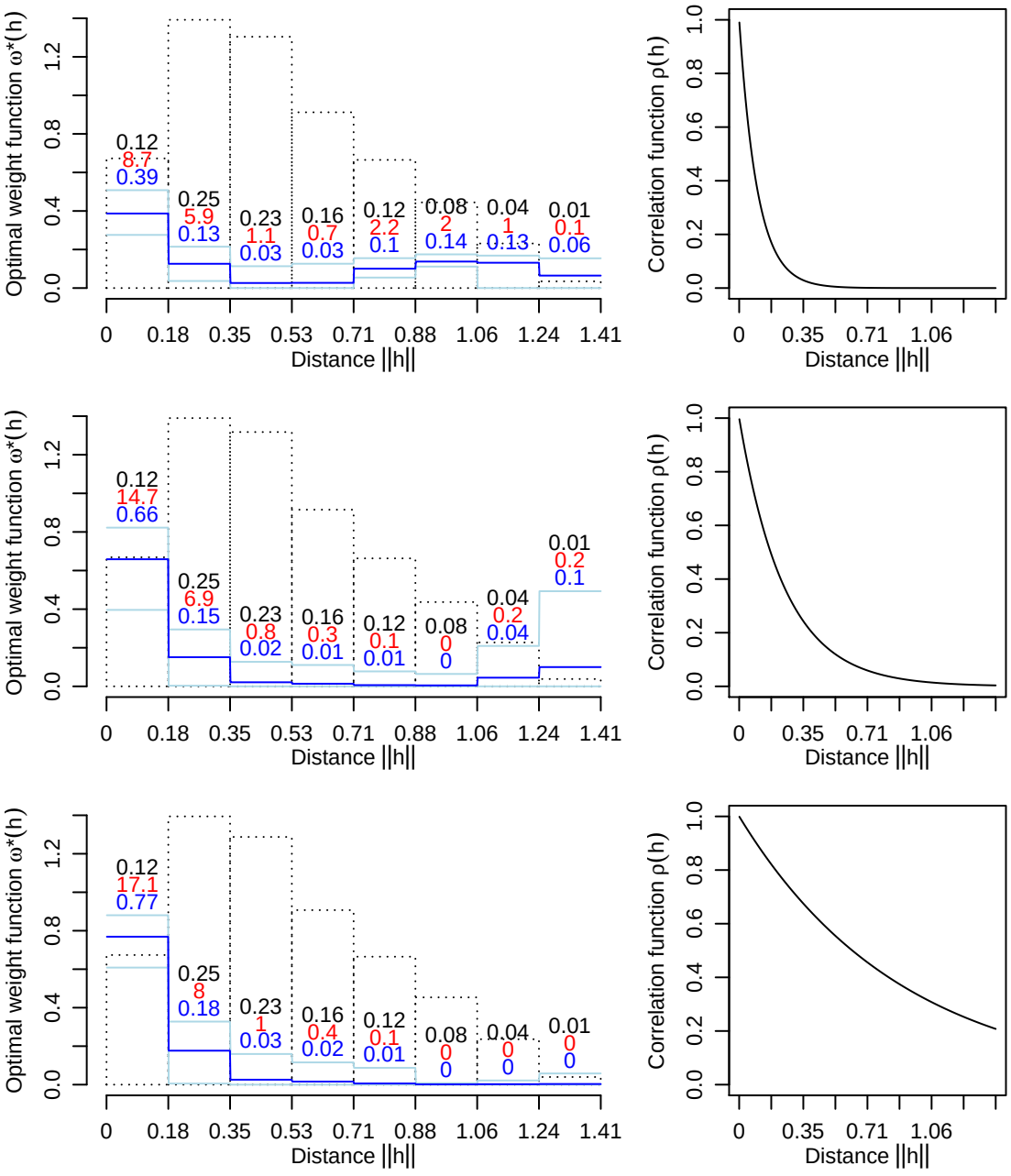

Figure 3.8: *Left*: Mean (blue) and 95%-confidence bands (light blue) of optimal weight functions $\omega^\star(h)$ resulting from 300 constrained minimizations (3.50), each based on 20 sites uniformly generated in $[0,0.5]\times[0,\sqrt{1.75}]$. The (dotted) histogram of distances is superimposed. The rows correspond to $\lambda = 0.1, 0.25, 0.9$ (from top to bottom). The numbers reported are the mean weight of a single pair (blue), the mean cumulated weight of all pairs in the distance class (red) and the expected proportion of pairs within the distance class (black). *Right*: Corresponding correlation functions $\rho(h)$.

ing to choose the knots non-regularly, yielding maybe a finer resolution for higher correlations, or to consider an adaptive configuration of knots which depends on the optimal weight function obtained at each optimization step. Finally, it would be natural to ask about the optimal way of *selecting* the pairs in the likelihood, that is, to consider binary weights $\omega_m \in \{0,1\}$ in (3.49) instead of $\omega_m \in \mathbb{R}_+$. This concern is, however, difficult to address in practice because the minimization (3.50) would have to be performed on a discrete set of values, which is not straightforward with classical optimization methods.

3.3.3 Max-stable models

Because the results of §3.3.2.1 for Gaussian processes do not apply directly to max-stable processes, we conduct two further simulation studies to assess the efficiency properties of maximum pairwise likelihood estimators in this framework. We first consider the multivariate logistic extreme-value model, for which we show how to perform approximate maximum *full* likelihood estimation, and compute the empirical efficiency of pairwise likelihoods in this context. We also discuss possible extensions to simple max-stable time series models. Second, we consider the Schlather model with random set in $\mathbb{R}$ and discuss the efficiency of several weighted pairwise likelihood estimators under three different estimation procedures.

3.3.3.1 Maximum likelihood estimation for the logistic model

The logistic extreme-value model in D dimensions, generalizing (1.30), has joint distribution

$$\Pr(Z_1 \leq z_1, \ldots, Z_D \leq z_D) = \exp\{-V(z_1, \ldots, z_D)\}, \tag{3.54}$$

where

$$V(z_1, \ldots, z_D) = \left(\sum_{d=1}^{D} z_d^{-1/\alpha} \right)^{\alpha}, \tag{3.55}$$

and $\alpha \in (0,1]$ is a dependence parameter: when $\alpha = 1$, the variables are independent and when $\alpha \to 0$, they become increasingly dependent. According to Stephenson (2009) and Fougères *et al.* (2009), a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_D)$ following model (3.54) admits the conditional representation

$$\Pr(Z_1 \leq z_1, \ldots, Z_D \leq z_D \mid S) = \prod_{d=1}^{D} \exp\left(-S z_d^{-1/\alpha}\right), \tag{3.56}$$

where $S \sim \mathrm{PS}(\alpha)$ is a positive α-stable random variate; recall §2.3.3. Hence, conditional on S, the vector $\boldsymbol{Z}$ is a set of independent Fréchet random variables with shape param-

eter $1/\alpha$ and scale parameter S^α. Thus, for n independent observations distributed according to $\boldsymbol{Z}$, denoted by $\boldsymbol{z}_i = (z_{i;1}, \ldots, z_{i;D})$, $i = 1, \ldots, n$, the log-likelihood is

$$\ell(\alpha) = \sum_{i=1}^{n} \log\left(\mathrm{E}\left[\prod_{d=1}^{D} \left\{ \frac{S_i}{\alpha} z_{i;d}^{-1/\alpha - 1} \exp\left(-S_i z_{i;d}^{-1/\alpha} \right) \right\} \right] \right), \tag{3.57}$$

where the expectation is with respect to $S_i \overset{\text{iid}}{\sim} \mathrm{PS}(\alpha)$, $i = 1, \ldots, n$. Since the latter is intractable, the log-likelihood may be approximated by

$$\ell(\alpha) \approx \ell_K(\alpha) = \sum_{i=1}^{n} \log\left[\frac{1}{K} \sum_{k=1}^{K} \prod_{d=1}^{D} \left\{ \frac{S_{i;k}}{\alpha} z_{i;d}^{-1/\alpha - 1} \exp\left(-S_{i;k} z_{i;d}^{-1/\alpha} \right) \right\} \right], \tag{3.58}$$

where $S_{i;k} \overset{\text{iid}}{\sim} \mathrm{PS}(\alpha)$, $i = 1, \ldots, n$, $k = 1, \ldots, K$. Hence, for a single evaluation of the log-likelihood, we need to simulate nK independent positive stable random variates, or equivalently $2nK$ uniform variables in $[0,1]$, using the algorithm (2.40). The larger K, the better the approximation. According to Gouriéroux & Monfort (1991), $K = K(n)$ should be such that $n^{1/2} K(n)^{-1} \to 0$, as $n \to \infty$, to ensure that the usual and simulated maximum likelihood estimators, maximizing (3.57) and (3.58) respectively, are asymptotically equivalent. In practice, if we let for example $K(n) = n$, then we need to simulate n^2 positive stable random variates for a single evaluation of the likelihood, which can be quite computationally expensive. Furthermore, there may be many values of S, which do not contribute much to the expectation on the right-hand side of (3.57). Hence, variance reduction techniques, and in particular importance sampling, are essential for a good performance of the simulated maximum likelihood estimator.

The argument inside the expectation in (3.57), when viewed as a function of S_i, is proportional to a gamma density with shape parameter $D+1$ and rate parameter $\sum_{d=1}^{D} z_{i;d}^{-1/\alpha}$. Let $q_{0.005;i}$ and $q_{0.995;i}$ denote its 0.5% and 99.5%-quantiles. An importance sampling-based maximum likelihood estimator may be constructed by sampling most of the random variates $S_{i;k}$, $k = 1, \ldots, K$, within the range $[q_{0.005;i}, q_{0.995;i}]$, for each $i = 1, \ldots, n$, and reweighting the contributions accordingly. Although not quite optimal, this solution was found to yield good results in practice. However, in high dimensions, it can be burdensome to compute the product inside the logarithm in (3.58) because it may be too small and out of the machine precision. Table 3.1 reports the relative errors when computing the simulated log-likelihood with or without importance sampling for $K = n, 4n, n^2$, assuming that the true values are well approximated by formula (3.58), using $K = n^4$ (without importance sampling). We can see that there is a huge improvement when importance sampling is used, especially in case of strong dependence (small α). Overall, the simulated maximum likelihood estimator based

Table 3.1: Absolute relative errors (%) for the computation of the log-likelihood function of n independent observations from the D-variate logistic extreme-value model with dependence parameter $\alpha = 0.1, \ldots, 0.9$, using the approximation (3.58) without/with importance sampling. We used $n = 10, 20$, $D = 20, 50$, and $K = n, 4n, n^2$. The values below are the means of 300 independent simulated log-likelihoods evaluated at the "true" parameter, each *conditioned* on a common set of observations generated from the same random seed. The cases where the use of importance sampling does not improve upon the "naive" method are highlighted in red.

	$n = 10$, $D = 20$			$n = 10$, $D = 50$		
$\alpha \setminus K$	n	$4n$	n^2	n	$4n$	n^2
0.1	1374/0.25	286/0.43	70/0.53	2194/0.24	644/0	156/0.07
0.2	181/0.20	19/0.07	3.8/0.05	195/0.09	27/0.02	5.1/0.01
0.3	61/0.13	4.6/0.05	0.87/0.01	66/0.04	5.2/0.02	0.94/0.03
0.4	22/0.14	1.7/0.02	0.40/0.01	25/0.05	1.7/0.01	0.32/0
0.5	8/0.15	0.63/0	0.21/0.02	10/0.06	0.64/0	0.15/0.01
0.6	3.5/0.16	0.30/0	0.09/0.04	4.4/0.06	0.35/0.02	0.08/0.04
0.7	1.5/0.21	0.20/0.01	0.11/0.02	1.8/0.06	0.20/0.01	0.06/0.03
0.8	0.62/0.25	0.09/0.02	0.03/0.02	0.85/0.11	0.12/0.01	0.04/0.04
0.9	0.26/0.56	0.08/0.26	0.02/0.1	0.24/0.15	0.04/0.02	0.02/0.02
	$n = 20$, $D = 20$			$n = 20$, $D = 50$		
$\alpha \setminus K$	n	$4n$	n^2	n	$4n$	n^2
0.1	1569/0.7	287/0.26	23/0.19	4517/0.76	1004/1.2	72/1.3
0.2	67/0.11	14/0.04	0.90/0.02	80/0.01	15/0.02	1.2/0.02
0.3	29/0.13	4.7/0.03	0.31/0	29/0.04	4.9/0	0.36/0.01
0.4	14/0.16	2.2/0.01	0.13/0.02	16/0.07	2.7/0.01	0.14/0
0.5	8.0/0.18	1.2/0.03	0.11/0.02	8.7/0.06	1.6/0.01	0.07/0.02
0.6	4.7/0.23	0.60/0.03	0.05/0.04	5.7/0.10	0.81/0	0.06/0.02
0.7	2.9/0.36	0.35/0.03	0.05/0.03	3.2/0.13	0.47/0.01	0.03/0.03
0.8	1.7/0.54	0.23/0.07	0.02/0.04	2.0/0.20	0.25/0.01	0.02/0.03
0.9	0.53/0.74	0.09/0.24	0.02/0.05	0.80/0.33	0.12/0.07	0.01/0.02

on importance sampling performs reasonably well, although the relative errors may not be very low when $\alpha \leq 0.1$ or $\alpha \geq 0.8$, and for $\alpha \geq 0.9$ (near-independence case) it is sometimes best to use the "naive" simulated maximum likelihood estimator.

Furthermore, Figure 3.9 illustrates that, unlike the approach based on importance sampling, the "naive" simulated log-likelihood may be very variable and unreliable when α is small, and thus that its maximization is burdensome.

We then compared simulated maximum likelihood estimators based on importance

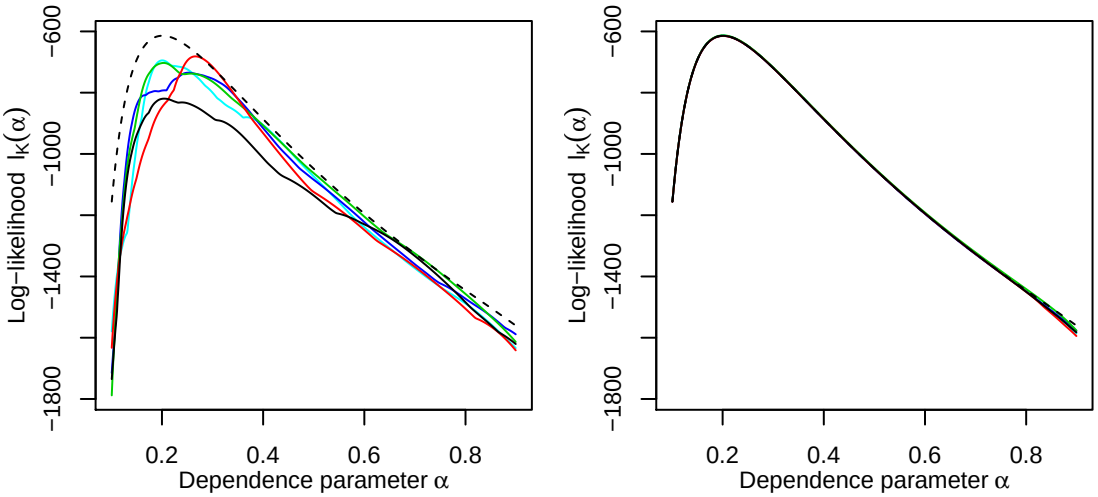

Figure 3.9: Five replications (solid lines) of the simulated log-likelihood (3.58) based on the naive approach (left) and importance sampling (right) with $K = 20$, for a dataset composed of $n = 20$ independent observations generated from model (3.54) with $\alpha = 0.2$ and $D = 50$. The dashed line is the true log-likelihood. All curves are almost superimposed in the right panel.

sampling and equally weighted pairwise likelihood estimators. Both are strongly consistent and asymptotically Gaussian, and Table 3.2 reports their empirical relative efficiencies calculated from 300 independent replicates, for $\alpha = 0.1, \ldots, 0.9$, $D = 20, 50, 100$ and $n = 10, 20, 50, 100$. For the simulated maximum likelihood estimator, we used $K = nD/5$, meaning that K is chosen between $4n$ and $2n^2$ for the selected values of n and D. The results show that the loss in efficiency of pairwise likelihood estimators may be substantial if the data are high-dimensional, and if dependence is weak. When α is small (strong dependence), the efficiencies are reasonably high, but as α approaches unity, they decrease and reach about 10% only, when $\alpha = 0.9$ and $D = 100$.

The logistic model (3.54), although appealing because it permits maximum likelihood estimation, is too rigid and simplistic to be realistic in applications. It can be generalized to the more flexible asymmetric logistic model (Tawn, 1988b), which has a very large number of parameters in high dimensions and so may overfit the data. However, several interesting submodels, which have analogue conditional representations in terms of positive stable variates, have been proposed by Fougères *et al.* (2009). One possible model for time series is based on a hidden autoregressive positive stable process. More precisely, let us consider a stream of independent positive α-stable random variables $S_t \overset{\text{iid}}{\sim} \text{PS}(\alpha)$ and, for some parameter $\rho \in [0, 1)$, let us define a hidden autoregressive "shock" process H_t, $t \in \mathbb{Z}$, by the implicit equation $H_t = \rho H_{t-1} + S_t$. It follows that $H_t = \sum_{s=0}^{\infty} \rho^t S_{t-s}$. Then, assuming that $\varepsilon_t \overset{\text{iid}}{\sim} \text{GEV}(1, \alpha, \alpha)$ is a white noise

Table 3.2: Empirical relative efficiency (%) of the equally weighted pairwise likelihood estimator with respect to the maximum full likelihood estimator, for $n = 10, 20, 50, 100$ independent data generated from the D-dimensional logistic extreme-value model (3.54) with $D = 20, 50, 100$ and dependence parameter $\alpha = 0.1, \ldots, 0.9$. For the computation of the full likelihood, we used the approximation (3.58) with importance sampling and $K = nD/5$. The empirical efficiencies below are based on 300 independent replicates.

	$n=10$			$n=20$			$n=50$			$n=100$		
$\alpha \setminus D$	20	50	100	20	50	100	20	50	100	20	50	100
0.1	81	67	65	73	82	67	78	72	69	70	82	59
0.2	65	70	59	64	57	53	70	61	45	78	55	58
0.3	70	44	36	74	45	39	60	56	43	66	52	33
0.4	62	41	28	62	43	28	64	43	31	53	41	23
0.5	40	32	18	48	32	19	58	33	20	53	30	22
0.6	46	25	13	46	24	11	44	27	14	43	25	15
0.7	35	18	10	40	21	10	40	19	11	45	23	13
0.8	34	16	9	35	18	10	40	17	11	33	14	8
0.9	24	18	10	36	19	8	27	18	9	32	12	8

process, the time series

$$Z_t = (1-\rho^\alpha) H_t^\alpha \varepsilon_t, \qquad t \in \mathbb{Z}, \tag{3.59}$$

is stationary max-stable with unit Fréchet margins, and the segment of observations recorded at times $t = 1, \ldots, n$, has an exponent measure with asymmetric logistic form, which may be written as

$$V(z_1, \ldots, z_n) = \sum_{i=1}^{n} (1-\rho^\alpha)^{I(i\neq 0)} \left\{ \sum_{j=0}^{n-i} \left(\frac{z_{i+j}}{\rho^{j\alpha}} \right)^{-1/\alpha} \right\}^\alpha,$$

where $I(\cdot)$ denotes the indicator function. Furthermore, the bivariate exponent measure for the variables at times 0 and $h > 0$ can be expressed as

$$V(z_0, z_h) = \left(z_0^{-1/\alpha} + \rho^h z_h^{-1/\alpha} \right)^\alpha + \frac{1-\rho^{\alpha h}}{z_h},$$

meaning that the bivariate extremal coefficient is $\theta_2(h) = \left(1+\rho^h\right)^\alpha + 1 - \rho^{\alpha h}$. Hence, complete independence is obtained when $\alpha = 1$ (for any ρ and h), when $\rho = 0$ (unless $\alpha = 0$, or $h = 0$), or as $h \to \infty$ (unless $\alpha = 0$), while the symmetric logistic dependence structure is recovered as $\rho \to 1$. Perfect dependence occurs for $h = 0$ or as $\alpha \to 0$; see Figure 3.10. In summary, α is a parameter of *global* dependence, while ρ determines the *range* of dependence. Simulations of this process are illustrated in Figure 3.11. We

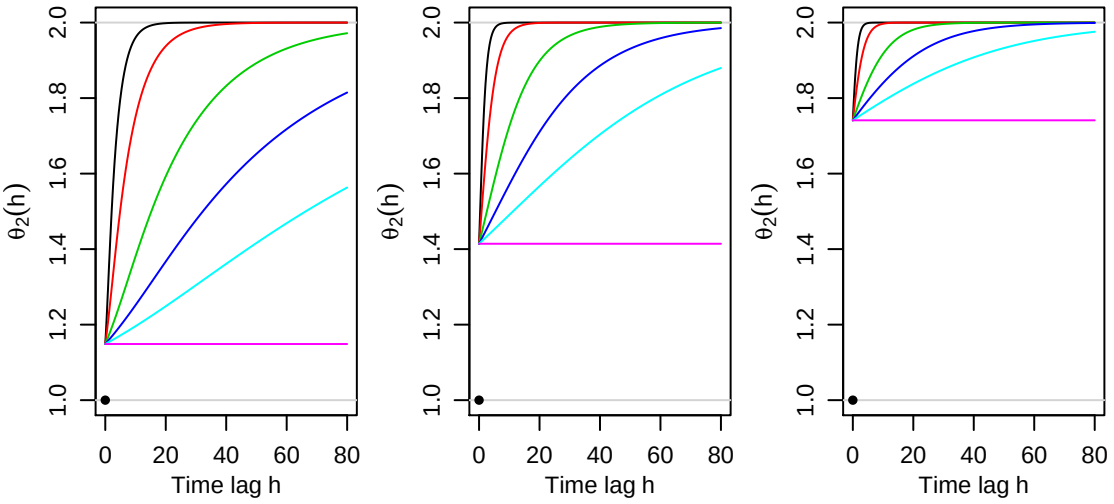

Figure 3.10: Bivariate extremal coefficient $\theta_2(h)$ for the logistic time series model (3.59) with $\alpha = 0.2$ (left), 0.5 (middle), 0.8 (right), and $\rho = 0.2$ (black), 0.5 (red), 0.8 (green), 0.9 (dark blue), 0.95 (light blue) and $\rho \to 1$ (purple). The black dots are the values at lag $h = 0$.

can see that the value of α determines the amount of noise in the series, whereas ρ is linked to the shock decays (i.e., the slopes of the "sawtooth" patterns in the plot of $\log(H_t)$).

Similarly to the symmetric logistic model, the log-likelihood for such a temporal process, based on the segment of observations $z_1, \ldots, z_n$, may be written as

$$\ell(\psi) = \log\left\{ \mathrm{E}\left(\prod_{t=1}^{n} \left[\frac{1}{\alpha} z_t^{-1/\alpha - 1} \exp\left\{ -H_t (1-\rho^\alpha)^{1/\alpha} z_t^{-1/\alpha} \right\} \right] \right) \right\}, \tag{3.60}$$

where $\psi = (\alpha, \rho)$ is the parameter vector and the expectation is with respect to the hidden α-stable variates, and it may be approximated by

$$\ell_K(\psi) = \log\left(\frac{1}{K} \sum_{k=1}^{K} \prod_{t=1}^{n} \left[\frac{1}{\alpha} z_t^{-1/\alpha - 1} \exp\left\{ -H_{t;k} (1-\rho^\alpha)^{1/\alpha} z_t^{-1/\alpha} \right\} \right] \right), \tag{3.61}$$

where $H_{t;k}$, $k = 1, \ldots, K$, are independent replicates of H_t. However, in practice, this approximation may be dubious and difficult to use for several reasons: First, (3.61) is an approximation to an n-fold integral using Monte Carlo components which may be very variable, so problems may arise for large n. Second, it is not obvious how to implement importance sampling in this context, because the process has a more complicated structure than the symmetric logistic model. Third, the product inside the logarithm in (3.61) may be very close to zero for moderate values of n, so it seems

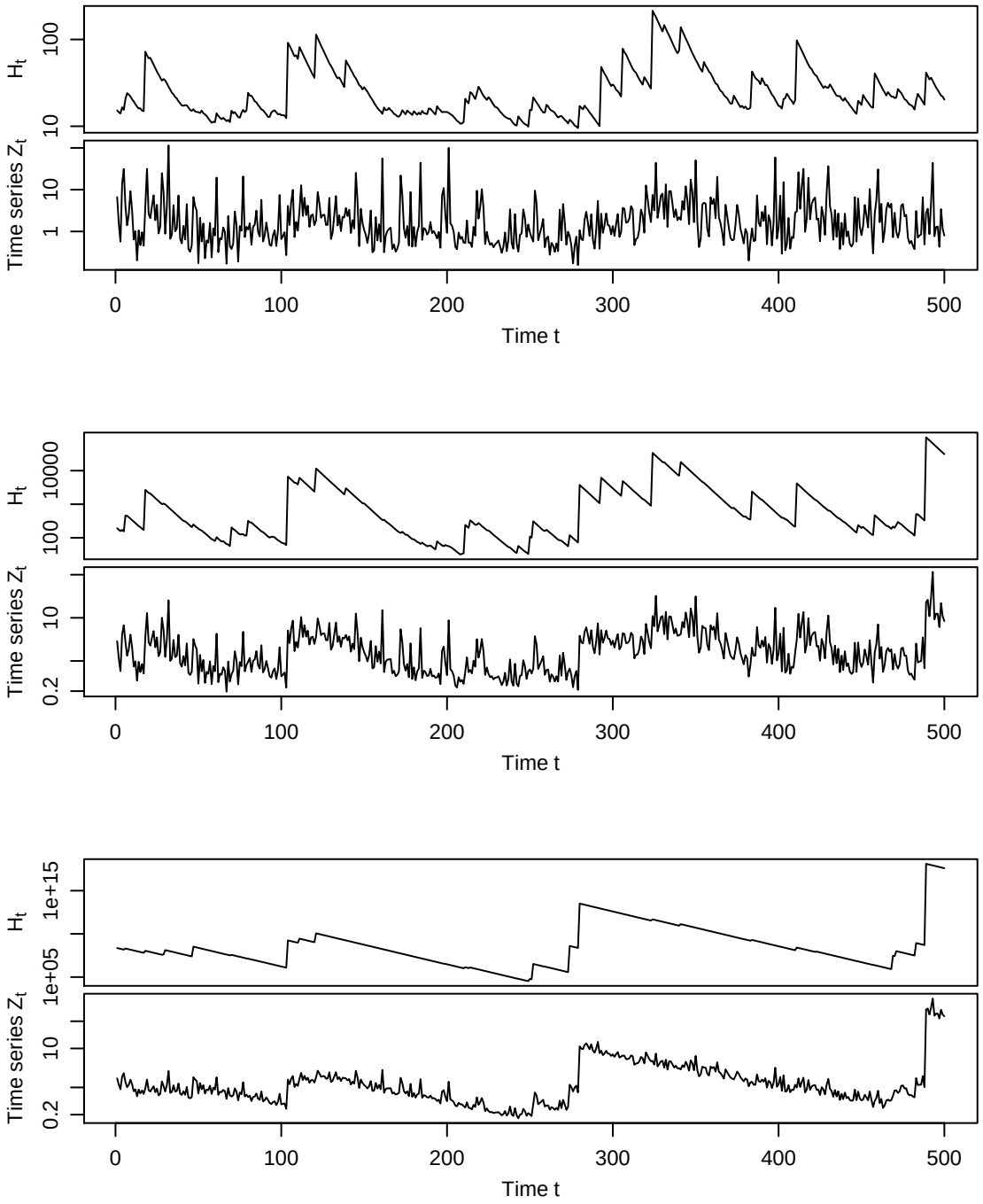

Figure 3.11: Three simulated time series from model (3.59) with $\rho = 0.9$ and $\alpha = 0.2$ (bottom panels), 0.5 (middle panels) and 0.8 (top panels), generated from the same random seed. The corresponding hidden autoregressive "shock" processes H_t are displayed above the time series.

that the "naive" computation of (3.61) is numerically infeasible when the record has more than a few hundred data points. Other techniques need to be used to perform inference. Below, we propose three possible alternatives: the first one is to use a Monte Carlo EM algorithm; the second one consists in embedding the model in a Bayesian framework, and in using standard MCMC methods for inference; and the third one, which works best in practice, is to view the model as a nonlinear non-Gaussian state-space model and to use simulation-based particle filters to fit it.

Inference based on a Monte Carlo EM algorithm. The classical EM (expectation maximization) algorithm (Dempster *et al.*, 1977; Gilks *et al.*, 1995, p.260) has found widespread interest for handling hidden (or missing) data. Suppose that the observed, respectively unobserved, data are denoted by $\boldsymbol{z}_o$, and $\boldsymbol{z}_u$. The EM algorithm may be of interest if $\ell(\psi; \boldsymbol{z}_o, \boldsymbol{z}_u)$, the log-likelihood based on the complete dataset $(\boldsymbol{z}_o, \boldsymbol{z}_u)$, is much simpler than $\ell(\psi; \boldsymbol{z}_o)$, the log-likelihood based solely on the observed data. Starting from some initial parameter vector $\psi^{(0)}$, the algorithm consists of iterating the two following successive steps until convergence:

- *E-step:* Compute $Q(\psi) = \mathrm{E}\{\ell(\psi; \boldsymbol{z}_o, \boldsymbol{z}_u)\}$, where the expectation is with respect to the distribution of $(\boldsymbol{z}_o, \boldsymbol{z}_u) \mid \boldsymbol{z}_o$, under the assumption that $\psi = \psi^{(r)}$.
- *M-step:* Maximize $Q(\psi)$ with respect to ψ, yielding the new parameter $\psi^{(r+1)}$.

According to Dempster *et al.* (1977), the value of the log-likelihood increases at each iteration. However, although appealing in this respect, the EM algorithm has some well-documented shortcomings: in particular, it can converge to local maxima and may be very sensitive to starting values. Furthermore, the expectation involved in the E-step may be intractable. To circumvent this issue, Chan & Ledolter (1995) have proposed a Monte Carlo EM algorithm, in which this expectation is substituted by its empirical counterpart, based on an ergodic sample from $(\boldsymbol{z}_o, \boldsymbol{z}_u) \mid \boldsymbol{z}_o$ (possibly generated using the well-known Metropolis-Hastings algorithm, see Hastings, 1970). Unfortunately, with this solution, the log-likelihood is random and its value does not necessarily increase at each iteration. Hence, the algorithm does not converge to a single value, and it may be difficult to find a good stopping criterion. Another drawback is that the Monte Carlo EM algorithm is computationally very demanding, because each iteration requires the simulation of a (perhaps large) Markov Chain, and it might be tricky to tune it. An alternative related approach is the so-called stochastic EM algorithm (Gilks *et al.*, 1995, p.261), in which the missing data are imputed at each step in a "clever" way.

For the max-stable time series model (3.59), the hidden process H_t is constructed from independent positive stable variates S_t, which may be generated from uniform

variables U_t and W_t; recall (2.40). Hence, let $\boldsymbol{z}_o = (z_1, \ldots, z_n)$ denote the observed data, and let $\boldsymbol{z}_u = (\boldsymbol{U}, \boldsymbol{W})$, where $\boldsymbol{U} = (U_1, \ldots, U_n)$ and $\boldsymbol{W} = (W_1, \ldots, W_n)$, denote the unobserved (hidden) data. The log-likelihood for the complete dataset may be written as

$$\begin{aligned} \ell(\psi; \boldsymbol{z}_o, \boldsymbol{z}_u) &= -n\log(\alpha) + \frac{n}{\alpha}\log\left(1-\rho^\alpha\right) + \sum_{t=1}^{n}\log\left(H_t\right) - \left(\frac{1}{\alpha}+1\right)\sum_{t=1}^{n}\log(z_t) \\ &\quad - \left(1-\rho^\alpha\right)^{1/\alpha}\sum_{t=1}^{n} H_t z_t^{-1/\alpha}, \end{aligned} \tag{3.62}$$

where $H_1 = (1-\rho^\alpha)^{-1/\alpha} S_1$, $H_t = \rho H_{t-1} + S_t$, $t = 2, \ldots, n$, and where

$$S_t = \left\{\frac{h(\pi U_t)}{-\log(W_t)}\right\}^{(1-\alpha)/\alpha}, \qquad h(\omega) = \left\{\frac{\sin(\alpha\omega)}{\sin(\omega)}\right\}^{1/(1-\alpha)} \frac{\sin\{(1-\alpha)\omega\}}{\sin(\alpha\omega)}.$$

Hence, the functional $Q(\psi)$ of the E-step may be expressed as

$$\begin{aligned} Q(\psi) &= -n\log(\alpha) + \frac{n}{\alpha}\log\left(1-\rho^\alpha\right) + \sum_{t=1}^{n}\mathrm{E}\left\{\log\left(H_t\right)\right\} - \left(\frac{1}{\alpha}+1\right)\sum_{t=1}^{n}\log(z_t) \\ &\quad - \left(1-\rho^\alpha\right)^{1/\alpha}\sum_{t=1}^{n}\mathrm{E}\left(H_t\right) z_t^{-1/\alpha}, \end{aligned} \tag{3.63}$$

where the expectations are with respect to the conditional distribution of $\boldsymbol{z}_u \mid \boldsymbol{z}_o$, under the assumption that $\psi = \psi^{(r)}$. Since this joint distribution is intractable, we need recourse to the Monte Carlo EM algorithm. The good news with (3.63) is that the product inside the logarithm in (3.60) has been transformed into a sum outside the logarithm, meaning that this method is expected to work better than the approximation (3.61) for longtime series. The bad news is that sampling from the distribution of $\boldsymbol{z}_u \mid \boldsymbol{z}_o$ is difficult and computationally intensive, which seems to preclude this approach in practice.

Bayesian inference. In order to perform Bayesian inference, prior distributions $\pi(\psi)$ reflecting our prior knowledge about the model parameters need to be assigned. The posterior distribution for ψ may be expressed as

$$\pi(\psi \mid \boldsymbol{z}_o) = \int \pi(\psi, \boldsymbol{z}_u \mid \boldsymbol{z}_o)\pi(\boldsymbol{z}_u \mid \boldsymbol{z}_o)\mathrm{d}\boldsymbol{z}_u$$

where $\boldsymbol{z}_o$ and $\boldsymbol{z}_u$ are respectively the observed and unobserved data as before, and where $\pi(\cdot)$ denotes a generic density component. Treating $\boldsymbol{z}_u$ as a hidden variable, an MCMC algorithm may be summarized as

(1) Start with an initial parameter $\psi^{(0)}$. Then, for $r = 0, \ldots, r_{\max}$, iterate the steps (2–3);

(2) Simulate from $\boldsymbol{z}_u \mid (\psi^{(r)}, \boldsymbol{z}_o)$;

(3) Simulate from $\psi \mid (\boldsymbol{z}_u, \boldsymbol{z}_o)$, yielding the new parameter $\psi^{(r+1)}$.

Since step (2) involves the simulation of $2n$ dependent hidden variables, it might be advisable to design efficient block updates (Fearnhead, 2011, p.515). Furthermore, since the above full conditionals are unknown, Gibbs sampling is not possible and Metropolis–Hastings updates are needed. The target density, namely $\pi(\psi, \boldsymbol{z}_u \mid \boldsymbol{z}_o)$, may be expressed as

$$\pi(\psi, \boldsymbol{z}_u \mid \boldsymbol{z}_o) \propto \pi(\boldsymbol{z}_o, \boldsymbol{z}_u \mid \psi)\pi(\psi).$$

In terms of log-densities, we have

$$\log\{\pi(\psi, \boldsymbol{z}_u \mid \boldsymbol{z}_o)\} \equiv \ell(\psi; \boldsymbol{z}_o, \boldsymbol{z}_u) + \log\{\pi(\psi)\},$$

where the first bit on the right-hand side comes from (3.62). In practice, it is common to choose independent priors for α and ρ, so that $\log\{\pi(\psi)\} = \log\{\pi(\alpha)\} + \log\{\pi(\rho)\}$. We end up with the following MCMC algorithm:

(1) Start with initial values $\psi^{(0)} = (\alpha^{(0)}, \rho^{(0)})$, and $\boldsymbol{z}_u^{(0)}$. Then, for $r = 0, \ldots, r_{\max}$, iterate the steps (2–4);

(2) Update the hidden variables (possibly in blocks) as follows:

   (i) Generate a candidate value $\boldsymbol{z}_u^\star$ from a proposal density $q(\boldsymbol{z}_u \mid \boldsymbol{z}_u^{(r)})$;

   (ii) Compute the log-acceptance probability

$$p_{\boldsymbol{z}_u} = \ell(\psi^{(r)}; \boldsymbol{z}_o, \boldsymbol{z}_u^\star) - \ell(\psi^{(r)}; \boldsymbol{z}_o, \boldsymbol{z}_u^{(r)}) + \log\{q(\boldsymbol{z}_u^{(r)} \mid \boldsymbol{z}_u^\star)\} - \log\{q(\boldsymbol{z}_u^\star \mid \boldsymbol{z}_u^{(r)})\};$$

   (iii) Simulate a random variable $X_{\boldsymbol{z}_u} \sim \text{Unif}(0,1)$;

   - If $X_{\boldsymbol{z}_u} \leq \exp(p_{\boldsymbol{z}_u})$, set $\boldsymbol{z}_u^{(r+1)} = \boldsymbol{z}_u^\star$;
   - otherwise, set $\boldsymbol{z}_u^{(r+1)} = \boldsymbol{z}_u^{(r)}$.

(3) Update the parameter α as follows:

   (i) Generate a candidate value $\alpha^\star$ from a proposal density $q(\alpha \mid \alpha^{(r)})$, and set $\psi^\star = (\alpha^\star, \rho^{(r)})$;

   (ii) Compute the log-acceptance probability

$$\begin{aligned} p_\alpha &= \ell(\psi^\star; \boldsymbol{z}_o, \boldsymbol{z}_u^{(r+1)}) - \ell(\psi^{(r)}; \boldsymbol{z}_o, \boldsymbol{z}_u^{(r+1)}) + \log\{\pi(\alpha^\star)\} - \log\{\pi(\alpha^{(r)})\} \\ &\quad + \log\{q(\alpha^{(r)} \mid \alpha^\star)\} - \log\{q(\alpha^\star \mid \alpha^{(r)})\}; \end{aligned}$$

(iii) Simulate a random variable $X_\alpha \sim \text{Unif}(0,1)$;

- If $X_\alpha \leq \exp(p_\alpha)$, set $\alpha^{(r+1)} = \alpha^\star$;
- otherwise, set $\alpha^{(r+1)} = \alpha^{(r)}$.

Set $\psi^{(r)} = (\alpha^{(r+1)}, \rho^{(r)})$.

(4) Update the parameter ρ as follows:

(i) Generate a candidate value $\rho^\star$ from a proposal density $q\big(\rho \mid \rho^{(r)}\big)$, and set $\psi^\star = (\alpha^{(r+1)}, \rho^\star)$;

(ii) Compute the log-acceptance probability

$$\begin{aligned} p_\rho &= \ell(\psi^\star; \boldsymbol{z}_o, \boldsymbol{z}_u^{(r+1)}) - \ell(\psi^{(r)}; \boldsymbol{z}_o, \boldsymbol{z}_u^{(r+1)}) + \log\{\pi\left(\rho^\star\right)\} - \log\{\pi\left(\rho^{(r)}\right)\} \\ &\quad + \log\{q\left(\rho^{(r)} \mid \rho^\star\right)\} - \log\{q\left(\rho^\star \mid \rho^{(r)}\right)\}; \end{aligned}$$

(iii) Simulate a random variable $X_\rho \sim \text{Unif}(0,1)$;

- If $X_\rho \leq \exp(p_\rho)$, set $\rho^{(r+1)} = \rho^\star$;
- otherwise, set $\rho^{(r+1)} = \rho^{(r)}$.

Set $\psi^{(r+1)} = (\alpha^{(r+1)}, \rho^{(r+1)})$.

The Bayesian approach is similar in spirit to the stochastic EM algorithm, except that prior distributions are given to the parameters. Compared to the frequentist approach, it allows more flexible modeling, and natural uncertainty assessment. However, in practice, we have found that the hyperparameters of the MCMC algorithm are difficult to tune to allow fast convergence of the Markov chain to its stationary distribution. Because there are many (correlated) hidden variables, which may also be highly correlated with the observed data, the mixing properties of the resulting random walks are poor, and blocking does not improve this much.

Inference using particle filters. Model (3.59) can also be formulated as a nonlinear non-Gaussian state-space model,

$$\begin{aligned} \textit{State equation:} \qquad H_t &= a(H_{t-1}, S_t), \\ \textit{Observation equation:} \qquad Z_t &= b(H_t, \varepsilon_t), \end{aligned} \tag{3.64}$$

in which $S_t \overset{\text{iid}}{\sim} \text{PS}(\alpha)$ and $\varepsilon_t \overset{\text{iid}}{\sim} \text{GEV}(1, \alpha, \alpha)$ are independent random innovations, H_t is the unobserved state at time t, Z_t is the observed noisy version of H_t and $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ are nonlinear transition functions defined, in our case, as

$$a(h,s) = \rho h + s, \qquad b(h,\varepsilon) = (1-\rho^\alpha) h^\alpha \varepsilon.$$

State-space models have been studied extensively, and the well-known Kalman filter permits an efficient treatment of the linear Gaussian case (see, e.g., Shumway & Stoffer, 2004, Chapter 6). Extensions beyond linearity and/or Gaussianity have been investigated by Kitagawa (1996), Durbin & Koopman (1997), Tanizaki & Mariano (1998), Kim *et al.* (1998), Einicke & White (1999), Tanizaki (2001), Durbin & Koopman (2002), Jungbacker & Koopman (2007) and many others. For example, the so-called *extended Kalman filter* (Einicke & White, 1999) handles nonlinear Gaussian state-space models using linear approximations of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. When the observation noise ε_t is non-Gaussian, its distribution may be approximated using Gaussian mixtures, and the Kalman filter may be applied to the approximate offset model; see, e.g., Kim *et al.* (1998), who use this idea to fit stochastic volatility models from a Bayesian perspective using efficient MCMC algorithms. Kitagawa (1996) has proposed a very simple particle filter and smoother, based on Monte Carlo simulations, which may be used to approximate the likelihood for a large variety of nonlinear non-Gaussian state-space models. His recursive algorithm can be adapted to model (3.64) as follows: let $g(x) = \alpha^{-1}x^{-1/\alpha-1}\exp\left(-x^{-1/\alpha}\right)$ denote the density of the noise $\varepsilon_t \sim \text{GEV}(1,\alpha,\alpha)$, and let $c(z,h) = z(1-\rho^\alpha)^{-1}h^{-\alpha}$ be such that $\epsilon_t = c(Z_t, H_t)$. Then,

(1) for $j = 1,\ldots,M$, simulate $S_{0;j} \sim \text{PS}(\alpha)$ and set $F_{0;j} = (1-\rho^\alpha)^{-1/\alpha}S_{0;j}$;

(2) for $i = 1,\ldots,n$, do:

   (i) for $j = 1,\ldots,M$, simulate $S_{i;j} \sim \text{PS}(\alpha)$ and set $P_{i;j} = a\left(F_{i-1;j}, S_{i;j}\right)$;

   (ii) for $j = 1,\ldots,M$, set $A_{i;j} = g\left\{c\left(Z_n, P_{i;j}\right)\right\} \times \left|\frac{\partial c(Z_n,P_{i;j})}{\partial Z_n}\right|$;

   (iii) for $j = 1,\ldots,M$, sample $F_{i;j}$ from $P_{i;1},\ldots,P_{i;M}$ with corresponding probabilities $A_{i;1}\left(\sum_{j=1}^M A_{i;j}\right)^{-1},\ldots,A_{i;M}\left(\sum_{j=1}^M A_{i;j}\right)^{-1}$;

(3) approximate the log likelihood by $\ell(\psi) \approx \sum_{i=1}^n \log\left(\sum_{j=1}^M A_{i;j}\right) - n\log(M)$.

A justification for such an algorithm can be found in Kitagawa (1996). The quality of the approximation to the likelihood is controlled by the size M of the Monte Carlo sample. The larger M, the better the approximation. Although this algorithm may be very intensive for large M, in principle it can easily be extended to other types of models for time series, which admit a state-space representation. Hence, for a broad variety of max-stable models with asymmetric logistic dependence and hidden Markov structure, inference based on the full (simulated) likelihood seems to be attainable.

In order to quantify the efficiency of pairwise likelihood estimation compared to full likelihood estimation in this context, we conducted a simple simulation study.

We simulated 300 independent segments of length $n = 500$ from model (3.64) with parameters $\alpha = 0.1, \ldots, 0.9$ and $\rho = 0.1, \ldots, 0.9$, and fitted the simulated time series using the maximum simulated full likelihood estimator $\widehat{\psi} = (\widehat{\alpha}, \widehat{\rho})$ stemming from Kitagawa (1996)'s algorithm with $M = 50000$, and the maximum pairwise likelihood estimator $\widehat{\psi}_{\mathcal{K}} = (\widehat{\alpha}_{\mathcal{K}}, \widehat{\rho}_{\mathcal{K}})$, maximizing a pairwise likelihood constructed from the contributions of pairs of observations (z_t, z_{t+h}), with time lag $h \in \mathcal{K}$. As in §3.3.2.1–3.3.2.2, we tested estimators corresponding to the sets of time lags

$$\mathcal{K}_a^K = \{1, \ldots, K\}, \qquad \mathcal{K}_b^K = \{b_k; k = 1, \ldots, K\}, \qquad \mathcal{K}_c^K = \{2^{k-1}; k = 1, \ldots, K\}, \tag{3.65}$$

where b_k is based on the Fibonacci sequence, and we took $K = 1, 3, 6, 9$. The 300 replicates were then used to compute the empirical variance-covariance matrices $\widehat{V}$ and $\widehat{V}_{\mathcal{K}}$, the empirical marginal efficiencies $\mathrm{RE}_\alpha = \widehat{\mathrm{var}}(\widehat{\alpha})/\widehat{\mathrm{var}}(\widehat{\alpha}_{\mathcal{K}})$ and $\mathrm{RE}_\rho = \widehat{\mathrm{var}}(\widehat{\rho})/\widehat{\mathrm{var}}(\widehat{\rho}_{\mathcal{K}})$, and the empirical global efficiency $\mathrm{RE}_\psi = \{\det(\widehat{V})/\det(\widehat{V}_{\mathcal{K}})\}^{1/2}$. Because the value chosen for M is relatively large compared to the sample size n, these efficiencies depend little on the Monte Carlo variability associated to $\widehat{\psi}$. The main results are representatively reported in Table 3.3.

The results reveal that the efficiency of pairwise likelihood estimators can be quite low when α is close to 0 or 1, or when ρ approaches unity. This contrasts with the results of Table 3.2 for the symmetric logistic model. However, in typical situations, the efficiencies are reasonably high, except maybe for estimators based on $\mathcal{K} = \{1\}$, suggesting that pairwise likelihood estimators based on additional distant pairs may be recommended for this model. For example, when $\alpha \approx 0.5$ and $\rho \approx 0.5$, the global efficiency reaches about 50% for the estimators based on $\mathcal{K}_a^6$, $\mathcal{K}_b^6$ and $\mathcal{K}_c^6$, but only about 25% for that based on $\mathcal{K} = \{1\}$. Overall, the former seem to behave appreciably better than the latter, but there is not much difference between $\mathcal{K}_a^6$ and $\mathcal{K}_b^6$ or $\mathcal{K}_c^K$, which contrasts with our previous results of Sections 3.3.2.1–3.3.2.2 for Gaussian models.

3.3.3.2 Simulation study for the Schlather model with random set

In Section 3.2.2, we introduced a censored maximum pairwise likelihood estimator based on threshold exceedances for spatio-temporal extremes. In this section, we assess its statistical efficiency for the Schlather model with random set (2.31), under different fitting procedures; this model is fitted to rainfall data in our application in Chapter 5. The weighting scheme that we consider here consists in including a pair $(y_{s_1;t_1}, y_{s_2;t_2})$ in the pairwise likelihood if $|t_2 - t_1| \in \mathcal{K}$, for some predetermined set of time lags $\mathcal{K}$. As we have seen in §3.3.2 for Gaussian models and in §3.3.3.1 for the max-stable asymmetric logistic model, the loss in efficiency is closely related to the

Table 3.3: Efficiency (%) of maximum pairwise likelihood estimators $\widehat{\psi}_{\mathscr{K}}$ relative to maximum simulated full likelihood estimators $\widehat{\psi}$ with $M = 50000$ and $n = 500$, based on 300 simulations of the time series model (3.64) with different values for the parameters α and ρ. The results are shown for typical parameter values and several choices of time lags $\mathscr{K}$. The numbers are respectively $\mathrm{RE}_\alpha/\mathrm{RE}_\rho/\mathrm{RE}_\psi$.

Varying α, fixed $\rho = 0.5$

$\alpha \setminus \mathscr{K}$	$\{1\}$	$\mathscr{K}_a^6$	$\mathscr{K}_b^6$	$\mathscr{K}_c^6$
0.1	47/15/11	53/37/28	49/37/33	47/38/35
0.2	63/18/22	66/46/45	65/45/48	64/45/48
0.3	69/14/25	68/43/50	67/42/51	72/44/52
0.4	63/12/25	68/42/48	67/40/48	71/42/51
0.5	57/9/23	67/44/52	66/41/51	67/40/51
0.6	53/9/24	79/56/61	77/50/59	73/46/57
0.7	48/7/22	75/56/62	77/50/61	71/48/58
0.8	42/6/21	76/52/62	75/41/56	69/38/54
0.9	20/9/17	42/43/48	33/36/40	31/32/39

Fixed $\alpha = 0.5$, varying ρ

$\rho \setminus \mathscr{K}$	$\{1\}$	$\mathscr{K}_a^6$	$\mathscr{K}_b^6$	$\mathscr{K}_c^6$
0.1	67/46/57	73/74/76	74/72/76	72/65/72
0.2	63/27/44	63/51/62	65/52/64	69/54/66
0.3	62/23/40	68/61/66	71/60/67	75/59/67
0.4	63/15/32	74/55/63	75/51/62	75/51/62
0.5	57/9/23	67/44/52	66/41/51	67/40/51
0.6	53/7/20	65/43/51	62/40/49	67/40/49
0.7	52/5/16	54/30/39	53/29/40	57/28/39
0.8	48/3/12	51/23/32	50/28/35	49/28/35
0.9	30/1/6	31/8/16	30/12/19	29/14/20

pairs that are included in the pairwise likelihood, i.e., to the choice of $\mathscr{K}$. Adding pairs might simultaneously increase the variability $K_C(\psi)$ of the composite score and the amount of composite information $J_C(\psi)$, so it is unclear how the selection of pairs acts on the asymptotic sandwich variance $J_C(\psi)^{-1}K_C(\psi)J_C(\psi)^{-1}$; recall §3.1.3. The amount of information contained in a single pair might be insufficient to counteract the increase of variability due to including it, so the choice of the optimal subset of pairs is not obvious. Some people have suggested the elimination of non-neighboring pairs (see §3.3.1 and Varin & Vidoni, 2005; Varin & Czado, 2010; Varin *et al.*, 2011), and we have found for Gaussian ARMA models that it might be better in some cases to consider a mixture of strongly correlated and weakly correlated pairs; recall §3.3.2.

Using the statistical software R (R Core Team, 2012), we simulated the Schlather

Table 3.4: Mean squared errors ($\times 1000$) for estimation of $\log\lambda$, the logarithm of the correlation range parameter, based on 1000 replications of the Schlather model, for different sets of pairs included in the pairwise likelihood, and μ known. There are three estimation procedures: margins known (MK); margins unknown, two-step approach (MU-2); margins unknown, one-step approach (MU-1).

	Number of time lags K								
	1	3		6			9		
Set $\mathcal{K}$	$\{1\}$	$\mathcal{K}_{a/b}^K$	$\mathcal{K}_c^K$	$\mathcal{K}_a^K$	$\mathcal{K}_b^K$	$\mathcal{K}_c^K$	$\mathcal{K}_a^K$	$\mathcal{K}_b^K$	$\mathcal{K}_c^K$
MK	19	21	21	26	24	22	29	24	23
MU-2	42	45	46	54	50	48	59	50	49
MU-1	37	41	42	48	45	44	52	46	44

model (2.31) on the time axis, taking $\mathcal{X} = [0, 10000]$, with random sets of the form $\mathcal{A} = [0, D]$, where $D = 24\delta$ and $\delta \sim \text{Beta}(10, 240/\mu - 10)$. The parameter μ corresponds to the mean length of the random set, which lies in the range $(0, 24)$, and we set $\mu = \text{E}(D) = 40/3 \simeq 13.3$. We chose an exponential correlation for the underlying Gaussian random field $\varepsilon(x)$, with range parameter $\lambda = 4$; the effective range is 12. We then transformed the simulated processes to the Student t_5 scale, so that the exceedances above some high threshold u are approximately $\text{GPD}(\sigma, \xi)$ with shape parameter $\xi = 0.2$ (Beirlant *et al.*, 2004, p. 59). The parameters were chosen to mimic rainfall data. The threshold u was set to the empirical 95% quantile, so that we have 500 exceedances contributing to the pairwise likelihood; in our application in Chapter 5, about 3000 exceedances were available at each station. A realization from this model is shown in the top panel of Figure 3.12. To assess the influence of the marginal estimation on the overall fit, we consider three estimation procedures: (i) estimation of the dependence parameters, treating the margins as known; (ii) a two-step approach, first estimating the marginal parameters by fitting the approximate GPD model, and then using the data thereby transformed to the unit Fréchet scale to estimate the dependence parameters; and (iii) a one-step approach, estimating marginal and dependence parameters simultaneously.

We first fixed the random set parameter μ to its true value, and estimated the logarithm of the range parameter, $\log\lambda$, with the threshold-based pairwise likelihood estimator, using the empirical 95% quantile threshold. Similarly to §3.3.2.1–3.3.2.2 and §3.3.3.1, we compared estimators corresponding to the three sets of time lags defined in (3.65), namely $\mathcal{K}_a^K$, for which all time lags are used up to some maximum time lag K, $\mathcal{K}_b^K$, based on the Fibonacci sequence, and $\mathcal{K}_c^K$, where the lags increase geometrically. We considered $K = 1, 3, 6, 9$ as before. Table 3.4 reports the mean squared errors (MSE) of these estimates based on 1000 realizations of the Schlather model.

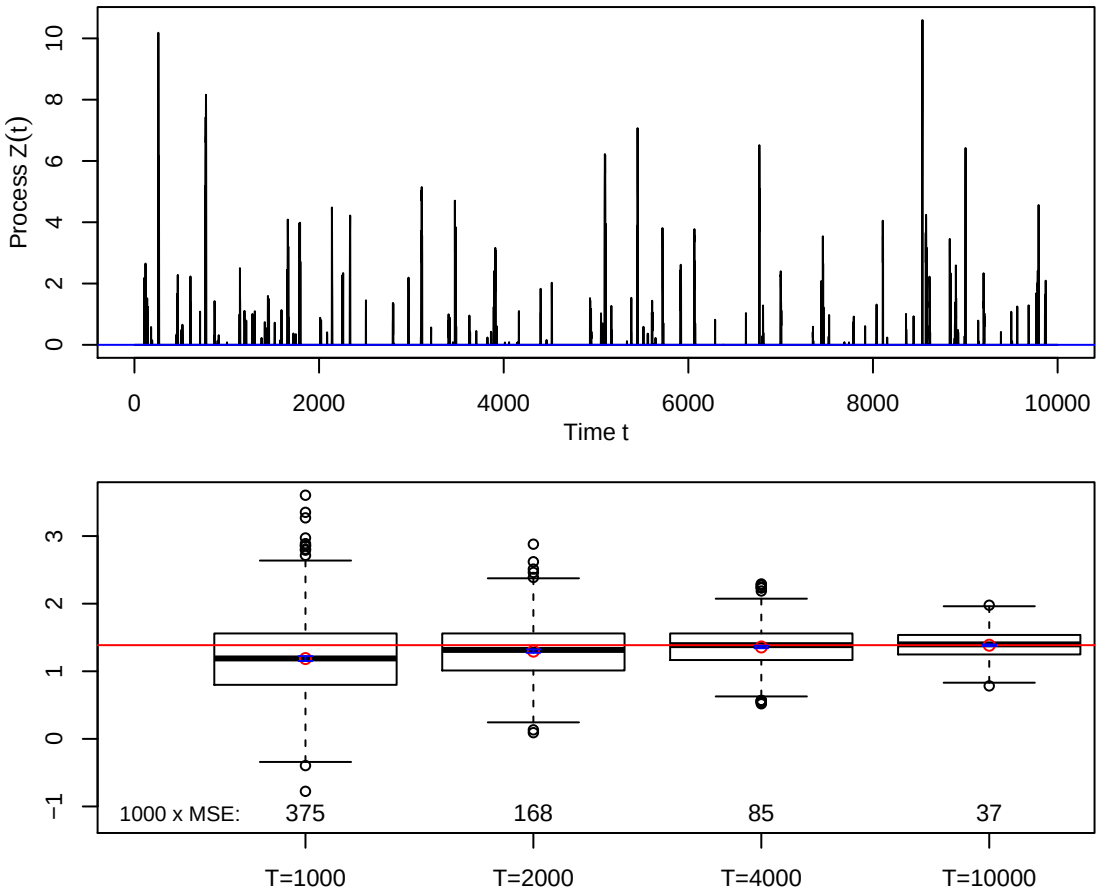

Figure 3.12: Simulated extreme rainfall process at a particular location, $Z(t)$, $t = 1,\ldots,T$. *Top*: Exceedances over the 95th percentile from a simulation of the Schlather model in time with beta distributed random sets. The marginal distribution of the exceedances is approximately GPD$(1.6, 0.2)$ and the correlation of the underlying Gaussian random process is exponential with range parameter $\lambda = 4$. *Bottom*: Boxplots (and mean squared errors) of the estimates of $\log\lambda$ (based on 1000 replications) using the censored pairwise likelihood estimator with pairs at lag 1 only, for an increasing number of observations T. The one-step estimator was used, fixing the random set parameters to their true values. The true value is shown by the horizontal line, and the average estimate is shown by the circle close to it.

With the random set parameter μ known, the MSE is minimized for $\mathcal{K} = \{1\}$, whatever the estimation procedure, corroborating the findings of §3.3.2 and Davis & Yau (2011) for AR(1) or MA(1) processes. Moreover, the MSE is systematically lower when $\mathcal{K}_b^6$ is used instead of $\mathcal{K}_a^6$, or when $\mathcal{K}_b^9$ is used instead of $\mathcal{K}_a^9$, even though the observations separated by more than 24 time units were independent. The same is true for $\mathcal{K}_c^K$. Thus, as for Gaussian models, the inclusion of some distant, less dependent, pairs

can improve inference significantly for fixed K, and marginal estimation does not influence the conclusions about the pairs that should be included in the pairwise likelihood.

The bottom panel of Figure 3.12 shows how the bias and variance of the dependence parameter estimator decrease as the number of observations T increases, confirming the theoretical results of Section 3.1.3 and Chapter 5.

We then estimated the correlation range parameter $\lambda > 0$ and the mean duration $\mu \in (0,24)$ of the random set simultaneously, using the censored maximum composite likelihood estimator and the threshold used above. The estimation of μ is more difficult, especially when only pairs at lag 1 are included, and in some cases its estimate reached the upper bound used in the optimization. Table 3.5 shows that with $\mathcal{K} = \{1\}$ this happens on 21%, 33% and 30% of occasions for known margins, the two-step estimator and the one-step estimator, respectively. This could be anticipated since the pairs at lag 1 are uninformative for the estimation of μ. When further lags are added, the difference between use of the set $\mathcal{K}_a^K$ and the other sets becomes striking, especially for $K = 6$. The upper bound for μ is not attained for $\mathcal{K}_b^6$ and $\mathcal{K}_c^6$, and the MSEs are two or three times lower than those for $\mathcal{K}_a^6$; λ is also better estimated. The estimators including distant pairs in the composite likelihood outperform those that do not or that use only the most dependent pairs. The same phenomenon is observed when $K = 9$, but the difference is less striking than for $K = 6$. In fact, the pairs at lags less than 6 are probably ineffective for estimation of the duration of sets that in this case last on average 13.3 time units, and so $\mathcal{K}_b^K$ or $\mathcal{K}_c^K$ are better choices than $\mathcal{K}_a^K$. To sum up, if K is fixed and not too large compared to the "true" independence range, then both estimators that include pairs at higher lags, $\mathcal{K}_b^K$ or $\mathcal{K}_c^K$, behave appreciably better than that based on $\mathcal{K}_a^K$, which does not.

As might be expected, the one-step estimation procedure outperforms the two-step procedure overall, though by an amount that depends on the choice of pairs; the differences are rather small for $\mathcal{K}_b^K$ and $\mathcal{K}_c^K$. The one-step approach performs relatively better for smaller samples or higher thresholds, but the procedures are essentially equivalent for samples of 3000 exceedances as used in our application in Chapter 5.

3.4 Summary

The contributions of this chapter, in which maximum composite (especially *pairwise*) likelihood estimators are studied in depth, consists of three main novelties.

First, in §3.2.2, we show how to fit extremal models using censored threshold-based

Table 3.5: Mean squared errors for the joint estimation of the mean duration μ of the random set (MSE_μ) and the logarithm of the range parameter $\log\lambda$ (MSE_λ), when different sets of pairs are included in the pairwise likelihood. The percentages of time when $\widehat{\mu}$ reaches its upper bound is also reported. We considered three estimation procedures: margins known (MK); margins unknown, two-step approach (MU-2); and margins unknown, one-step approach (MU-1). This simulation is based on 1000 replications of the Schlather model.

		Number of time lags K								
		1	3		6			9		
	Set $\mathcal{K}$	$\{1\}$	$\mathcal{K}_{a/b}^K$	$\mathcal{K}_c^K$	$\mathcal{K}_a^K$	$\mathcal{K}_b^K$	$\mathcal{K}_c^K$	$\mathcal{K}_a^K$	$\mathcal{K}_b^K$	$\mathcal{K}_c^K$
MK	$10^3 \times \text{MSE}_\lambda$	28	28	24	24	23	22	24	23	22
	MSE_μ	22.5	16.0	10.9	6.9	2.1	2.8	3.3	2.2	2.9
	Bound (%)	21	10	6	1	0	0	0	0	0
MU-2	$10^3 \times \text{MSE}_\lambda$	67	70	66	62	48	47	52	49	48
	MSE_μ	27.3	20.9	17.5	9.4	2.1	2.7	3.5	2.3	2.7
	Bound (%)	33	22	10	2	0	0	0	0	0
MU-1	$10^3 \times \text{MSE}_\lambda$	58	60	56	53	45	45	48	47	46
	MSE_μ	24.7	18.4	15.2	8.5	2.1	2.6	3.3	2.3	2.7
	Bound (%)	30	19	8	1	0	0	0	0	0

pairwise likelihoods, and hence propose a novel efficient approach to perform inference for max-stable and related complicated models.

Second, in §3.3, the efficiency of pairwise likelihoods is assessed under various scenarios. In particular, theoretical relative efficiencies for pairwise likelihood estimators have been derived in the case of Gaussian AR(1) and MA(1) models, and extensive simulations have been performed for more complicated ARMA(p,q) models. We have shown that for autoregressive-type processes, the selection of pairs to be included in the pairwise likelihood is crucial (though not obvious), whereas for moving average-type processes, it does not matter, so far as the parameters remain identifiable, and that the loss in efficiency may be very substantial in this case. We have found that in most cases, a good rule of thumb, if a fixed number of pairs have to be included, is to consider a mixture of strongly dependent and weakly dependent pairs. We have also investigated optimal weighting for Gaussian processes, and have come up with similar conclusions. In the max-stable framework, we have assessed, using simulations, the efficiency of pairwise likelihood estimators for the Schlather model with random set, which we fit in our application in Chapter 5.

Third, in §3.3.3.1, we have shown how to perform inference based on the full likelihood for the max-stable asymmetric logistic model, and have assessed the loss in efficiency

of pairwise likelihood estimators in this context. We have also proposed a novel max-stable time series model, and have shown how to fit it using a stochastic version of the EM algorithm, Bayesian methods or simulation-based particle filters.

In summary, this chapter provides a broad, though not exhaustive, overview of the efficiency of maximum pairwise likelihood estimators, and sheds some light on the selection of pairs. In Chapter 4, we focus on the max-stable Brown–Resnick process, and compute the efficiency gains of higher-dimensional marginal likelihood estimators.

4 Composite likelihood estimation for the Brown–Resnick process

Max-stable processes are useful for the statistical modeling of spatial extreme events; recall Chapter 2. No finite parametrization of such processes exists, but the spectral representation (2.20) aids in constructing models. In a 1990 University of Surrey technical report, R. L. Smith proposed a max-stable model based on deterministic storm profiles, which has become popular because it is simple, readily interpreted and easily simulated; but unfortunately it is too inflexible for realistic situations, recall §2.3.2.1. Another popular model, the Brown–Resnick process (see §2.3.2.3), is based on intrinsically stationary log-Gaussian processes, can handle a wide range of dependence structures, and often provides a better fit to data; see, for example, Davison *et al.* (2012) or Jeon (2012). In particular, the isotropic Brown–Resnick model constructed from fractional Brownian motions has received much attention, owing to its nice theoretical properties (Kabluchko *et al.*, 2009). The Smith model is obtained by taking a Brown–Resnick process with variogram $2\gamma(h) = h^{\mathrm{T}}\Sigma^{-1}h$ for some covariance matrix Σ, corresponding after an affine transformation to taking the smoothness parameter $\alpha = 2$, whereas Davison *et al.* (2012) found that $1/2 < \alpha < 1$ for the rainfall data they examined.

Likelihood inference for max-stable models is difficult, since only the bivariate density functions are known in most cases, and pairwise marginal likelihood is typically used (recall §3.2), even though the efficiency of the resulting estimators is usually inferior to that of classical maximum likelihood estimators, recall §3.3. This raises the question whether some other approach to inference would be preferable. Genton *et al.* (2011) derived the general form of the likelihood function for the Smith model, indexed by $\mathbb{R}^d$ in dimension $D \leq d+1$, and showed that large efficiency gains can arise when fitting it using triplewise, rather than pairwise, likelihood. In this chapter, we extend their investigation to the Brown–Resnick process and show that for rougher models, which are more realistic than those considered by Genton *et al.* (2011), the efficiency

gains are much less striking. Thus, pairwise likelihood inference provides a good compromise between statistical and computational efficiency in many applications.

In §4.4, we also explore higher-order composite likelihoods, and investigate their loss in efficiency compared to the maximum full likelihood estimator.

Most of the work presented below has been published in Huser & Davison (2013a).

4.1 Brown–Resnick process constructed from fractional Brownian motions

4.1.1 Definition and properties

The Brown–Resnick process (Brown & Resnick, 1977; Kabluchko *et al.*, 2009) is a stationary max-stable process that may be represented as $Z(x) = \sup_{i \geq 1} W_i(x)/P_i$, $x \in \mathscr{X} \subset \mathbb{R}^d$, where $0 < P_1 < P_2 < \cdots$ are the points of a unit rate Poisson process on $\mathbb{R}_+$ and the $W_i(x)$ are independent replicates of the random process $W(x) = \exp\{\varepsilon(x) - \gamma(x)\}$; recall §2.20 and §2.3.2.3. Here $\varepsilon(x)$ is an intrinsically stationary Gaussian random field with semi-variogram $\gamma(h)$, where h is the spatial lag, and $\varepsilon(0) = 0$ almost surely. In particular, when the variogram equals $2\gamma(h) = (\|h\|/\lambda)^\alpha$, for some range parameter $\lambda > 0$ and smoothness parameter $\alpha \in (0,2]$, the random field $\varepsilon(x)$ is a so-called fractional Brownian motion, and the resulting max-stable field $Z(x)$ is isotropic. Kabluchko *et al.* (2009) provided strong underpinning for this process by showing that under mild conditions, the Brown–Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$ is essentially the only isotropic limit of properly rescaled maxima of Gaussian processes (recall Theorem 55 and the paragraph thereafter). The special case $\alpha = 2$ corresponds to the isotropic Smith model (recall §2.3.2.1) since the variogram may be expressed as $2\gamma(h) = h^{\mathrm{T}}\Sigma^{-1}h$, where $\Sigma = \mathrm{diag}(\lambda^2, \lambda^2)$ is a covariance matrix. The parameter vector is denoted by $\psi = (\lambda, \alpha)$.

In Section §2.3.1, we have seen that the full distribution of $Z(x)$ at the set of sites $\mathscr{D} \subset \mathscr{X}$ is

$$\Pr\{Z(x) \leq z(x), x \in \mathscr{D}\} = \exp\left(-\mathrm{E}\left[\sup_{x \in \mathscr{D}}\left\{\frac{W(x)}{z(x)}\right\}\right]\right),$$

where the exponent measure function $V_{\mathscr{D}}\{z(x)\} = \mathrm{E}\left[\sup_{x \in \mathscr{D}}\{W(x)/z(x)\}\right]$ must satisfy certain constraints. It follows that the univariate margins of $Z(x)$ equal $\exp(-1/z)$, for $z > 0$, and for $\mathscr{D} = \{x_1, x_2\}$ the exponent measure of the Brown–Resnick process is

$$V_{\mathscr{D}}(z_1, z_2) = \frac{1}{z_1}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_1}{z_2}\right)\right\} + \frac{1}{z_2}\Phi\left\{\frac{a}{2} - \frac{1}{a}\log\left(\frac{z_2}{z_1}\right)\right\}, \tag{4.1}$$

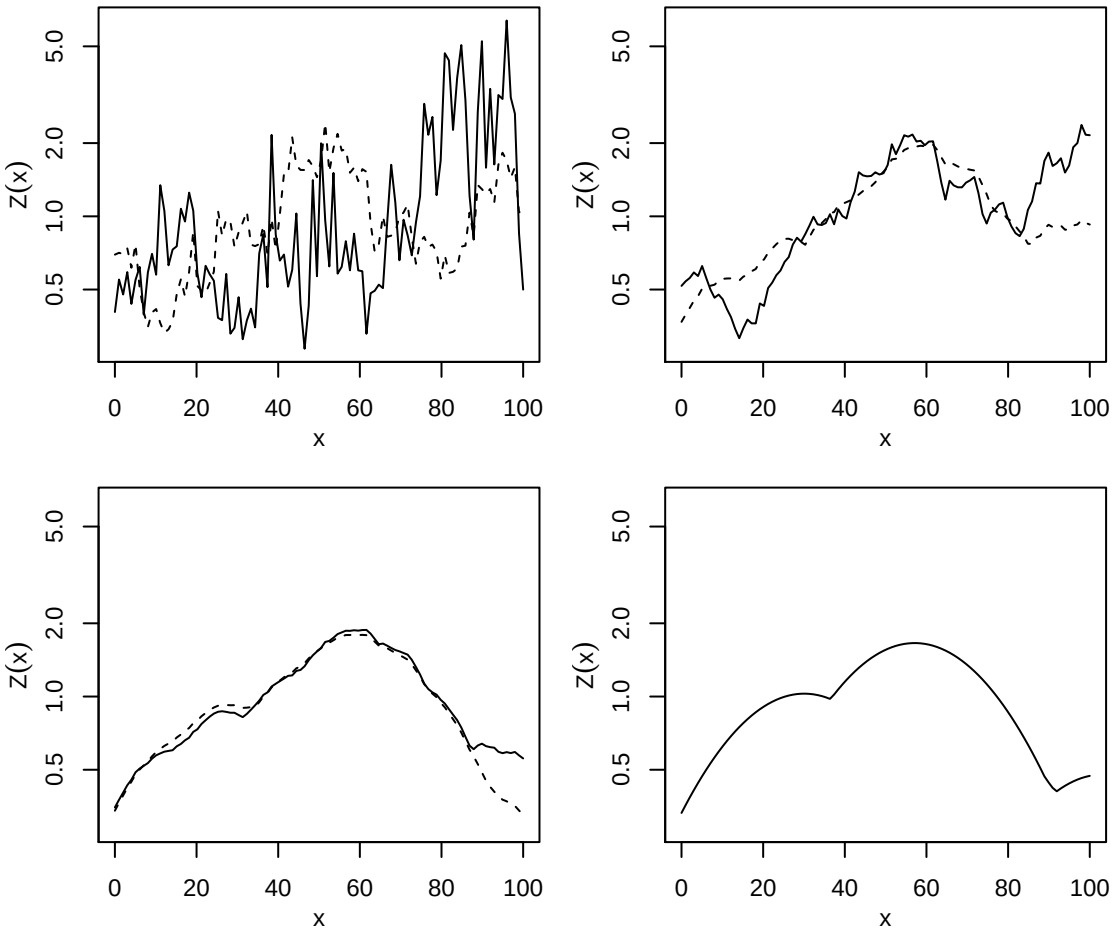

Figure 4.1: Seven simulated Brown–Resnick processes in one dimension, i.e., $d = 1$, with variogram $2\gamma(h) = (\|h\|/28)^{\alpha}$, and different smoothness parameters. *Top left*: $\alpha = 0.5$ (solid), 1 (dashed); *top right*: $\alpha = 1.5$ (solid), 1.9 (dashed); *bottom left*: $\alpha = 1.95$ (solid), 1.98 (dashed); *bottom right*: $\alpha = 2$, which corresponds to the isotropic Smith model. The same random seed was used in all seven cases.

where $z_i = z(x_i)$, $i = 1,2$, $a = \{2\gamma(x_2 - x_1)\}^{1/2}$ and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. In this case expression (4.1) boils down to the Hüsler & Reiss (1989) model for bivariate extremes. The bivariate marginal density functions $g(z_1, z_2; \psi)$ are easily expressed using derivatives of (4.1); see Appendix B.1.1.

Figure 4.1 shows how the variogram influences the smoothness of the max-stable process. In particular, when α equals 2, the isotropic Smith model is recovered, and the storm shapes are deterministic, taking the form of Gaussian densities.

4.1.2 Inference based on pairwise likelihood

Since full likelihood is impractical for max-stable processes, inference is usually performed using pairwise likelihood. Suppose that n independent replicates of a Brown–Resnick process with variogram $2\gamma(h)$ depending on the parameter vector ψ are observed at S sites in $\mathscr{X} \subset \mathbb{R}^d$, and let $z_{s,i}$ denote the value of the ith process at the sth site. Following the considerations in Section 3.2.1, one can make inference based on the log-pairwise likelihood

$$\ell_2(\psi) = \sum_{i=1}^{n} \sum_{s_1=1}^{S-1} \sum_{s_2=s_1+1}^{S} \log g(z_{s_1,i}, z_{s_2,i}; \psi),$$

and the corresponding maximum pairwise likelihood estimator $\widehat{\psi}_2$ is consistent and asymptotically Gaussian as n increases, but suffers from a loss in efficiency compared to the maximum full likelihood estimator; recall §3.1.3. In the following sections, we investigate the potential efficiency gains when higher-order marginal likelihoods are used for this process.

4.2 Derivation of the likelihood

4.2.1 Exponent measure

As we have seen, the joint distribution is determined by the exponent measure. For the sake of clarity, we start with the trivariate case, leaving the derivation of the joint distribution in $D > 3$ dimensions for the end of this section.

4.2.1.1 Case $D = 3$

Let $\mathscr{D} = \{x_1, x_2, x_3\}$ be any set of three points in the state space $\mathscr{X}$. For compactness of notation we write $z_1 = z(x_1)$, $W_1 = W(x_1)$, $\varepsilon_1 = \varepsilon(x_1)$, $\gamma_1 = \gamma(x_1)$, $\gamma_{1;2} = \gamma(x_1 - x_2)$, etc. By definition of the Brown–Resnick process, one has $\varepsilon(0) = 0$ almost surely, so it is easy to see that $c_{i;i} = \operatorname{var}(\varepsilon_i) = 2\gamma_i$ and that $c_{i;j} = \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \gamma_i + \gamma_j - \gamma_{i;j}$, for $i, j = 1, 2, 3$. Furthermore, the following equivalences hold:

$$\begin{aligned} W_1/z_1 > W_2/z_2 &\iff \log W_1 - \log z_1 > \log W_2 - \log z_2 \\ &\iff \varepsilon_1 - \gamma_1 - \log z_1 > \varepsilon_2 - \gamma_2 - \log z_2 \\ &\iff \varepsilon_1 > \varepsilon_2 + b_{1;2}, \end{aligned}$$

where $b_{1;2} = \gamma_1 - \gamma_2 + \log(z_1/z_2)$. Similarly, $W_1/z_1 > W_3/z_3$ if and only if $\varepsilon_1 > \varepsilon_3 + b_{1;3}$, where $b_{1;3} = \gamma_1 - \gamma_3 + \log(z_1/z_3)$. Thanks to (2.22), we may write

$$V_{\mathscr{D}}(z_1, z_2, z_3) = \mathrm{E}\left\{\max\left(\frac{W_1}{z_1}, \frac{W_2}{z_2}, \frac{W_3}{z_3}\right)\right\} = \frac{I_1}{z_1} + \frac{I_2}{z_2} + \frac{I_3}{z_3},$$

say, where

$$I_1 = \mathrm{E}\left\{W_1 I\left(\frac{W_1}{z_1} > \frac{W_2}{z_2}, \frac{W_1}{z_1} > \frac{W_3}{z_3}\right)\right\}$$

and so forth. Now provided $x_1 \neq 0$ and with $w_i = \exp(\varepsilon_i - \gamma_i)$ and using ϕ to denote Gaussian densities, possibly multivariate, we have

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(\varepsilon_1 - \gamma_1) I\left(\frac{w_1}{z_1} > \frac{w_2}{z_2}, \frac{w_1}{z_1} > \frac{w_3}{z_3}\right) \phi(\varepsilon_1, \varepsilon_2, \varepsilon_3) \mathrm{d}\varepsilon_1 \mathrm{d}\varepsilon_2 \mathrm{d}\varepsilon_3 \\ &= \int_{-\infty}^{\infty} \exp(\varepsilon_1 - \gamma_1)\phi(\varepsilon_1) \int_{-\infty}^{\varepsilon_1 - b_{1;2}} \int_{-\infty}^{\varepsilon_1 - b_{1;3}} \phi(\varepsilon_2, \varepsilon_3 \mid \varepsilon_1) \mathrm{d}\varepsilon_3 \mathrm{d}\varepsilon_2 \mathrm{d}\varepsilon_1 \qquad (4.2) \\ &= \int_{-\infty}^{\infty} \frac{1}{(4\pi\gamma_1)^{1/2}} \exp\{-(\varepsilon_1 - 2\gamma_1)^2/(4\gamma_1)\} K(\varepsilon_1) \mathrm{d}\varepsilon_1, \end{aligned}$$

say, where $K(\varepsilon_1)$ denotes the inner double integral in (4.2), and thus

$$I_1 = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp(-\xi^2/2) K\left\{(2\gamma_1)^{1/2}\xi + 2\gamma_1\right\} \mathrm{d}\xi = \mathrm{E}_\xi\left[K\left\{(2\gamma_1)^{1/2}\xi + 2\gamma_1\right\}\right],$$

where $\xi \sim \mathscr{N}(0,1)$. As the joint distribution of $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ is trivariate normal with zero mean and covariance matrix $C = (c_{i;j})$, the properties of the multivariate normal distribution imply that the joint density of $\varepsilon_2, \varepsilon_3$ conditional on ε_1 is $\mathscr{N}_2(\mu_{2,3|1}, C_{2,3|1})$, where

$$\mu_{2,3|1} = \begin{pmatrix} c_{1;2}\varepsilon_1/c_{1;1} \\ c_{1;3}\varepsilon_1/c_{1;1} \end{pmatrix}, \quad C_{2,3|1} = \begin{pmatrix} c_{2;2} - c_{1;2}^2/c_{1;1} & c_{2;3} - c_{1;2}c_{1;3}/c_{1;1} \\ c_{2;3} - c_{1;2}c_{1;3}/c_{1;1} & c_{3;3} - c_{1;3}^2/c_{1;1} \end{pmatrix}.$$

Therefore, conditional on ξ, we have

$$\begin{aligned} K\left\{(2\gamma_1)^{1/2}\xi + 2\gamma_1\right\} &= \int_{-\infty}^{(2\gamma_1)^{1/2}\xi + 2\gamma_1 - b_{1;2}} \int_{-\infty}^{(2\gamma_1)^{1/2}\xi + 2\gamma_1 - b_{1;3}} \phi\left\{\varepsilon_2, \varepsilon_3 \mid \varepsilon_1 = (2\gamma_1)^{1/2}\xi + 2\gamma_1\right\} \mathrm{d}\varepsilon_3 \mathrm{d}\varepsilon_2 \\ &= \Pr\big[Z_1 \leq (2\gamma_1)^{1/2}\xi + 2\gamma_1 - b_{1;2} - c_{1;2}\{(2\gamma_1)^{1/2}\xi + 2\gamma_1\}/c_{1;1}, \\ &\qquad Z_2 \leq (2\gamma_1)^{1/2}\xi + 2\gamma_1 - b_{1;3} - c_{1;3}\{(2\gamma_1)^{1/2}\xi + 2\gamma_1\}/c_{1;1} \,\big|\, \xi\big], \end{aligned}$$

where Z_1 and Z_2 form a bivariate normal random variable with zero mean, and covariance matrix $C_{2,3|1}$. Integrating over ξ, we get

$$\mathrm{E}_\xi\left[K\left\{(2\gamma_1)^{1/2}\xi + 2\gamma_1\right\}\right] = \Pr\big\{Z_1 + \xi(-\gamma_1 + \gamma_2 - \gamma_{1;2})/(2\gamma_1)^{1/2} \leq -b_{1;2} + \gamma_1 - \gamma_2 + \gamma_{1;2},$$

$$
\begin{aligned}
& Z_2 + \xi(-\gamma_1 + \gamma_3 - \gamma_{1;3})/(2\gamma_1)^{1/2} \leq -b_{1;3} + \gamma_1 - \gamma_3 + \gamma_{1;3}\big\} \\
&= \Pr\big(Y_1 \leq -b_{1;2} - \gamma_1 - \gamma_2 + \gamma_{1;2}, Y_2 \leq -b_{1;3} - \gamma_1 - \gamma_3 + \gamma_{1;3}\big) \\
&= \Pr\big\{Y_1 \leq \gamma_{1;2} - \log(z_1/z_2), Y_2 \leq \gamma_{1;3} - \log(z_1/z_3)\big\}, \qquad (4.3)
\end{aligned}
$$

where (Y_1, Y_2) is a bivariate normal vector with zero mean and covariance matrix

$$
\Omega_1 = \begin{pmatrix} 2\gamma_{1;2} & \gamma_{1;2} + \gamma_{1;3} - \gamma_{2;3} \\ \gamma_{1;2} + \gamma_{1;3} - \gamma_{2;3} & 2\gamma_{1;3} \end{pmatrix}.
$$

The right-hand side of equation (4.3) yields

$$
I_1 = \Phi_2\{\eta(z_1, z_2), \eta(z_1, z_3); R_1\}, \qquad (4.4)
$$

where $\Phi_2(\cdot, \cdot; R)$ denotes the bivariate normal cumulative distribution function with zero mean, correlation R and unit variance, $R_1 = (\gamma_{1;2} + \gamma_{1;3} - \gamma_{2;3})/\{2(\gamma_{1;2}\gamma_{1;3})^{1/2}\}$ and $\eta(z_i, z_j) = (2\gamma_{i;j})^{1/2}/2 - \log(z_i/z_j)/(2\gamma_{i;j})^{1/2}$, $i, j = 1, 2, 3$. The case $x_1 = 0$ can be treated separately and turns out to give the same result. By interchanging the labels, I_2 and I_3 are derived similarly. Hence, the triplewise exponent measure may be expressed as

$$
\begin{aligned}
V_{\mathscr{D}}(z_1, z_2, z_3) \quad = \quad & \frac{1}{z_1}\Phi_2\left\{\eta(z_1, z_2), \eta(z_1, z_3); R_1\right\} + \frac{1}{z_2}\Phi_2\left\{\eta(z_2, z_1), \eta(z_2, z_3); R_2\right\} \\
& + \frac{1}{z_3}\Phi_2\left\{\eta(z_3, z_1), \eta(z_3, z_2); R_3\right\}, \qquad (4.5)
\end{aligned}
$$

where the function $\eta(\cdot, \cdot)$ has been defined earlier, and

$$
R_1 = \frac{\gamma_{1;2} + \gamma_{1;3} - \gamma_{2;3}}{2(\gamma_{1;2}\gamma_{1;3})^{1/2}}, \quad R_2 = \frac{\gamma_{1;2} + \gamma_{2;3} - \gamma_{1;3}}{2(\gamma_{1;2}\gamma_{2;3})^{1/2}}, \quad R_3 = \frac{\gamma_{1;3} + \gamma_{2;3} - \gamma_{1;2}}{2(\gamma_{1;3}\gamma_{2;3})^{1/2}}.
$$

The function $\Phi_2(\cdot, \cdot; R)$ can be rapidly computed using the efficient algorithms based on quasi-random sampling methods developed by Genz (1992, 1993); Genz & Bretz (2002).

Expressions (4.4) and (4.5) and its counterparts hold if $|R_k| \neq 1$, $k = 1, 2, 3$, which is always true when the variogram is $2\gamma(h) = (\|h\|/\lambda)^\alpha$ and $\alpha < 2$. However, if $\alpha = 2$ and the sites x_1, x_2 and x_3 form a degenerate simplex in $\mathbb{R}^d$, then $R_k = \pm 1$, $k = 1, 2, 3$. If $d = 1$, the simplex is always degenerate. In dimension $d \geq 2$, certain configurations of points may also be problematic, for example if the sites x_1, x_2, x_3 lie on a linear subset of $\mathbb{R}^2$. This will lead to problems when the sites of $\mathscr{D}$ form a grid.

4.2.1.2 Case $D > 3$

For $D > 3$, the exponent measure may be written as

$$V(z_1, \ldots, z_D) = \frac{I_1}{z_1} + \cdots + \frac{I_D}{z_D},$$

where, for each $k = 1, \ldots, D$,

$$I_k = \mathrm{E}\left\{ W_k I\left(\frac{W_k}{z_k} \geq \frac{W_s}{z_s}, s = 1, \ldots, N \right) \right\} = \mathrm{E}_\xi \left[K_k \left\{ (2\gamma_k)^{1/2} \xi + 2\gamma_k \right\} \right]. \tag{4.6}$$

The quantities involved in the right-hand side of equation (4.6) are defined as

$$\xi \sim \mathcal{N}(0,1), \quad K_k(x) = \int_{-\infty}^{x - b_{-k}} \phi(\varepsilon_{-k} \mid \varepsilon_k = x) \mathrm{d}\varepsilon_{-k}, \quad k = 1, \ldots, D,$$

where ε_{-k} represents the $(D-1)$-dimensional vector $(\varepsilon_1, \ldots, \varepsilon_D)$ with the kth component removed, and where b_{-k} is the $(D-1)$-dimensional vector whose sth component equals $\gamma_k - \gamma_s + \log(z_k / z_s)$, $s = 1, \ldots, D; s \neq k$. Moreover, using the same computations as those above, equation (4.3) becomes

$$I_k = \Pr\{ Y_s \leq \gamma_{k;s} - \log(z_k / z_s); s = 1, \ldots, D, j \neq k \},$$

where the $(D-1)$-dimensional vector of Y_s's has a joint Gaussian distribution with $\mathrm{E}(Y_s) = 0$, $\mathrm{var}(Y_s) = 2\gamma_{k;s}$ and $\mathrm{cov}(Y_s, Y_t) = \gamma_{k;s} + \gamma_{k;t} - \gamma_{s;t}$, from which we get

$$I_k = \Phi_{D-1}(\eta_k; R_k),$$

where η_k is the $(D-1)$-dimensional vector with sth component $\eta(z_k, z_s)$, $s = 1, \ldots, D; s \neq k$, the function $\eta(\cdot, \cdot)$ being defined in §4.2.1.1, $\Phi_D(\cdot; R)$ denotes the cumulative distribution function of the D-variate normal distribution function with zero mean, unit variance and correlation matrix R, and R_k is the $(D-1) \times (D-1)$ correlation matrix whose (s,t)th entry is $(\gamma_{k;s} + \gamma_{k;t} - \gamma_{s;t}) / \{2(\gamma_{k;s}\gamma_{k;t})^{1/2}\}$, $s, t = 1, \ldots, D; s, t \neq k$. Thus,

$$V_{\mathscr{D}}(z_1, \ldots, z_D) = \sum_{k=1}^{D} \frac{1}{z_k} \Phi_{D-1}(\eta_k; R_k). \tag{4.7}$$

This result holds if the correlation matrices R_k are invertible, which is always true when $2\gamma(h) = (\|h\| / \lambda)^\alpha$ and $\alpha < 2$. However, in the special case $\alpha = 2$, i.e., the isotropic Smith model, if the sites $x_1, \ldots, x_D$ form a degenerate simplex in $\mathbb{R}^d$, then the determinants of the correlation matrices equal zero and the result fails. If $D > d + 1$, the simplex is always degenerate (Genton *et al.*, 2011). Moreover, if $\alpha \approx 2$, so that the Brown–Resnick

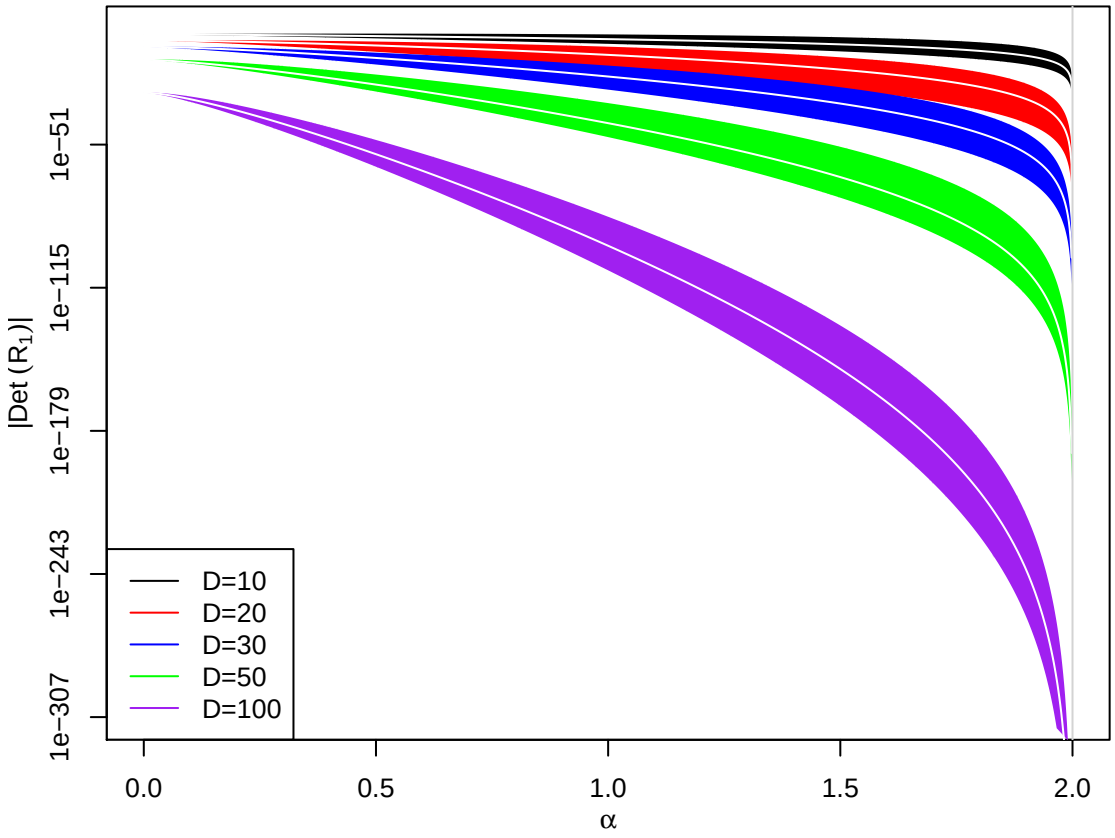

Figure 4.2: Determinant of the correlation matrix R_1 in expression (4.7) using the isotropic variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$, against the smoothness parameter $\alpha \in (0,2]$ when the range parameter λ equals 100, for $d = 2$ and $D = 10$ (black), 20 (red), 30 (blue), 50 (green) and 100 (purple). The colored areas correspond to 95% confidence regions, while the white lines denote the medians based on 50 simulated locations in $[0,100]^2$. In this illustration, the matrix R_1 is almost singular for $\alpha \approx 2$, and exactly singular for $\alpha = 2$.

process is rather smooth, and especially for large D, the correlation matrices could be numerically singular; this problem is illustrated in Figure 4.2.

We recover the results of Section 4.2.1.1 when $D = 3$ and those of Genton *et al.* (2011) when the variogram is $2\gamma(h) = h^\mathrm{T}\Sigma^{-1}h$ for some covariance matrix Σ.

4.2.2 Density

In principle the full density can be obtained by differentiation of the cumulative distribution function $G(z_1,\ldots,z_D) = \exp\{-V_{\mathscr{D}}(z_1,\ldots,z_D)\}$ with respect to $z_1,\ldots,z_D$, but the number of terms grows very fast as D increases, so direct likelihood inference

seems infeasible except for small D. Hence this section gives explicit formulae for the density function of the Brown–Resnick process in dimension $D = 3$.

In this case, the exponent measure may be written as $V_{\mathscr{D}}(z_1, z_2, z_3) = I_1/z_1 + I_2/z_2 + I_3/z_3$, where $I_k = \Phi_2\{x_k(z_1, z_2, z_3), y_k(z_1, z_2, z_3); R_k\}$ for some differentiable functions x_k and y_k of z_1, z_2, z_3, $k = 1, 2, 3$; see equation (4.7). Therefore, since the trivariate distribution is $G(z_1, z_2, z_3) = \exp\{-V_{\mathscr{D}}(z_1, z_2, z_3)\}$, the density $g(z_1, z_2, z_3)$ is

$$\begin{aligned} g(z_1, z_2, z_3) &= \frac{\partial^3}{\partial z_1 \partial z_2 \partial z_3} \exp\{-V_{\mathscr{D}}(z_1, z_2, z_3)\} && (4.8) \\ &= (-V_{123} + V_1 V_{23} + V_2 V_{13} + V_3 V_{12} - V_1 V_2 V_3) \exp(-V), && (4.9) \end{aligned}$$

where the derivatives $V_1 = \partial V_{\mathscr{D}}(z_1, z_2, z_3)/\partial z_1$, etc., are given by expressions such as

$$\begin{aligned} V_1 &= -z_1^{-2} I_1 + z_1^{-1} \frac{\partial I_1}{\partial z_1} + z_2^{-1} \frac{\partial I_2}{\partial z_1} + z_3^{-1} \frac{\partial I_3}{\partial z_1}, \\ V_{12} &= -z_1^{-2} \frac{\partial I_1}{\partial z_2} + z_1^{-1} \frac{\partial^2 I_1}{\partial z_1 \partial z_2} - z_2^{-2} \frac{\partial I_2}{\partial z_1} + z_2^{-1} \frac{\partial^2 I_2}{\partial z_1 \partial z_2} + z_3^{-1} \frac{\partial^2 I_3}{\partial z_1 \partial z_2}, \\ V_{123} &= -z_1^{-2} \frac{\partial^2 I_1}{\partial z_2 \partial z_3} + z_1^{-1} \frac{\partial^3 I_1}{\partial z_1 \partial z_2 \partial z_3} - z_2^{-2} \frac{\partial^2 I_2}{\partial z_1 \partial z_3} + z_2^{-1} \frac{\partial^3 I_2}{\partial z_1 \partial z_2 \partial z_3} \\ &\quad - z_3^{-2} \frac{\partial^2 I_3}{\partial z_1 \partial z_2} + z_3^{-1} \frac{\partial^3 I_3}{\partial z_1 \partial z_2 \partial z_3}. \end{aligned}$$

By the chain rule, and writing $x_k = x_k(z_1, z_2, z_3)$, $y_k = y_k(z_1, z_2, z_3)$ for simplicity, we have for $k, s, t, u = 1, 2, 3$ that

$$\begin{aligned} I_k &= \Phi_2(x_k, y_k; R_k), \\ \frac{\partial}{\partial z_s} I_k &= \frac{\partial}{\partial x_k} \Phi_2\left(x_k, y_k; R_k\right) \frac{\partial x_k}{\partial z_s} + \frac{\partial}{\partial y_k} \Phi_2\left(x_k, y_k; R_k\right) \frac{\partial y_k}{\partial z_s}, \\ \frac{\partial^2}{\partial z_s \partial z_t} I_k &= \frac{\partial^2}{\partial x_k^2} \Phi_2(x_k, y_k; R_k) \frac{\partial x_k}{\partial z_s} \frac{\partial x_k}{\partial z_t} + \frac{\partial^2}{\partial y_k^2} \Phi_2(x_k, y_k; R_k) \frac{\partial y_k}{\partial z_s} \frac{\partial y_k}{\partial z_t} \\ &\quad + \frac{\partial^2}{\partial x_k \partial y_k} \Phi_2(x_k, y_k; R_k) \left(\frac{\partial x_k}{\partial z_s} \frac{\partial y_k}{\partial z_t} + \frac{\partial x_k}{\partial z_t} \frac{\partial y_k}{\partial z_s} \right) \\ &\quad + \frac{\partial}{\partial x_k} \Phi_2(x_k, y_k; R_k) \frac{\partial^2 x_k}{\partial z_s \partial z_t} + \frac{\partial}{\partial y_k} \Phi_2(x_k, y_k; R_k) \frac{\partial^2 y_k}{\partial z_s \partial z_t}, \\ \frac{\partial^3}{\partial z_s \partial z_t \partial z_u} I_k &= \frac{\partial^3}{\partial x_k^3} \Phi_2(x_k, y_k; R_k) \frac{\partial x_k}{\partial z_s} \frac{\partial x_k}{\partial z_t} \frac{\partial x_k}{\partial z_u} \\ &\quad + \frac{\partial^3}{\partial x_k^2 \partial y_k} \Phi_2(x_k, y_k; R_k) \left(\frac{\partial x_k}{\partial z_s} \frac{\partial x_k}{\partial z_t} \frac{\partial y_k}{\partial z_u} + \frac{\partial x_k}{\partial z_s} \frac{\partial x_k}{\partial z_u} \frac{\partial y_k}{\partial z_t} + \frac{\partial x_k}{\partial z_t} \frac{\partial x_k}{\partial z_u} \frac{\partial y_k}{\partial z_s} \right) \end{aligned}$$

$$
\begin{aligned}
&+\frac{\partial^3}{\partial x_k \partial y_k^2}\Phi_2(x_k,y_k;R_k)\left(\frac{\partial x_k}{\partial z_s}\frac{\partial y_k}{\partial z_t}\frac{\partial y_k}{\partial z_u}+\frac{\partial x_k}{\partial z_t}\frac{\partial y_k}{\partial z_s}\frac{\partial y_k}{\partial z_u}+\frac{\partial x_k}{\partial z_u}\frac{\partial y_k}{\partial z_s}\frac{\partial y_k}{\partial z_t}\right)\\
&+\frac{\partial^3}{\partial y_k^3}\Phi_2(x_k,y_k;R_k)\frac{\partial y_k}{\partial z_s}\frac{\partial y_k}{\partial z_t}\frac{\partial y_k}{\partial z_u}\\
&+\frac{\partial^2}{\partial x_k^2}\Phi_2(x_k,y_k;R_k)\left(\frac{\partial^2 x_k}{\partial z_s \partial z_t}\frac{\partial x_k}{\partial z_u}+\frac{\partial^2 x_k}{\partial z_s \partial z_u}\frac{\partial x_k}{\partial z_t}+\frac{\partial^2 x_k}{\partial z_t \partial z_u}\frac{\partial x_k}{\partial z_s}\right)\\
&+\frac{\partial^2}{\partial x_k \partial y_k}\Phi_2(x_k,y_k;R_k)\Big(\frac{\partial^2 x_k}{\partial z_s \partial z_t}\frac{\partial y_k}{\partial z_u}+\frac{\partial^2 x_k}{\partial z_s \partial z_u}\frac{\partial y_k}{\partial z_t}+\frac{\partial^2 x_k}{\partial z_t \partial z_u}\frac{\partial y_k}{\partial z_s}\\
&\qquad\qquad+\frac{\partial x_k}{\partial z_s}\frac{\partial^2 y_k}{\partial z_t \partial z_u}+\frac{\partial x_k}{\partial z_t}\frac{\partial^2 y_k}{\partial z_s \partial z_u}+\frac{\partial x_k}{\partial z_u}\frac{\partial^2 y_k}{\partial z_s \partial z_t}\Big)\\
&+\frac{\partial^2}{\partial y_k^2}\Phi_2(x_k,y_k;R_k)\left(\frac{\partial^2 y_k}{\partial z_s \partial z_t}\frac{\partial y_k}{\partial z_u}+\frac{\partial^2 y_k}{\partial z_s \partial z_u}\frac{\partial y_k}{\partial z_t}+\frac{\partial^2 y_k}{\partial z_t \partial z_u}\frac{\partial y_k}{\partial z_s}\right)\\
&+\frac{\partial}{\partial x_k}\Phi_2(x_k,y_k;R_k)\frac{\partial^3 x_k}{\partial z_s \partial z_t \partial z_u}+\frac{\partial}{\partial y_k}\Phi_2(x_k,y_k;R_k)\frac{\partial^3 y_k}{\partial z_s \partial z_t \partial z_u}.
\end{aligned}
$$

The derivatives of the bivariate normal cumulative distribution function are easily derived as

$$
\begin{aligned}
\frac{\partial}{\partial x}\Phi_2(x,y;\rho) &= \phi(x)\Phi\left\{\frac{y-\rho x}{(1-\rho^2)^{1/2}}\right\},\\
\frac{\partial^2}{\partial x^2}\Phi_2(x,y;\rho) &= -\phi(x)x\Phi\left\{\frac{y-\rho x}{(1-\rho^2)^{1/2}}\right\}-\rho\,\phi_2(x,y;\rho),\\
\frac{\partial^2}{\partial x \partial y}\Phi_2(x,y;\rho) &= \phi_2(x,y;\rho),\\
\frac{\partial^3}{\partial x^3}\Phi_2(x,y;\rho) &= (x-1)\phi(x)\Phi\left\{\frac{y-\rho x}{(1-\rho^2)^{1/2}}\right\}+\rho\,\phi_2(x,y;\rho)\left(-x^2+x+\frac{x-\rho y}{1-\rho^2}\right),\\
\frac{\partial^3}{\partial x^2 \partial y}\Phi_2(x,y;\rho) &= -\phi_2(x,y;\rho)\frac{x-\rho y}{1-\rho^2},
\end{aligned}
$$

with the others defined by symmetry, and the non-zero derivatives of $x_k(z_1,z_2,z_3)$ and $y_k(z_1,z_2,z_3)$ with respect to z_1,z_2,z_3 are given for $n=1,2,\ldots$ by

$$
\begin{aligned}
\frac{\partial^n x_1}{\partial z_1^n} &= (n-1)!(-z_1)^{-n}\gamma_{12}^{-1/2}, & \frac{\partial^n x_1}{\partial z_2^n} &= -(n-1)!(-z_2)^{-n}\gamma_{12}^{-1/2},\\
\frac{\partial^n y_1}{\partial z_1^n} &= (n-1)!(-z_1)^{-n}\gamma_{13}^{-1/2}, & \frac{\partial^n y_1}{\partial z_3^n} &= -(n-1)!(-z_3)^{-n}\gamma_{13}^{-1/2},\\
\frac{\partial^n x_2}{\partial z_1^n} &= -(n-1)!(-z_1)^{-n}\gamma_{12}^{-1/2}, & \frac{\partial^n x_2}{\partial z_2^n} &= (n-1)!(-z_2)^{-n}\gamma_{12}^{-1/2}\\
\frac{\partial^n y_2}{\partial z_2^n} &= (n-1)!(-z_2)^{-n}\gamma_{23}^{-1/2}, & \frac{\partial^n y_2}{\partial z_3^n} &= -(n-1)!(-z_3)^{-n}\gamma_{23}^{-1/2},
\end{aligned}
$$

$$\frac{\partial^n x_3}{\partial z_1^n} = -(n-1)!(-z_1)^{-n}\gamma_{13}^{-1/2}, \qquad \frac{\partial^n x_3}{\partial z_3^n} = (n-1)!(-z_3)^{-n}\gamma_{13}^{-1/2}$$
$$\frac{\partial^n y_3}{\partial z_2^n} = -(n-1)!(-z_2)^{-n}\gamma_{23}^{-1/2}, \qquad \frac{\partial^n y_3}{\partial z_3^n} = (n-1)!(-z_3)^{-n}\gamma_{23}^{-1/2}.$$

4.3 Efficiency gains of the triplewise likelihood approach

4.3.1 Inference based on triplewise likelihood

As before, suppose that we have observed n independent copies of a Brown–Resnick process at S sites in $\mathscr{X} \subset \mathbb{R}^d$, and that the unknown parameters are summarized in the vector ψ. Using the same notation as in §4.1.2, the calculations of the previous section permit inference based on the log-triplewise likelihood

$$\ell_3(\psi) = \sum_{i=1}^{n} \sum_{s_1=1}^{S-2} \sum_{s_2=s_1+1}^{S-1} \sum_{s_3=s_2+1}^{S} \log g(z_{s_1,i}, z_{s_2,i}, z_{s_3,i}; \psi),$$

where g is the density stemming from equation (4.8). As for pairwise likelihood, the resulting maximum triplewise likelihood estimator $\widehat{\psi}_3$ is strongly consistent and asymptotically Gaussian, as $n \to \infty$. Although asymptotically less efficient than the maximum likelihood estimator, $\widehat{\psi}_3$ is expected to be less variable than $\widehat{\psi}_2$, owing to the additional information in the trivariate terms used for the fitting.

4.3.2 Simulation study

Since $\widehat{\psi}_3$ might be expected to perform better than $\widehat{\psi}_2$, the question of their relative statistical efficiency arises. In order to study this for random fields with different smoothness properties, we consider the isotropic semi-variogram $\gamma(h) = (\|h\|/\lambda)^\alpha$, $\lambda > 0, 0 < \alpha \leq 2$, which corresponds to Brown–Resnick processes built from fractional Brownian motions. We consider the seven smoothness scenarios $\alpha = 0.5, 1, 1.5, 1.9, 1.95, 1.98, 2$, the last being equivalent to the Smith model. For each scenario we consider three levels of spatial dependence, with range parameters $\lambda = 14, 28, 42$, broadly corresponding to the three cases $\sigma_{11} = \sigma_{22} = 10, 20, 30$ in Genton *et al.* (2011); Figure 4.3 displays the corresponding true extremal coefficient curves. Using the R package *SpatialExtremes* (Ribatet, 2011), we simulated n independent copies of the Brown–Resnick process with variogram $2\gamma(h)$ at the same set of S random sites uniformly generated in $[0, 100]^2$, and computed the estimates $\widehat{\psi}_2 = \{\log(\widehat{\lambda}_2), \widehat{\alpha}_2\}$ and $\widehat{\psi}_3 = \{\log(\widehat{\lambda}_3), \widehat{\alpha}_3\}$, the latter based on the expressions given in Section 4.2.2. Such simulated datasets and random locations were generated 300 times and the resulting estimates were used to compute

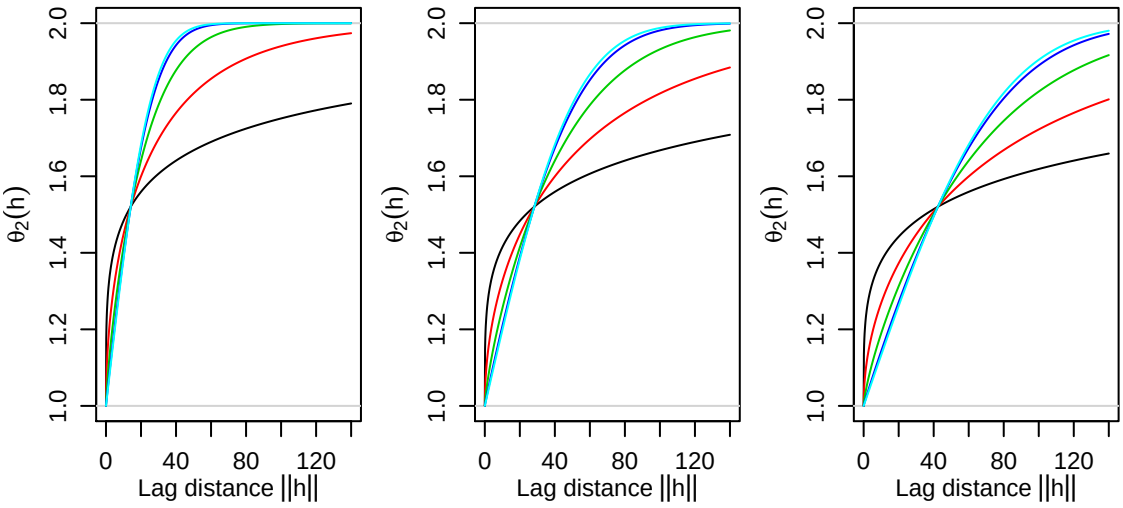

Figure 4.3: Bivariate extremal coefficients $\theta_2(h)$ for the Brown–Resnick model with range parameter $\lambda = 14$ (left), 28 (middle) and 42 (right), and smoothness parameter $\alpha = 0.5$ (black), 1 (red), 1.5 (green), 1.9 (dark blue) and 2 (light blue). The cases $\alpha = 1.95$ and 1.98 (not shown) are almost identical to the curves for $\alpha = 2$, which correspond to the Smith model.

empirical covariance matrices $\widehat{C}_2$ and $\widehat{C}_3$ for $\widehat{\psi}_2$ and $\widehat{\psi}_3$, the empirical marginal relative efficiencies $\mathrm{RE}_\lambda = \widehat{\mathrm{var}}\{\log(\widehat{\lambda}_3)\}/\widehat{\mathrm{var}}\{\log(\widehat{\lambda}_2)\}$ and $\mathrm{RE}_\alpha = \widehat{\mathrm{var}}(\widehat{\alpha}_3)/\widehat{\mathrm{var}}(\widehat{\alpha}_2)$, and the empirical global efficiency $\mathrm{RE}_\psi = \{\det(\widehat{C}_3)/\det(\widehat{C}_2)\}^{1/2}$.

4.3.2.1 Comparison of efficiencies for increasing n and fixed S

In a first simulation setting, the number of sites was fixed to $S = 20$, and the number of replicates of the process was set to $n = 5$, 10, 20 and 50.

These efficiencies are reported in Table 4.1. For rough processes, with $\alpha = 0.5, 1, 1.5$, maximum pairwise likelihood estimation has efficiency of at least 70%, and often closer to 90%, relative to the use of triples, and the efficiencies depend little on n. For smooth processes, with $\alpha = 1.9, 1.95, 1.98, 2$, the efficiency of pairwise likelihood estimation can be markedly lower, and decreases rapidly as n increases. In particular, when $\alpha = 2$, i.e., for the Smith model, observations on the same storm profile at four different sites completely determine the profile in $\mathbb{R}^2$ and thus the underlying variogram; as for triples, they provide non-negligible information about it. Hence, the triplewise estimator is much more efficient compared to the pairwise one, explaining the dramatic drop in relative efficiency observed when $\alpha \approx 2$. This behavior is more striking either when the range parameter λ is big or for large n, since in either case it

Table 4.1: Efficiency (%) of maximum pairwise likelihood estimators relative to maximum triplewise likelihood estimators for $n = 5, 10, 20, 50$ replicates, based on 300 simulations of the Brown–Resnick process with semi-variogram $\gamma(h) = (\|h\|/\lambda)^\alpha$ observed at 20 random sites in $[0, 100]^2$. The numbers in each cell are respectively $\mathrm{RE}_\lambda/\mathrm{RE}_\alpha/\mathrm{RE}_\psi$.

	$n = 5$			$n = 10$		
$\alpha \setminus \lambda$	14	28	42	14	28	42
0.5	83/89/86	89/93/91	87/93/91	94/95/94	90/93/92	93/94/93
1.0	96/92/94	97/84/90	98/88/92	96/89/93	93/90/93	95/85/90
1.5	87/81/83	93/72/79	89/67/74	89/77/82	91/71/81	89/69/78
1.9	79/81/80	72/60/61	74/56/58	84/76/79	76/48/54	66/35/47
1.95	77/80/78	67/54/54	72/54/53	76/75/74	64/46/51	60/38/43
1.98	73/80/77	63/62/58	55/42/46	70/67/66	56/38/39	49/22/29
2.0	74/80/76	61/59/52	53/48/44	64/74/68	42/39/38	26/11/16
	$n = 20$			$n = 50$		
$\alpha \setminus \lambda$	14	28	42	14	28	42
0.5	94/94/93	92/93/93	92/95/95	92/92/92	91/97/94	89/92/91
1.0	94/89/91	96/87/92	94/86/92	93/84/88	95/85/91	95/90/95
1.5	88/77/82	90/68/78	88/69/76	92/77/84	90/65/76	87/69/77
1.9	79/60/67	74/36/47	66/28/39	75/48/58	69/22/35	62/18/32
1.95	73/60/64	59/24/35	50/15/26	73/44/55	54/11/22	48/8/17
1.98	68/56/60	49/22/29	38/7/16	68/42/51	40/5/12	33/2/7
2.0	62/65/63	20/6/11	16/3/6	38/30/33	6/0/1	1/0/0

is then more likely that a single storm profile will be observed at three sites.

In the Appendix, Figure C.1 suggests that in a typical situation, with $\lambda = 28$ and $\alpha = 1$, $\widehat{\psi}_2$ and $\widehat{\psi}_3$ both estimate ψ consistently as $n \to \infty$, whereas Figure C.2 illustrates the spectacular efficiency improvement of $\widehat{\psi}_3$ compared to $\widehat{\psi}_2$ when $\alpha = 2$.

4.3.2.2 Comparison of efficiencies for increasing S and fixed n

In a second simulation setting, we fixed the number of independent copies to $n = 20$, and let the number of sites take the values $S = 10, 20, 30, 50$.

The results are reported in Table 4.2. They show that when the process is rather rough, that is for $\alpha = 0.5, 1, 1.5$, the efficiencies depend little on the number of sites S, but that when $\alpha = 1.9, 1.95, 1.98, 2$ they decrease rapidly as S increases. Analogously to the first simulation setting, when S is larger, more triples observed on the same storm profile

Table 4.2: Efficiency (%) of maximum pairwise likelihood estimators relative to maximum triplewise likelihood estimators for $n = 20$ replicates, based on 300 simulations of the Brown–Resnick process with semi-variogram $\gamma(h) = (\|h\|/\lambda)^\alpha$ observed at $S = 10, 20, 30, 50$ random sites in $[0, 100]^2$. The numbers in each cell are respectively $\mathrm{RE}_\lambda/\mathrm{RE}_\alpha/\mathrm{RE}_\psi$.

	$S = 10$			$S = 20$		
$\alpha \setminus \lambda$	14	28	42	14	28	42
0.5	97/96/96	93/92/92	94/97/95	95/94/93	93/96/95	93/96/94
1.0	94/85/90	95/84/89	96/86/91	94/85/90	95/89/93	95/90/93
1.5	88/83/88	92/64/74	91/64/74	92/78/85	91/68/76	89/69/77
1.9	80/79/82	70/50/55	66/30/42	79/66/71	71/41/48	66/30/40
1.95	76/78/79	61/40/47	55/20/29	72/54/61	60/29/35	54/20/32
1.98	77/85/83	48/29/37	40/10/20	67/51/56	50/18/27	37/7/16
2.0	75/75/75	42/37/36	26/14/15	55/62/56	24/19/21	11/0/2
	$S = 30$			$S = 50$		
$\alpha \setminus \lambda$	14	28	42	14	28	42
0.5	91/93/92	90/92/92	89/95/92	89/93/91	88/93/91	89/94/92
1.0	98/86/92	95/84/90	92/87/92	96/84/90	94/87/93	93/90/93
1.5	94/81/86	92/70/78	89/72/79	96/77/84	90/69/81	88/67/79
1.9	85/56/64	71/34/44	65/25/40	83/46/57	70/30/40	61/27/36
1.95	73/42/53	57/22/31	55/15/25	76/47/56	59/22/33	54/17/25
1.98	66/42/49	45/14/24	39/6/13	65/39/47	44/10/18	33/3/10
2.0	54/50/50	24/9/12	9/0/2	47/39/41	15/4/7	5/0/1

are likely to occur, so the "super-efficiency" of the triplewise likelihood estimator when $\alpha = 2$ has more impact in finite samples.

4.3.2.3 Further comments

Figure 4.4 shows that the relevance of the limiting Gaussian distribution of $\widehat{\psi}_3$ is questionable when $\alpha = 2$: the log triplewise likelihood is very asymmetric even for $n = 50$, whereas it is much more nearly quadratic when α is smaller. Inference based on profile marginal likelihood might thus be advisable when α is thought to be close to 2, even though classical likelihood theory does not apply in this setting. Numerical issues may be encountered when $\alpha \approx 2$, due to the sharp drop in the likelihood as the range parameter exceeds its true value, and in experiments we have found that the computation often breaks down.

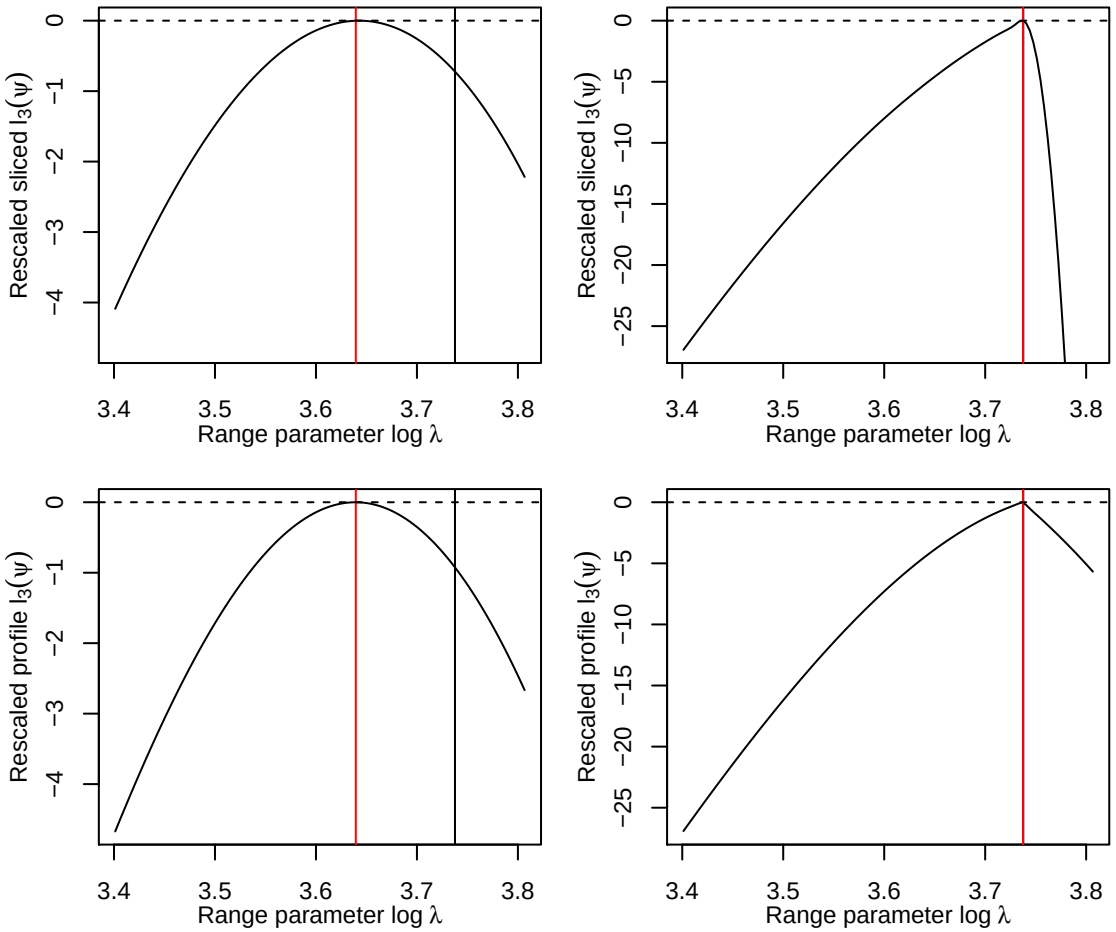

Figure 4.4: Sliced (top row) and profile (bottom row) triplewise log-likelihoods for the range parameter $\log(\lambda)$, shifted to have maximum at zero and scaled by the factor $K = \binom{S}{2}^{-1}$, for two datasets generated from a Brown–Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^{\alpha}$, where $\alpha = 1$ (left column) and $\alpha = 2$ (right column). In the upper plots, α is held fixed to its true value, while in the lower plots, α corresponds to the maximum triplewise likelihood estimator computed for given λ. The true value $\log(\lambda) \approx 3.74$ is represented by the vertical black line. The vertical red line, which corresponds to the maximum triplewise likelihood estimator, coincides with the black line in the right panels. The processes were simulated at the same 20 random sites in $[0,100]^2$, with $n = 50$ replicates, using the same random seed.

4.4 Inference using the occurrence times of extreme events

We now study the loss in efficiency of higher-order marginal likelihoods for the estimation of Brown-Resnick processes, based on the limiting Poisson process representation of threshold exceedances.

4.4.1 Stephenson–Tawn likelihood

Let Ω_D denote the set of all possible partitions of $\{1,\dots,D\}$. Since the joint distribution of max-stable processes may be written as $\exp\{-V_{\mathscr{D}}(z_1,\dots,z_D)\}$, for some exponent measure $V_{\mathscr{D}}$ with $\mathscr{D} = \{x_1,\dots,x_D\}$, it can be shown that the classical full likelihood in dimension D involves $B_D = |\Omega_D|$ terms, completely impractical for large D. Recently, Stephenson & Tawn (2005) and Wadsworth & Tawn (2013) showed how to use the occurrence times of extreme events, if available, to perform efficient inference using the full likelihood in arbitrary dimension, recall (1.35). More precisely, let $\Pi \in \Omega_D$ be a partition that indicates which elements of the componentwise maxima $(z_1,\dots,z_D)$ occurred together. For example, if $D = 3$, and the maxima at sites 1 and 3 occur simultaneously, but at different times from that at site 2, then the partition is $\Pi = \{\{1,3\},\{2\}\}$. Furthermore, let π_j, $j = 1,\dots,|\Pi|$, denote the elements of Π, and let V_{π_j} be the partial differentiation of the exponent measure $V_{\mathscr{D}}(z_1,\dots,z_D)$ with respect to the indices in π_j. Then Wadsworth & Tawn (2013) show that the contribution of an observation $(z_1,\dots,z_D)$ to the limiting Poisson process likelihood is

$$g(\boldsymbol{z};\psi) = \exp\{-V_{\mathscr{D}}(z_1,\dots,z_D)\}\prod_{j=1}^{|\Pi|}\{-V_{\pi_j}(z_1,\dots,z_D)\}, \tag{4.10}$$

thus corresponding to a single element of Ω_D only. To compute expression (4.10), the partial derivatives of potentially all orders of the function $V_{\mathscr{D}}$ are needed. Dombry *et al.* (2013) derived the mean intensity of the limiting Poisson process of exceedances associated to the Brown–Resnick process, that is the full derivative $-V_{1:D}$, and Wadsworth & Tawn (2013) thence deduced, by integration, the partial derivatives involved in (4.10). Using the shorthand notation $i:j$ to denote the set $\{i,\dots,j\}$, they showed that for each $k = 1,\dots,D$, the partial derivative of $V_{\mathscr{D}}$ with respect to the arguments $z_1,\dots,z_k$ may be expressed as

$$\begin{aligned}
-V_{1:k}(\boldsymbol{z}) \;=\; & \frac{\Phi_{D-k}\left\{\log(\boldsymbol{z}_{k+1:D}) - \boldsymbol{\mu};\Gamma\right\}}{(2\pi)^{(k-1)/2}|\Sigma_{1:k}|^{1/2}(\mathbf{1}_k^T\boldsymbol{q}_k)^{1/2}\prod_{i=1}^k z_i} \\
& \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}_k^T\Sigma_{1:k}^{-1}\boldsymbol{\gamma}_k - \frac{\boldsymbol{\gamma}_k^T\boldsymbol{q}_k\boldsymbol{q}_k^T\boldsymbol{\gamma}_k - 2\boldsymbol{\gamma}_k^T\boldsymbol{q}_k + 1}{\mathbf{1}_k^T\boldsymbol{q}_k}\right)\right\} \\
& \times \exp\left[-\frac{1}{2}\left\{\log(\boldsymbol{z}_{1:k}^T)A_{1:k}\log(\boldsymbol{z}_{1:k}) + 2\log(\boldsymbol{z}_{1:k}^T)\left(\Sigma_{1:k}^{-1}\boldsymbol{\gamma}_k + \frac{\boldsymbol{q}_k - \boldsymbol{q}_k\boldsymbol{q}_k^T\boldsymbol{\gamma}_k}{\mathbf{1}_k^T\boldsymbol{q}_k}\right)\right\}\right]
\end{aligned}$$

where $\Phi_{D-k}(\cdot;\Gamma)$ is the cumulative distribution function of a $(D-k)$-dimensional Gaussian variate with covariance matrix Γ, and $\Phi_0(\cdot;\Gamma) = 1$ by convention, where the matrix Σ has elements $\gamma(x_i) + \gamma(x_j) - \gamma(x_i - x_j)$ and $\Sigma_{1:k}$ is the matrix derived from Σ

corresponding to the indices $\{1,\ldots,k\}$, where

$$\begin{aligned}
\boldsymbol{z}_{1:k} &= (z_1,\ldots,z_k), & \boldsymbol{z}_{k+1:D} &= (z_{k+1},\ldots,z_D),\\
\boldsymbol{\gamma} &= \{\gamma(x_1),\ldots,\gamma(x_D)\}, & \boldsymbol{\gamma}_k &= \{\gamma(x_1),\ldots,\gamma(x_k)\},\\
\mathbf{1} &= (1,\ldots,1)\in\mathbb{R}^D, & \mathbf{1}_k &= (1,\ldots,1)\in\mathbb{R}^k,\\
\boldsymbol{q} &= \Sigma^{-1}\mathbf{1}, & \boldsymbol{q}_k &= \Sigma_{1:k}^{-1}\mathbf{1}_k,\\
A &= \Sigma^{-1}-\boldsymbol{q}\boldsymbol{q}^T/\mathbf{1}^T\boldsymbol{q}, & A_{1:k} &= \Sigma_{1:k}^{-1}-\boldsymbol{q}_k\boldsymbol{q}_k^T/\mathbf{1}_k^T\boldsymbol{q}_k,
\end{aligned}$$

and where, denoting by I_k the $k\times k$ identity matrix and $0_{i,j}$ a matrix full of zeros of dimension $i\times j$, we have

$$\begin{aligned}
M_{10} &= \begin{pmatrix} I_k \\ 0_{D-k,k}\end{pmatrix}\in\mathbb{R}^{D\times k}, & M_{01} &= \begin{pmatrix} 0_{k,D-k} \\ I_{D-k}\end{pmatrix}\in\mathbb{R}^{D\times(D-k)},\\
\Gamma &= \left(M_{01}^T A M_{01}\right)^{-1},\\
\boldsymbol{\mu} &= -\Gamma\left\{M_{01}^T A M_{10}\log(\boldsymbol{z}_{1:k})+M_{01}^T\left(\Sigma^{-1}\boldsymbol{\gamma}+\frac{\boldsymbol{q}-\boldsymbol{q}\boldsymbol{q}^T\boldsymbol{\gamma}}{\mathbf{1}^T\boldsymbol{q}}\right)\right\}.
\end{aligned}$$

By combining these formulae with expression (4.10), one can perform inference for the Brown–Resnick process using a full likelihood constructed from the Poisson process representation for extremes. Wadsworth & Tawn (2013) also show how to use this result to make inference based on a full likelihood for large threshold exceedances. Moreover, from the Bayesian perspective, it seems entirely feasible to adapt these methods to sample from the full posterior distribution using MCMC algorithms.

The full derivative of the function $V_{\mathscr{D}}$ is also known for the Schlather process (Dombry *et al.*, 2013) and the extremal-t process (results to appear in the forthcoming Ph.D thesis of E. Thibaud) but the partial derivatives of $V_{\mathscr{D}}$ for these models are more difficult to obtain; the extension of this methodology to other types of models is a current domain of research.

For our purposes, that is, the study of marginal likelihoods for max-stable processes, let us define the D-dimensional marginal Stephenson–Tawn log-likelihood as

$$\ell_D(\psi)=\sum_{i=1}^{n}\sum_{s_1=1}^{S-D+1}\cdots\sum_{s_D=s_{D-1}+1}^{S}\left[-V_{\mathscr{D}}(z_{s_1,i},\ldots,z_{s_D,i})+\sum_{j=1}^{|\Pi^i_{s_1,\ldots,s_D}|}\log\{-V_{\pi_j}(z_{s_1,i},\ldots,z_{s_D,i})\}\right], \tag{4.11}$$

where $\Pi^i_{s_1,\ldots,s_D}$ is the partition of $\{s_1,\ldots,s_D\}$ that indicates which elements of the componentwise maxima $(z_{s_1,i},\ldots,z_{s_D,i})$ occurred simultaneously, and the π_js denote its elements. In particular, when $D=S$ (the number of sites), we recover the full

Stephenson–Tawn log-likelihood. Because $\ell_D(\psi)$ is a composite likelihood, the corresponding maximum composite likelihood estimator $\widehat{\psi}_D$ obeys the standard theory; recall (3.10).

4.4.2 Relative efficiencies of marginal likelihood estimators

In order to study the loss in efficiency of maximum marginal likelihood estimators for max-stable models, we conducted a simulation study, focusing on the composite likelihood constructed from Stephenson–Tawn contributions (4.11), which can be computed in arbitrary dimensions. We generated $n = 20$ independent componentwise maxima from a Brown–Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$, $\lambda > 0$, $\alpha \in (0,2]$, at $S = 10$ fixed spatial sites uniformly distributed in $[0,100]^2$. For each of these block maxima, we kept track of the corresponding partitions indicating which of these extreme events occurred together. We then fitted the Brown–Resnick model to these simulated data and estimated the parameter vector $\psi = \{\log(\lambda), \alpha\}$ by means of the D-dimensional composite likelihood (4.11), for $D = 2,3,\ldots,10$, the last corresponding to the full Stephenson–Tawn likelihood. We repeated this procedure 300 times, generating new locations and datasets, and then used the resulting estimates $\widehat{\psi}_D = \{\log(\widehat{\lambda}_D), \widehat{\alpha}_D\}$ to compute the empirical covariance matrix $\widehat{C}_D$, the empirical marginal relative efficiencies $\mathrm{RE}_\lambda = \widehat{\mathrm{var}}\{\log(\widehat{\lambda}_S)\}/\widehat{\mathrm{var}}\{\log(\widehat{\lambda}_D)\}$ and $\mathrm{RE}_\alpha = \widehat{\mathrm{var}}(\widehat{\alpha}_S)/\widehat{\mathrm{var}}(\widehat{\alpha}_D)$ and the empirical global relative efficiency $\mathrm{RE}_\psi = \{\det(\widehat{C}_S)/\det(\widehat{C}_D)\}^{1/2}$. In order to study these estimators in a wide range of different situations, we used $\lambda = 14, 28, 42$ (from short to long-range dependence) and $\alpha = 0.5, 1, 1.5, 1.9, 1.95, 1.98$ (from rough to smooth processes).

Figure 4.5 reports the results (see also Figures D.1 and D.2 in the appendix), giving a broad idea about the loss of efficiency of maximum marginal likelihood estimators in an extreme-value context. Not surprisingly, the estimator based on the full likelihood beats the other estimators constructed from lower-marginal densities, and the efficiencies are increasing with the dimension. However, the results for $\lambda = 14$ and $\alpha \geq 1.9$ need to be interpreted with care, for these cases are difficult to handle, and the simulated max-stable random fields may actually be poor approximations to the targeted processes; indeed, in these situations, we simulated the Brown-Resnick processes based on 10^8 random storms (recall (2.20) and the following discussion), and still in about 0.5% of occasions the maximum was attained by one of the last 10% of storms, suggesting that this number of storms is maybe not large enough to cover all significant contributions to the maximum. By contrast, when $\alpha = 0.5$, only about 1800 storm replicates would be sufficient to produce similar characteristics. It is also worth noting that for very smooth processes, with $\alpha = 1.98$, the computation of the

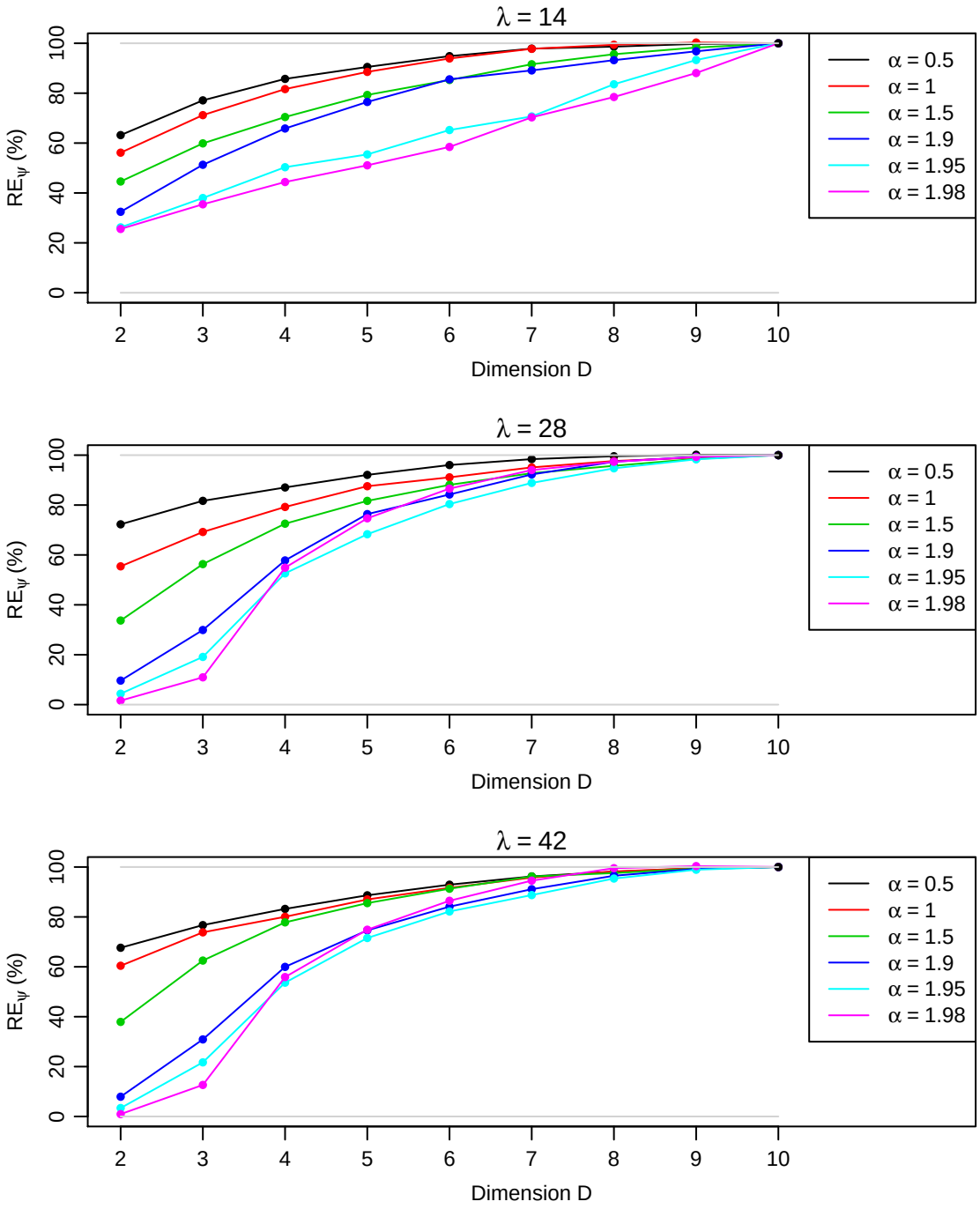

Figure 4.5: Global relative efficiency RE_ψ of the maximum D-dimensional marginal Stephenson–Tawn likelihood estimator (with respect to the maximum full likelihood estimator), based on 300 replications of the Brown-Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$. We used $\lambda = 14$ (top), 28 (middle), 42 (bottom), and $\alpha = 0.5, 1, 1.5, 1.9, 1.95, 1.98$ (different colors).

maximum likelihood estimator was not free of numerical issues in about 2% of cases for $\lambda = 14$.

Apart from these numerical limitations, Figure 4.5 shows that the relative efficiency of the composite Stephenson–Tawn likelihood estimators appears to be higher for rougher processes, and this difference is more pronounced for processes with longer-range dependence. In typical real applications (which usually have $\alpha \leq 1$), pairwise likelihoods are good enough to be useful. Hence, pairwise likelihood estimators turn out to have several advantages concerning the variogram estimation of Brown–Resnick processes: they can be quickly computed, they are robust against misspecification of higher-order interactions and finally, they offer quite high statistical efficiency for rough processes.

Furthermore, we can also see that the four-dimensional composite likelihood estimator has more than 79% relative efficiency when $\alpha \leq 1$, more than 70% when $\alpha \leq 1.5$, and lower but reasonable values when $1.5 < \alpha \leq 2$. Hence, the higher-order interactions do not add much information for the estimation of dependence parameters, especially when the process is rough. It would be interesting to see how these results generalize to weighted composite likelihoods, or when the numbers of replicates n and of sites S are larger.

4.5 Discussion and extensions

In this chapter, we provide explicit expressions (4.5) and (4.7) for the exponent measure of the Brown–Resnick process in arbitrary dimensions, on which likelihood inference may be based. Use of triplewise likelihood rather than pairwise likelihood to fit these models can lead to an efficiency gain of up to 30% for rough processes, and much more if the process is very smooth. This augments the results of Genton *et al.* (2011), which show huge efficiency gains associated to high order composite likelihoods for the Smith model. Our more general results confirm those of Genton *et al.* (2011) for the Smith model, but in the more realistic setting when the process is rough, the small improvement afforded by the triplewise approach is probably not worth the additional computational and coding effort, particularly as issues of numerical precision may then arise if the number of sites is large. In principle it is possible to compute the full likelihood for the Brown–Resnick process in high dimensions, but the number of terms in the likelihood and the need to compute high-dimensional multivariate normal distribution functions in numerically near-singular cases seem to preclude this in practice.

When the occurrence times of extreme events are available, Wadsworth & Tawn (2013)

explain how the full likelihood can be markedly simplified, and hence made computable in practice. We have used these recent results to investigate the loss of efficiency of composite likelihood methods in an extreme-value context, supplementing in some sense our study of §3.3.3.1. We show that the pairwise likelihood estimator performs reasonably well relative to the maximum full likelihood estimator, especially for rough processes, and unlike the latter, is free of numerical problems in large dimensions when the process is too smooth. However, if possible, the maximum full likelihood estimator should always be used, for it provides the most efficient estimator as $n \to \infty$ among the class of unbiased estimators, and outperforms lower-dimensional composite likelihood estimators in high-dimensional applications.

It would be interesting to know whether the efficiency results given here generalize to weighted marginal composite likelihoods (Varin *et al.*, 2011). The best choice of subsets of sites is related to the separate topic of optimal design for likelihood estimation. Also, since the full likelihood is available when the individual observations are known, it seems entirely feasible to embed the Brown–Resnick model into a Bayesian framework, using the standard MCMC methods to perform inference. The actual question is whether this would be possible for other types of model, such as the Schlather or more generally the extremal-t models. Ongoing work by the EPFL PhD candidate Emeric Thibaud seems to indicate that this is the case. Furthermore, threshold-based inference based on the full likelihood seems to be possible, promising even higher efficiency gains for this class of processes.

5 Real case study: Space-time modeling of extreme rainfall

All previous chapters converge to this chapter, where we tie together geostatistics and statistics of extremes to construct asymptotically valid space-time models for extremes. The spatio-temporal aspect of this modeling is novel, though related work includes Davis & Mikosch (2008) and Davis *et al.* (2013a,b). In Section 5.1, we briefly recall useful notions from the background chapters 1–2. In Section 5.2, inference based on pairwise likelihood is tackled in a framework where spatial and temporal dependence are present and we prove the strong consistency, as well as the asymptotic normality, of our maximum pairwise likelihood estimator under mild mixing conditions. Our result generalizes the established literature (recall §3.1.3) and complements Davis & Yau (2011)'s results, stated for linear time series models. In Section 5.3, we describe an application to space-time modeling of extreme rainfall in Switzerland and construct extremal models able to capture the space-time interactions. In particular, we extend the random set model proposed by Davison & Gholamrezaee (2012) to the space-time framework and show that it provides a reasonable fit to our data. To our knowledge, this is the first space-time application of max-stable processes to model high threshold exceedances. Finally, in Section 5.4, we give some concluding discussion and some ideas for future extensions. A summary of the work presented here has been published in Huser & Davison (2013b).

5.1 Threshold modeling for extremes

By contrast to block maximum approaches, models based on threshold exceedances permit to use more data for inference and to have a more sophisticated treatment of temporal dependence, which is crucial for risk management purposes. In the following sections, we describe marginal (§5.1.1) and dependence (§5.1.2) models for large threshold exceedances.

5.1.1 Marginal modeling

As explained in §1.1, for a large class of stationary processes $\{Y_t\}$, $t \in \mathbb{Z}$, the distribution of renormalized block maxima can be well approximated by the generalized extreme-value (GEV) distribution (1.4), $G(y) = \exp\left[-\{1+\xi(y-\mu)/\sigma\}^{-1/\xi}\right]$, defined on the set $\{y \in \mathbb{R} : 1+\xi(y-\mu)/\sigma > 0\}$, with $\mu \in \mathbb{R}$, $\sigma > 0$, $\xi \in \mathbb{R}$ and with the value for $\xi = 0$ interpreted as the limit when $\xi \to 0$. Similarly, provided the result for maxima holds and u is a high threshold, the conditional distribution of $Y_t - u$, suitably rescaled and conditional on $Y_t > u$, may be approximated by the generalized Pareto distribution, GPD(τ, ξ) (Davison & Smith, 1990). Its distribution function is

$$H(y) = 1 - \left(1 + \frac{\xi}{\tau} y\right)_+^{-1/\xi}, \qquad y > 0,$$

where the scale parameter is linked to that of the GEV distribution by $\tau = \sigma + \xi(u-\mu)$, and the shape parameter ξ is the same as for G. Therefore, the distribution F of Y_t can be approximated by

$$\tilde{F}(y) = \begin{cases} \widehat{F}(y), & y \le u; \\ 1 - \widehat{\zeta}_u \left\{1 + \widehat{\xi}(y-u)/\widehat{\tau}\right\}^{-1/\widehat{\xi}}, & y > u, \end{cases} \tag{5.1}$$

where $\widehat{F}(y)$ is the empirical distribution function of the sample $Y_1, \ldots, Y_n$, $\widehat{\zeta}_u$ is the estimated probability of exceeding the threshold u and $\widehat{\tau}$ and $\widehat{\xi}$ are estimates of τ and ξ. The transformation $t(y) = -1/\log \tilde{F}(y)$ therefore approximately standardizes the observations to have the unit Fréchet distribution $\exp(-1/y)$, for $y > 0$. The choice of threshold u is typically made informally using diagnostic plots, sometimes aided by theoretical arguments (recall §1.1.2.5), though it can be vexing for nonstationary data; see Scarrott & MacDonald (2012) and Northrop & Jonathan (2011) and its discussion.

A spatial model for the margins may be obtained by letting the GPD parameters vary with spatial covariates, e.g., as (log-)linear/logit functions of altitude, or annual mean precipitation (recall §2.2 and see Cooley *et al.*, 2007, Thibaud *et al.*, 2013). However, in the data analysis performed in §5.3, separate marginal models are used at each station because our interest resides in extremal dependence rather than marginal aspects.

Joint modeling of extremes is crucial for realistic assessment of spatial risks, and the next section describes models for spatial or spatio-temporal extremes, with margins previously transformed to the unit Fréchet scale.

5.1.2 Dependence modeling based on max-stable processes

Several models may be used to describe spatial dependence among maxima or threshold exceedances at distinct locations. Models based on the Gaussian copula (recall §2.2 and 2.4.1) have been used extensively in classical geostatistics, but have serious limitations for the modeling of extremes: they appear to be rather inflexible and, more importantly, they are asymptotically independent (recall the application in §2.7 and the discussion therein). Alternatively, in §2.3, we have seen that under mild conditions, max-stable processes are justified models for spatial, or spatio-temporal, extremes, and in §2.3.2, different parametric families of max-stable processes have been given. Although initially constructed for modeling spatial maxima of properly renormalized random processes, max-stable models may also be fitted to exceedances above high thresholds, because the dependence structure at high levels is essentially the same as that of maxima of independent replicates of the "baseline process"; recall §2.3.4. Under mild conditions (see Theorem 52), stationary max-stable processes with unit Fréchet margins $Z(x)$, $x \in \mathscr{X}$, may be represented as

$$Z(x) = \sup_{i \geq 1} W_i(x)/P_i, \tag{5.2}$$

where the P_is are the points of a unit rate Poisson process on $\mathbb{R}_+$ and the W_is are independent replicates of a stationary positive random process $W(x)$ with unit mean. Their finite-dimensional joint distributions may be expressed as

$$\Pr\{Z(x_1) \leq z_1, \ldots, Z(x_D) \leq z_D\} = \exp\{-V_{\mathscr{D}}(z_1, \ldots, z_D)\}, \tag{5.3}$$

where $V_{\mathscr{D}}$ is the exponent measure corresponding to a set of sites $\mathscr{D} = \{x_1, \ldots, x_D\}$, and the extremal coefficient, defined as $\theta_D(\boldsymbol{h}) = V_{\mathscr{D}}(1, \ldots, 1)$, may be used as a measure of extremal dependence among the variables $Z(x_1), \ldots, Z(x_D)$; recall (2.19).

When spatial dependence is thought to weaken at higher levels, asymptotic independence models may be preferred (recall §2.4 and the discussion in §2.7) and so-called inverted max-stable processes provide useful alternatives to the rigid Gaussian copula model. The finite-dimensional joint survival distributions for these processes may be expressed as

$$\Pr\{Z(x_1) > z_1, \ldots, Z(x_D) > z_D\} = \exp\left[-V_{\mathscr{D}}\{s(z_1), \ldots, s(z_D)\}\right], \tag{5.4}$$

where $V_{\mathscr{D}}$ is the exponent measure of some max-stable process, and $s(z) = -1/\log\{1 - \exp(-1/z)\}$ is a site-wise transformation; recall (2.49). For this class of processes, $\theta_D(\boldsymbol{h}) = D$, and the coefficient of tail dependence $\eta(h)$ may be useful to measure the rate of the upper tail decay towards independence; recall §1.2.4.1 and §2.4.

In Section 5.2, we address pairwise likelihood inference for these extremal models, and discuss the asymptotic properties of the resulting estimator in a space-time context. In Section 5.3, we shall build and fit space-time models for extreme rainfall in Switzerland based on max-stable and inverted max-stable models.

5.2 Inference

Recent advances have shown how inference may be performed for some classes of max-stable models, based on a full likelihood (Wadsworth & Tawn, 2013), but this is still unclear for other types of max-stable families and inverted max-stable models. In Chapter 3, we have investigated the usefulness of pairwise likelihoods in this context, and have shown that the loss in efficiency of maximum pairwise likelihood estimators may be controlled by considering a quite small subset of significant pairs. In the following sections, we extend the discussion of §3.2 to the case of temporal dependence, and show that the classical asymptotic results of §3.1.3 remain valid under mild mixing conditions in time.

From now on, we shall assume that the non-degenerate space-time random process $Z(s,t)$, $x=(s,t)\in\mathscr{X}=\mathscr{S}\times\mathscr{T}$, has been observed at S monitoring stations and at times $1,\ldots,T$, that is at $N=ST$ locations in $\mathscr{X}$. For simplicity of notation we let $Z_{s,t}$ denote the value recorded at the sth station at time t.

5.2.1 Censored pairwise likelihood approach

Assuming that some extremal model (either max-stable or inverted max-stable) provides a suitable fit above a high threshold u, we can follow (3.19) and consider the censored threshold-based log-pairwise likelihood

$$\ell_{\mathscr{K}}(\psi)=\sum_{t=1}^{T}\sum_{h\in\mathscr{K}_t}\sum_{s_1=1}^{S}\sum_{s_2=1}^{S}\{1-I(s_1\geq s_2 \text{ and } h=0)\}\log p_u\left(Z_{s_1,t},Z_{s_2,t+h};\psi\right), \tag{5.5}$$

with corresponding maximum pairwise likelihood estimator

$$\widehat{\psi}_{\mathscr{K}}=\arg\max_{\psi\in\Psi}\ell_{\mathscr{K}}(\psi), \tag{5.6}$$

where $\Psi\subset\mathbb{R}^p$ is the parameter space, where $\mathscr{K}_t=\{h\in\mathscr{K}:h\leq T-t\}$ and $\mathscr{K}\subset\mathbb{N}\cup\{0\}$ is a finite collection of time lags, with p_u given by equation (3.18), the pairwise distribution G stemming for example from (5.3) or (5.4), and where $I(\cdot)$ is the indicator function. Hence, expression (5.5) corresponds to summing the contributions of pairs

of observations across all stations, at predefined time lags (within the set $\mathcal{K}$). For example, with $\mathcal{K} = \{0, 1, \ldots, K\}$ for $K < \infty$, all space-time pairwise contributions up to a maximum time lag K are considered, and if $K = T - 1$, (5.5) reduces to the full pairwise likelihood. As already mentioned in Chapter 3, the associated computational burden can be reduced and statistical efficiency gained by taking another subset $\mathcal{K}$. Any non-empty finite set can essentially be used, but by considering the sets $\mathcal{K}_a^K, \mathcal{K}_b^K, \mathcal{K}_c^K$ defined in (3.65), we have found that the last two often perform better than $\mathcal{K}_a^K$ under various scenarios and assumptions.

5.2.2 Asymptotics under mixing conditions

Davison & Gholamrezaee (2012) and Padoan *et al.* (2010) use pairwise likelihood for inference on max-stable processes, regarding annual maxima at distinct locations as independent, which implies that the maximum pairwise likelihood estimator is strongly consistent and asymptotically normal; recall §3.1.3. In the case of spatio-temporal extremes, the asymptotic normality of the estimator $\widehat{\psi}_{\mathcal{K}}$ in (5.6) stems from a central limit theorem for stationary time series applied to the composite score $U(\psi) = \partial \ell_{\mathcal{K}}(\psi)/\partial\psi = \sum_{t=1}^{T} u_t(\psi)$, where $u_t(\psi)$ is the derivative of the rightmost triple sums in equation (5.5) with respect to ψ. However, as the elements $u_t(\psi)$ are generally correlated over time t, we need an additional mixing condition in order for classical asymptotics to hold. A mild sufficient condition is that the process $Z(s,t)$ be temporally α-mixing, along with a condition on the rate at which the mixing coefficients $\alpha(n)$ must decay, ensuring that the correlation vanishes sufficiently fast at infinity.

Definition 56 (α-mixing)**.** *According to Definition 1.6 of Bradley (2007a), a time series Z_t, $t \in \mathbb{Z}$, is* α-mixing *with coefficients $\alpha(n)$, or* strongly-mixing, *if*

$$\alpha(n) = \sup |\Pr(\mathscr{A} \cap \mathscr{B}) - \Pr(\mathscr{A}) \Pr(\mathscr{B})| \to 0, \qquad n \to \infty,$$

where the supremum is over all sets $\mathscr{A} \in \mathscr{F}_{-\infty}^{t}$ and $\mathscr{B} \in \mathscr{F}_{t+n}^{\infty}$, with $t \in \mathbb{Z}$, using the notation $\mathscr{F}_{t_1}^{t_2}$ to denote the σ-algebra generated by the random variables $\{Z_t : t_1 \leq t \leq t_2\}$.

A space-time process $Z(s,t)$, $(s,t) \in \mathscr{X} = \mathscr{S} \times \mathscr{T}$, is temporally α-mixing *with coefficients $\alpha(n)$ if for all $s \in \mathscr{S}$, for all sequences $t_n \subset \mathscr{T}$ such that $t_n/n \to C > 0$ as $n \to \infty$, the time series $\{Z(s,t_n) : n \in \mathbb{N}\}$ is α-mixing with coefficients $\alpha_s(n)$ and where $\sup_{s \in \mathscr{S}} \alpha_s(n) \leq \alpha(n) \to 0$ as $n \to \infty$.*

With the α-mixing condition, also known as the *strong mixing* condition, two events

become more and more independent as their time lag increases. In particular, all m-dependent processes are α-mixing. In order to demonstrate Theorem 58, whose proof relies on the theory of estimating equations, we need first to introduce regularity conditions for composite likelihoods (see also Davison, 2003, p.118).

Definition 57 (regularity conditions for composite likelihoods)**.** *A log-composite likelihood $\ell_C(\psi)$ with component densities $g(\boldsymbol{z};\psi)$, $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^D$, $\psi \in \Psi \subset \mathbb{R}^p$, is said to be* regular *if it satisfies the following properties:*

- *The true value ψ_0 of ψ is interior to the parameter space Ψ, which has finite dimension and is compact.*
- *If $\ell_C(\psi^1) = \ell_C(\psi^2)$, then $\psi^1 = \psi^2$, meaning that ψ is* identifiable *from the underlying component densities.*
- *There exists a neighborhood $\mathcal{N}$ of ψ_0 within which the first three derivatives of the log-composite likelihood with respect to ψ exist almost surely, and for $k, l, m = 1, \ldots, p$, $\mathrm{E}\left\{|\partial^3 \ell_C(\psi)/(\partial\psi_k \partial\psi_l \partial\psi_m)|\right\}$ is uniformly bounded for $\psi \in \mathcal{N}$.*
- *Within $\mathcal{N}$, the sensitivity and variability matrices $J_C(\psi) = \mathrm{E}\left\{\partial^2 \ell_C(\psi)/(\partial\psi\partial\psi^T)\right\}$ and $K_C(\psi) = \mathrm{var}\left\{\partial\ell_C(\psi)/\partial\psi\right\}$ are finite and invertible, and the sandwich matrix $V_C(\psi) = J_C^{-1}(\psi) K_C(\psi) J_C^{-1}(\psi)$ is positive definite.*
- *For any measurable subset $A \subset \mathcal{Z}$ that does not depend upon ψ, the order of differentiation and integration can be interchanged in*

$$\int_A \frac{\partial g(\boldsymbol{z};\psi)}{\partial\psi}\mathrm{d}\boldsymbol{z} = \frac{\partial}{\partial\psi}\int_A g(\boldsymbol{z};\psi)\mathrm{d}\boldsymbol{z}, \qquad \int_A \frac{\partial^2 g(\boldsymbol{z};\psi)}{\partial\psi\partial\psi^T}\mathrm{d}\boldsymbol{z} = \frac{\partial^2}{\partial\psi\partial\psi^T}\int_A g(\boldsymbol{z};\psi)\mathrm{d}\boldsymbol{z}.$$

Theorem 58. *Let $Z(s,t)$ be a strictly stationary spatio-temporal extremal process, for which the log-pairwise likelihood (5.5) is regular (see Definition 57). Moreover, assume that $Z(s,t)$ is temporally α-mixing with coefficients $\alpha(n)$, and that for the true underlying parameter $\psi_0 \in \Psi$, there exists some $\delta > 0$ such that $\mathrm{E}\left\{|u_1(\psi_0)|^{2+\delta}\right\} < \infty$ and $\sum_{n\geq 1} |\alpha(n)|^{\delta/(2+\delta)} < \infty$. Then, the following conclusions hold:*

i) $\widehat{\psi}_{\mathcal{K}}$ is strongly consistent, meaning $\widehat{\psi}_{\mathcal{K}} \to \psi_0$ with probability one as $T \to \infty$;

ii) $\widehat{\psi}_{\mathcal{K}}$ is asymptotically Gaussian; more precisely,

$$T^{1/2} k(\psi_0)^{-1/2} j(\psi_0)(\widehat{\psi}_{\mathcal{K}} - \psi_0) \to \mathcal{N}\left(0, I_p\right),$$

in distribution as $T \to \infty$, where I_p is the $p \times p$ identity matrix and where $j(\psi)$ and $k(\psi)$ are respectively the renormalized sensitivity and variability matrices, that is,

$$j(\psi) = \mathrm{E}\left\{-\frac{\partial}{\partial\psi}u_1(\psi)\right\}, \tag{5.7}$$

$$\begin{aligned} k(\psi) &= T^{-1}\mathrm{var}\left\{\sum_{t=1}^{T} u_t(\psi)\right\} \qquad (5.8)\\ &= \mathrm{E}\{u_1(\psi)u_1(\psi)^T\} + \sum_{t=1}^{T-1}\left(1-\frac{t}{T}\right)\left[\mathrm{E}\left\{u_1(\psi)u_{t+1}(\psi)^T\right\} + \mathrm{E}\left\{u_{t+1}(\psi)u_t(\psi)^T\right\}\right]\\ &\to \mathrm{E}\{u_1(\psi)u_1(\psi)^T\} + \sum_{t=1}^{\infty}\left[\mathrm{E}\left\{u_1(\psi)u_{t+1}(\psi)^T\right\} + \mathrm{E}\left\{u_{t+1}(\psi)u_t(\psi)^T\right\}\right] < \infty, \end{aligned}$$

as $T \to \infty$.

Proof. For notational simplicity, we give the proof in the case where the parameter ψ is scalar, but the argument extends to the vector case. We follow the lines of the proof given in Davison (2003, p.122–125) for the usual maximum (full) likelihood estimator.

i) **Consistency**. Let $\overline{\ell}_{\mathcal{K}}(\psi) = T^{-1}\ell_{\mathcal{K}}(\psi) = T^{-1}\sum_{t=1}^{T}\ell_t(\psi)$, with $\ell_t(\psi)$ corresponding to the rightmost triple sums in equation (5.5). Using the linearity of the expectation, and Jensen's inequality, we have that

$$\begin{aligned} &\mathrm{E}\left\{\overline{\ell}_{\mathcal{K}}(\psi) - \overline{\ell}_{\mathcal{K}}(\psi_0)\right\} = \frac{1}{T}\mathrm{E}\left\{\ell_{\mathcal{K}}(\psi) - \ell_{\mathcal{K}}(\psi_0)\right\}\\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{h\in\mathcal{K}_t}\sum_{s_1=1}^{S}\sum_{s_2=1}^{S}(1 - I\{s_1 \geq s_2 \text{ and } h=0\})\,\mathrm{E}\left[\log\left\{\frac{p_u\left(Z_{s_1,t}, Z_{s_2,t+h};\psi\right)}{p_u\left(Z_{s_1,t}, Z_{s_2,t+h};\psi_0\right)}\right\}\right] \qquad (5.9)\\ &\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{h\in\mathcal{K}_t}\sum_{s_1=1}^{S}\sum_{s_2=1}^{S}(1 - I\{s_1 \geq s_2 \text{ and } h=0\})\log\left[\mathrm{E}\left\{\frac{p_u\left(Z_{s_1,t}, Z_{s_2,t+h};\psi\right)}{p_u\left(Z_{s_1,t}, Z_{s_2,t+h};\psi_0\right)}\right\}\right] = 0, \end{aligned}$$

the last equality being true because

$$\begin{aligned} \mathrm{E}\left\{\frac{p_u\left(z_1,z_2;\psi\right)}{p_u\left(z_1,z_2;\psi_0\right)}\right\} &= \iint \frac{p_u\left(z_1,z_2;\psi\right)}{p_u\left(z_1,z_2;\psi_0\right)}g(z_1,z_2;\psi_0)\mathrm{d}z_1\mathrm{d}z_2\\ &= \int_0^u\int_0^u \frac{G\left(u,u;\psi\right)}{G\left(u,u;\psi_0\right)}g(z_1,z_2;\psi_0)\mathrm{d}z_1\mathrm{d}z_2\\ &\quad + \int_0^u\int_u^\infty \frac{G_2\left(u,z_2;\psi\right)}{G_2\left(u,z_2;\psi_0\right)}g(z_1,z_2;\psi_0)\mathrm{d}z_1\mathrm{d}z_2\\ &\quad + \int_u^\infty\int_0^u \frac{G_1\left(z_1,u;\psi\right)}{G_1\left(z_1,u;\psi_0\right)}g(z_1,z_2;\psi_0)\mathrm{d}z_1\mathrm{d}z_2 \end{aligned}$$

$$
\begin{aligned}
&+\int_u^\infty\int_u^\infty \frac{g\left(z_1,z_2;\psi\right)}{g\left(z_1,z_2;\psi_0\right)} g(z_1,z_2;\psi_0)\mathrm{d}z_1\mathrm{d}z_2 \\
= \;& G\left(u,u;\psi\right)+\left\{\exp(-1/u)-G(u,u;\psi)\right\} \\
&+\left\{\exp(-1/u)-G(u,u;\psi)\right\}+\left\{1-2\exp(-1/u)+G(u,u;\psi)\right\} \\
= \;& 1.
\end{aligned}
$$

Hence, $\mathrm{E}\left\{\overline{\ell}_{\mathcal{K}}(\psi)-\overline{\ell}_{\mathcal{K}}(\psi_0)\right\}\leq 0$, and the inequality is strict unless $\psi=\psi_0$ because $\ell_{\mathcal{K}}(\psi)$ is assumed to be regular (second regularity condition).

Now, because the process $Z(s,t)$ is stationary and temporally α-mixing, the sequence $\{\ell_t(\psi)-\ell_t(\psi_0)\}$, $t=1,2,\ldots$, is also stationary and α-mixing with coefficients $\alpha'(n)=\alpha(n-\max\mathcal{K})$. This implies that $\{\ell_t(\psi)-\ell_t(\psi_0)\}$ is also *ergodic*, and in particular that the strong law of large numbers applies (Bradley, 2007a, pp.48–51), that is

$$
\overline{\ell}_{\mathcal{K}}(\psi)-\overline{\ell}_{\mathcal{K}}(\psi_0) \quad\rightarrow\quad \mathrm{E}\left\{\ell_1(\psi)-\ell_1(\psi_0)\right\}=-D(g_\psi,g_{\psi_0}), \tag{5.10}
$$

as $T\rightarrow\infty$ with probability one, where the right-hand side of expression (5.10),

$$
\begin{aligned}
-D(g_\psi,g_{\psi_0}) \;=\; & \sum_{h\in\mathcal{K}}\sum_{s_1=1}^{S}\sum_{s_2=1}^{S}\left(1-I\{s_1\geq s_2 \text{ and } h=0\}\right) \\
& \times\int\log\left\{\frac{p_u(z_{s_1,1},z_{s_2,1+h};\psi)}{p_u(z_{s_1,1},z_{s_2,1+h};\psi_0)}\right\}p_u(z_{s_1,1},z_{s_2,1+h};\psi_0)\mathrm{d}z_{s_1,1}\mathrm{d}z_{s_2,1+h},
\end{aligned}
$$

corresponds to a (negated) composite Kullback–Leibler discrepancy between the models g_ψ and g_{ψ_0} above the level u, which is strictly negative unless $\psi=\psi_0$. Therefore, the function $h(\psi)=\overline{\ell}_{\mathcal{K}}(\psi)-\overline{\ell}_{\mathcal{K}}(\psi_0)$ satisfies

- $h(\psi_0)=0$,
- For all $\epsilon>0$, $h(\psi_0-\epsilon)\rightarrow -D(g_{\psi_0-\epsilon},g_{\psi_0})<0$, as $T\rightarrow\infty$,
- For all $\epsilon>0$, $h(\psi_0+\epsilon)\rightarrow -D(g_{\psi_0+\epsilon},g_{\psi_0})<0$, as $T\rightarrow\infty$.

Hence, this implies that there exists a T' such that for all $T>T'$, $h(\psi)$ (and therefore also $\ell_{\mathcal{K}}(\psi)$) has a local maximum in the interval $(\psi_0-\epsilon,\psi_0+\epsilon)$. Because this occurs with probability one for all $\epsilon>0$, the maximum composite likelihood estimator $\widehat{\psi}_{\mathcal{K}}$ is strongly consistent.

ii) **Asymptotic normality**. First, notice that by definition of the censored pairwise density $p_u(z_1,z_2;\psi)$ and because the order of differentiation and integration may be

interchanged (fifth regularity condition), we have

$$\begin{aligned}
\mathrm{E}\left\{\frac{\partial}{\partial\psi}\log p_u\left(z_1,z_2;\psi_0\right)\right\} &= \iint \frac{\frac{\partial}{\partial\psi}p_u\left(z_1,z_2;\psi_0\right)}{p_u\left(z_1,z_2;\psi_0\right)} g\left(z_1,z_2;\psi_0\right)\mathrm{d}z_1\mathrm{d}z_2 \\
&= \frac{\partial}{\partial\psi}G\left(u,u;\psi_0\right) + \int_u^\infty \frac{\partial}{\partial\psi}G_2\left(u,z_2;\psi_0\right)\mathrm{d}z_2 \\
&\quad + \int_u^\infty \frac{\partial}{\partial\psi}G_1\left(z_1,u;\psi_0\right)\mathrm{d}z_1 + \int_u^\infty\int_u^\infty \frac{\partial}{\partial\psi}g\left(z_1,z_2;\psi_0\right)\mathrm{d}z_1\mathrm{d}z_2 \\
&= \frac{\partial}{\partial\psi}\Big[G\left(u,u;\psi_0\right) + \left\{\exp(-1/u) - G(u,u;\psi_0)\right\} \\
&\quad + \left\{\exp(-1/u) - G(u,u;\psi_0)\right\} + \left\{1 - 2\exp(-1/u) + G(u,u;\psi_0)\right\}\Big] \\
&= 0.
\end{aligned}$$

Therefore, it follows that

$$\mathrm{E}\{u_t(\psi_0)\} = \sum_{h\in\mathscr{K}_t}\sum_{s_2=1}^{S}\sum_{s_1=1}^{S}\left(1 - I\{s_1 \geq s_2 \text{ and } h=0\}\right)\underbrace{\mathrm{E}\left\{\frac{\partial}{\partial\psi}\log p_u\left(Z_{s_1,t}, Z_{s_2,t+h};\psi_0\right)\right\}}_{=0} = 0,$$

and we also have that $\mathrm{E}\{U(\psi_0)\} = \mathrm{E}\left\{\sum_{t=1}^T u_t(\psi_0)\right\} = 0$. The variance of $U(\psi_0)$ renormalized by T is (Shumway & Stoffer, 2004, p.510)

$$\begin{aligned}
T^{-1}\mathrm{var}\{U(\psi_0)\} &= \mathrm{E}\{u_1(\psi_0)^2\} + 2\sum_{t=1}^{T-1}\left(1 - \frac{t}{T}\right)\mathrm{E}\{u_1(\psi_0)u_{t+1}(\psi_0)\} \\
&\to \mathrm{E}\{u_1(\psi_0)^2\} + 2\sum_{t=1}^{\infty}\mathrm{E}\{u_1(\psi_0)u_{t+1}(\psi_0)\}, \quad T\to\infty,
\end{aligned}$$

if the sum converges absolutely. Now, as $\widehat{\psi}_{\mathscr{K}}$ is the maximum pairwise likelihood estimator, second-order Taylor expansion of $u_t(\widehat{\psi}_{\mathscr{K}})$ around the true parameter ψ_0 gives

$$0 = \sum_{t=1}^T u_t(\widehat{\psi}_{\mathscr{K}}) \doteq \sum_{t=1}^T\left\{u_t(\psi_0) + \frac{\partial}{\partial\psi}u_t(\psi_0)(\widehat{\psi}_{\mathscr{K}} - \psi_0)\right\},$$

which gives, up to a term of the order $O_p\left\{\left(\widehat{\psi}_{\mathscr{K}} - \psi_0\right)^2\right\}$, that

$$\widehat{\psi}_{\mathscr{K}} \doteq \psi_0 + \left\{\sum_{t=1}^T h_t(\psi_0)\right\}^{-1}\sum_{t=1}^T u_t(\psi_0) = \psi_0 + H(\psi_0)^{-1}U(\psi_0), \tag{5.11}$$

where $h_t(\psi_0) = -\partial u_t(\psi_0)/\partial\psi$ and $H(\psi_0) = \sum_{t=1}^T h_t(\psi_0)$ is the composite observed information, provided the latter is invertible. Moreover, since the process $Z(s,t)$ is

assumed to be temporally α-mixing with coefficients $\alpha(n)$, the time series $\{u_t(\psi_0)\}$, $t = 1, 2, \ldots$, is also α-mixing with coefficients $\alpha'(n) = \alpha(n - \max \mathscr{K})$. Hence,

$$\alpha'(n) \to 0, \qquad \sum_{n \geq 1} |\alpha'(n)|^{\delta/(2+\delta)} < \infty,$$

with the same $\delta > 0$ (given in the theorem assumptions). These results, along with the stationarity assumption and $\mathrm{E}\left(|u_1(\psi_0)|^{2+\delta}\right) < \infty$, ensure that the central limit theorem 10.7 of Bradley (2007a) applies, and thus the infinite sum $k_\infty(\psi_0) = \mathrm{E}\left\{u_1(\psi_0)^2\right\} + 2\sum_{t=1}^{\infty} \mathrm{E}\left\{u_1(\psi_0)\, u_{t+1}(\psi_0)\right\}$ converges, i.e., $k_\infty(\psi_0) < \infty$, and

$$T^{-1/2} U(\psi_0) \xrightarrow{D} \mathscr{N}\left\{0, k_\infty(\psi_0)\right\}, \qquad T \to \infty,$$

where $\xrightarrow{D}$ denotes convergence in distribution. Therefore, returning to equation (5.11), and by definition of $j(\psi)$, by the law of large numbers, by Slutsky's theorem and thanks to the regularity conditions, we get

$$\begin{aligned}
T^{1/2}(\widehat{\psi}_{\mathscr{K}} - \psi_0) &\doteq T^{1/2} H(\psi_0)^{-1} U(\psi_0) \\
&= \left\{T^{-1} H(\psi_0)\right\}^{-1} \left\{T^{-1/2} U(\psi_0)\right\} \\
&\xrightarrow{D} j(\psi_0)^{-1} \mathscr{N}\left\{0, k_\infty(\psi_0)\right\}, \qquad T \to \infty \\
&\overset{D}{=} \mathscr{N}\left\{0, j(\psi_0)^{-1} k_\infty(\psi_0)\, j(\psi_0)^{-1}\right\},
\end{aligned}$$

where $\overset{D}{=}$ denotes equality in distribution. But $k_\infty(\psi_0)$ is the asymptotic variance of the score, renormalized by T. Hence, the result is proved. □

This result shows that the standard asymptotic normality result for composite likelihoods (recall §3.1.3) still holds under mild conditions for moderately temporally dependent processes. In fact, Theorem 58 could certainly be extended to weaker mixing conditions similar to the $\Delta(u_n)$ condition in (1.12), under which only high threshold exceedances are constrained. Furthermore, the asymptotic variance is of sandwich form, as is standard for misspecified models.

If the process $Z(s,t)$ were instead assumed to be Gaussian, and hence not max-stable, and if the pairwise likelihood were defined in terms of the marginal bivariate normal densities, then the moment condition of the theorem, that is $\mathrm{E}\left\{|u_1(\psi_0)|^{2+\delta}\right\} < \infty$, would be automatically satisfied for all $\delta > 0$, and thus the mixing condition would reduce to $\sum_{n \geq 1} |\alpha(n)|^{1-\epsilon} < \infty$, for some $\epsilon > 0$. Similar results were obtained by Davis & Yau (2011), who established the asymptotic normality and the strong consistency of the maximum consecutive pairwise likelihood estimator for ARMA models, under a condition on the autocorrelation function, and treated certain long-memory models.

Moreover, Davis *et al.* (2013b) extend Theorem 58 to slightly more general assumptions, and obtain similar results.

5.2.3 Variance estimation

Variance estimation for $\widehat{\psi}_{\mathcal{K}}$ was discussed in §3.1.5. However, it is more difficult in the space-time framework, owing to the complicated form of the sandwich matrices in equations (5.7) and (5.8). The log-pairwise likelihood is formed by summing the pairwise contributions for the time lags in the set $\mathcal{K}$ and across all S stations, so a single evaluation of the log-pairwise likelihood requires $\mathrm{O}(T|\mathcal{K}|S^2)$ operations, and the computation of (5.8) is yet more intensive.

The temporal dependence of the data suggests that block bootstrap or jackknife methods can be used. In our application we apply a block bootstrap, treating rainfall data from different summers as independent. We resample the summers with replacement and use replicates of $\widehat{\psi}_{\mathcal{K}}$ to estimate its variability. Fortunately, the replicates can be computed in parallel. The bootstrap was originally developed for independent data, and its consistency is discussed in Davison & Hinkley (1997, §2.6), for example. When the data are dependent, but can be decomposed into mutually independent blocks (e.g., distinct summers), the bootstrap may be applied to the latter.

5.3 Data analysis

Although the proposed methodology is very general and could be applied to many types of data, we give here an illustration to rainfall data, kindly provided by *MétéoSuisse*, the Swiss Federal Office of Meteorology and Climatology. In §5.3.1, we do some basic exploratory analysis, including fitting of marginal distributions, assessment of the stationarity assumption, and empirical measure of space-time dependence. In §5.3.2, new max-stable and asymptotically independence models for space-time extremes are fitted, the first of which includes a random set element "playing the role of clouds", and the others are Lagragian versions of classical Eulerian models. In §5.3.3, we compare the different model fits and in §5.3.4, we summarize the results and give some perspectives for future extensions.

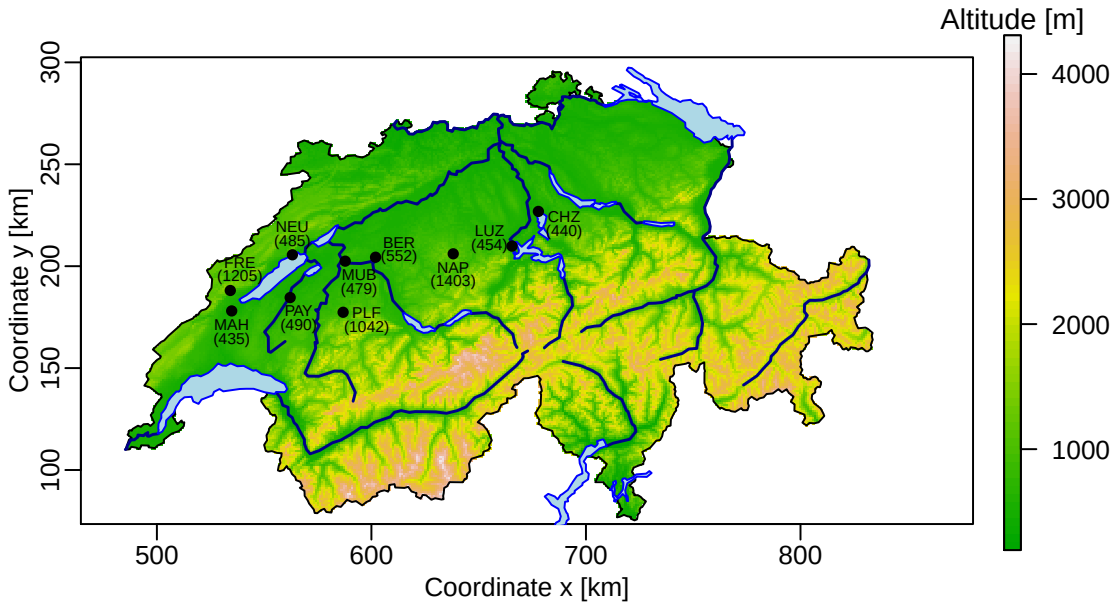

Figure 5.1: Topographic map of Switzerland, showing the location and altitude of the monitoring stations used. Their elevations are all close to 500m above mean sea level (amsl), except for three stations (FRE, NAP, PLF) at about 1000m amsl. The x and y axes use the Swiss coordinate system. The closest stations (FRE, MAH) are 10km apart and the most distant ones (CHZ, MAH) are 151km apart.

5.3.1 Exploratory analysis

5.3.1.1 Description of the dataset

The dataset used for our application is composed of hourly rainfall measurements (mm) recorded from 1981 to 2007 at ten monitoring stations in western Switzerland, and rounded to the closest tenth of millimeter. Figure 5.1 illustrates the location and topography of the area of study. All stations are located between the Alps and the Jura mountains, in the so-called *plateau* or *midland* region, and their altitudes are similar. Only the periods from midnight on June 21st to 11 pm on September 20th were considered, these summers being treated as mutually independent. The entire dataset comprises 503988 measurements, with up to 59616 data points per station.

The rainfall time series are shown in Figure 5.2. About 89% of the measurements equal zero, but they do not influence our study of extremes because we focus on exceedances over the 95%-threshold (that is, around the median of the non-zero measurements). Table 5.1 reports basic statistics about the data and reveals that the density of (strictly positive) precipitation is very asymmetric and right-skewed, with minima of 0.1mm, medians of about 0.5mm and maxima reaching about 50mm.

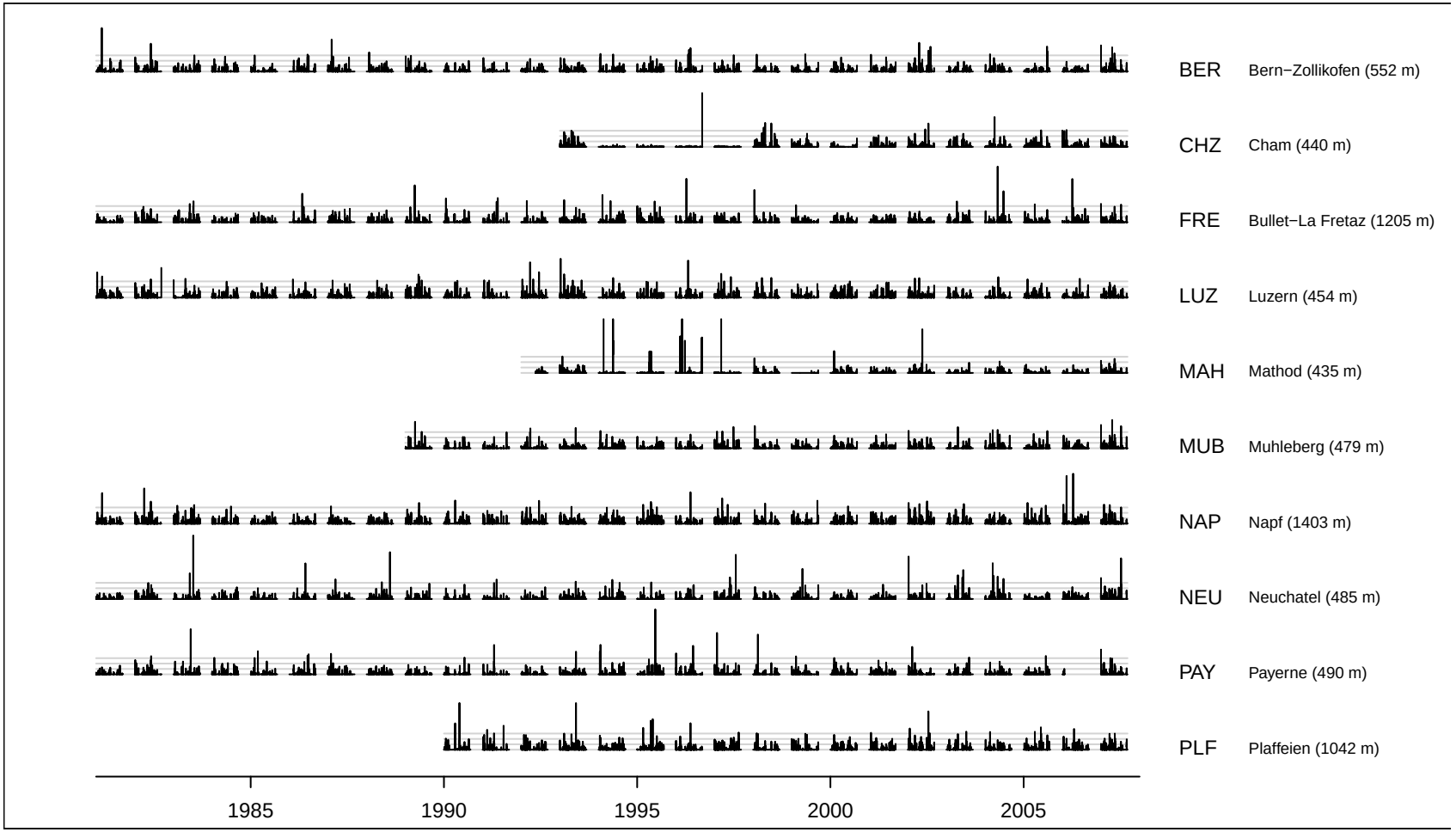

Figure 5.2: Summer hourly rainfall data (mm) at ten monitoring stations. The light grey lines show $0, 5, 10, 15$ mm. The univariate thresholds used for transformation to the unit Fréchet scale are the 95%-quantiles, ranging from 0.3 – 1.3 mm depending on the station. The gaps indicate that summers were treated as independent from one year to the next.

Table 5.1: Statistics for the rainfall time series recorded at the 10 monitoring stations. The percentage of non-zero measurements, and empirical $p-$quantiles for the non-zero data with $p = 0, 0.25, 0.5, 0.75, 0.90, 1$, are reported for each station.

Station	non-zero (%)	Min.	1st Qu.	Med.	3st Qu.	$p = 90\%$	Max.
Bern-Zollikofen	10.4	0.1	0.2	0.6	1.7	3.5	40.4
Cham	10.7	0.1	0.2	0.4	1.3	2.9	50.0
Bullet-La Frétaz	11.3	0.1	0.2	0.6	1.6	3.3	51.7
Luzern	13.3	0.1	0.2	0.6	1.6	3.6	36.1
Mathod	8.3	0.1	0.1	0.4	1.2	2.4	50.0
Muhleberg	11.2	0.1	0.2	0.5	1.4	3.1	26.3
Napf	14.6	0.1	0.2	0.7	1.9	3.8	46.4
Neuchâtel	9.4	0.1	0.2	0.6	1.6	3.3	59.0
Payerne	9.0	0.1	0.2	0.6	1.5	3.3	60.4
Plaffeien	12.0	0.1	0.2	0.7	1.9	4.0	43.4

In fact, the distribution of the data consists in a mixture between a point mass at zero (which we do not model) and a continuous density for non-zero measurements, which we model above high thresholds; see Figure 5.3.

5.3.1.2 Marginal distributions

The marginal model (5.1) was fitted separately to the rainfall time series in Figure 5.2, where the thresholds were the empirical 95% quantiles of each series, under the false assumption of independence in time. Table 5.2 reports the estimated parameters (with standard errors computed using a block bootstrap with yearly blocks), and Figure 5.3 compares, for each station, a kernel density estimation of the non-zero rainfall data and the fitted GPD density. Although there are minor discrepancies for some stations (e.g., MAH, NEU or PAY), the agreement is good overall, and the fitted curves seem to capture the tail behavior quite well. These small differences suggest that higher thresholds might be better, for example as the 70–80% empirical percentiles of the non-zero measurements. However, quantile-quantile plots displayed in Figure 5.4 show satisfactory agreement between the empirical and fitted quantiles, except for the station MAH. Furthermore, these slight mismatches could be due to the inability of the kernel density estimation to handle discretized data, in regions where the density is steep. Hence, because our main interest is the modeling of extremal dependence (see §5.3.2), rather than the marginal behavior, we assume that these marginal fits are good enough for our illustration.

Due to the size of the dataset at each station, the margins were fitted with small

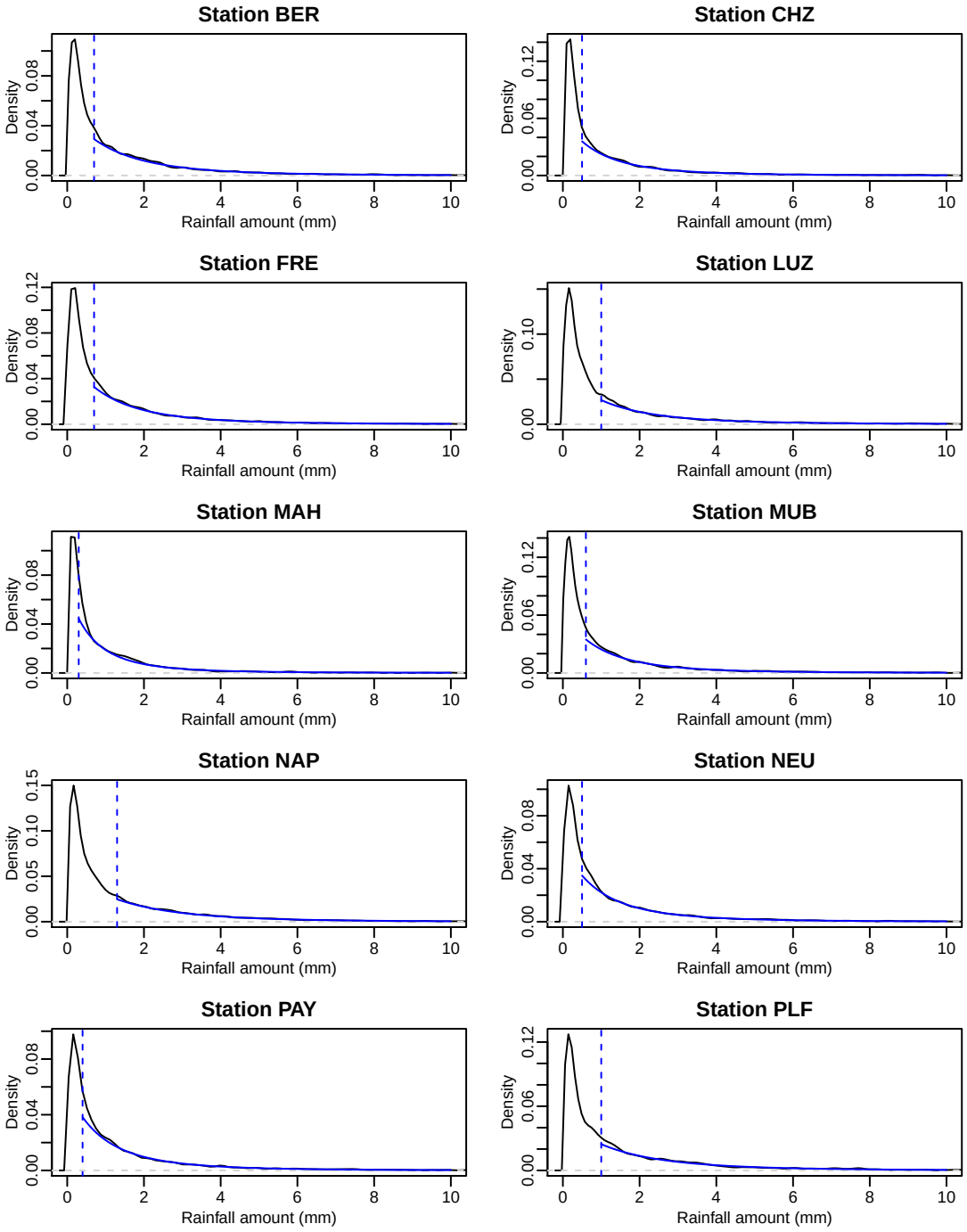

Figure 5.3: For each station: kernel density estimation of the rainfall data, scaled by the proportion of non-zero measurements (solid black line), compared to the GPD fitted to the exceedances over the empirical 0.95-quantile, scaled by 0.05 (solid blue line). The vertical dashed blue lines are the marginal thresholds used for fitting.

Table 5.2: Estimated parameters (with corresponding standard errors based on 300 block bootstrap replicates) resulting from fitting the marginal model (5.1) to each rainfall time series. The threshold exceedance probability is set to $\zeta_u = 0.05$, and the threshold u is estimated to the empirical $(1-\zeta_u)$-quantile. The quantities $\widehat{\tau}$ and $\widehat{\xi}$ are the maximum likelihood estimates of the scale and shape parameters under the false assumption of independence, $\widehat{\theta}$ is the intervals estimator (1.21) for the extremal index, $\widehat{\tau}_{\text{decl}}$ and $\widehat{\xi}_{\text{decl}}$ are the maximum likelihood estimates computed after a declustering based on $\widehat{\theta}$ and $\widehat{\tau}_{\text{MS}}$ and $\widehat{\xi}_{\text{MS}}$ are maximum pairwise likelihood estimates obtained by fitting a max-stable process to the exceedances over u.

Station	$\widehat{\tau}$	$\widehat{\xi}$	$\widehat{\tau}_{\text{decl}}$	$\widehat{\xi}_{\text{decl}}$	$\widehat{\tau}_{\text{MS}}$	$\widehat{\xi}_{\text{MS}}$
BER	1.61 (0.06)	0.22 (0.02)	2.12 (0.11)	0.23 (0.04)	1.62 (0.06)	0.28 (0.03)
CHZ	1.31 (0.18)	0.28 (0.06)	1.59 (0.30)	0.35 (0.11)	1.30 (0.18)	0.37 (0.07)
FRE	1.50 (0.05)	0.24 (0.02)	1.85 (0.09)	0.30 (0.03)	1.51 (0.05)	0.31 (0.03)
LUZ	1.74 (0.08)	0.25 (0.03)	2.27 (0.11)	0.25 (0.03)	1.76 (0.08)	0.31 (0.03)
MAH	0.99 (0.14)	0.37 (0.09)	1.07 (0.23)	0.56 (0.14)	1.00 (0.15)	0.43 (0.10)
MUB	1.39 (0.08)	0.28 (0.04)	1.90 (0.13)	0.28 (0.05)	1.39 (0.08)	0.35 (0.04)
NAP	1.96 (0.07)	0.16 (0.02)	2.43 (0.12)	0.20 (0.04)	1.98 (0.07)	0.21 (0.02)
NEU	1.36 (0.06)	0.30 (0.03)	1.65 (0.10)	0.37 (0.04)	1.36 (0.06)	0.37 (0.04)
PAY	1.30 (0.06)	0.31 (0.02)	1.61 (0.09)	0.38 (0.04)	1.29 (0.06)	0.38 (0.03)
PLF	1.95 (0.12)	0.21 (0.04)	2.47 (0.15)	0.24 (0.05)	1.96 (0.12)	0.27 (0.04)

Station	u	$\widehat{\theta}$
BER	0.7 (0.06)	0.24 (0.01)
CHZ	0.5 (0.16)	0.15 (0.03)
FRE	0.7 (0.06)	0.24 (0.01)
LUZ	1.0 (0.07)	0.28 (0.01)
MAH	0.3 (0.08)	0.16 (0.05)
MUB	0.6 (0.08)	0.24 (0.01)
NAP	1.3 (0.09)	0.27 (0.01)
NEU	0.5 (0.06)	0.24 (0.01)
PAY	0.4 (0.06)	0.24 (0.01)
PLF	1.0 (0.09)	0.25 (0.01)

variability, and all estimated shape parameters $\widehat{\xi}$ are significantly positive, and close to 0.2, which is standard in hydrology. This suggests that the upper tail has a power law decay, and that it has no finite upper bound.

Another concern is the presence of temporal dependence that may affect the estimation of marginal parameters. To assess this, we estimated the extremal index using the intervals estimator (1.21) proposed by Ferro & Segers (2003), we identified clusters of extremes, and then we fitted the marginal models based on cluster maxima. The

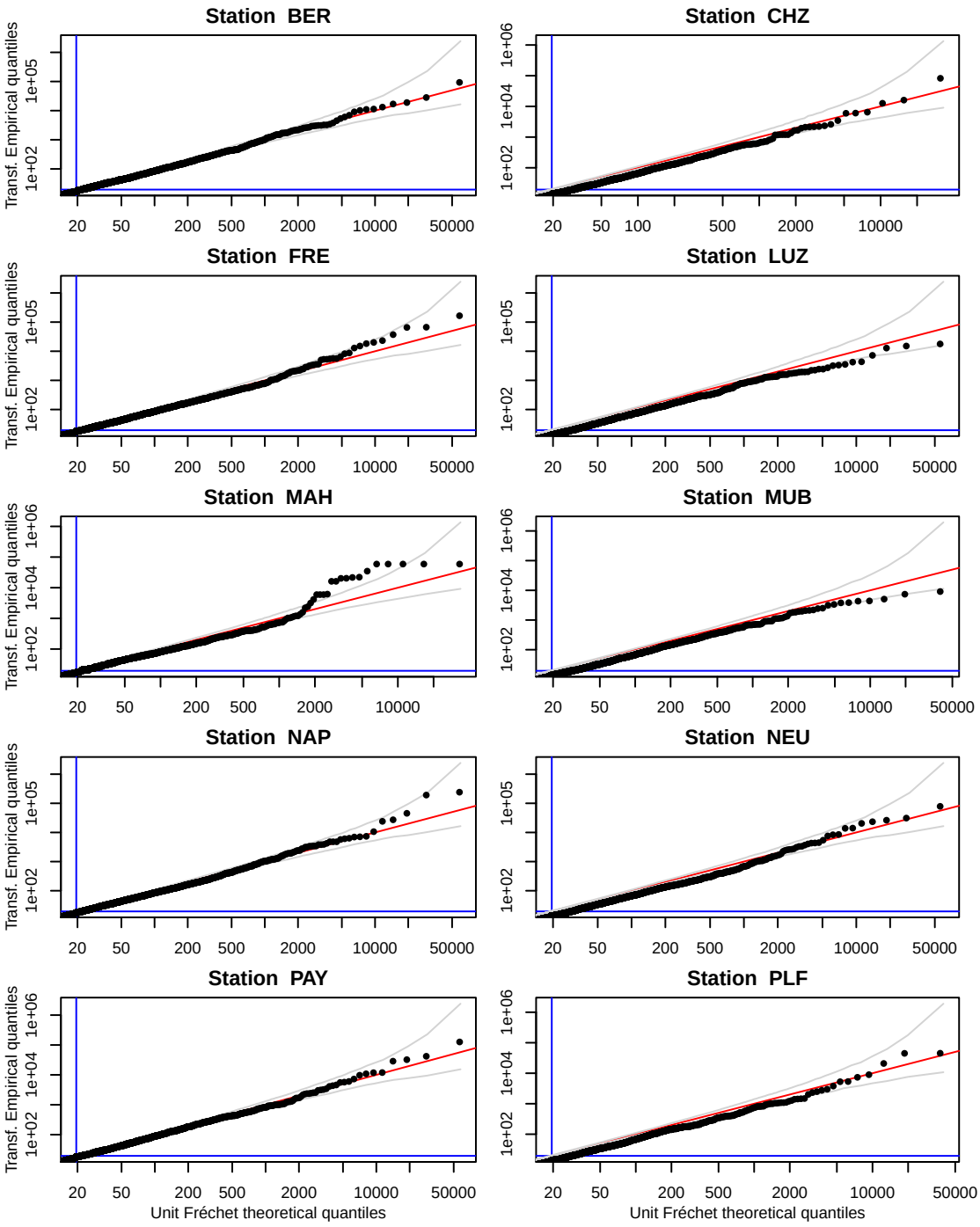

Figure 5.4: Quantile-quantile plots of each rainfall time series transformed to the unit Fréchet scale using model (5.1) and ignoring temporal dependence. The red diagonal lines indicate a perfect fit, and the grey lines are 95% confidence bands for *independent* unit Fréchet variables. The blue lines are the thresholds used for the marginal fits.

values are reported in Table 5.2. Although the value for the shape parameter ξ should not differ with or without temporal dependence (recall Proposition 18), it nevertheless changes quite substantially for some stations, owing to the smaller dataset used for its estimation. Regarding the scale parameter τ, its declustering-based estimate increases slightly, but systematically, with respect to its "naive" counterpart ignoring dependence. The uncertainty of declustering-based estimators is almost 1.5–2 larger than that of the "naive" estimators. The extremal index is estimated to 0.15–0.25 depending on the location, suggesting that extreme events tend to exceed the threshold in clusters of size about 4–6 on average; recall §1.1.2.4.

Since the marginal estimates are known to be quite sensitive to the declustering scheme, we also fitted the temporal max-stable model of §3.3.3.2 (that is, the Schlather model with beta random sets) to the exceedances over the 95%-quantile for each series, using the censored maximum pairwise likelihood estimator (5.6) with $\mathcal{K} = \{1,2,3,5,8,13,21\}$. We jointly estimated the GPD parameters and the underlying correlation parameter, fixing the random set parameters to sensible values. The results are reported in Table 5.2. Surprisingly, the estimates for τ are very close to the naive estimates (ignoring dependence). As for ξ, the estimates are generally closer to the corresponding declustering-based estimates.

From the results in Table 5.2, it is difficult to tell whether temporal dependence should be accounted for in order to transform our data to the unit Fréchet scale, but since it barely influences the subsequent fit of extremal dependence models (recall the results of §3.3.3.2), we choose to trust the naive approach. In principle, it would be better to consider a full space-time model, linking marginal and dependence models together, but since the dependence models proposed in §5.3.2 are already tricky to fit on their own, we prefer to use, for simplicity, the two-step estimating procedure advocated in §3.3.3.2, whose performance was found to be comparable to the one-step approach. We therefore independently transformed the rainfall series to the unit Fréchet scale, following §5.1.1, and using the naive estimates $\widehat{\tau}$ and $\widehat{\xi}$. In §5.3.2, we focus on the modeling of extremal dependence, rather than on the marginal behavior, and we fit spatio-temporal models to the transformed data, assuming known marginals.

5.3.1.3 Stationarity

In order to assess the assumption of temporal stationarity, we re-estimated the marginal parameters by fitting model (5.1) separately to each summer, using the same thresholds as above. The results are displayed in Figure 5.5. The confidence intervals computed from the yearly data almost always contain the parameter estimates based on the full dataset (with the 27 summers). When the uncertainty is taken into

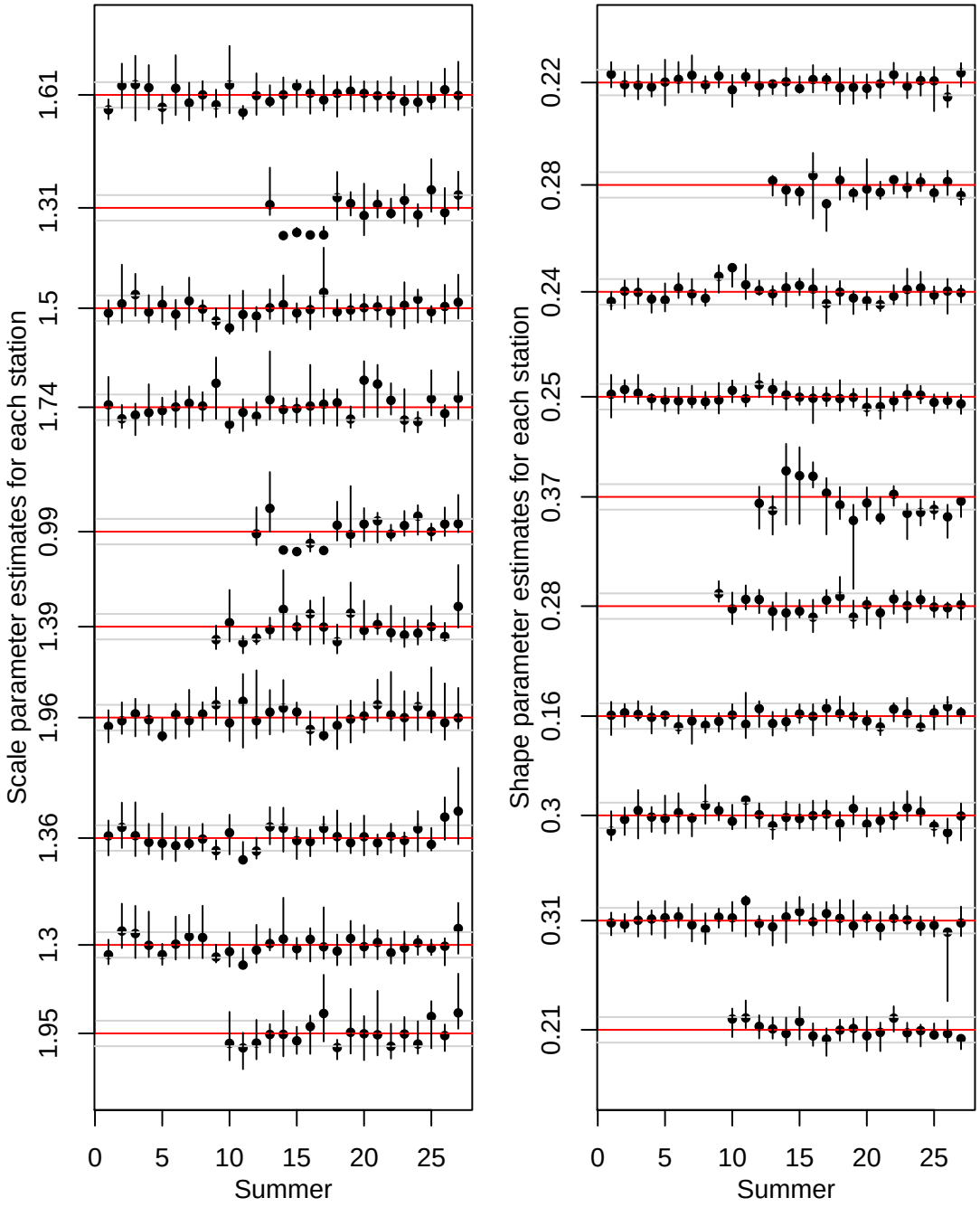

Figure 5.5: Scale (left) and shape (Right) parameters estimated by fitting model (5.1) separately to each summer for the time series in Figure 5.2 (top: BER to bottom: PLF). The vertical segments are 95% confidence intervals computed from a block bootstrap with weekly blocks. The horizontal lines correspond to estimates obtained from the full fit (red) $\pm\Delta$ (grey), with $\Delta = 0.5$ for the scale, and $\Delta = 0.3$ for the shape.

account, the parameter estimates seem to be quite stable in time, suggesting that the stationarity assumption above the 95%-quantile is plausible.

Regarding spatial stationarity, as the monitoring stations are located in the same geographic region and at similar altitudes (see Figure 5.1), the local climate is likely to be very similar. The basic statistics in Table 5.1 confirm this. Furthermore, the marginal GPD parameter estimates in Table 5.2 are also quite homogenous, except for the station MAH, for which the data have a strange behavior around years 1994–1997 (see Figure 5.2). Hence, this suggests that, globally, our dataset is more or less stationary over space. It would have been interesting to consider a spatial model with, for example, one shape parameter for the whole region of study and site-wise scale parameters (possibly depending on covariates such as elevation), but this goes beyond the scope of the present data analysis.

5.3.1.4 Spatio-temporal dependence

Figure 5.6 shows empirical space-time pairwise extremal coefficients for a subset of 5 stations at different time lags, based on a censored version of the naive Schlather–Tawn estimator (1.66); see Figure E.1 to see this for all stations. There is evidence of significant spatial and temporal dependence between the different series. Panel $(1,1)$ shows the temporal extremal coefficients at Bern-Zollikofen (BER); it starts with the value 1 (complete dependence at lag 0), and tends smoothly to the value 2 (independence) as the time lag increases. This pattern repeats itself for the other stations. The off-diagonal panels represent extremal coefficients for the different pairs of stations, and hence display space-time interactions. For example, Panel $(1,4)$, in the 1st row and 4th column, displays the extremal coefficients between the rainfall time series at Luzern (LUZ) at time t and the rainfall time series at Bern-Zollikofen (BER) at time $t+h_t$, for $h_t = 0,1,\ldots,24$. Panel $(4,1)$ reverses the roles of the stations. The extremal coefficient functions differ for the panels, showing that the orientation of the stations matters. The extremal coefficient decreases at lags 1 or 2 when the stations are west-east oriented: during the summer months, western Switzerland is governed by dominant winds from the west or north-west, so that the clouds tend to discharge their rain first in the west. The same rainfall event could therefore be recorded by two distant monitoring stations at a lag of 1 or 2 hours, depending on their location and on the wind velocity. Consequently, extremal dependence might be higher at lag 1 or 2 than at lag 0. A good model for the data should be able to capture such features.

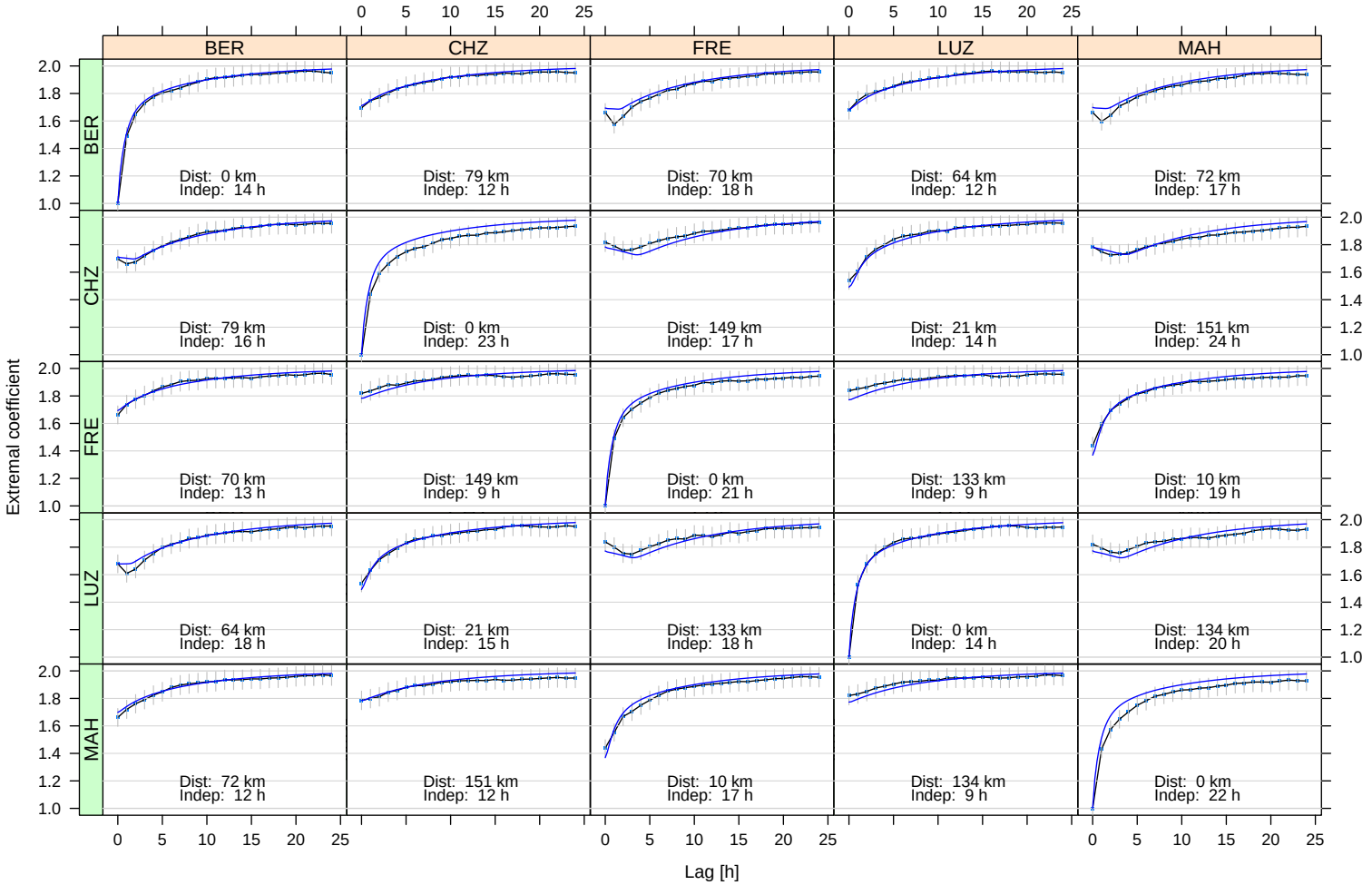

Figure 5.6: Empirical and model-based pairwise extremal coefficients (2.19), $\theta_2(h_t, h_s)$, for five stations. The black lines join the empirical extremal coefficients found using the censored Schlather–Tawn estimator (1.66) at the 0.95-quantile threshold, the vertical grey segments being 95% confidence intervals (assuming independence in time). The blue lines correspond to the extremal coefficient curves derived from the fitted model in §5.3.2.1. The panel at the rth row and cth column shows the extremal coefficients between Z_t^{c} and $Z_{t+h_t}^{\mathrm{r}}$, for $h_t = 0, 1, 2, \ldots, 24$. "Dist" stands for the distance between stations, and "Indep" is the time needed to reach independence (the first lag h_t for which the value $\theta_2(h_t, h_s) = 2$ lies within the confidence interval).

5.3.2 Modeling of extremal dependence

We now tackle space-time modeling of extremal dependence based on max-stable (5.3) and asymptotic independence (5.4) models. In §5.3.2.1, we first discuss the construction and fitting of a max-stable model based on (2.30) for the rainfall data transformed to the unit Fréchet scale, and then in §5.3.2.2, we consider alternative max-stable models. In §5.3.2.3, we fit inverted max-stable models in order to assess asymptotic independence, and in §5.3.3, we compare all these models using the CLIC* (recall §3.1.4) and graphical diagnostics.

5.3.2.1 Schlather model with random set

Model construction. The Schlather model (2.28) is a non-ergodic max-stable process, and is therefore unsuitable for the modeling of our rainfall data, which are independent at relatively short space-time lags. By contrast, its extension with a random set element (2.30) can capture independence, and may be useful to this end.

Letting $x = (s,t) \in \mathscr{X} = \mathscr{S} \times \mathscr{T} \subset \mathbb{R}^2_+$ denote a generic location in space and time, the Schlather model with random set is constructed from (5.2) with

$$W_i(s,t) \propto \max\{0, \varepsilon_i(s,t)\} I_{\mathscr{A}_i}\{(s,t) - X_i\}, \qquad (s,t) \in \mathscr{S} \times \mathscr{T}, \tag{5.12}$$

where $\varepsilon_i(s,t)$ are independent replicates of a Gaussian random field with space-time correlation function $\rho(h_s, h_t)$, $I_{\mathscr{A}}(\cdot)$ is the indicator function of a compact random set $\mathscr{A} \subset \mathscr{S} \times \mathscr{T}$, the $\mathscr{A}_i$ are independent replicates of $\mathscr{A}$, and the X_i are points of a unit rate Poisson process on $\mathscr{S} \times \mathscr{T}$, independent of the ε_i.

The Gaussian random field is supposed to model the short-range behavior of the process within single storms, so it is important to have a correlation function that can flexibly capture space-time interactions. As a simple but fairly flexible possibility, we used the nonseparable space-time correlation function (2.10) from Gneiting (2002),

$$\rho(h_s, h_t) = \frac{1}{\left\{1 + \left(\frac{|h_t|}{\lambda_t}\right)^{\alpha_t}\right\}^{1+d\varrho/2}} \exp\left[-\frac{\left(\frac{\|h_s\|}{\lambda_s}\right)^{\alpha_s}}{\left\{1 + \left(\frac{|h_t|}{\lambda_t}\right)^{\alpha_t}\right\}^{\varrho\alpha_s/2}}\right], \tag{5.13}$$

where h_s and h_t are lags in space and time, $\lambda_s, \lambda_t > 0$ are spatial and temporal scale parameters, $\alpha_s, \alpha_t \in (0,2]$ are spatial and temporal shape parameters, $d = 2$ is the spatial dimension, and $\varrho \in [0,1]$ is a separability parameter. Davis *et al.* (2013a) show that this class of covariance functions satisfies a natural smoothness property at the origin, directly linked to the smoothness of the random field, and is therefore suitable

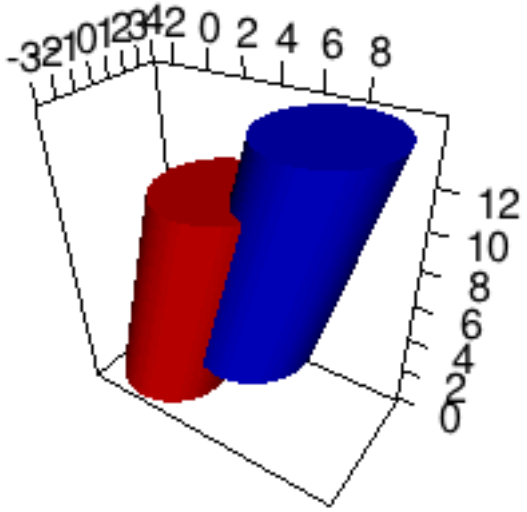

Figure 5.7: Illustration of the random set element $\mathscr{A}$ in space $\mathscr{S}$ (horizontal plane) and time $\mathscr{T}$ (vertical axis). The storms are conceptualized as random disks with a random radius moving at a random velocity for a random duration. The red tilted cylinder represents a realization $\mathscr{A}$ of such a storm in $\mathscr{S} \times \mathscr{T}$, and the blue one is $\mathscr{A} + (h_s, h_t)$, for a given spatio-temporal lag vector (h_s, h_t). The coefficient $\delta(h_s, h_t)$ needed for the fitting is the expected volume of intersection between the two cylinders.

for the modeling of physical processes such as rainfall.

The random set $\mathscr{A}$ is interpreted as a random storm having a finite extent, which enables the model to capture complete independence. Conceptualizing storms as disks of random radius R moving at a random velocity V during a random lifetime L starting from a random position, the storm extent $\mathscr{A}$ in space and time becomes a tilted cylinder in $\mathscr{S} \times \mathscr{T}$, with a truncated Gaussian process inside; see Figure 5.7. For tractability we assume that $R \sim \text{Gamma}(m_R/k_R, k_R)$ (with mean m_R km), $V \sim \mathscr{N}_2(m_V, \Omega)$ (km/hour) and $L \sim \text{Gamma}(m_L/k_L, k_L)$ (with mean m_L hour). Furthermore, we parametrize the mean velocity as the vector $m_V = \{\|V\|\cos(\nu), \|V\|\sin(\nu)\}^T$, where $\nu \in (-\pi, \pi]$ is the angle of the main winds with respect to West-East direction, and the covariance of V as $\Omega = \|V\|^2\omega^2 I_2$, where I_2 is the 2×2 identity matrix. The factor ω is a dispersion parameter.

Other max-stable models are also considered in §5.3.2.2, but are outperformed by the model described above; see Section 5.3.3.

Model fitting. In order to fit the space-time model described above to the exceedances over the 95%-quantile, we maximized the censored pairwise likelihood (5.5), which requires the specification of bivariate contributions only. Following (5.3) and §B.1.2, the bivariate distribution for our model may be expressed as $\exp\{-V_{\mathscr{D}}(z_1, z_2)\}$ with exponent measure

$$V_{\mathscr{D}}(z_1, z_2) = \left(\frac{1}{z_1} + \frac{1}{z_2}\right)\left\{1 - \frac{\delta(h_s, h_t)}{2}\left(1 - \left[1 - 2\frac{\{\rho(h_s, h_t) + 1\} z_1 z_2}{(z_1 + z_2)^2}\right]^{1/2}\right)\right\}, \qquad (5.14)$$

where $\delta(h_s, h_t) = \mathrm{E}[|\mathscr{A} \cap \{(h_s, h_t) + \mathscr{A}\}|]/\mathrm{E}(|\mathscr{A}|)$ is the normalized expected volume of overlap between the random set $\mathscr{A}$ and itself shifted by the space-time lag (h_s, h_t).

The computation of the coefficient $\delta(h_s, h_t)$ for any $(h_s, h_t) \in \mathscr{K}$, is not trivial and may be quite intensive. Several mild approximations, some analytical calculations and a single one-dimensional finite integration yield a good approximation to $\delta(h_s, h_t)$, which is then used in computing the pairwise likelihood; see Appendix F.

After some exploratory analysis, we fixed $k_L = 9$ and $k_R = 0.4$, since these parameters are difficult to estimate; the model then has five parameters for the correlation function, and five for the random set. Due to the complexity of the problem, we split the estimation procedure into two steps: we first estimate $\lambda_s, \lambda_t, \varrho, m_R, \|V\|, m_L$, with the other parameters held fixed, and then all ten parameters together, with the former estimates as starting values. We always use the pairwise likelihood estimator (5.6). Confidence intervals are calculated by the block bootstrap described in Section 5.2.3 using yearly blocks. Based on the results in Chapter 3, in particular Section 3.3.3.2, we include the pairs at lags in $\mathscr{K} = \{0, 1, 2, 3, 5, 8, 13, 21\}$ in the pairwise log likelihood, a single evaluation of which involves contributions for about $T|\mathscr{K}|S^2 = 50000 \times 8 \times 10^2 = 40$ million pairs—the full pairwise likelihood would have 7 billion pairs, completely impractical for inference purposes! We coded the pairwise likelihood in C, parallelized the work on 8 CPUs, and fitted the model using the Nelder–Mead optimization routine in R. Even though there is a large amount of data and our model is very complex, a full fit took only about 10 minutes. Uncertainty assessment was based on 300 bootstrap replicates. The results are presented in Table 5.3.

The estimated mean speed of the dominant winds is 39.6km/hr and the estimated angle is about 14° in the Argand diagram, which seem reasonable when compared to radar images of precipitation for the same region and time of year. This means that the clouds are likely to move in a rough east-north-easterly direction, in agreement with the summer climate in Western Switzerland. However, as the estimated angle coincides more or less with the main orientation of our monitoring stations and as the information along the perpendicular axis is likely to be small, one should interpret

Table 5.3: Parameter estimates and 95%-confidence intervals from fitting our random set model to the rainfall data. Uncertainty assessment was based on 300 block bootstrap replicates, using yearly blocks.

				Estimate	Conf. interval
Correlation	Scale	Space	$\widehat{\lambda}_s$ (km)	37.9	(31.7,43.1)
		Time	$\widehat{\lambda}_t$ (hr)	2.2	(1.6,2.4)
	Shape	Space	$\widehat{\alpha}_s$	0.93	(0.89,1.09)
		Time	$\widehat{\alpha}_t$	1.45	(1.33,1.65)
	Separability		$\widehat{\varrho}$	1.00	(0.43,1.00)
Random Set	Lifetime	Mean	$\widehat{m}_L$ (hr)	111	(99,120)
		Shape	k_L	9	(—)
	Radius	Mean	$\widehat{m}_R$ (km)	68	(61,80)
		Shape	k_R	0.4	(—)
	Velocity	Absolute speed	$\|\widehat{V}\|$ (km/hr)	39.6	(37.2,46.3)
		Angle	$\widehat{\nu}$ (rad)	0.24	(0.09,0.29)
		Dispersion	$\widehat{\omega}$ (km/hr)	0.12	(0.08,0.12)

it with care.

The mean duration and mean radius of a storm are estimated as 111hr and 68km. Given the shape parameters, one half of the clouds have durations of over 107hr and a radius of over 25km. These estimates seem rather large. However, as shown in Figure 5.8, the pairwise likelihood for the duration parameter is almost flat in its right tail, and that for the radius parameter is somewhat asymmetric. With a speed of about 40km/hr on average, a cloud moves across the region of study in less than 4hr, so it is hard to estimate these parameters based on the available observations. Data collected at a larger number of monitoring stations in a wider region of study would give more reliable conclusions. Hence the bootstrap confidence intervals for m_L, $(99,120)$, and for m_R, $(61,80)$, seem optimistically narrow.

The correlation parameters appear to be better estimated. In particular, the separability parameter ϱ reaches its upper bound and its 95% confidence interval does not include zero, suggesting that the data are highly nonseparable and that ϱ tries to capture this. The fitted correlation function, displayed in the bottom right panel of Figure 2.2, appears plausible.

Model checking. Figure 5.6 compares empirical estimates of the pairwise extremal coefficients with their model-based counterparts for a subset of 5 representative stations (see also Figure E.1). There is a good agreement overall, but the fitted model often provides less extremal dependence at lag 1 than is present in the data. This

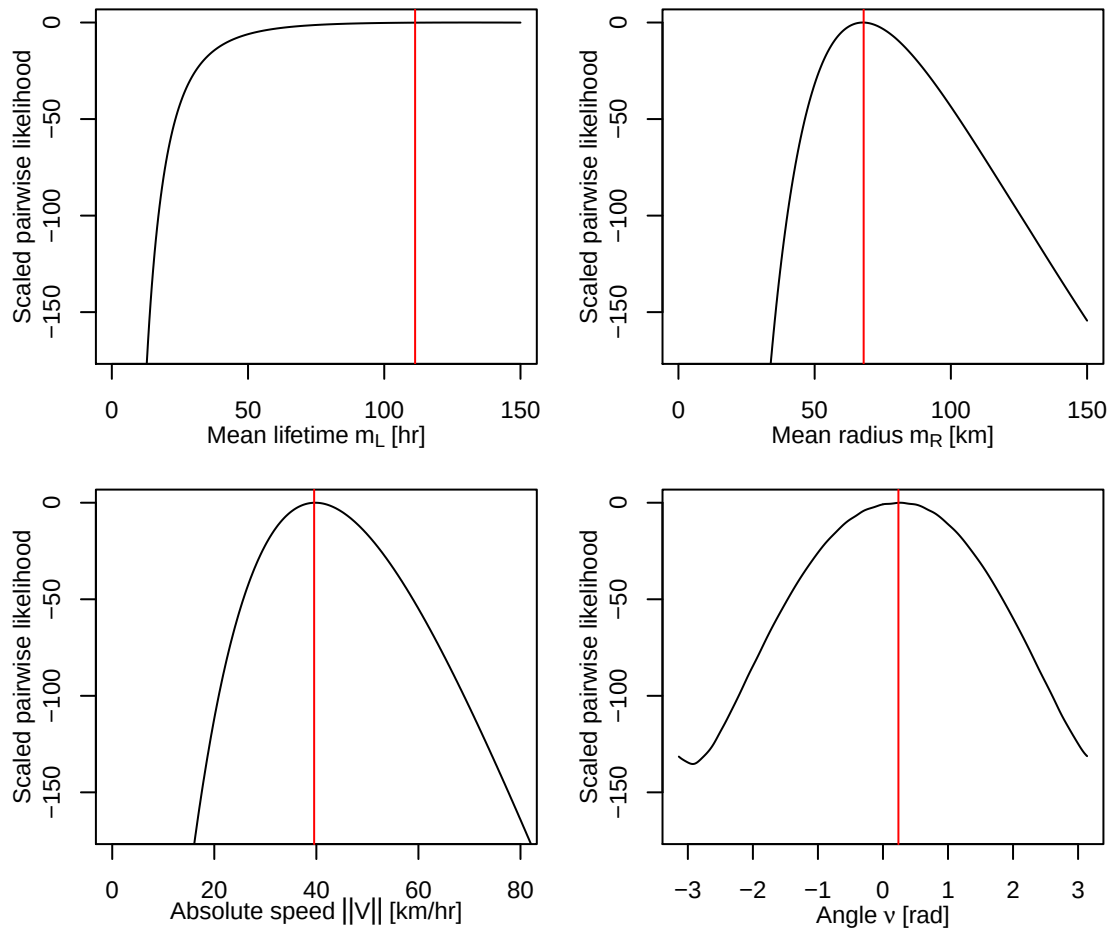

Figure 5.8: Slice pairwise likelihoods around the maximum pairwise likelihood estimate $\widehat{\psi}_{\mathcal{K}}$, shifted to have maximum at zero, and scaled to be comparable to maximum likelihood under independence. *Top left*: Slice with respect to the mean lifetime m_L; *top right*: Slice with respect to the mean radius m_R; *bottom left*: Slice with respect to the absolute speed $\|V\|$; *bottom right*: Slice with respect to the angle of the main winds ν.

lack of fit at short time lags might be explained either by a lack of flexibility due to the (conceptually) simplistic model that we used or by optimization difficulties. The diagonal plots, showing the marginal temporal dependence of the extremes, show a good fit. The small differences at Cham (CHZ) or Mathod (MAH) may be due to nonstationarity or because data at those monitoring stations seem unreliable; see Figure 5.2. The left top panel of Figure 5.9 shows pairwise extremal coefficients $\theta_2(h_s, h_t)$ in (5.3) for all pairs of stations and $h_t \in \mathcal{K}$.

As the model was fitted using pairs of observations, one might wonder whether it can capture higher-order interactions. We therefore computed the trivariate extremal coef-

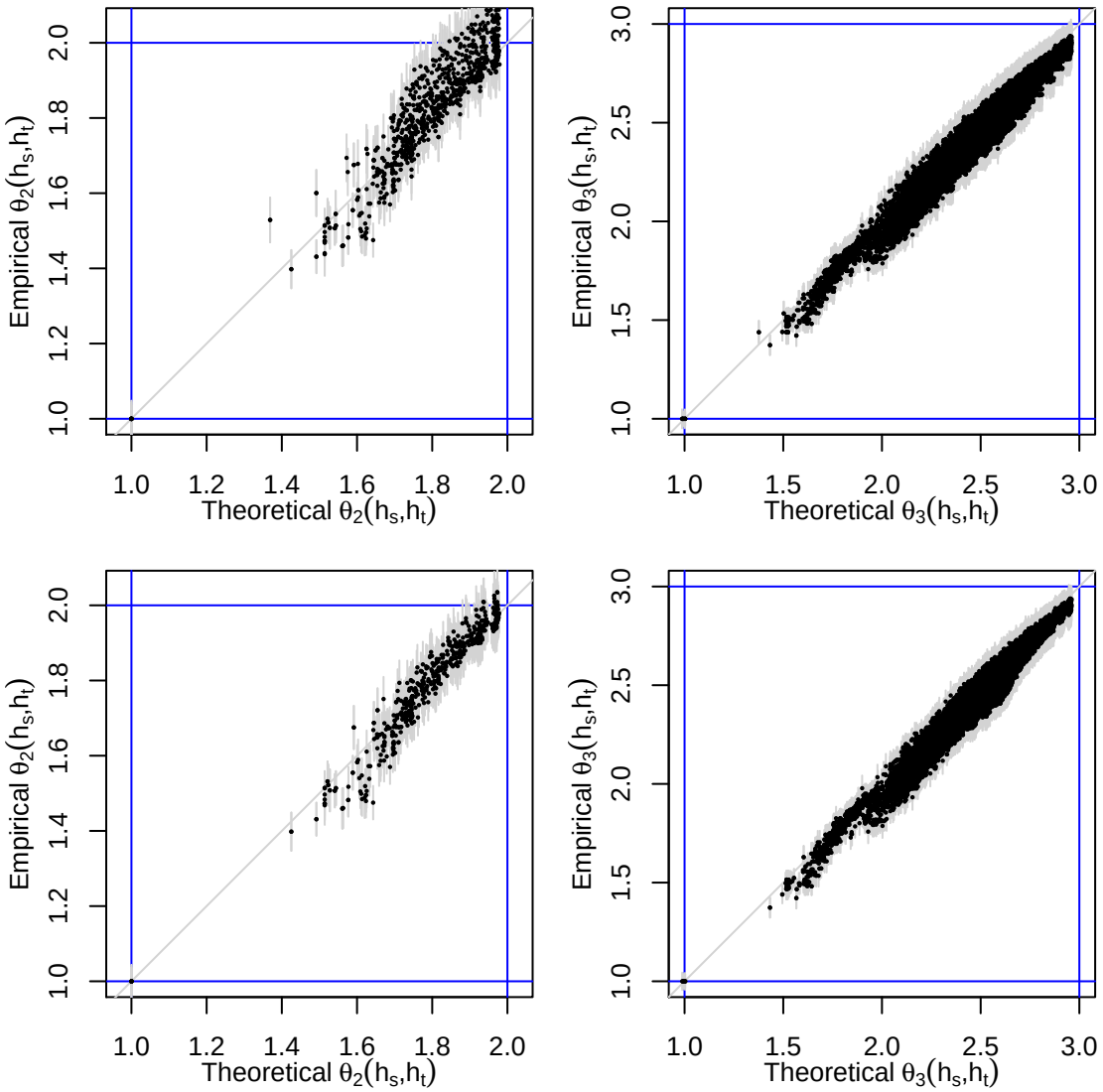

Figure 5.9: *Top row*: Comparison of empirical estimates of pairwise (left column) and triplewise (right column) extremal coefficients at the fitted lags for the rainfall data with their model-based counterparts. The light-grey vertical lines are 95% confidence intervals. A perfect agreement would place all points on the grey diagonal line. *Bottom row*: Same diagnostic plots without the points involving the stations at Cham (CHZ) and Mathod (MAH).

ficients $\theta_3(\boldsymbol{h}_s, \boldsymbol{h}_t)$ (see Appendix G) and found good agreement between nonparametric estimates of trivariate extremal coefficients and their model-based counterparts; see the top right panel of Figure 5.9. It seems that the trivariate interactions are fairly well modeled, though there is strong dependence among their estimates. The biggest discrepancies are from stations CHZ (Cham) and MAH (Mathod), but without these

stations the points lie quite close to the diagonal; see the bottom panels of Figure 5.9.

In order to assess the sensitivity of the results to initial conditions, we re-fitted the model with different starting values. The results were sometimes fairly different, but with similar bivariate properties and with almost the same pairwise likelihood. Consequently, we believe that some parameters are likely to play a similar role, giving rise to identifiability issues.

Simulation of the fitted model. Simulation of the max-stable model (5.12) in space and time is very intensive because it requires the generation of a large number of Gaussian random fields at many space-time locations; recall §2.3. For example, if it is needed to simulate the fitted random field at an hourly scale at the 10 monitoring stations during a single summer (that is, 92 days), the number of space-time locations equals $N = 22080$. And this number would be even much bigger if we had to consider a finer spatial grid.

Exact simulation of Gaussian random fields at N sites relies on the inversion of an $N \times N$ covariance matrix, which is usually performed using a Cholesky decomposition and the inversion of the resulting triangular matrix. But these operations are currently too time- and memory-consuming to be applied in such a high-dimensional problem. To circumvent this issue, approximate simulation methods for Gaussian processes may be used, an example of which is the so-called *turning bands* method (Matheron, 1973; Schlather, 1999). If the random set is stochastically small compared to the simulation region, another possibility is to simulate first the random sets, and then the Gaussian random fields *within* the realized sets. By contrast to the classical method, this approach involves the inversion of multiple covariance matrices of lower dimensions.

As an illustration, Figures H.1–H.10 in the appendix show one simulation of the fitted max-stable model (with unit Fréchet margins) on a 50×25 grid covering our region of study during 10 hours, based on the exact simulation method for Gaussian processes.

5.3.2.2 Alternative max-stable models

We also fitted models based on the Brown–Resnick process (2.32) proposed by Brown & Resnick (1977); Kabluchko *et al.* (2009), which, unlike the Schlather process, can capture full independence without a random set component. This is a stationary max-stable process that may be represented in the space-time framework as in equation (5.2), where the $W_i(s,t)$ are independent replicates of the random process $W(s,t) = \exp\{\varepsilon(s,t) - \gamma(s,t)\}$ and $\varepsilon(s,t)$ is an intrinsically stationary Gaussian random field with space-time semi-variogram $\gamma(h_s, h_t)$, with $\varepsilon(0,0) = 0$ almost surely. The bivariate

exponent measure for this model is

$$V_{\mathscr{D}}(z_1,z_2)=\frac{1}{z_1}\Phi\left\{\frac{a}{2}-\frac{1}{a}\log\left(\frac{z_1}{z_2}\right)\right\}+\frac{1}{z_2}\Phi\left\{\frac{a}{2}-\frac{1}{a}\log\left(\frac{z_2}{z_1}\right)\right\},$$

where $a=\left\{2\gamma(h_s,h_t)\right\}^{1/2}$, $h_s=s_2-s_1$ is the spatial lag, $h_t=t_2-t_1$ is the temporal lag, and where $\Phi(x)$ is the standard normal cumulative distribution function. Inference can be made similarly to the model (5.12), using the threshold-based censored pairwise likelihood estimator (5.6).

Four different space-time semi-variograms were considered:

(i) Model 1: $\gamma(h_s,h_t)=\sigma^2\left\{1-\rho(h_s,h_t)\right\}$, where the correlation function $\rho(h_s,h_t)$ is defined in (5.13), and σ^2 determines the global "amount" of extremal dependence. There are 6 parameters: $\psi=(\sigma^2,\lambda_s,\lambda_t,\alpha_s,\alpha_t,\rho)$;

(ii) Model 2: $\gamma(h_s,h_t)=(\|h_s-h_tV\|/\lambda)^{\alpha}$, where $V=(\|V\|\cos(v),\|V\|\sin(v))^T$ is the wind velocity treated as constant by contrast to model (5.12), with $\|V\|$ being the wind speed and $v\in(-\pi,\pi]$ its direction, $\lambda>0$ is a range parameter and $\alpha\in(0,2]$ is a smoothness parameter. This model has 4 parameters: $\psi=(\lambda,\alpha,\|V\|,v)$;

(iii) Model 3: $\gamma(h_s,h_t)=\left(h^T\Sigma^{-1}h\right)^{\alpha/2}$, where $h=h_s-h_tV$, and Σ is a 2×2 covariance matrix, parametrized in terms of its correlation $\rho=\Sigma_{12}\Sigma_{11}^{-1/2}\Sigma_{22}^{-1/2}$. The standard deviations $\Sigma_{11}^{1/2},\Sigma_{22}^{1/2}$ are range parameters, and ρ is an anisotropy parameter. There are 6 parameters: $\psi=(\Sigma_{11}^{1/2},\Sigma_{22}^{1/2},\rho,\alpha,\|V\|,v)$;

(iv) Model 4: $\gamma(h_s,h_t)=\left(\|h_s-h_tV\|^2/\lambda_s^2+|h_t|^2/\lambda_t^2\right)^{\alpha/2}$, where $\lambda_1,\lambda_2>0$ are range parameters capturing spatial and temporal dependence decays. This model has 5 parameters: $\psi=(\lambda_s,\lambda_t,\alpha,\|V\|,v)$.

Model 1 is the counterpart of model (5.12) without the random set element, whereas Models 2–3 are based on fractional Brownian motions satisfying the Taylor hypothesis, i.e., $\gamma(0,h_t)=\gamma(h_tV,0)$. The latter two models are constructed by considering Lagrangian versions of motion-invariant spatial variograms. Finally, Model 4 is similar to Model 2 with an additional temporal component, in order to better capture the different spatial and temporal behaviors. This model may be also be expressed as Model 3, with $h=(h_s-h_tV,h_t)\in\mathbb{R}^3$ and Σ being a 3×3 block diagonal matrix.

The estimated parameters with 95%-confidence intervals are reported in Table 5.4, and Figure 5.10 displays empirical and fitted space-time extremal coefficients for five representative stations (see also Figure E.1).

Table 5.4: Parameter estimates and 95%-confidence intervals from fitting the alternative Brown-Resnick models to the rainfall data. Uncertainty assessment was based on 300 block bootstrap replicates, using yearly blocks. The difference in CLIC* with respect to model (5.12) is also reported.

Model	Estimated dependence parameters $\widehat{\psi}_{\mathcal{K}}$			CLIC*
1	$\widehat{\sigma}^2 = 10.0\ (8.7, 13.1)$	$\widehat{\lambda}_s = 1297\ (708, 3395)$	$\widehat{\lambda}_t = 20.6\ (14.1, 36.7)$	800
	$\widehat{\varrho} = 0.89\ (0.43, 1.00)$	$\widehat{\alpha}_s = 0.54\ (0.47, 0.65)$	$\widehat{\alpha}_t = 0.98\ (0.89, 1.06)$	
2	$\widehat{\lambda} = 14.9\ (12.9, 17.9)$	$\widehat{\alpha} = 0.57\ (0.54, 0.60)$		587
	$\|\widehat{V}\| = 34.6\ (29.8, 41.9)$	$\widehat{\nu} = -0.63\ (-1.03, -0.59)$		
3	$\widehat{\Sigma}_{11}^{1/2} = 24.8\ (16.6, 26.7)$	$\widehat{\Sigma}_{22}^{1/2} = 41.8\ (28.4, 43.6)$	$\widehat{\rho} = 0.86\ (0.69, 0.89)$	821
	$\widehat{\alpha} = 0.58\ (0.56, 0.60)$	$\|\widehat{V}\| = 68.9\ (47.5, 74.5)$	$\widehat{\nu} = -1.80\ (-1.86, -1.56)$	
4	$\widehat{\lambda}_s = 16.4\ (14.8, 18.8)$	$\widehat{\lambda}_t = 1.02\ (0.83, 1.61)$	$\widehat{\alpha} = 0.60\ (0.58, 0.64)$	599
	$\|\widehat{V}\| = 20.2\ (17.5, 22.6)$	$\widehat{\nu} = 0.11\ (0.00, 0.26)$		

The fits appear to be reasonable overall, except for Model 1, which sometimes completely mismatches the pairwise diagnostics at short time lags, because of its intrinsic spatial and temporal isotropy. The estimated smoothness parameters are comparable and usually quite close to 0.6, meaning that the realized random fields are very rough. The estimated wind speed varies between about 20km/hr and 69km/hr, which seems odd a priori, but this parameter does not have the same interpretation for the different models.

Since Models 2–4 seem to have similar performances, but Model 2 has fewer parameters, the latter is probably the best max-stable alternative proposed to model (5.12). However, all these models are by far outperformed by the random set model (5.12); see the details in §5.3.3. This may be due to the fact that Models 2–4 treat the wind velocity as fixed, whereas model (5.12) puts a rather flexible distribution on it, or because Models 2–4 assume a single smoothness parameter for space and time, whereas model (5.12) has two distinct parameters.

5.3.2.3 Asymptotic independence models

As an extension of the present work, we also investigated whether asymptotic independence models provide a better fit. To assess asymptotic independence, we considered the inverted max-stable models constructed from the random set model (5.12), and from Models 1–4. Their survival distribution is (5.4), where the underlying exponent measure corresponds to that of the corresponding max-stable model.

The estimated parameters with 95%-confidence intervals are reported in Table 5.5, and Figure 5.11 displays empirical and fitted space-time coefficients of tail dependence

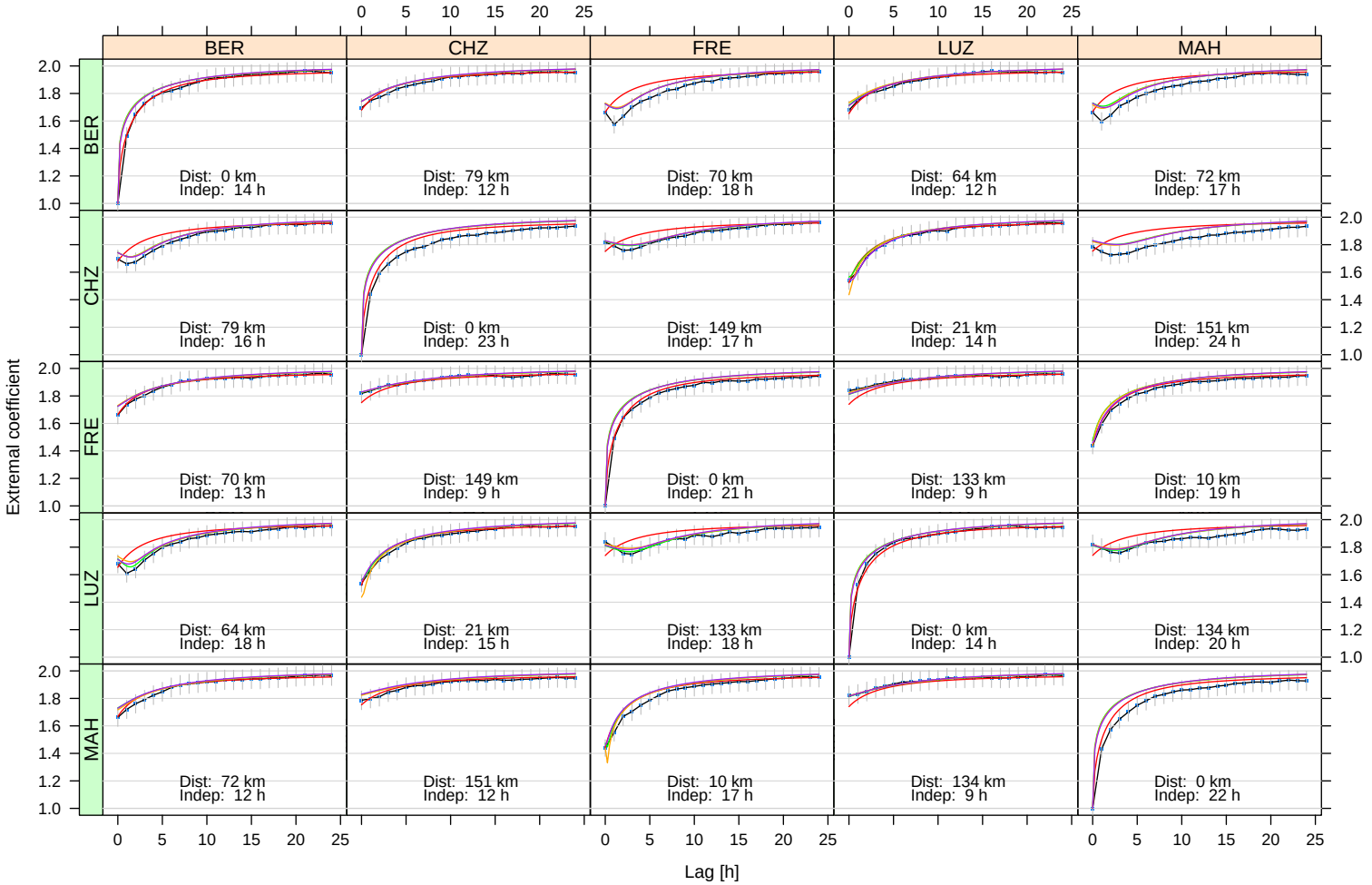

Figure 5.10: Empirical and model-based pairwise extremal coefficients (2.19), $\theta_2(h_t, h_s)$, for five stations. The black lines join the empirical extremal coefficients found using the censored Schlather–Tawn estimator (1.66) at the 0.95-quantile threshold, the vertical grey segments being 95% confidence intervals (assuming independence in time). The solid curves are the fitted extremal coefficients for Models 1 (red), 2 (green), 3 (orange) and 4 (purple). For the rest, see the caption of Figure 5.6.

for five representative stations. Figure E.2 shows this for all stations. Although the fits appear to be better than for max-stable processes (see §5.3.3 for more details), the parameters of inverted max-stable models considered here are much harder to interpret. Moreover, even though the pairwise diagnostics for the inverted model based on (5.12) seem quite reasonable, its estimated parameters are suspicious, and it is difficult to check its validity in higher dimensions.

Furthermore, there is a clear trade-off between having a small range parameter in a max-stable process, and a large one in an asymptotic independence process. In particular, for the inverted model based on (5.12), the range parameters are of order 10^4 for λ_s and 10^{10} for λ_t, and it is very likely that the pairwise likelihood is asymmetric and right-skewed around its maximum.

For the models based on the Brown-Resnick process, the estimated wind speed is similar for max-stable and inverted max-stable models. By contrast, the estimated wind velocity for the fitted models based on random sets is weaker, but more variable, when the process is asymptotically independent.

5.3.3 Model comparison

Model selection was performed by minimizing the composite likelihood information criterion, $\text{CLIC} = -2\ell_{\mathcal{K}}(\widehat{\psi}_{\mathcal{K}}) + \text{tr}\{j(\widehat{\psi}_{\mathcal{K}})^{-1}k(\widehat{\psi}_{\mathcal{K}})\}$, an analogue of the Akaike information criterion (AIC) in a composite likelihood framework, where $j(\psi)$ and $k(\psi)$ are the matrices defined in (5.7) and (5.8). We considered a variant, CLIC*, which is scaled to be comparable with AIC for independent data; recall §3.1.4. The matrix $k(\widehat{\psi}_{\mathcal{K}})$ has a complicated form, so we estimated the product $j(\widehat{\psi}_{\mathcal{K}})^{-1}k(\widehat{\psi}_{\mathcal{K}})$ by right-multiplication of the covariance matrix $V(\widehat{\psi}_{\mathcal{K}}) \approx j(\widehat{\psi}_{\mathcal{K}})^{-1}k(\widehat{\psi}_{\mathcal{K}})j(\widehat{\psi}_{\mathcal{K}})^{-1}/T$ found using the block bootstrap by the Hessian matrix $J(\widehat{\psi}_{\mathcal{K}}) = Tj(\widehat{\psi}_{\mathcal{K}})$ estimated by finite differences.

Tables 5.4–5.5 report the differences in CLIC* for the different max-stable and inverted max-stable models fitted in §5.3.2.1–5.3.2.3, with respect to the max-stable model (5.12). Based on this criterion, asymptotic independence models are greatly preferable to max-stable models for our data. Figures 5.12–5.13 display nonparametric estimates of the coefficients $\chi_h(u)$ and $\overline{\chi}_h(u)$ (2.56) for four models at five representative stations (see also Figures E.3–E.4), and provide stronger support for inverted max-stable models, though the latter are not entirely adequate at very high thresholds either. Hence, this confirms the discrepancy observed in the CLIC* values, and suggests that extreme rainfall may be better modeled using asymptotic independence models at finite thresholds. Furthermore, Figure 5.13 reveals that our models cannot capture the

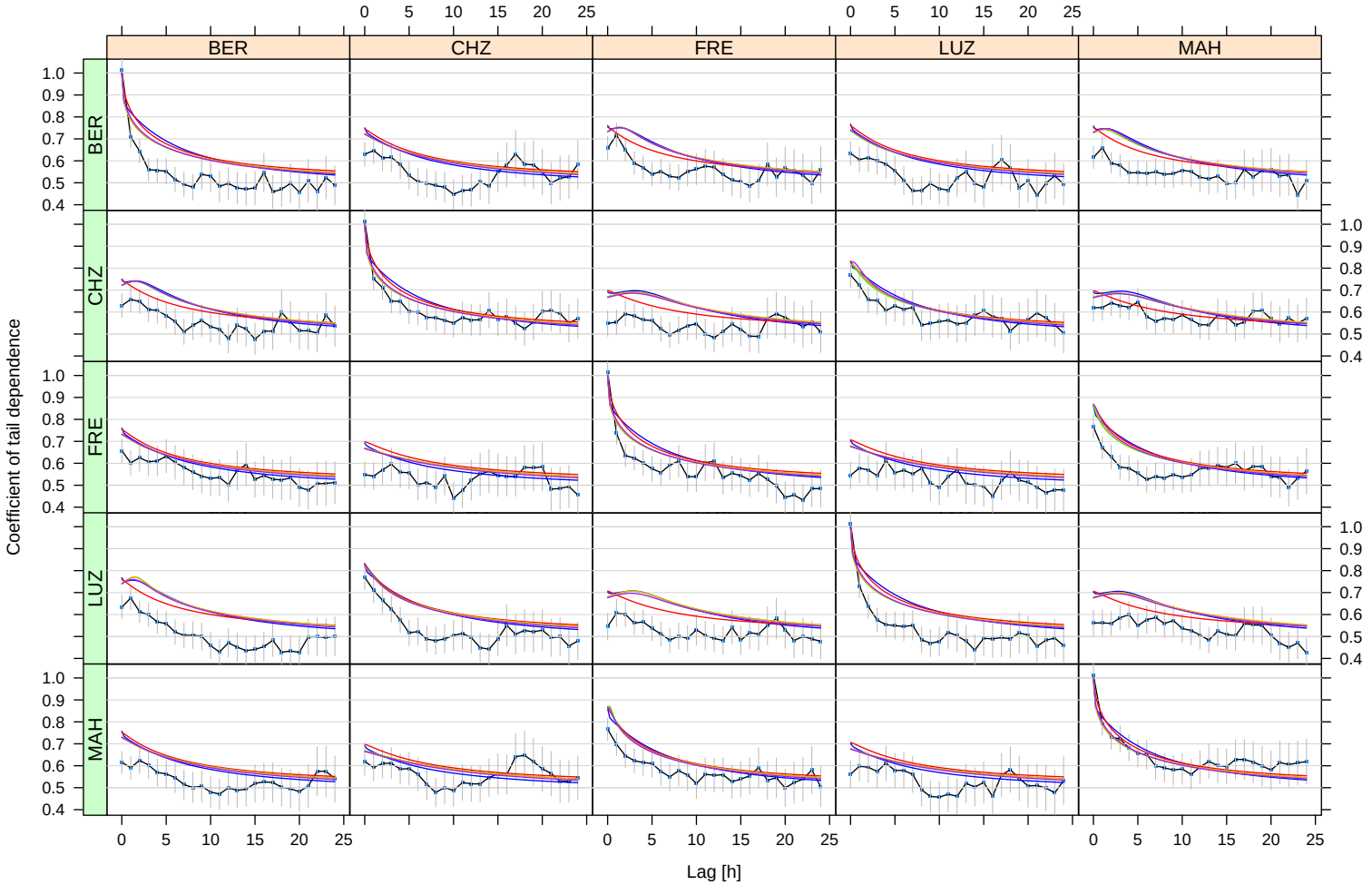

Figure 5.11: Empirical and model-based coefficients of tail dependence (recall §2.5), $\eta(h_t, h_s)$, for five stations. The black lines join the empirical coefficients of tail dependence found using the Hill estimator (1.67) at the 0.95-quantile threshold, the vertical grey segments being 95% confidence intervals (assuming independence in time). The solid curves are the fitted coefficients of tail dependence for the inverted max-stable models corresponding to model (5.12) (blue), and Models 1 (red), 2 (green), 3 (orange) and 4 (purple).

Table 5.5: Parameter estimates and 95%-confidence intervals from fitting inverted max-stable models corresponding to model (5.12) and Models 1–4, to the rainfall data. Uncertainty assessment was based on 300 block bootstrap replicates, using yearly blocks. The difference in CLIC* with respect to the max-stable model (5.12) is also reported. In the first column, "I" stands for "Inverted".

Model	Estimated dependence parameters $\widehat{\psi}_{\mathcal{K}}$			CLIC*
(5.12)-I	$\widehat{\lambda}_s \doteq 29 \times 10^3$ $(10^3, 10^6)$	$\widehat{\lambda}_t \doteq 19 \times 10^9$ $(10^4, 10^{13})$	$\widehat{\varrho} = 0.99$ $(0, 1)$	-1014
	$\widehat{\alpha}_s = 0.40$ $(0.25, 0.69)$	$\widehat{\alpha}_t = 0.15$ $(0.11, 0.37)$	$\|\widehat{V}\| = 26.2$ $(16.8, 30.3)$	
	$\widehat{m}_L = 84.2$ $(39.8, 145.1)$	$\widehat{m}_R = 72.3$ $(54.0, 83.5)$	$\widehat{\nu} = 0.03$ $(-0.10, 0.21)$	
	$k_L = 9$ (—)	$k_R = 0.4$ (—)	$\widehat{\omega} = 0.54$ $(0.35, 0.77)$	
1-I	$\widehat{\sigma}^2 = 6.6$ $(5.2, 11.0)$	$\widehat{\lambda}_s = 2335$ $(1388, 4376)$	$\widehat{\lambda}_t = 49.8$ $(37.3, 87.6)$	-643
	$\varrho = 1$ (—)	$\widehat{\alpha}_s = 0.82$ $(0.77, 0.94)$	$\widehat{\alpha}_t = 1.17$ $(1.11, 1.21)$	
2-I	$\widehat{\lambda} = 172$ $(140, 214)$	$\widehat{\alpha} = 0.87$ $(0.84, 0.92)$		-729
	$\|\widehat{V}\| = 33.4$ $(28.7, 40.2)$	$\widehat{\nu} = -0.62$ $(-0.71, -0.59)$		
3-I	$\widehat{\Sigma}_{11}^{1/2} = 192$ $(166, 229)$	$\widehat{\Sigma}_{22}^{1/2} = 217$ $(184, 448)$	$\widehat{\rho} = -0.33$ $(-0.59, 0.76)$	-661
	$\widehat{\alpha} = 0.87$ $(0.85, 0.92)$	$\|\widehat{\widetilde{V}}\| = 43.3$ $(40.1, 66.9)$	$\widehat{\nu} = -0.64$ $(-1.66, -0.59)$	
4-I	$\widehat{\lambda}_s = 164$ $(141, 186)$	$\widehat{\lambda}_t = 5.5$ $(5.0, 5.9)$	$\widehat{\alpha} = 0.94$ $(0.90, 0.98)$	-767
	$\|\widehat{V}\| = 19.2$ $(17.3, 21.6)$	$\widehat{\nu} = 0$ $(-0.32, 0.13)$		

decay in the coefficient $\overline{\chi}_h(u)$, as the threshold u increases; however, to our current knowledge, no spatial model proposed in the literature is able to capture this special tail behavior, and though not ideal, the use of inverted max-stable models appears to be the best we can do for the moment. As an idea for future work, it would be interesting to construct flexible extremal models that can fit a wider class of tail behaviors, and to extend this to hybrid models (2.51). In practice, though max-stable models may not fit very well at finite thresholds, they provide conservative bounds for joint probabilities of extreme events, which may be useful for risk assessment.

The CLIC* values also suggest that the random set models are better than the alternative Brown-Resnick processes, even though they have more parameters. Radar images of summer rain fields show areas where there is heavy rain and others where there is none; see Figure 1. The random set models capture this, but the space-time Brown–Resnick processes do not.

5.3.4 Summary

In our space-time application, we have considered max-stable and inverted max-stable processes for the modeling of extreme hourly rainfall measurements recorded in western Switzerland.

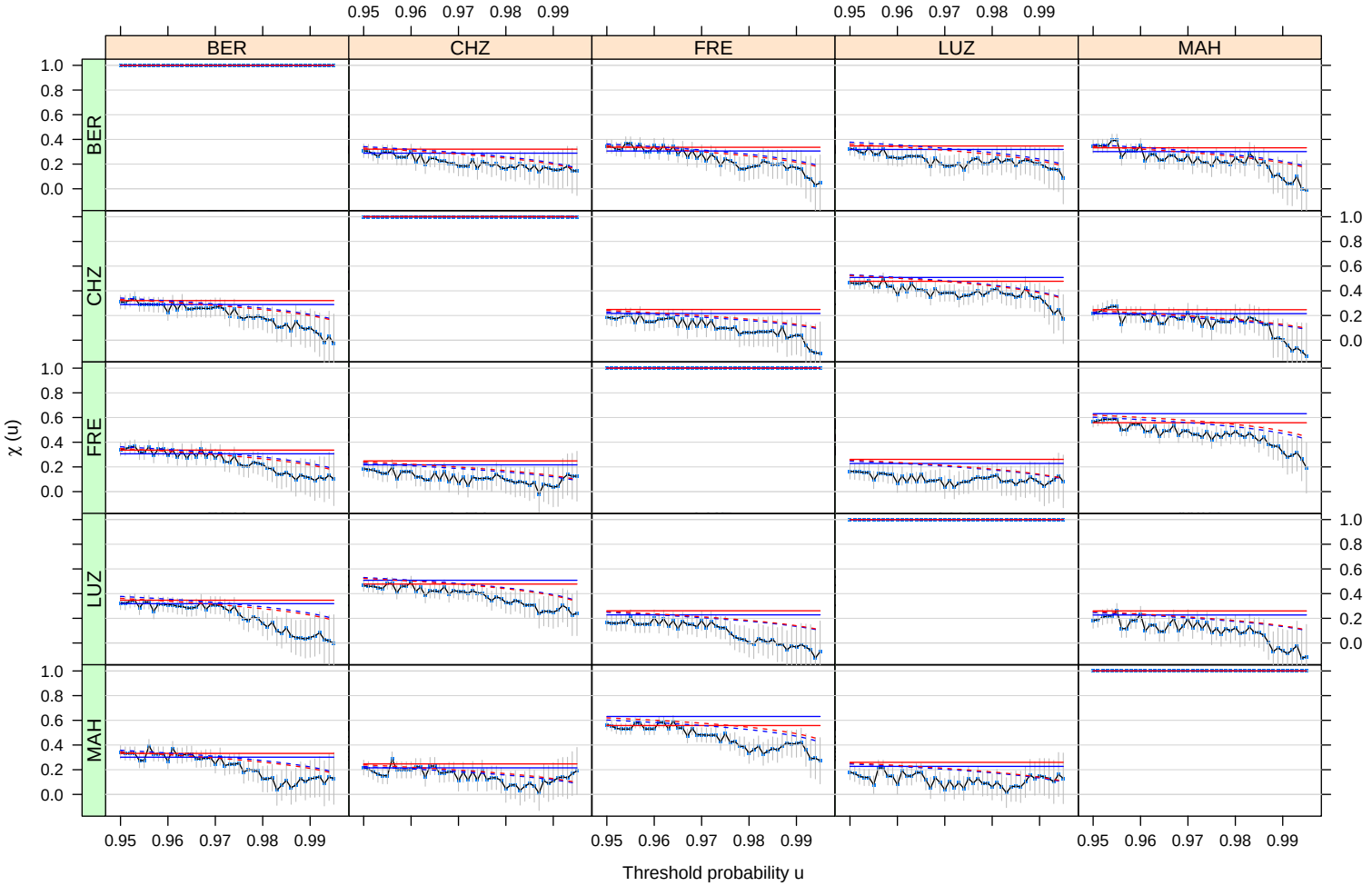

Figure 5.12: Empirical and model-based coefficients $\chi_{(h_t,h_s)}(u)$ (recall §2.56) for five stations. The black lines join the empirical rank-based estimates for $h_t = 0$, the vertical grey segments being 95% confidence intervals (assuming independence in time). The solid curves are the fitted coefficients for the max-stable models corresponding to model (5.12) (blue), and Model 1 (red), whereas the dashed curves are the coefficients for the corresponding inverted max-stable models.

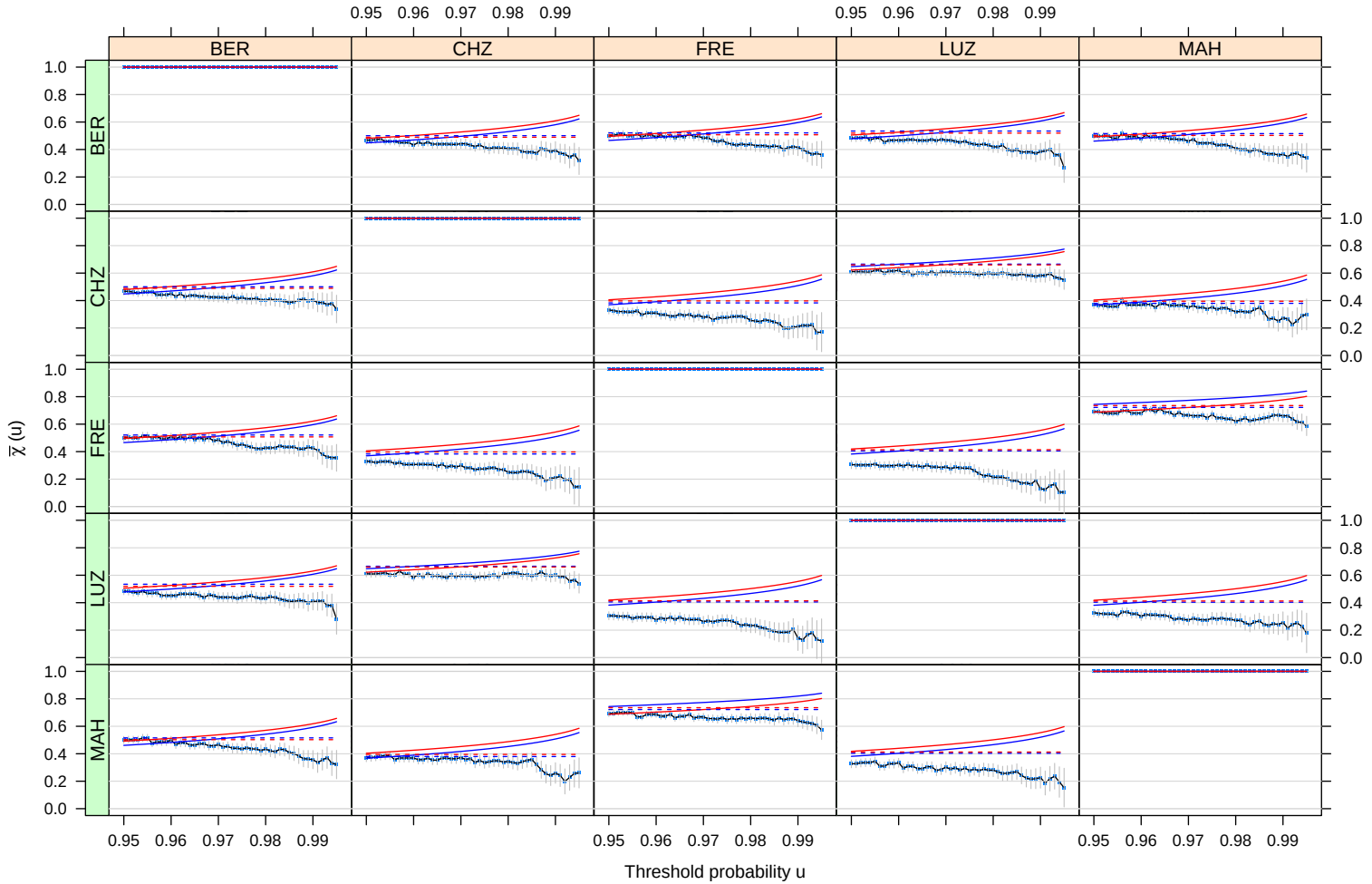

Figure 5.13: Empirical and model-based coefficients $\overline{\chi}_{(h_t,h_s)}(u)$ (recall §2.56) for five stations. The black lines join the empirical rank-based estimates for $h_t = 0$, the vertical grey segments being 95% confidence intervals (assuming independence in time). The solid curves are the fitted coefficients for the max-stable models corresponding to model (5.12) (blue), and Model 1 (red), whereas the dashed curves are the coefficients for the corresponding inverted max-stable models.

The comparison between empirical and fitted extremal coefficients in Figures 5.6 and 5.10 suggests that the proposed max-stable processes are reasonable candidates within the class of max-stable processes. Similarly, although the confidence bands in Figure 5.11 quite often do not contain the fitted coefficients of tail dependence (probably because temporal dependence was ignored for the computation of empirical diagnostics), the fits of inverted max-stable processes appear to be acceptable within the class of asymptotic independence models. Because all models were fitted using a censored pairwise likelihood at the 95%-quantile, it is natural to wonder whether they remain valid at higher thresholds. The coefficients $\chi_h(u)$ and $\overline{\chi}_h(u)$ in Figure 5.12–5.13 show that all the models proposed are not flexible enough to capture adequately the "true" joint probabilities of extreme events at extremely high quantiles. Asymptotic independence models appear to be best, which is confirmed by the composite likelihood information criterion (CLIC), but it is difficult to tell whether these models are more suitable because the data are truly asymptotically independent, or because they are simply more flexible at finite thresholds. Furthermore, the data could be asymptotically independent in time (at moderate time lags), but asymptotically dependent in space (at short distances), in which case model selection would be biased towards inverted max-stable models if many time lags are used for estimation. By contrast to asymptotic independence models, max-stable processes tend to overestimate the joint probabilities of rare events, and predict a positive probability of occurrence of simultaneous extremes in the limit, which may be desirable for risk assessment purposes. It would be interesting to fit hybrid models (2.51), which can capture different tail behaviors at different spatio-temporal distances, and to extend the existing max-stable models to flexible asymptotic dependence models (for example based on Student t processes).

Among all fitted models, the best candidate for our data appears to be constructed from a random set element, according to the CLIC and more informal graphical diagnostics. Although highly idealized, such an extreme rainfall model is fairly complex, and estimation and uncertainty assessment are demanding. Moreover, model checking is tricky, due to the computational burden that it requires. Spatio-temporal return levels, such as extreme quantiles of the cumulative rainfall in a certain region during a certain time period, may be derived using extensive simulations. To this end, it would be necessary to build a spatial model for the margins, for example letting the marginal parameters τ, ξ, ζ_u in (5.1) depend upon covariates such as elevation, and to specify a dependence model under the threshold (Thibaud *et al.*, 2013). Constructing spatio-temporal dependence models that continuously link distinct submodels below and above a predetermined threshold is not obvious and is an idea for future research.

For simplicity, the marginal models considered above specify different GPD models

at each monitoring station. Our results revealed that the rainfall data are in fact more or less stationary and heavy-tailed, with shape parameters reaching about 0.2. Considering a single shape parameter for the whole region of study would therefore enable borrowing of strength across locations, and thus would give more accurate marginal parameter estimates.

5.4 Discussion and perspectives

The work described above proposes inference for space-time extremes using a censored pairwise likelihood, and illustrates this by fitting several max-stable and asymptotic independence models for extreme rainfall; there are clear possibilities for extension to other phenomena, such as temperatures or wind speeds. 'Dynamic' space-time modeling of extremes thus seems to be feasible; complex models can be consistently fitted using composite censored likelihood based on threshold exceedances. However, the large amount of data involved and the consequent usefulness of parallel computation underline the advantages of access to substantial computing resources when tackling such problems.

Wadsworth & Tawn (2013) show how to perform inference based on a censored full likelihood for Brown-Resnick processes, and recent promising developments suggest that this is also possible for extremal-t processes (recall §2.3.2.4), which extend the Schlather and Brown-Resnick models. In addition to the potential large gain in efficiency, such approaches would permit more natural treatment of the uncertainty and model comparison, avoiding the need to estimate the complicated sandwich matrices in (5.7) and (5.8). Full likelihood estimation also avoids the need to select the pairs in the pairwise likelihood, but the treatment of missing data is more awkward. Though it seems entirely feasible to apply these ideas to max-stable processes, it is, however, much less obvious to see how they might be adapted to inverted max-stable processes. This is an area of future research.

Another point concerns nonstationarity. All space-time models for extremes considered in this chapter are stationary, but in practice, data often depart from this assumption. Blanchet & Davison (2012) propose to capture nonstationarity in the margins by adding linear regression terms in the parameters, and to deal with anisotropy in the dependence structure by deforming the plane elliptically. Alternatively, Anderes & Stein (2008) consider quasiconformal maps, which are smooth nonparametric deformations of the plane, in order to flexibly model nonstationarity for Gaussian processes. It would be interesting to pursue this idea and to adapt these methods to spatial extremes.

Conclusion and future work

The research presented in this thesis has two main orientations. The first aspect relates to the creation of new models for spatio-temporal extremes, based on max-stable processes and which can be applied to environmental data, such as hourly rainfall. This issue is tackled in Chapters 2 and 5. The second aspect concerns the development of inference methods for these models, and the assessment of their performance in terms of relative efficiency. This is mostly addressed in Chapters 3 and 4, though Chapters 1 and 5 also contain some new results.

In Chapter 1, we studied different estimators for bivariate extremes, some based on block maxima and others based on threshold exceedances, and we compared their root mean squared errors using simulations. In order to simplify matters, we considered the (symmetric) extreme-value logistic distribution, where the degree of dependence is controlled by a single parameter. We found that the quality of these estimators is worse when the data are more independent, and that threshold-based estimators usually beat those based on block maxima, because they use more data. More interestingly, we have also found that for thresholds often chosen in practice (e.g., 95%–98% empirical quantiles), the censored estimator used in our application in Chapter 5 outperforms its natural competitors. Unlike the latter, the former has a rather small mis-specification bias, and its variability is quite well controlled. By analytical calculations, we have also investigated its theoretical efficiency loss due to marginal censoring, which confirmed us that a careful choice of threshold is essential. It would be interesting to know whether these conclusions extend to other dependence structures, and to multivariate extreme-value distributions of higher orders.

In Chapter 2, we have shown how to build models for spatial (or spatio-temporal) extremes, using two main ingredients: existing geostatistical tools, and a spectral representation for max-stable processes. We have also seen that, in practice, asymptotic independence models (which are not max-stable) might fit better at subasymptotic levels, and that it is difficult to discriminate between these types of tail behavior. Extrapolation beyond the recorded data rests entirely on assumptions, so careful model

checking is crucial for risk assessment. We illustrated this with two datasets recorded at the same set of monitoring stations, namely cumulative rainfall and temperature minima, and concluded that asymptotic independent processes provide a better fit for the former, while max-stable processes fit the latter better. As an extension, it would be interesting to fit hybrid models, which are flexible intermediates between the two aforementioned classes of extremal dependence models, or to specify a spatial model for the margins in order to draw maps of return levels.

Chapter 3 contributes to the composite likelihood literature. We explained how to perform inference for max-stable processes and related asymptotic independence models using weighted pairwise likelihoods, and saw that a careful selection of pairs to include in the pairwise likelihood may improve computational and statistical efficiencies. By considering simple time series models, we were able to draw two main informal conclusions: first, we should retain as few pairs as possible, so far as the parameters remain identifiable from the bivariate densities; second, for a fixed number of pairs, the best option, in most cases, is to consider a mixture between strongly dependent pairs and weakly correlated ones. These results were obtained by theoretical calculations for Gaussian AR(1) and MA(1) models, and by simulations for more general ARMA and max-stable models. We also investigated the best weighting strategy for Gaussian processes, in terms of minimal asymptotic variance of the pairwise likelihood estimator, and came up with similar conclusions. In this chapter, we have also shown how to perform full likelihood inference for the multivariate extreme-value logistic model, based on accurate simulated approximations to the likelihood. To this end, variance reduction techniques (in particular importance sampling) are essential. Using similar ideas, we developed a simple max-stable time series model with asymmetric logistic dependence structure, for which full likelihood inference is attainable using particle filters, and assessed the relative efficiency of pairwise likelihood estimators in this context. Again, we found that the inclusion of some distant pairs results in a significant increase in efficiency. Following Reich & Shaby (2012), this time series model generalizes to similar spatial models, and it does not seem infeasible to adapt our methods to the latter.

In Chapter 4, we investigated the gain in efficiency of triplewise likelihood estimators, compared to pairwise ones, and also assessed the relative efficiency of higher-order composite likelihoods, in the case of the (max-stable) Brown–Resnick process. By deriving the joint distribution in arbitrary dimensions, we showed, using simulations, that for rough processes, the efficiency gains are minor compared to the additional computational and coding effort of the triplewise likelihood approach. Unless the process of interest is very smooth, which is rarely the case with environmental applications, and assuming the unknown parameters are identifiable from the bivariate

densities, pairwise likelihood estimators appear to be a good solution for inference. Although the selection of triples is less obvious than that of pairs, it would be interesting to see how these results extend to weighted composite likelihoods. Furthermore, recent developments suggest that it is entirely feasible to repeat this whole study for extremal-t processes, which generalize Brown–Resnick processes.

In Chapter 5, we proposed a novel censored threshold-based pairwise likelihood estimator for the estimation of max-stable processes, and demonstrated its asymptotic normality and strong consistency under mild conditions. Moreover, in this chapter, we also constructed useful models for space-time extremes, and fitted them to hourly rainfall data, using the methodology developed in the earlier chapters. We considered max-stable models, among which the "best" appeared to be based on space-time random sets —"moving clouds", and we also considered asymptotic independence models, which seemed to fit better overall at the 95% threshold. This analysis could be extended in several respects. First, the margins were fitted independently from each other and from the dependence model. It would be natural to consider a full model that links the margins to the dependence structure, and to estimate the parameters at once. Second, it would be interesting to consider models based on the extremal-t process, which appears to be quite flexible. Third, recent work suggests that it is possible to perform inference based on the full likelihood for Brown-Resnick or extremal-t models, so more efficient inference could be made by treating blocks (e.g., summers or weeks) of correlated observations as a whole. Finally, following similar ideas, it seems possible to embed our space-time models into a Bayesian framework and to fit them using standard MCMC methods, which would permit flexible modeling, efficient inference and natural uncertainty assessment —though at a high computational cost.

To summarize, the methods developed in the present thesis allow one to fit complicated models for extreme events, which include but are not limited to rainfall. "Dynamic modeling" of such phenomena above high thresholds is feasible, but there is room for many future improvements and extensions.

A Performance of various estimators for the bivariate extreme-value logistic model

Assuming unit Fréchet marginals, the bivariate logistic extreme value distribution is

$$G(z_1, z_2) = \exp\left\{-\left(z_1^{-1/\alpha} + z_2^{-1/\alpha}\right)^{\alpha}\right\}, \qquad z_1, z_2 > 0,$$

where $\alpha \in (0,1]$ is a dependence parameter, recall (1.30) and Figure 1.5, and the Archimedean copula with generator $\varphi(t) = (t^{-1} - 1)^{1/\alpha}$ is known to belong to its max-domain of attraction.

In order to assess the performance of various sub-asymptotic estimators for α, we conducted a simple simulation study (details can be found in Section 1.2.2.2), and estimated empirically the bias, the standard error and the root mean squared error (RMSE) of each estimator. The results are reported in Tables A.1, A.2 and A.3.

Results show that the censored threshold-based estimators, and especially $\widehat{\alpha}_7$, perform much better than the non-censored or partially censored ones, or those based on block maxima.

Table A.1: Negative empirical bias (with standard errors in brackets), multiplied by 1000, of the estimators $\widehat{\alpha}_{\text{naive}}$, $\widehat{\alpha}_B$, $\widehat{\alpha}_{ST}$, $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\alpha}_5$, $\widehat{\alpha}_6$, $\widehat{\alpha}_7$ and $\widehat{\alpha}_8$ of Section 1.2.2.2, computed from 300 bivariate datasets of size $n = 10000$ generated independently from the joint density (1.51).

Method	Estimator		Value of α								
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Naive	$\widehat{\alpha}_{\text{naive}}$		29.8 (0)	59 (0.1)	87.7 (0.1)	116.2 (0.1)	144.1 (0.2)	172.1 (0.2)	199.5 (0.3)	226.6 (0.3)	252.1 (0.3)
Block-maxima	$\widehat{\alpha}_B$	$b = 20$	0 (1.1)	1 (2.2)	−4 (3.2)	5.2 (4.2)	−1 (4.9)	0 (6.3)	1.3 (7.1)	13.7 (7)	16.3 (6.4)
	$\widehat{\alpha}_{ST}$		0.6 (1)	−0.1 (2)	−6.8 (2.8)	5.3 (3.5)	−1.7 (4.1)	−0.4 (4.4)	4.4 (4.6)	4.6 (4.5)	−4.4 (3.7)
	$\widehat{\alpha}_B$	$b = 50$	0.1 (0.7)	−0.8 (1.5)	−1.5 (2.1)	6 (2.6)	3.9 (3.5)	3.7 (3.8)	−1.6 (4.5)	7.9 (4.6)	4.5 (4.4)
	$\widehat{\alpha}_{ST}$		0 (0.7)	−0.8 (1.3)	−0.6 (1.8)	5.6 (2.2)	3.7 (2.7)	5.6 (2.8)	−0.1 (2.9)	11.7 (2.8)	3.1 (2.3)
	$\widehat{\alpha}_B$	$b = 100$	0.1 (0.5)	−0.5 (1)	−0.3 (1.4)	4.7 (1.9)	−0.5 (2.3)	4.6 (2.8)	4.2 (3.1)	8.7 (3.3)	1.9 (3.3)
	$\widehat{\alpha}_{ST}$		0.1 (0.5)	−0.4 (0.9)	1.2 (1.2)	5.2 (1.5)	4.1 (1.7)	8.7 (2)	7.9 (2)	15.3 (1.9)	13.6 (1.6)
Threshold-based	$\widehat{\alpha}_1$	$p = 0.9$	5.5 (0.1)	12.8 (0.2)	24.4 (0.2)	42.3 (0.3)	68.7 (0.3)	105.1 (0.3)	150.5 (0.3)	205.7 (0.2)	266.8 (0.2)
	$\widehat{\alpha}_2$		2.9 (0.1)	6.7 (0.3)	13.1 (0.3)	24.9 (0.4)	43.4 (0.4)	72.6 (0.3)	110.9 (0.3)	160.6 (0.3)	217.5 (0.2)
	$\widehat{\alpha}_3$		3.8 (0.1)	10.9 (0.2)	22.7 (0.3)	41.9 (0.3)	69 (0.3)	105.7 (0.3)	150.6 (0.3)	205.2 (0.2)	265.5 (0.2)
	$\widehat{\alpha}_4$		3.6 (0.1)	10.1 (0.2)	21.4 (0.3)	40.2 (0.3)	67.4 (0.3)	104.6 (0.3)	150.3 (0.3)	205.7 (0.2)	267 (0.2)
	$\widehat{\alpha}_5$		3.4 (0.1)	8.6 (0.2)	15.6 (0.4)	26 (0.4)	37.2 (0.5)	51.8 (0.5)	67 (0.5)	86.1 (0.5)	104.6 (0.5)
	$\widehat{\alpha}_6$		3.1 (0.1)	7.4 (0.3)	13.1 (0.4)	21.6 (0.4)	31.2 (0.5)	44.1 (0.5)	58.7 (0.5)	76.9 (0.5)	96.8 (0.5)
	$\widehat{\alpha}_7$		2.7 (0.1)	5.7 (0.3)	9.1 (0.4)	14.4 (0.5)	19.3 (0.5)	26.2 (0.5)	32.7 (0.6)	41.3 (0.6)	47.5 (0.6)
	$\widehat{\alpha}_8$		2.7 (0.1)	5.7 (0.3)	8.9 (0.4)	13.9 (0.5)	18.6 (0.5)	25.2 (0.6)	31.8 (0.6)	39.6 (0.6)	46.3 (0.6)
	$\widehat{\alpha}_1$	$p = 0.95$	2.9 (0.1)	6.6 (0.3)	12.7 (0.4)	24.5 (0.4)	42.7 (0.4)	71.8 (0.3)	109.5 (0.4)	159.3 (0.3)	215.9 (0.3)
	$\widehat{\alpha}_2$		1.6 (0.2)	2.9 (0.4)	6.6 (0.5)	14.2 (0.6)	25.9 (0.5)	48.2 (0.5)	79.9 (0.4)	123 (0.4)	175.4 (0.3)
	$\widehat{\alpha}_3$		2.1 (0.2)	5.2 (0.3)	11.8 (0.4)	24.8 (0.5)	43.7 (0.4)	73.2 (0.4)	111 (0.4)	160.7 (0.3)	217.2 (0.3)
	$\widehat{\alpha}_4$		2 (0.2)	4.8 (0.3)	11.2 (0.4)	23.9 (0.5)	43.1 (0.4)	72.9 (0.4)	111.8 (0.3)	162 (0.3)	219.5 (0.3)
	$\widehat{\alpha}_5$		1.7 (0.2)	3.4 (0.4)	7.3 (0.5)	13.7 (0.6)	18.3 (0.7)	27 (0.7)	34.5 (0.8)	45.9 (0.7)	54.3 (0.7)
	$\widehat{\alpha}_6$		1.6 (0.2)	2.9 (0.4)	6.1 (0.5)	11.4 (0.7)	15.3 (0.7)	22.6 (0.8)	30.7 (0.8)	41.1 (0.7)	51 (0.7)
	$\widehat{\alpha}_7$		1.4 (0.2)	2 (0.4)	4.1 (0.5)	7.8 (0.7)	9.2 (0.7)	13.8 (0.8)	16.8 (0.8)	22.9 (0.8)	24.8 (0.7)
	$\widehat{\alpha}_8$		1.4 (0.2)	2 (0.4)	3.9 (0.5)	7.4 (0.7)	8.7 (0.7)	12.7 (0.8)	16.6 (0.8)	21.5 (0.8)	24.3 (0.7)
	$\widehat{\alpha}_1$	$p = 0.98$	1.2 (0.2)	2.1 (0.4)	5.1 (0.6)	12 (0.7)	21.7 (0.6)	42.1 (0.6)	71.7 (0.5)	113.4 (0.4)	163.5 (0.4)
	$\widehat{\alpha}_2$		0.9 (0.3)	1.7 (0.6)	1.5 (0.8)	7 (1)	13.1 (0.9)	27.8 (0.9)	52.4 (0.7)	88.1 (0.6)	135 (0.4)
	$\widehat{\alpha}_3$		1.1 (0.3)	2.4 (0.5)	5.4 (0.7)	12.7 (0.7)	23.1 (0.6)	44.1 (0.6)	74.3 (0.5)	117 (0.4)	168.5 (0.4)
	$\widehat{\alpha}_4$		1 (0.3)	2.2 (0.5)	5 (0.7)	12.4 (0.7)	22.9 (0.7)	44 (0.6)	75.2 (0.5)	118.1 (0.4)	170.6 (0.3)
	$\widehat{\alpha}_5$		0.6 (0.3)	1 (0.6)	1.4 (0.9)	5.2 (1)	6.3 (1.1)	10.8 (1.2)	13.1 (1.2)	20.7 (1.1)	21.4 (1)
	$\widehat{\alpha}_6$		0.5 (0.3)	0.8 (0.6)	0.7 (0.9)	4.2 (1)	4.7 (1.2)	8.1 (1.3)	11.4 (1.3)	17.5 (1.2)	20.1 (1)
	$\widehat{\alpha}_7$		0.4 (0.3)	0.4 (0.6)	0.1 (0.9)	2.9 (1)	2.6 (1.1)	5.5 (1.2)	5.9 (1.3)	11.3 (1.2)	9.4 (1)
	$\widehat{\alpha}_8$		0.4 (0.3)	0.4 (0.6)	−0.2 (0.9)	2.6 (1)	2 (1.2)	4 (1.3)	5.5 (1.3)	9.3 (1.2)	9 (1)

Table A.2: Empirical standard error (with standard errors in brackets), multiplied by 1000, of the estimators $\widehat{\alpha}_{\text{naive}}$, $\widehat{\alpha}_B$, $\widehat{\alpha}_{ST}$, $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\alpha}_5$, $\widehat{\alpha}_6$, $\widehat{\alpha}_7$ and $\widehat{\alpha}_8$ of Section 1.2.2.2, computed from 300 bivariate datasets of size $n = 10000$ generated independently from the joint density (1.51).

Method	Estimator		Value of α								
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Naive	$\widehat{\alpha}_{\text{naive}}$		0.7 (0)	1.3 (0)	2 (0)	2.5 (0)	3.3 (0)	3.7 (0)	4.7 (0)	5 (0)	5.7 (0)
Block-maxima	$\widehat{\alpha}_B$	$b = 20$	19.2 (0.2)	37.4 (0.4)	55.9 (0.8)	73.1 (1.2)	85 (1.5)	109.8 (2.2)	123.7 (2.4)	121.9 (2.3)	110.5 (2.1)
	$\widehat{\alpha}_{ST}$		18.1 (0.1)	34.7 (0.4)	49.3 (0.6)	61.1 (0.9)	71.9 (1.2)	75.6 (1.4)	79.5 (1.3)	77.6 (1.3)	64.8 (1)
	$\widehat{\alpha}_B$	$b = 50$	11.7 (0.1)	25.4 (0.2)	36.5 (0.3)	44.9 (0.6)	60.5 (0.9)	65.6 (0.9)	78.2 (1.3)	80.2 (1.3)	75.9 (1.1)
	$\widehat{\alpha}_{ST}$		11.4 (0.1)	22.1 (0.2)	30.5 (0.3)	37.7 (0.5)	46 (0.6)	48.8 (0.6)	50.3 (0.6)	48.5 (0.6)	40.2 (0.5)
	$\widehat{\alpha}_B$	$b = 100$	8.3 (0)	16.9 (0.1)	24.7 (0.2)	33 (0.3)	39.9 (0.5)	47.9 (0.6)	54.2 (0.8)	57.5 (0.8)	57.8 (0.8)
	$\widehat{\alpha}_{ST}$		8 (0)	15.5 (0.1)	21.1 (0.2)	26.1 (0.2)	30 (0.3)	35 (0.4)	34.5 (0.4)	33.2 (0.3)	28.2 (0.3)
Threshold-based	$\widehat{\alpha}_1$	$p = 0.9$	1.7 (0)	3.1 (0)	4.2 (0)	4.9 (0)	4.9 (0)	4.6 (0)	4.7 (0)	4.3 (0)	4.1 (0)
	$\widehat{\alpha}_2$		2.5 (0)	4.5 (0)	5.9 (0)	6.9 (0)	6.7 (0)	5.7 (0)	5.7 (0)	4.6 (0)	4.1 (0)
	$\widehat{\alpha}_3$		2.3 (0)	3.7 (0)	4.8 (0)	5.3 (0)	5.2 (0)	4.7 (0)	4.7 (0)	4.3 (0)	4 (0)
	$\widehat{\alpha}_4$		2.3 (0)	3.8 (0)	4.8 (0)	5.3 (0)	5.1 (0)	4.6 (0)	4.3 (0)	4 (0)	3.8 (0)
	$\widehat{\alpha}_5$		2.4 (0)	4.3 (0)	6.2 (0)	7.5 (0)	8.3 (0)	8.6 (0.1)	9.2 (0.1)	8.9 (0.1)	9.4 (0)
	$\widehat{\alpha}_6$		2.4 (0)	4.4 (0)	6.4 (0)	7.7 (0)	8.6 (0)	8.8 (0.1)	9 (0.1)	9.3 (0.1)	9.2 (0.1)
	$\widehat{\alpha}_7$		2.4 (0)	4.4 (0)	6.5 (0)	7.9 (0)	8.9 (0)	9.3 (0.1)	9.9 (0.1)	9.9 (0.1)	10.4 (0.1)
	$\widehat{\alpha}_8$		2.5 (0)	4.5 (0)	6.6 (0)	8.1 (0)	9.2 (0)	9.7 (0.1)	10 (0.1)	10.7 (0.1)	10.7 (0.1)
	$\widehat{\alpha}_1$	$p = 0.95$	2.5 (0)	4.5 (0)	6.2 (0)	7 (0)	6.9 (0)	6 (0)	6.3 (0)	5.2 (0)	4.8 (0)
	$\widehat{\alpha}_2$		3.5 (0)	6.7 (0)	8.7 (0)	10.3 (0.1)	9.2 (0.1)	8.8 (0.1)	7.5 (0)	6.2 (0)	5.3 (0)
	$\widehat{\alpha}_3$		3.2 (0)	5.7 (0)	6.7 (0)	7.9 (0)	7.1 (0)	6.4 (0)	6.3 (0)	5.3 (0)	4.9 (0)
	$\widehat{\alpha}_4$		3.2 (0)	5.8 (0)	6.8 (0)	7.9 (0)	6.9 (0)	6.3 (0)	6 (0)	5 (0)	4.6 (0)
	$\widehat{\alpha}_5$		3.4 (0)	6.5 (0)	9 (0)	11.1 (0.1)	11.7 (0.1)	13 (0.1)	13 (0.1)	12.6 (0.1)	12 (0.1)
	$\widehat{\alpha}_6$		3.5 (0)	6.6 (0)	9.2 (0.1)	11.3 (0.1)	12.2 (0.1)	13.5 (0.1)	13.1 (0.1)	12.9 (0.1)	12 (0.1)
	$\widehat{\alpha}_7$		3.5 (0)	6.6 (0)	9.2 (0.1)	11.4 (0.1)	12.2 (0.1)	13.6 (0.1)	13.6 (0.1)	13.3 (0.1)	12.5 (0.1)
	$\widehat{\alpha}_8$		3.5 (0)	6.6 (0)	9.4 (0.1)	11.6 (0.1)	12.7 (0.1)	14.1 (0.1)	13.8 (0.1)	13.9 (0.1)	13 (0.1)
	$\widehat{\alpha}_1$	$p = 0.98$	3.9 (0)	7.4 (0)	10.6 (0.1)	11.8 (0.1)	10.5 (0.1)	10.2 (0.1)	8.7 (0)	7.3 (0)	6.2 (0)
	$\widehat{\alpha}_2$		5.7 (0)	10.9 (0.1)	14.6 (0.1)	16.5 (0.1)	15.6 (0.1)	14.9 (0.1)	12.2 (0.1)	9.5 (0.1)	6.7 (0)
	$\widehat{\alpha}_3$		5.2 (0)	9.3 (0.1)	11.8 (0.1)	12.9 (0.1)	11.2 (0.1)	10.9 (0.1)	9.3 (0.1)	7.5 (0)	6.3 (0)
	$\widehat{\alpha}_4$		5.2 (0)	9.4 (0.1)	12 (0.1)	12.9 (0.1)	11.4 (0.1)	10.8 (0.1)	9 (0)	7.2 (0)	6 (0)
	$\widehat{\alpha}_5$		5.6 (0)	11 (0.1)	15.3 (0.1)	17.6 (0.1)	19 (0.2)	20.8 (0.2)	21.4 (0.2)	19.8 (0.2)	16.7 (0.1)
	$\widehat{\alpha}_6$		5.6 (0)	11.1 (0.1)	15.6 (0.1)	17.8 (0.1)	20 (0.2)	21.7 (0.2)	22 (0.2)	20.5 (0.2)	16.8 (0.1)
	$\widehat{\alpha}_7$		5.6 (0)	11 (0.1)	15.5 (0.1)	17.8 (0.1)	19.3 (0.2)	21.2 (0.2)	21.8 (0.2)	20.2 (0.2)	17 (0.1)
	$\widehat{\alpha}_8$		5.6 (0)	11.1 (0.1)	15.7 (0.1)	18 (0.1)	20.3 (0.2)	22.1 (0.2)	22.5 (0.2)	21.1 (0.2)	17.4 (0.1)

Table A.3: Empirical root mean squared error (with standard errors in brackets), multiplied by 1000, of the estimators $\widehat{\alpha}_{\text{naive}}$, $\widehat{\alpha}_B$, $\widehat{\alpha}_{ST}$, $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$, $\widehat{\alpha}_4$, $\widehat{\alpha}_5$, $\widehat{\alpha}_6$, $\widehat{\alpha}_7$ and $\widehat{\alpha}_8$ of Section 1.2.2.2, computed from 300 bivariate datasets of size $n = 10000$ generated independently from the joint density (1.51).

Method	Estimator		Value of α								
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Naive	$\widehat{\alpha}_{\text{naive}}$		29.8 (0)	59 (0.1)	87.7 (0.1)	116.3 (0.1)	144.2 (0.2)	172.2 (0.2)	199.6 (0.3)	226.6 (0.3)	252.2 (0.3)
Block-maxima	$\widehat{\alpha}_B$	$b = 20$	19.2 (0.2)	37.4 (0.5)	56.1 (1.2)	73.2 (1.6)	85 (1.6)	109.8 (2.2)	123.7 (2.5)	122.7 (3.3)	111.7 (3.4)
	$\widehat{\alpha}_{ST}$		18.1 (0.2)	34.7 (0.4)	49.7 (1.2)	61.3 (1.4)	71.9 (1.4)	75.6 (1.4)	79.6 (1.7)	77.8 (1.7)	65 (1.4)
	$\widehat{\alpha}_B$	$b = 50$	11.7 (0.1)	25.4 (0.4)	36.5 (0.5)	45.3 (1.1)	60.6 (1.2)	65.7 (1.3)	78.2 (1.4)	80.5 (1.9)	76.1 (1.6)
	$\widehat{\alpha}_{ST}$		11.4 (0.1)	22.1 (0.3)	30.6 (0.4)	38.1 (1)	46.2 (1)	49.2 (1.1)	50.3 (0.6)	49.9 (1.5)	40.3 (0.8)
	$\widehat{\alpha}_B$	$b = 100$	8.3 (0.1)	16.9 (0.2)	24.7 (0.3)	33.3 (0.8)	39.9 (0.5)	48.1 (1.1)	54.4 (1.2)	58.1 (1.5)	57.8 (1)
	$\widehat{\alpha}_{ST}$		8 (0.1)	15.5 (0.2)	21.1 (0.3)	26.7 (0.7)	30.3 (0.7)	36 (1.1)	35.4 (1)	36.6 (1.3)	31.3 (1.1)
Threshold-based	$\widehat{\alpha}_1$	$p = 0.9$	5.8 (0.1)	13.2 (0.2)	24.7 (0.2)	42.6 (0.3)	68.9 (0.3)	105.2 (0.3)	150.6 (0.3)	205.7 (0.2)	266.8 (0.2)
	$\widehat{\alpha}_2$		3.8 (0.1)	8.1 (0.2)	14.3 (0.3)	25.9 (0.4)	43.9 (0.4)	72.9 (0.3)	111.1 (0.3)	160.7 (0.3)	217.6 (0.2)
	$\widehat{\alpha}_3$		4.4 (0.1)	11.5 (0.2)	23.2 (0.3)	42.2 (0.3)	69.2 (0.3)	105.8 (0.3)	150.7 (0.3)	205.3 (0.2)	265.5 (0.2)
	$\widehat{\alpha}_4$		4.3 (0.1)	10.8 (0.2)	22 (0.3)	40.5 (0.3)	67.6 (0.3)	104.7 (0.3)	150.4 (0.3)	205.8 (0.2)	267.1 (0.2)
	$\widehat{\alpha}_5$		4.2 (0.1)	9.6 (0.2)	16.8 (0.3)	27.1 (0.4)	38.2 (0.5)	52.5 (0.5)	67.6 (0.5)	86.5 (0.5)	105 (0.5)
	$\widehat{\alpha}_6$		4 (0.1)	8.6 (0.2)	14.5 (0.3)	23 (0.4)	32.4 (0.5)	45 (0.5)	59.4 (0.5)	77.4 (0.5)	97.2 (0.5)
	$\widehat{\alpha}_7$		3.6 (0.1)	7.3 (0.2)	11.1 (0.3)	16.5 (0.4)	21.3 (0.5)	27.8 (0.5)	34.2 (0.6)	42.5 (0.6)	48.7 (0.6)
	$\widehat{\alpha}_8$		3.6 (0.1)	7.2 (0.2)	11.1 (0.3)	16.1 (0.4)	20.7 (0.5)	26.9 (0.5)	33.4 (0.6)	41.1 (0.6)	47.6 (0.6)
	$\widehat{\alpha}_1$	$p = 0.95$	3.8 (0.1)	8 (0.2)	14.1 (0.3)	25.4 (0.4)	43.3 (0.4)	72 (0.3)	109.7 (0.4)	159.3 (0.3)	216 (0.3)
	$\widehat{\alpha}_2$		3.9 (0.1)	7.3 (0.2)	10.9 (0.4)	17.6 (0.5)	27.5 (0.5)	49 (0.5)	80.3 (0.4)	123.2 (0.4)	175.4 (0.3)
	$\widehat{\alpha}_3$		3.8 (0.1)	7.7 (0.3)	13.6 (0.4)	26 (0.4)	44.3 (0.4)	73.4 (0.4)	111.2 (0.4)	160.8 (0.3)	217.2 (0.3)
	$\widehat{\alpha}_4$		3.8 (0.1)	7.5 (0.3)	13 (0.4)	25.2 (0.4)	43.7 (0.4)	73.2 (0.4)	111.9 (0.3)	162.1 (0.3)	219.6 (0.3)
	$\widehat{\alpha}_5$		3.8 (0.1)	7.3 (0.3)	11.6 (0.4)	17.6 (0.6)	21.8 (0.6)	29.9 (0.7)	36.9 (0.7)	47.6 (0.7)	55.6 (0.7)
	$\widehat{\alpha}_6$		3.8 (0.1)	7.2 (0.2)	11 (0.4)	16.1 (0.6)	19.6 (0.6)	26.4 (0.7)	33.4 (0.7)	43.1 (0.7)	52.4 (0.7)
	$\widehat{\alpha}_7$		3.7 (0.1)	6.9 (0.2)	10.1 (0.3)	13.8 (0.5)	15.3 (0.5)	19.4 (0.7)	21.7 (0.7)	26.4 (0.7)	27.8 (0.7)
	$\widehat{\alpha}_8$		3.7 (0.1)	6.9 (0.2)	10.2 (0.3)	13.8 (0.5)	15.4 (0.6)	19 (0.7)	21.6 (0.7)	25.6 (0.7)	27.5 (0.7)
	$\widehat{\alpha}_1$	$p = 0.98$	4.1 (0.1)	7.7 (0.2)	11.7 (0.4)	16.8 (0.6)	24.2 (0.6)	43.3 (0.6)	72.3 (0.5)	113.6 (0.4)	163.6 (0.4)
	$\widehat{\alpha}_2$		5.8 (0.1)	11 (0.3)	14.7 (0.3)	18 (0.6)	20.3 (0.7)	31.5 (0.8)	53.8 (0.7)	88.6 (0.5)	135.2 (0.4)
	$\widehat{\alpha}_3$		5.3 (0.1)	9.6 (0.3)	13 (0.4)	18.1 (0.6)	25.7 (0.6)	45.4 (0.6)	74.9 (0.5)	117.2 (0.4)	168.6 (0.4)
	$\widehat{\alpha}_4$		5.3 (0.1)	9.7 (0.3)	13 (0.4)	17.9 (0.6)	25.5 (0.6)	45.4 (0.6)	75.8 (0.5)	118.4 (0.4)	170.7 (0.3)
	$\widehat{\alpha}_5$		5.6 (0.1)	11 (0.2)	15.4 (0.3)	18.4 (0.6)	20 (0.6)	23.5 (0.8)	25.1 (0.9)	28.6 (1)	27.1 (0.9)
	$\widehat{\alpha}_6$		5.6 (0.1)	11.1 (0.2)	15.6 (0.2)	18.3 (0.5)	20.6 (0.6)	23.2 (0.8)	24.8 (0.9)	27 (1)	26.2 (0.9)
	$\widehat{\alpha}_7$		5.6 (0.1)	11 (0.1)	15.5 (0.1)	18 (0.4)	19.5 (0.4)	21.9 (0.6)	22.6 (0.7)	23.2 (0.8)	19.4 (0.7)
	$\widehat{\alpha}_8$		5.6 (0.1)	11.1 (0.1)	15.7 (0.2)	18.2 (0.4)	20.4 (0.4)	22.5 (0.6)	23.2 (0.7)	23.1 (0.8)	19.6 (0.7)

B Pairwise margins for max-stable and asymptotic independence models

Recall Sections 2.3.2 and 2.4, where stationary parametric max-stable models and asymptotic independence models are discussed.

In this section, we use $\Phi(\cdot)$ and $\phi(\cdot)$ to denote, respectively, the cumulative distribution function (CDF) and the probability density function (PDF) of a standard normal random variable. In two dimensions, the CDF and PDF of a bivariate normal random vector with zero mean, unit variance and correlation ρ are denoted by $\Phi_2(\cdot;\rho)$ and $\phi_2(\cdot;\rho)$. Furthermore, $T_\nu(\cdot)$ and $t_\nu(\cdot)$ are used, respectively, for the CDF and PDF of a Student-t random variable with ν degrees of freedom.

B.1 Max-stable models

All max-stable models with unit Fréchet margins have bivariate cumulative distribution functions of the form

$$G(z_1, z_2) = \exp\{-V_{\mathscr{D}}(z_1, z_2)\},$$

where $V_{\mathscr{D}}$ is the underlying exponent measure for $\mathscr{D} = \{x, x+h\} \subset \mathscr{X}$. In the following lines, we shall drop the subscript $\mathscr{D}$ for simplicity. The censored contributions involve the partial derivatives of $G(z_1, z_2)$ with respect to z_1 and z_2, that is,

$$\begin{aligned}
\frac{\partial}{\partial z_1} G(z_1, z_2) &= -V_1(z_1, z_2)\exp\{-V(z_1, z_2)\},\\
\frac{\partial}{\partial z_2} G(z_1, z_2) &= -V_2(z_1, z_2)\exp\{-V(z_1, z_2)\},\\
\frac{\partial^2}{\partial z_1 \partial z_2} G(z_1, z_2) &= \{V_1(z_1, z_2)V_2(z_1, z_2) - V_{12}(z_1, z_2)\}\exp\{-V(z_1, z_2)\},
\end{aligned}$$

where we have written $V_1(z_1,z_2) = \partial V(z_1,z_2)/\partial z_1$, and so forth. In the next sections, we give the explicit forms for V, V_1, and V_{12} for the mainstream max-stable models. The function V_2 can be deduced from V_1 by symmetry of the arguments, interchanging z_1 and z_2.

B.1.1 Smith and Brown–Resnick models

The Smith model can be viewed as a special case of the Brown–Resnick model, so the same expressions hold for both models. For notational simplicity, let $f(z_1,z_2) = a/2 - \log(z_1/z_2)/a$, where $a = \sqrt{2\gamma(h)}$ for the Brown–Resnick, $\gamma(h)$ being the underlying semi-variogram, and $a = \sqrt{h^T\Sigma^{-1}h}$ for the Smith model, Σ being the underlying variance-covariance matrix. Then we have

$$\begin{aligned}
V(z_1,z_2) &= \frac{1}{z_1}\Phi\left\{f(z_1,z_2)\right\} + \frac{1}{z_2}\Phi\left\{f(z_2,z_1)\right\},\\
V_1(z_1,z_2) &= -\frac{1}{z_1^2}\left[\Phi\left\{f(z_1,z_2)\right\} + \frac{\phi\left\{f(z_1,z_2)\right\}}{a}\right] + \frac{1}{az_1z_2}\phi\left\{f(z_2,z_1)\right\},\\
V_{12}(z_1,z_2) &= -\frac{1}{az_1z_2}\left[\frac{\phi\left\{f(z_1,z_2)\right\}}{z_1}\left\{1-\frac{f(z_1,z_2)}{a}\right\} + \frac{\phi\left\{f(z_2,z_1)\right\}}{z_2}\left\{1-\frac{f(z_2,z_1)}{a}\right\}\right].
\end{aligned}$$

B.1.2 Schlather model with or without random set

The Schlather model without a random set can be viewed as a special case of the Schlather model with random set, so the same expressions hold for both models. We have

$$\begin{aligned}
V(z_1,z_2) &= \left(\frac{1}{z_1}+\frac{1}{z_2}\right)\left(1-\frac{\delta(h)}{2}\left[1-\frac{1}{z_1+z_2}\left\{z_1^2-2z_1z_2\rho(h)+z_2^2\right\}^{1/2}\right]\right),\\
V_1(z_1,z_2) &= \frac{1}{z_1^2}\left\{\frac{\delta(h)}{2}-1\right\} + \frac{\delta(h)}{2}\left\{\frac{\rho(h)}{z_1}-\frac{z_2}{z_1^2}\right\}\left\{z_1^2-2z_1z_2\rho(h)+z_2^2\right\}^{-1/2},\\
V_{12}(z_1,z_2) &= -\frac{\delta(h)}{2}\left\{1-\rho(h)^2\right\}\left\{z_1^2-2z_1z_2\rho(h)+z_2^2\right\}^{-3/2}.
\end{aligned}$$

In the expressions above, $\rho(h)$ is the underlying correlation function and $\delta(h)$ is the expected volume of overlap between the random set $\mathscr{A}$ and $\mathscr{A}+h$. For the Schlather model without random set, we set $\delta(h) \equiv 1$.

B.1.3 Extremal-t model

For simplicity, let $b = \sqrt{(\nu+1)/\{1-\rho(h)^2\}}$ and $f(z_1,z_2) = b\left\{(z_2/z_1)^{1/\nu} - \rho(h)\right\}$, where $\rho(h)$ is the underlying correlation function. Then we have

$$\begin{aligned}
V(z_1,z_2) &= \frac{1}{z_1}T_{\nu+1}\left\{f(z_1,z_2)\right\} + \frac{1}{z_2}T_{\nu+1}\left\{f(z_2,z_1)\right\},\\
V_1(z_1,z_2) &= -\frac{1}{z_1^2}T_{\nu+1}\left\{f(z_1,z_2)\right\},\\
V_{12}(z_1,z_2) &= -\frac{b}{\nu}t_{\nu+1}\left\{f(z_1,z_2)\right\}\frac{z_2^{1/\nu-1}}{z_1^{1/\nu+2}}\\
&= -\frac{\{1-\rho(h)^2\}^{(\nu+1)/2}}{\nu\sqrt{\pi}}\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}(z_1z_2)^{1/\nu-1}\left\{z_1^{2/\nu} - 2(z_1z_2)^{1/\nu}\rho(h) + z_2^{2/\nu}\right\}^{-(\nu+2)/2},
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function.

B.2 Asymptotic independence models

B.2.1 Gaussian copula model

The bivariate cumulative distribution function of the Gaussian copula model with unit Fréchet margins is

$$G(z_1,z_2) = \Phi_2\left\{s(z_1),s(z_2);\rho(h)\right\}, \tag{B.1}$$

where $s(z) = \Phi^{-1}\{\exp(-1/z)\}$, $z>0$. The partial derivatives of $G(z_1,z_2)$ with respect to z_1 and z_2 may be written as

$$\begin{aligned}
\frac{\partial}{\partial z_1}G(z_1,z_2) &= s'(z_1)\frac{\partial}{\partial z_1}\Phi_2\left\{s(z_1),s(z_2);\rho(h)\right\},\\
\frac{\partial}{\partial z_2}G(z_1,z_2) &= s'(z_2)\frac{\partial}{\partial z_2}\Phi_2\left\{s(z_1),s(z_2);\rho(h)\right\},\\
\frac{\partial^2}{\partial z_1\partial z_2}G(z_1,z_2) &= s'(z_1)s'(z_2)\phi_2\left\{s(z_1),s(z_2);\rho(h)\right\},
\end{aligned}$$

where

$$s'(z) = \frac{\exp(-1/z)}{z^2\phi[\Phi\{\exp(-1/z)\}]},$$

and

$$\begin{aligned}\frac{\partial}{\partial z_1}\Phi_2\left(z_1,z_2;\rho\right) &= \phi(z_1)\Phi\left\{\frac{z_2-\rho z_1}{(1-\rho^2)^{1/2}}\right\},\\ \frac{\partial}{\partial z_2}\Phi_2\left(z_1,z_2;\rho\right) &= \phi(z_2)\Phi\left\{\frac{z_1-\rho z_2}{(1-\rho^2)^{1/2}}\right\}.\end{aligned}$$

B.2.2 Inverted max-stable models

The cumulative distribution function of inverted max-stable models can be expressed in terms of the corresponding max-stable process as

$$G(z_1,z_2) = -1+\exp(-1/z_1)+\exp(-1/z_2)+\tilde{G}\{s(z_1),s(z_2)\},$$

where $\tilde{G}$ is the bivariate distribution function of the latent max-stable process, and $s(z) = -1/\log\{1-\exp(-1/z)\}$, $z>0$. Let $\tilde{G}_1,\tilde{G}_2,\tilde{G}_{12}$ denote the partial derivatives of $\tilde{G}$ with respect to z_1 and/or z_2. The partial derivatives of $G(z_1,z_2)$ with respect to z_1 and z_2 are

$$\begin{aligned}\frac{\partial}{\partial z_1}G(z_1,z_2) &= \exp(-1/z_1)z_1^{-2}+s'(z_1)\tilde{G}_1\{s(z_1),s(z_2)\},\\ \frac{\partial}{\partial z_2}G(z_1,z_2) &= \exp(-1/z_2)z_2^{-2}+s'(z_2)\tilde{G}_2\{s(z_1),s(z_2)\},\\ \frac{\partial^2}{\partial z_1\partial z_2}G(z_1,z_2) &= s'(z_1)s'(z_2)\tilde{G}_{12}\{s(z_1),s(z_2)\},\end{aligned}$$

where

$$s'(z) = \left\{1-\exp(-1/z)\right\}^{-1}z^{-2}\left[\log\left\{1-\exp(-1/z)\right\}\right]^{-2}.$$

B.2.3 Hybrid models

Hybrid models are defined in terms of a max-stable model, with bivariate CDF $G^1(z_1,z_2)$, an asymptotically independent model, with bivariate CDF $G^2(z_1,z_2)$, and a mixture proportion $a\in[0,1]$ which determines the importance of each submodel. Their bivariate margins and partial derivatives may be expressed as

$$\begin{aligned}G(z_1,z_2) &= G^1\left(\frac{z_1}{a},\frac{z_2}{a}\right)G^2\left(\frac{z_1}{1-a},\frac{z_2}{1-a}\right),\\ \frac{\partial}{\partial z_1}G(z_1,z_2) &= \frac{1}{a}G^1_1\left(\frac{z_1}{a},\frac{z_2}{a}\right)G^2\left(\frac{z_1}{1-a},\frac{z_2}{1-a}\right)+\frac{1}{1-a}G^1\left(\frac{z_1}{a},\frac{z_2}{a}\right)G^2_1\left(\frac{z_1}{1-a},\frac{z_2}{1-a}\right),\\ \frac{\partial}{\partial z_2}G(z_1,z_2) &= \frac{1}{a}G^1_2\left(\frac{z_1}{a},\frac{z_2}{a}\right)G^2\left(\frac{z_1}{1-a},\frac{z_2}{1-a}\right)+\frac{1}{1-a}G^1\left(\frac{z_1}{a},\frac{z_2}{a}\right)G^2_2\left(\frac{z_1}{1-a},\frac{z_2}{1-a}\right),\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial z_1 \partial z_2} G(z_1, z_2) &= \frac{1}{a^2} G^1_{12}\left(\frac{z_1}{a}, \frac{z_2}{a}\right) G^2\left(\frac{z_1}{1-a}, \frac{z_2}{1-a}\right) + \frac{1}{a(1-a)} G^1_1\left(\frac{z_1}{a}, \frac{z_2}{a}\right) G^2_2\left(\frac{z_1}{1-a}, \frac{z_2}{1-a}\right) \\ &+ \frac{1}{a(1-a)} G^1_2\left(\frac{z_1}{a}, \frac{z_2}{a}\right) G^2_1\left(\frac{z_1}{1-a}, \frac{z_2}{1-a}\right) + \frac{1}{(1-a)^2} G^1\left(\frac{z_1}{a}, \frac{z_2}{a}\right) G^2_{12}\left(\frac{z_1}{1-a}, \frac{z_2}{1-a}\right),\end{aligned}$$

where the subscripts of G^1 and G^2 denote the partial derivatives with respect to z_1 and/or z_2. Different classes of models can be obtained by choosing special max-stable and asymptotic independence families.

C Consistency and efficiency of pairwise and triplewise likelihood estimators for the Brown–Resnick process

In Section 4.3.2, we study the efficiency of the maximum triplewise likelihood estimator $\widehat{\psi}_3$ compared to the pairwise counterpart $\widehat{\psi}_2$ for the Brown–Resnick process. In our simulation study, we simulated n independent replicates of the isotropic Brown–Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$, $\lambda > 0$, $\alpha \in (0,2]$, and compared the empirical variability of $\widehat{\psi}_3$ and $\widehat{\psi}_2$ for different values of range parameter λ, smoothness parameter α, number of replicates n and number of sites S.

Figure C.1 suggests that in a typical situation, with $\lambda = 28$, $\alpha = 1$ and $S = 20$, both estimators $\widehat{\psi}_2$ and $\widehat{\psi}_3$ estimate ψ consistently as $n \to \infty$, confirming the theoretical results of Section 3.1.3, whereas Figure C.2 illustrates the dramatic efficiency improvement of $\widehat{\psi}_3$ compared to $\widehat{\psi}_2$ when $\alpha = 2$.

Appendix C. Consistency and efficiency of pairwise and triplewise likelihood estimators for the Brown–Resnick process

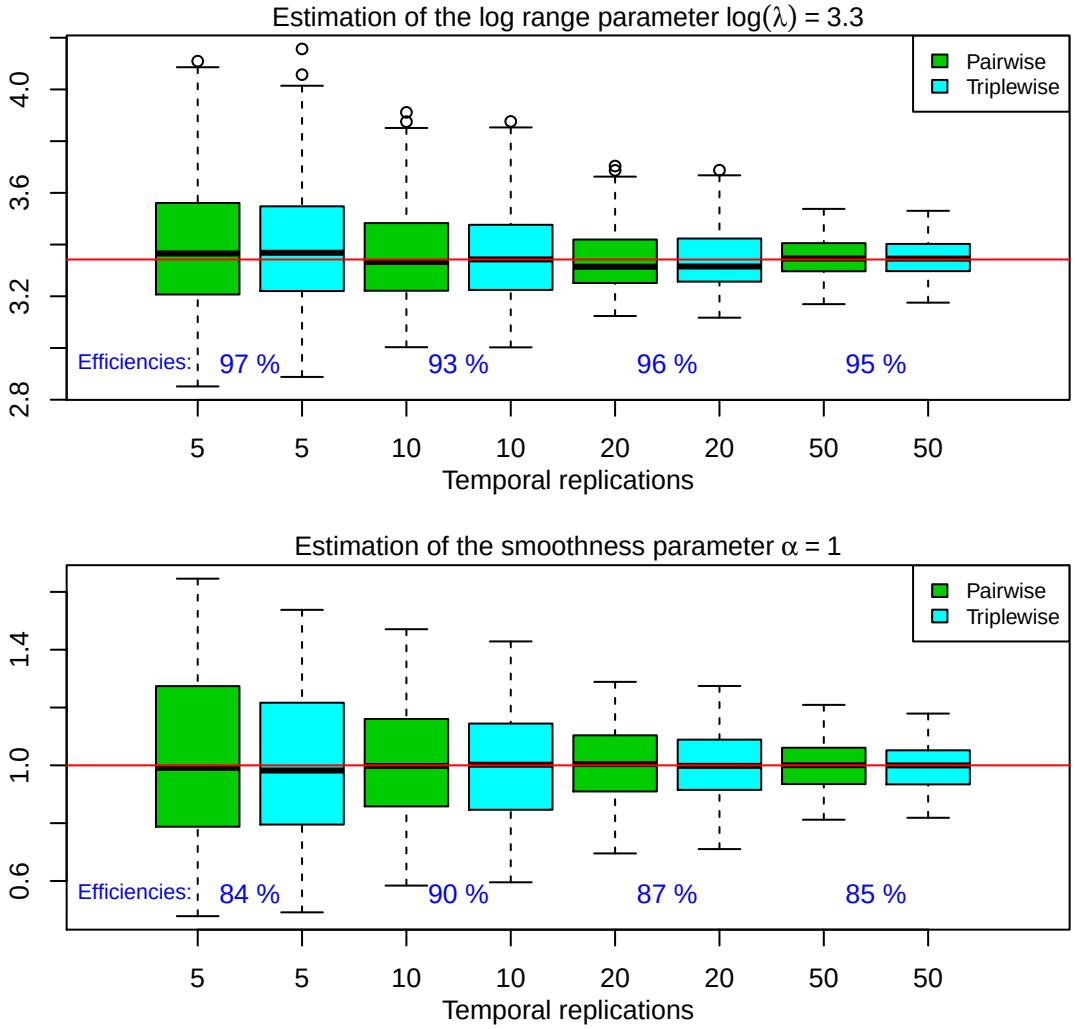

Figure C.1: Boxplots of the 300 independent estimates of the log-range parameter (top) and smooth parameter (bottom), as the number of temporal replicates n increases. Green and turquoise boxes correspond respectively to $\widehat{\psi}_2$ and $\widehat{\psi}_3$. The horizontal red lines correspond to the true values $\log(\lambda) \approx 3.3$ (that is $\lambda = 28$) and $\alpha = 1$. The relative efficiencies $\widehat{\mathrm{RE}}_\lambda$ and $\widehat{\mathrm{RE}}_\alpha$ are also reported.

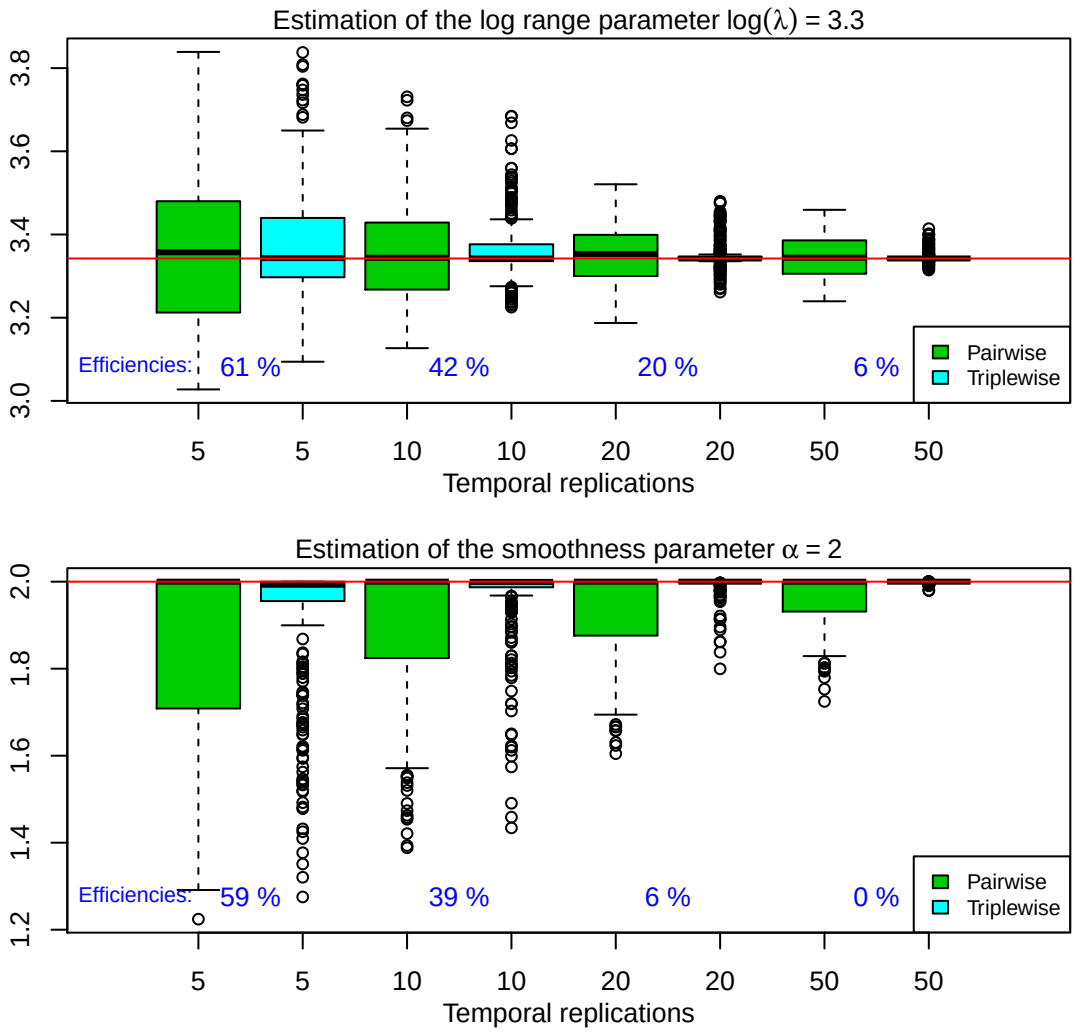

Figure C.2: Boxplots of the 300 independent estimates of the log-range parameter (top) and smoothness parameter (bottom), as the number of temporal replicates n increases. Green and turquoise boxes correspond respectively to $\widehat{\psi}_2$ and $\widehat{\psi}_3$. The horizontal red lines correspond to the true values $\log(\lambda) \approx 3.3$ (that is $\lambda = 28$) and $\alpha = 2$ (Smith model). The relative efficiencies $\widehat{\mathrm{RE}}_\lambda$ and $\widehat{\mathrm{RE}}_\alpha$ are also reported.

D Performance of composite Stephenson–Tawn likelihood estimators for the Brown–Resnick process

In Section 4.4.2, we conduct a simulation study to shed some light on the loss of efficiency of composite likelihood methods in an extreme value framework. For that purpose, we simulated $n = 20$ independent Brown–Resnick processes with variogram $2\gamma(h) = (\|h\|/\lambda)^{\alpha}$, $\lambda > 0$, $\alpha \in (0,2]$ at $S = 10$ random sites in $[0,100]^2$ and estimated the variogram parameters with the composite Stephenson–Tawn likelihood estimator maximizing (4.11).

We repeated this procedure 300 times, with different simulation settings, namely $\lambda = 14, 28, 42$ and $\alpha = 0.5, 1, 1.5, 1.9, 1.95, 1.98$, and considered D-dimensional marginal likelihood estimators with $D = 2, 3, \ldots, 10$. Figures D.1 and D.2 report the marginal relative efficiencies for λ and α respectively, while Figure 4.5 reports the global relative efficiency.

Appendix D. Performance of composite Stephenson–Tawn likelihood estimators for the Brown–Resnick process

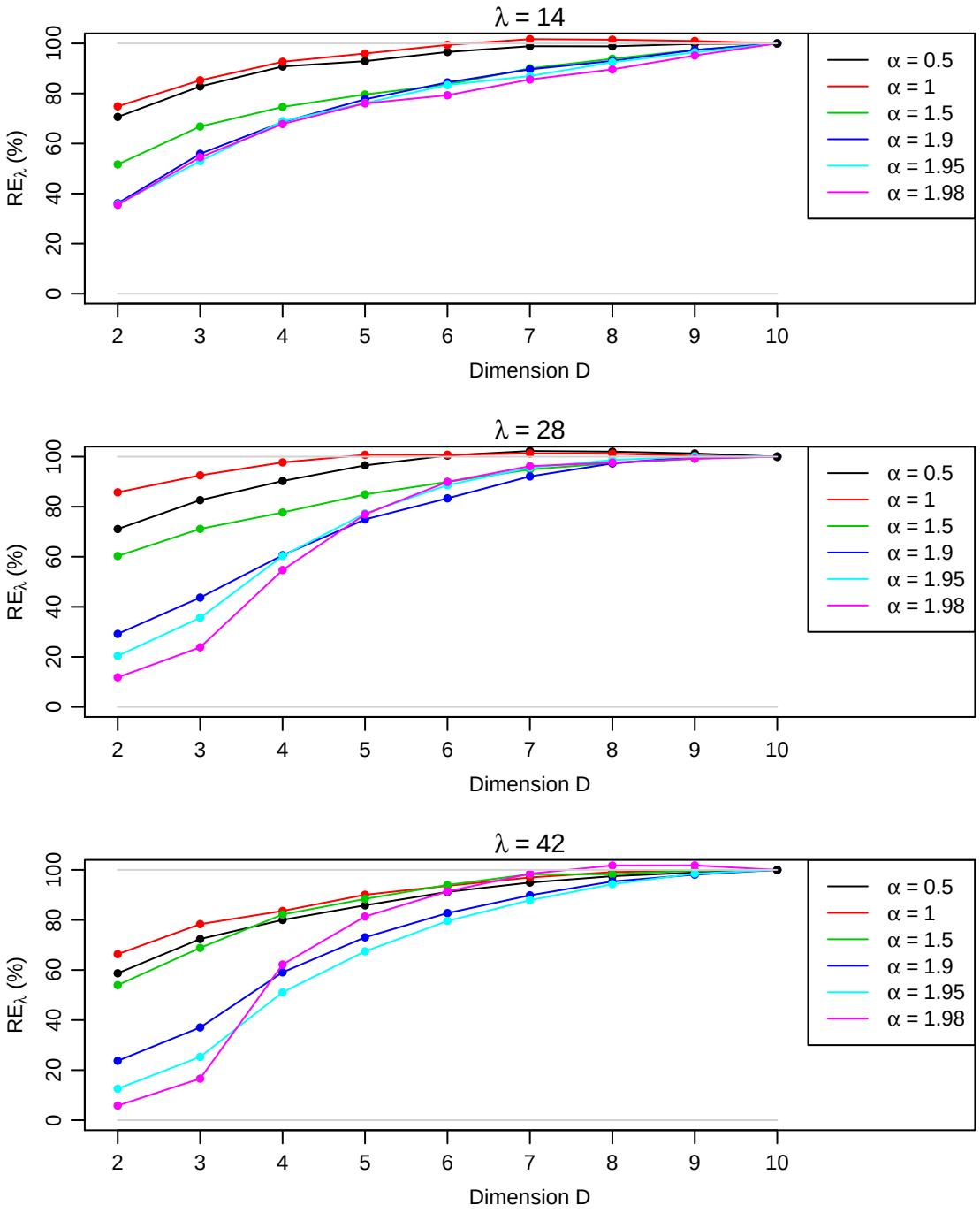

Figure D.1: Marginal relative efficiency RE_λ of the maximum D-dimensional marginal Stephenson–Tawn likelihood estimator (with respect to the maximum full likelihood estimator), based on 300 replications of the Brown-Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$. We used $\lambda = 14$ (top), 28 (middle), 42 (bottom), and $\alpha = 0.5, 1, 1.5, 1.9, 1.95, 1.98$ (different colours).

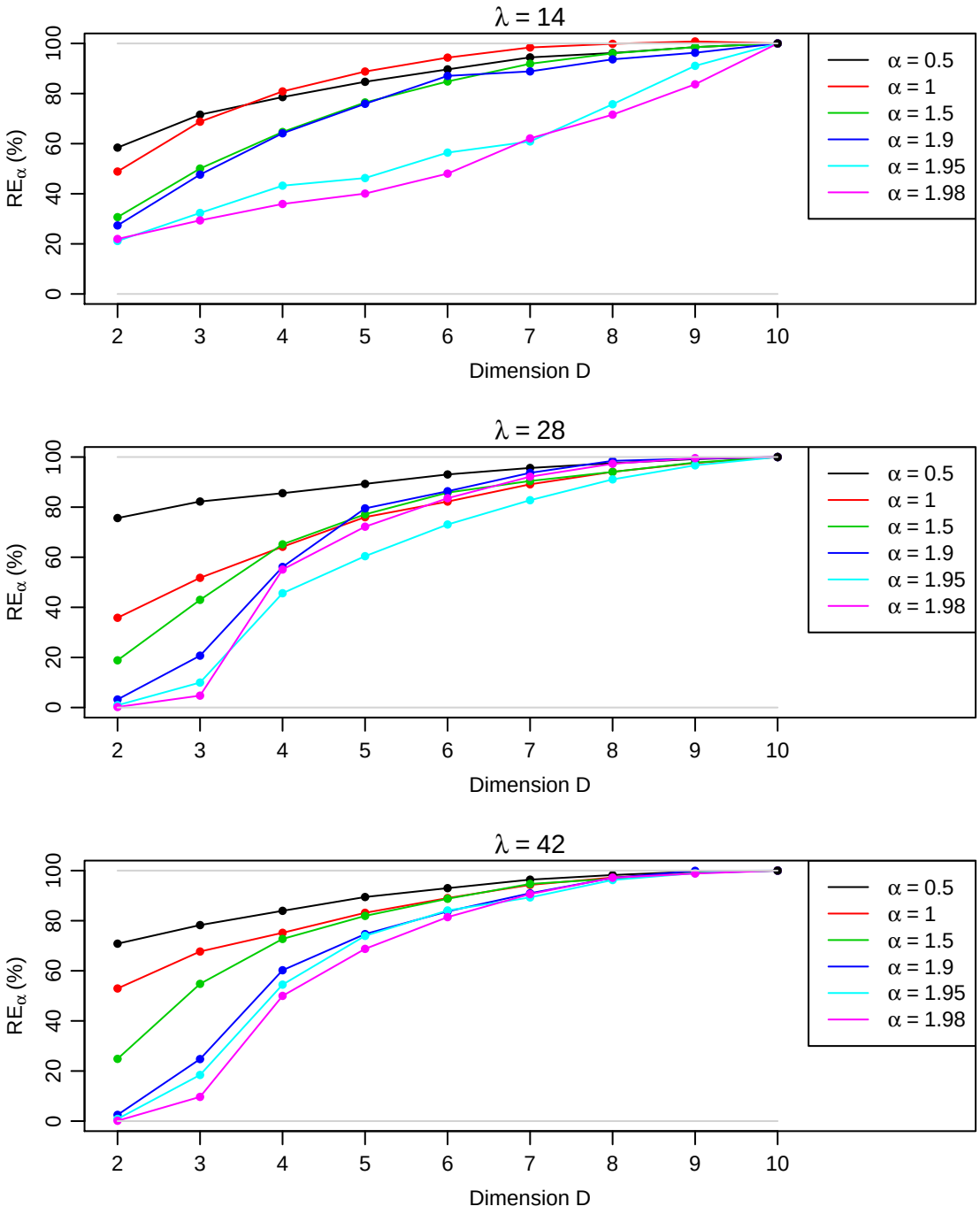

Figure D.2: Marginal relative efficiency RE_α of the maximum D-dimensional marginal Stephenson–Tawn likelihood estimator (with respect to the maximum full likelihood estimator), based on 300 replications of the Brown-Resnick process with variogram $2\gamma(h) = (\|h\|/\lambda)^\alpha$. We used $\lambda = 14$ (top), 28 (middle), 42 (bottom), and $\alpha = 0.5, 1, 1.5, 1.9, 1.95, 1.98$ (different colours).

E Additional diagnostic plots of extremal dependence for the rainfall data

In Section 5.3.2, we fit extremal dependence models to the rainfall data in Figure 5.2. In this appendix, we show additional empirical and fitted diagnostic plots, namely space-time extremal coefficients, coefficients of tail dependence, and coefficients χ and $\overline{\chi}$ for all pairs for stations. In each Figure, the black lines join the empirical diagnostics, whereas the colored curves correspond to the fitted diagnostics for model (5.12) (blue), and Models 1 (red), 2 (green), 3 (orange) and 4 (purple) from §5.3.2.2.

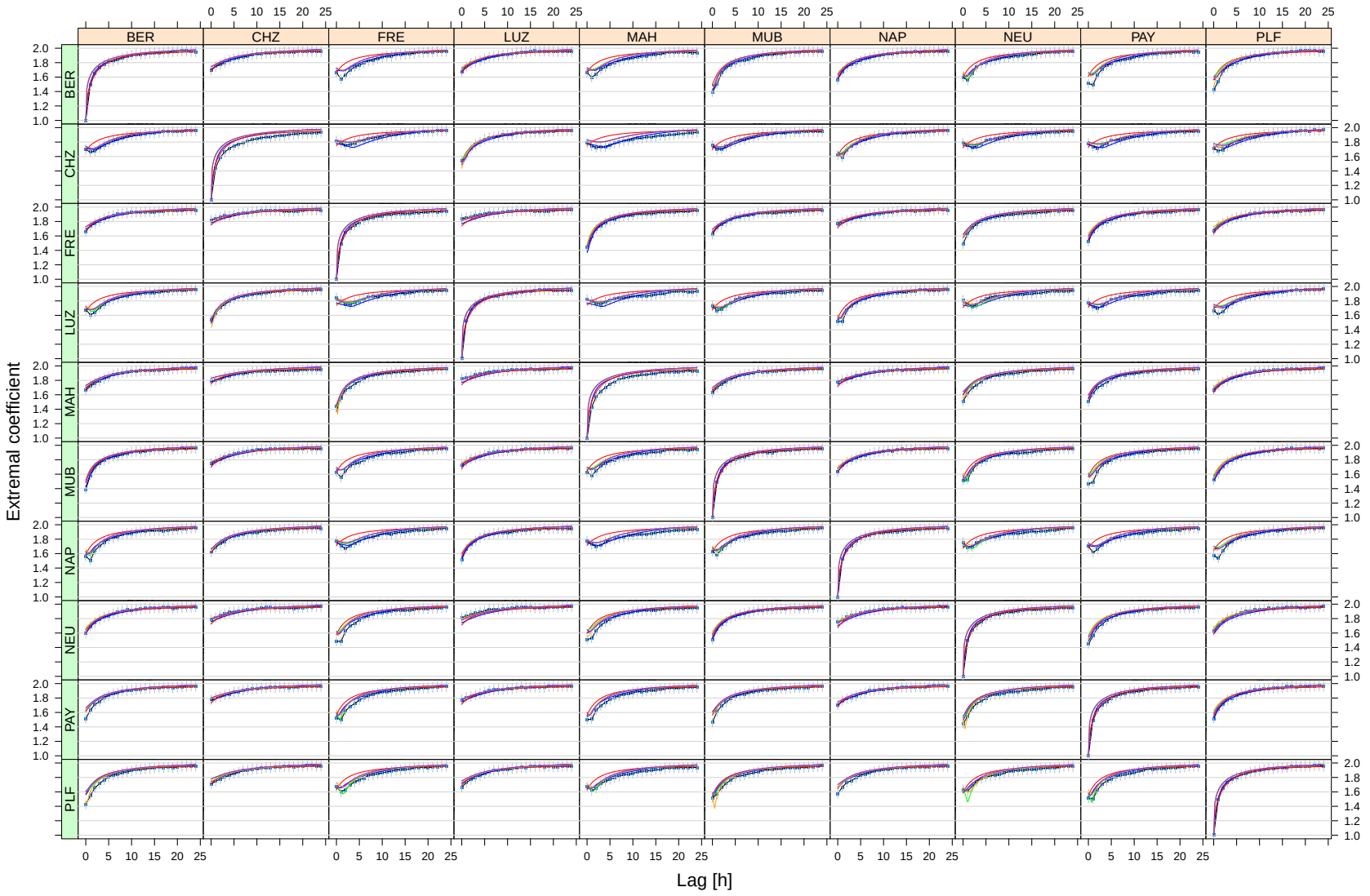

Figure E.1: Extremal coefficients. See the caption of Figure 5.6.

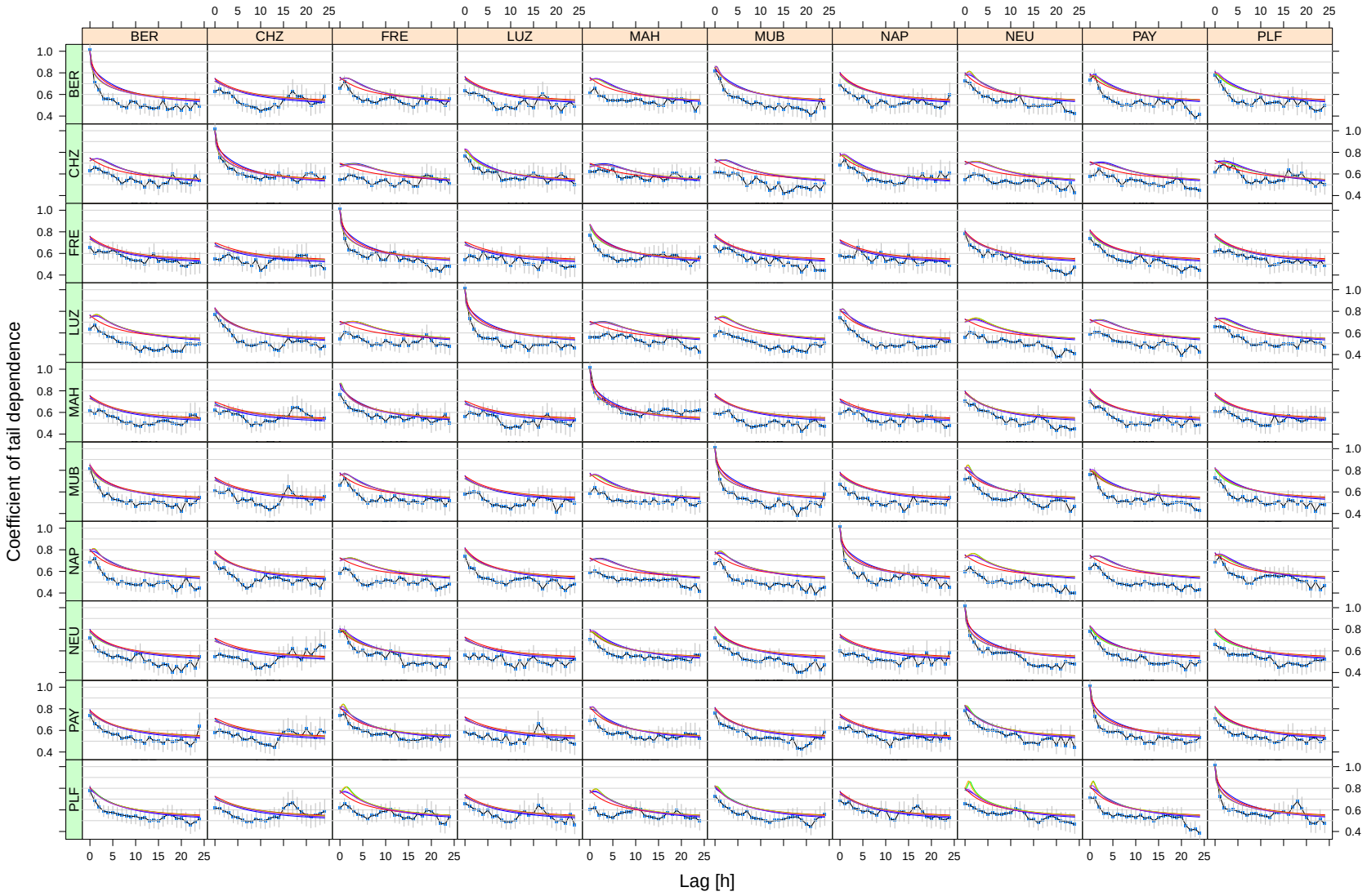

Figure E.2: Coefficients of tail dependence. See the caption of Figure 5.11.

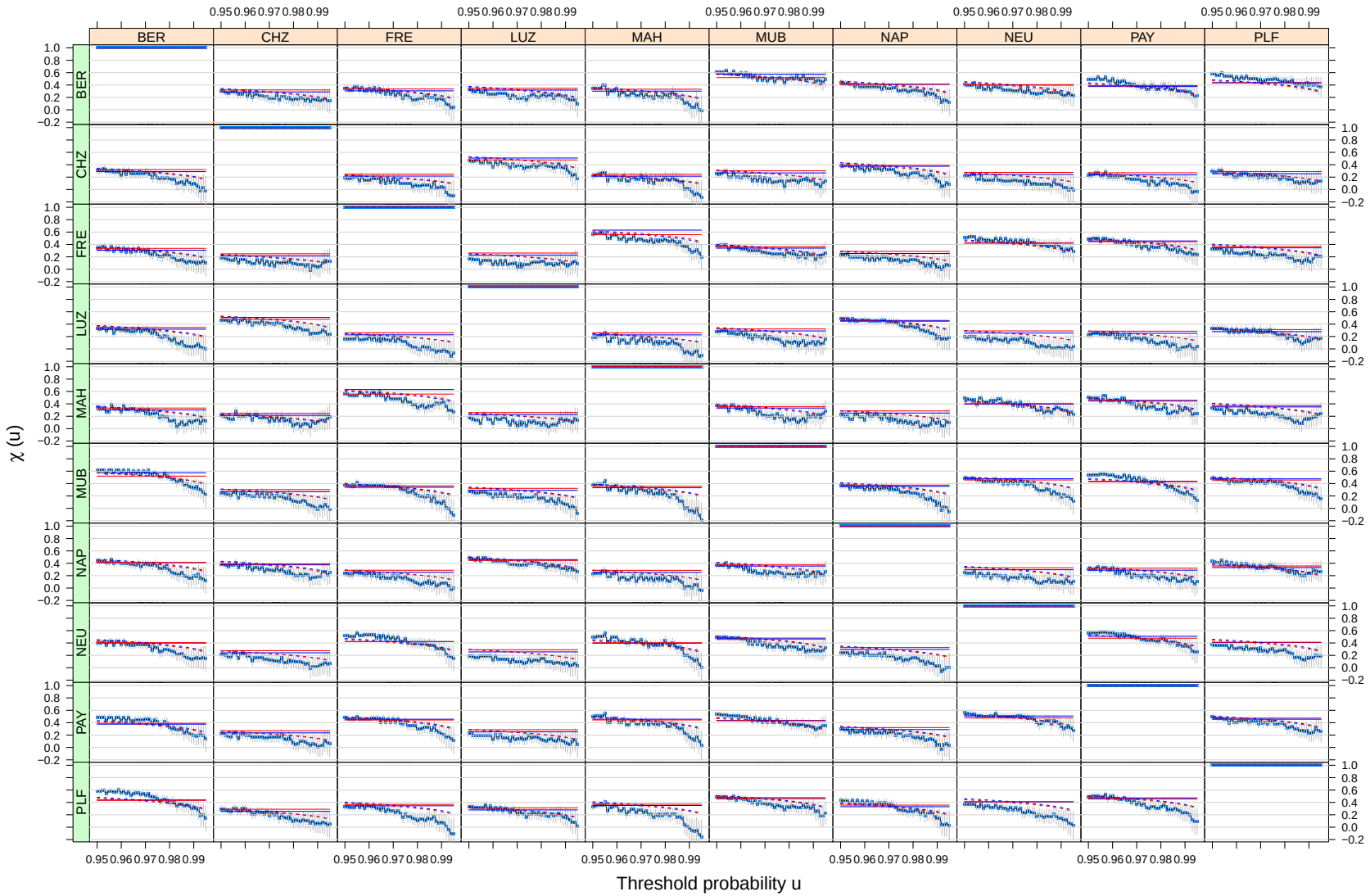

Figure E.3: Coefficients $\chi_{(h_t,h_s)}(u)$. See the caption of Figure 5.12.

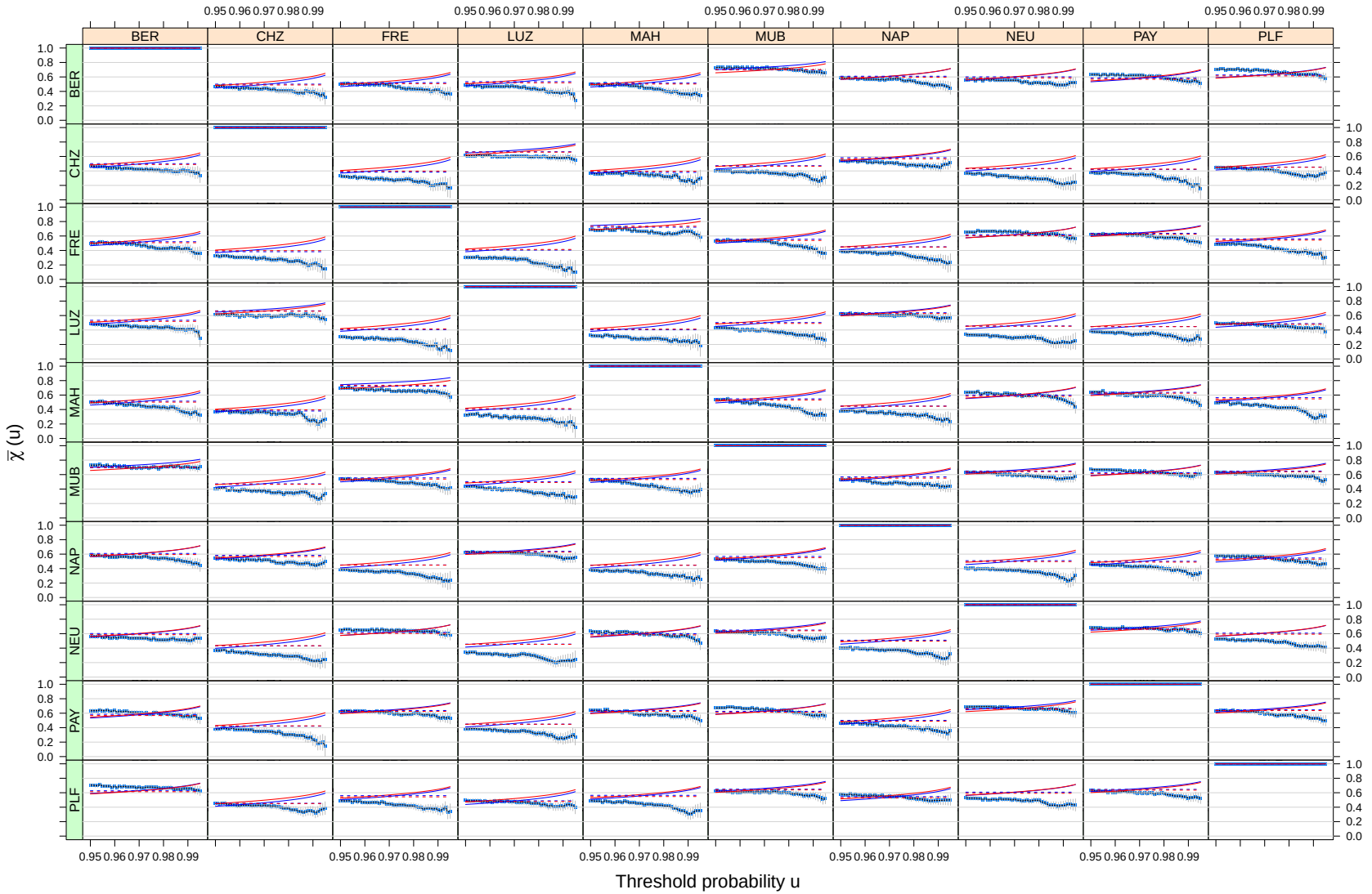

Figure E.4: Coefficients $\overline{\chi}_{(h_t,h_s)}(u)$. See the caption of Figure 5.13.

F Computation of the volume of overlap $\delta(h_s, h_t)$

In §5.3.2.1, the coefficient $\delta(h_s, h_t)$ is defined as $\mathrm{E}[|\mathscr{A} \cap \{(h_s, h_t) + \mathscr{A}\}|]/\mathrm{E}(|\mathscr{A}|)$, where $\mathscr{A}$ is a random tilted cylinder in $\mathscr{X} = \mathscr{S} \times \mathscr{T} = \mathbb{R}^2 \times \mathbb{R}_+$ (see Figure 5.7), and $(h_s, h_t) \in \mathscr{X}$. The random set $\mathscr{A}$ is interpreted as a "cloud" in $\mathscr{X}$ may be defined in terms of its radius R, its lifetime L and its velocity $V = (V_1, V_2)$; recall §5.3.2.1.

If the cylinder were vertical (zero wind velocity), the volume of overlap would simply be the product of the area of overlap between two discs distant by $\|h_s\|$ and the corresponding height, the storm duration minus h_t. Moreover, a good linear approximation to the area of overlap of two discs of radius R distant by $\|h_s\|$ is $\pi R^2 \max\{0, 1 - \|h_s\|/(2R)\}$ (Davison & Gholamrezaee, 2012). Therefore, for a vertical cylinder $\mathscr{A}$, $|\mathscr{A} \cap \{(h_s, h_t) + \mathscr{A}\}|$ can be approximated by

$$\pi R^2 \left(1 - \frac{\|h_s\|}{2R}\right)_+ (L - h_t)_+,$$

where $a_+ = \max(0, a)$. When the cloud is moving, giving a tilted cylinder, a geometric argument shows that in the general case, the volume of overlap is transformed to

$$|\mathscr{A} \cap \{(h_s, h_t) + \mathscr{A}\}| \doteq \pi R^2 \left(1 - \frac{d^*}{2R}\right)_+ (L - h_t)_+,$$

where $d^* = \left[\|h_s\|^2 + h_t^2(V_1^2 + V_2^2) - 2\|h_s\| h_t \{V_1 \cos(\nu) + V_2 \sin(\nu)\}\right]^{1/2}$, and where $\nu = \arctan(h_{s;1}/h_{s;2})$ is the angle between the stations with respect to a reference axis in the West-East direction. Careful checking suggests that this provides adequate approximations for the values of h_s and h_t and the parameter values used in the pairwise likelihood in our application.

In order to compute the coefficient $\delta(h_s, h_t)$, which depends upon the spatial distance $\|h_s\|$, the temporal lag $|h_t|$ and the orientation of the stations ν, we need to obtain

the expected volume of overlap $\mathrm{E}[|\mathscr{A} \cap \{(h_s, h_t) + \mathscr{A}\}|]/\mathrm{E}(|\mathscr{A}|)$, by putting tractable distributions on R, L, and $V = (V_1, V_2)$. We choose to set

(a) $R \sim \text{Gamma}(m_R/k_R, k_R)$ (with mean m_R km),

(b) $V \sim \mathscr{N}_2(m_V, \Omega)$ (km/hour), with $m_V = (m_1, m_2)^T$ and $\Omega = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 \end{pmatrix}$,

(c) $L \sim \text{Gamma}(m_L/k_L, k_L)$ (with mean m_L hours),

and we assume that R, L and V are mutually independent. To compute this expectation, note first that $(L-h_t)_+$ can be integrated out analytically. Second, by conditioning on V, it is possible to integrate over R as well. We can then reduce the full computation to this single expectation with respect to $V = (V_1, V_2)$:

$$\begin{aligned} \delta(h_s, h_t) &= \mathrm{E}_V\left\{\Pr\left(G_{m_R/k_R;k_R+2} > d^*/2\right) - \frac{d^* k_R}{2(k_R+1)m_R}\Pr\left(G_{m_R/k_R;k_R+1} > d^*/2\right)\right\} \\ &\quad \times \underbrace{\left\{\Pr\left(G_{m_L/k_L;k_L+1} > h_t\right) - \frac{h_t}{m_L}\Pr\left(G_{m_L/k_L;k_L} > h_t\right)\right\}}_{=C}, \end{aligned} \tag{F.1}$$

where $G_{\theta;k}$ is a gamma random variable with scale parameter θ and shape parameter k; its mean equals $m = \theta k$. The expectation in (F.1) does not have a closed form, but it can be remarkably well approximated by a function of the form $\exp\left[-a\left\{(V_1-\mu_1)^2 + (V_2-\mu_2)^2\right\}\right]$, where a is a real number that does not depend upon $V = (V_1, V_2)$ and can be estimated with a few points by least squares, and where $\mu_1 = \|h_s\|\cos(v)/h_t$ and $\mu_2 = \|h_s\|\sin(v)/h_t$. Therefore, we have

$$\begin{aligned} \delta(h_s, h_t) &\approx C \times \mathrm{E}_V\left(\exp\left[-a\left\{(V_1-\mu_1)^2 + (V_2-\mu_2)^2\right\}\right]\right) \\ &= C\int_{\mathbb{R}^2} \exp\left[-a\left\{(v_1-\mu_1)^2 + (v_2-\mu_2)^2\right\}^{1/2}\right] \\ &\quad \times \frac{C}{2\pi\det(\Omega)^{1/2}} \exp\left\{-\frac{1}{2}(v_1-m_1, v_2-m_2)\Omega^{-1}(v_1-m_1, v_2-m_2)^T\right\} \mathrm{d}v_1\mathrm{d}v_2 \\ &= \frac{C}{2\pi\det(\Omega)^{1/2}}\int_{\mathbb{R}^2} \exp\Big[-a\left\{(v_1-\mu_1)^2 + (v_2-\mu_2)^2\right\}^{1/2} \\ &\quad -\frac{}{2\det(\Omega)}\left\{(v_1-m_1)^2\sigma_2^2 - 2(v_1-m_1)(v_2-m_2)\sigma_1\sigma_2\rho_{12} + (v_2-m_2)^2\sigma_1^2\right\}\Big] \mathrm{d}v_1\mathrm{d}v_2 \\ &= \frac{C}{2\pi\det(\Omega)^{1/2}}\int_0^{2\pi}\mathrm{d}\xi\int_{\mathbb{R}_+} r\exp\left[-ar - \frac{1}{2\det(\Omega)}\left\{r^2 a(\xi) + r\,b(\xi) + c(\xi)\right\}\right]\mathrm{d}r \qquad \text{(F.2)} \\ &= \frac{C}{(2\pi)^{1/2}}\int_0^{2\pi}\frac{1}{\sqrt{a(\xi)}}\exp\left[-\frac{1}{2\sigma(\xi)^2}\left\{\frac{c(\xi)}{a(\xi)} - \mu(\xi)^2\right\}\right]\mathrm{d}\xi \end{aligned}$$

$$
\begin{aligned}
& \times \int_{\mathbb{R}_+} r \frac{1}{\sqrt{2\pi}\sigma(\xi)} \exp\left[-\frac{1}{2}\left\{\frac{r-\mu(\xi)}{\sigma(\xi)}\right\}^2\right] \mathrm{d}r \\
= \; & \frac{C}{2\pi}\int_0^{2\pi} \frac{1}{\sqrt{a(\xi)}} \exp\left[-\frac{1}{2\sigma(\xi)^2}\left\{\frac{c(\xi)}{a(\xi)} - \mu(\xi)^2\right\}\right] \\
& \times \left(\sigma(\xi)\exp\left\{-\frac{1}{2}\frac{\mu(\xi)^2}{\sigma(\xi)^2}\right\} + \sqrt{2\pi}\mu(\xi)\left[1-\Phi\left\{-\frac{\mu(\xi)}{\sigma(\xi)}\right\}\right]\right)\mathrm{d}\xi, \qquad \text{(F.3)}
\end{aligned}
$$

where $\Phi(\cdot)$ is the normal cumulative distribution function and

$$
\begin{aligned}
a(\xi) &= \cos^2(\xi)\sigma_2^2 + \sin^2(\xi)\sigma_1^2 - 2\cos(\xi)\sin(\xi)\sigma_1\sigma_2\rho_{12}, \\
b(\xi) &= 2\cos(\xi)(\mu_1 - m_1)\sigma_2^2 + 2\sin(\xi)(\mu_2 - m_2)\sigma_1^2 - 2\cos(\xi)(\mu_2 - m_2)\sigma_1\sigma_2\rho_{12} \\
&\quad -2\sin(\xi)(\mu_1 - m_1)\sigma_1\sigma_2\rho_{12}, \\
c(\xi) &= (\mu_1 - m_1)^2\sigma_2^2 + (\mu_2 - m_2)^2\sigma_1^2 - 2(\mu_1 - m_1)(\mu_2 - m_2)\sigma_1\sigma_2\rho_{12}, \\
\mu(\xi) &= -\frac{b(\xi)}{2a(\xi)} - \frac{a\det(\Omega)}{a(\xi)}, \\
\sigma(\xi) &= \sqrt{\det(\Omega)/|a(\xi)|}, \\
\det(\Omega) &= \sigma_1^2\sigma_2^2(1-\rho_{12}).
\end{aligned}
$$

Expression (F.2) was computed with a straightforward change of variables $v_1 = r\cos(\xi) + \mu_1$, $v_2 = r\sin(\xi) + \mu_2$, and expression (F.3) stems from the properties of the normal cumulative distribution function. Since the integral (F.3) is impossible to handle analytically, we can use a finite approximation to estimate $\delta(h_s, h_t)$, based on 100 points equi-spaced in the interval $[0, 2\pi]$. The approximation seems to be adequate when $\sigma_1^2, \sigma_2^2 > 5$, which we impose in the R optimization routine.

G Trivariate extremal coefficients for model (5.12)

From equation (2.22), we know that the multivariate extremal coefficient in dimension D is

$$\theta_D(\boldsymbol{h}) = V_{\mathscr{D}}(1,\ldots,1) = \mathrm{E}\left[\max_{i=1,\ldots,D}\{W(x_i)\}\right].$$

This takes values between 1 and D, ranging from complete dependence to asymptotic independence. Therefore, the extremal coefficient of order $D = 3$ is

$$\theta_3(\boldsymbol{h}) = \mathrm{E}\left[\max\{W(x_1), W(x_2), W(x_3)\}\right],$$

where, for model (5.12), $W(x) \propto \max\{0, \varepsilon(x)\}\, I_{\mathscr{A}}(x - X)$, $x = (s,t) \in \mathscr{X} = \mathscr{S} \times \mathscr{T}$, $\varepsilon(x)$ being an isotropic Gaussian random field with zero mean, unit variance and correlation function $\rho(\cdot)$ and $I_{\mathscr{A}}(\cdot)$ being the indicator that the point $x - X$ belongs to a random set $\mathscr{A}$ (where X is a Poisson process of unit rate in $\mathscr{X}$). The proportionality constant is such that $W(x)$ has mean 1, so it must be

$$\begin{aligned}\frac{1}{\mathrm{E}\left[\max\{0,\varepsilon(x)\}\, I_{\mathscr{A}}(x-X)\right]} &= \frac{1}{\mathrm{E}\left[\max\{0,\varepsilon(x)\}\right]\mathrm{E}\{I_{\mathscr{A}}(x-X)\}} \\ &= \frac{\sqrt{2\pi}}{\Pr(x - X \in \mathscr{A})} = \frac{\sqrt{2\pi}|\mathscr{X}|}{\mathrm{E}\{|(x-\mathscr{A})\cap\mathscr{X}|\}} \approx \frac{\sqrt{2\pi}|\mathscr{X}|}{\mathrm{E}\{|\mathscr{A}|\}}.\end{aligned}$$

Below we use for simplicity the notation $W_1 = W(x_1)$, $\varepsilon_1 = \varepsilon(x_1)$, $I_1 = I_{\mathscr{A}}(x_1 - X)$, $I_{1;2;-} = I\{x_1 - X \in \mathscr{A}$ and $x_2 - X \in \mathscr{A}$ and $x_3 - X \notin \mathscr{A}\}$ and so forth. Then, assuming that x_1, x_2, x_3 are not too distant from each other and that $\mathscr{X}$ is large compared to $\mathscr{A}$, so that $\Pr(x_i - X \in \mathscr{A})$ are similar for $i = 1,2,3$, the required extremal coefficient is

$$\begin{aligned}\theta_3(\boldsymbol{h}) &= \mathrm{E}\{\max(W_1, W_2, W_3)\} \\ &= \frac{\sqrt{2\pi}}{\Pr(x_1 - X \in \mathscr{A})}\mathrm{E}\{\max(0, \varepsilon_1 I_1, \varepsilon_2 I_2, \varepsilon_3 I_3)\}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{2\pi}}{\Pr(x_1 - X \in \mathscr{A})} \Big[\mathrm{E}\left\{\max(0, \varepsilon_1, \varepsilon_2, \varepsilon_3)\, I_{1;2;3}\right\} + \mathrm{E}\left\{\max(0, \varepsilon_1, \varepsilon_2)\, I_{1;2;-}\right\} \\
&\quad + \mathrm{E}\left\{\max(0, \varepsilon_1, \varepsilon_3)\, I_{1;-;3}\right\} + \mathrm{E}\left\{\max(0, \varepsilon_2, \varepsilon_3)\, I_{-;2;3}\right\} \\
&\quad + \mathrm{E}\left\{\max(0, \varepsilon_1)\, I_{1;-;-}\right\} + \mathrm{E}\left\{\max(0, \varepsilon_2)\, I_{-;2;-}\right\} + \mathrm{E}\left\{\max(0, \varepsilon_3)\, I_{-;-;3}\right\} \Big] \\
&= \Pr(x_2 - X \in \mathscr{A}, x_3 - X \in \mathscr{A} \mid x_1 - X \in \mathscr{A}) \sqrt{2\pi} \mathrm{E}\{\max(0, \varepsilon_1, \varepsilon_2, \varepsilon_3)\} \\
&\quad + \Pr(x_2 - X \in \mathscr{A}, x_3 - X \notin \mathscr{A} \mid x_1 - X \in \mathscr{A}) \sqrt{2\pi} \mathrm{E}\{\max(0, \varepsilon_1, \varepsilon_2)\} \\
&\quad + \Pr(x_2 - X \notin \mathscr{A}, x_3 - X \in \mathscr{A} \mid x_1 - X \in \mathscr{A}) \sqrt{2\pi} \mathrm{E}\{\max(0, \varepsilon_1, \varepsilon_3)\} \\
&\quad + \Pr(x_1 - X \notin \mathscr{A}, x_3 - X \in \mathscr{A} \mid x_2 - X \in \mathscr{A}) \sqrt{2\pi} \mathrm{E}\{\max(0, \varepsilon_2, \varepsilon_3)\} \\
&\quad + \Pr(x_2 - X \notin \mathscr{A}, x_3 - X \notin \mathscr{A} \mid x_1 - X \in \mathscr{A}) \\
&\quad + \Pr(x_1 - X \notin \mathscr{A}, x_3 - X \notin \mathscr{A} \mid x_2 - X \in \mathscr{A}) \\
&\quad + \Pr(x_1 - X \notin \mathscr{A}, x_2 - X \notin \mathscr{A} \mid x_3 - X \in \mathscr{A}).
\end{aligned}
$$

The expression $\sqrt{2\pi}\mathrm{E}\{\max(0, \varepsilon_1, \varepsilon_2, \varepsilon_3)\}$ above is the trivariate extremal coefficient for the Schlather model without random sets, and can either be calculated analytically (Nikoloulopoulos *et al.*, 2009; Opitz, 2013) or evaluated quickly and accurately by simulation, whereas $\sqrt{2\pi}\mathrm{E}\left\{\max\left(0, \varepsilon_i, \varepsilon_j\right)\right\}$ is the bivariate extremal coefficient between station i and station j, and can be computed analytically with the bivariate exponent measure $V_{\mathscr{D}}(1,1)$, where $\mathscr{D} = \{x_i, x_j\}$.

If the compact state space $\mathscr{X}$ is large compared to $\mathscr{A}$, the probabilities above correspond to the normalized expected volumes of overlap of three sets $\mathscr{A}$ centered at x_1, x_2 and x_3. For example, denoting the complement of a set A by A^c,

$$
\begin{aligned}
\Pr(x_2 - X \in \mathscr{A}, x_3 - X \in \mathscr{A} \mid x_1 - X \in \mathscr{A}) &\approx \mathrm{E}\{|\mathscr{A} \cap \{\mathscr{A} + (x_2 - x_1)\} \cap \{\mathscr{A} + (x_3 - x_1)\}|\} / \mathrm{E}(|\mathscr{A}|), \\
\Pr(x_2 - X \in \mathscr{A}, x_3 - X \notin \mathscr{A} \mid x_1 - X \in \mathscr{A}) &\approx \mathrm{E}[|\mathscr{A} \cap \{\mathscr{A} + (x_2 - x_1)\} \cap \{\mathscr{A} + (x_3 - x_1)\}^c|] / \mathrm{E}(|\mathscr{A}|).
\end{aligned}
$$

For given radius R, lifetime L and velocity V, the random set is fixed and the volume of overlap can be calculated analytically (the `R` code is available from the author upon request). Simulation can then be used to compute the expectation of such random quantities.

The same approach could be used to compute extremal coefficients at a higher order D, though it would be painful to compute all the areas of overlap between D discs.

H Simulation of the fitted max-stable model (5.12) in space and time

The spatio-temporal Schlather model with random set (5.12) was fitted in §5.3.2.1 to the rainfall data plotted in Figure 5.2. As an illustration, we show here a simulation of the fitted dependence model (with unit Fréchet margins) on a 50×25 spatial grid with coordinates in the rectangle $[500,750] \times [150,250]$ (km), covering our monitoring stations (black dots), for 10 time points. The number of space-time locations equals $50 \times 25 \times 10 = 12500$. The simulation was performed following the approximate algorithm for max-stable processes in §2.3.1, using exact simulations of Gaussian processes. The big red disc represents a heavy rainstorm moving in the north-east direction.

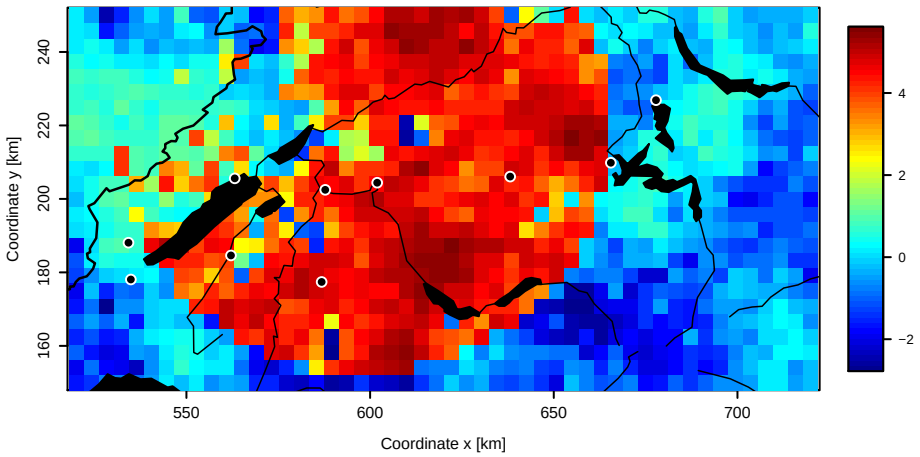

Figure H.1: Simulated spatial field, time $t = 1$.

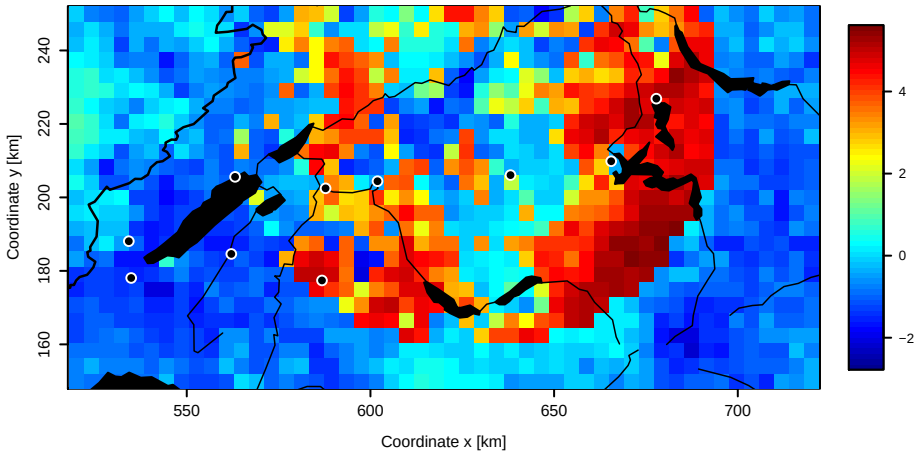

Figure H.2: Simulated spatial field, time $t = 2$.

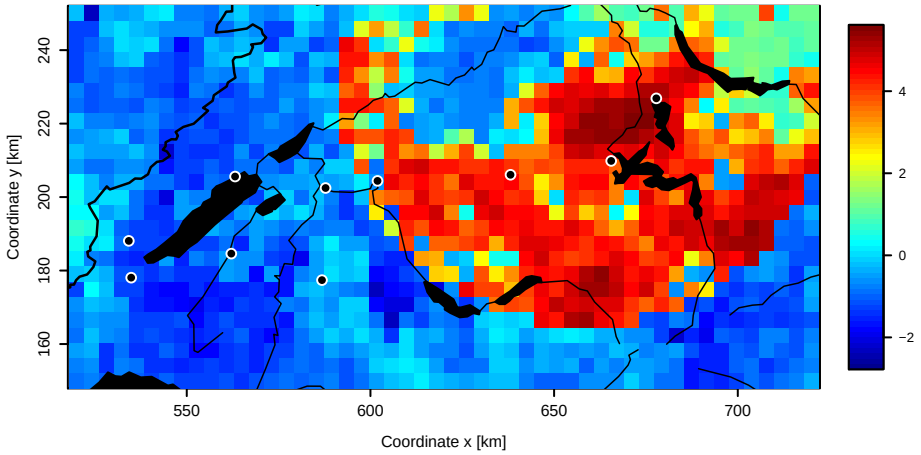

Figure H.3: Simulated spatial field, time $t = 3$.

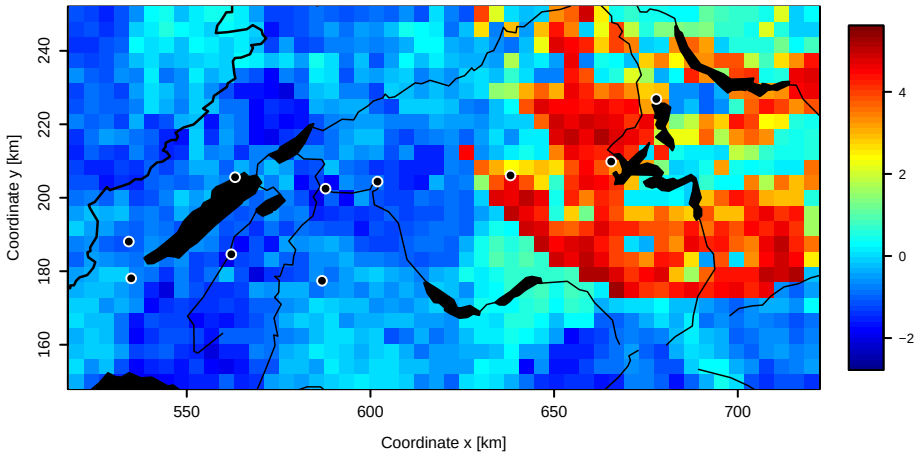

Figure H.4: Simulated spatial field, time $t = 4$.

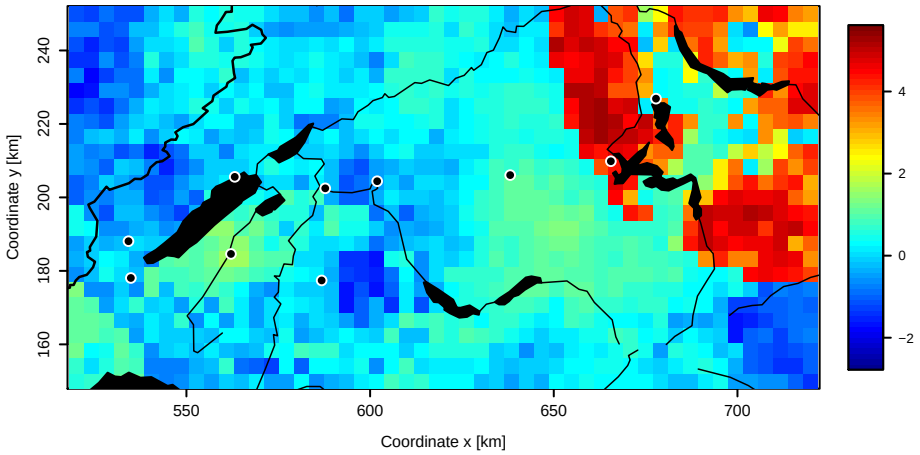

Figure H.5: Simulated spatial field, time $t = 5$.

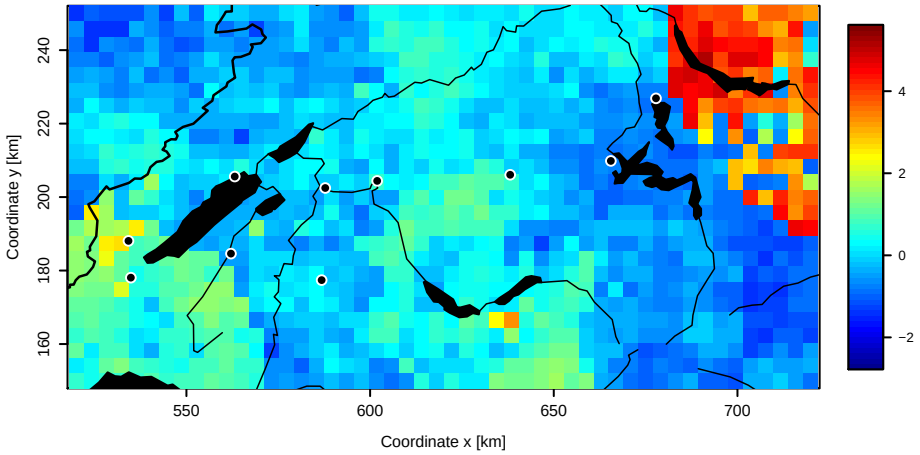

Figure H.6: Simulated spatial field, time $t = 6$.

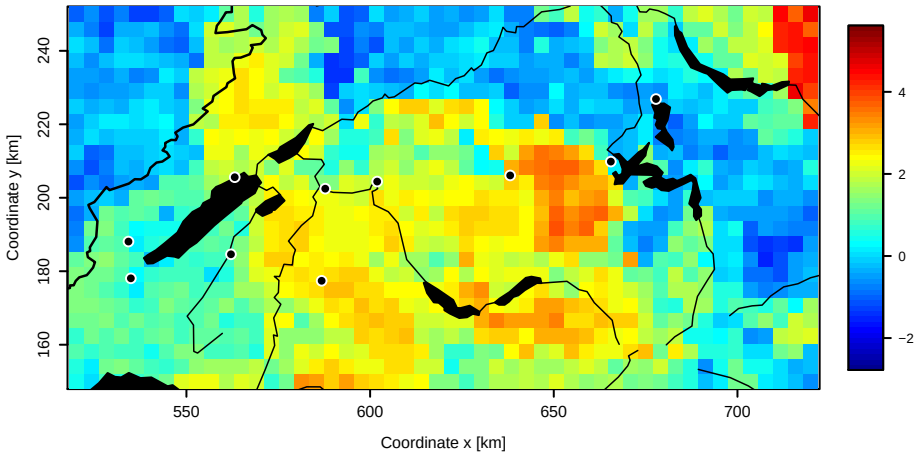

Figure H.7: Simulated spatial field, time $t = 7$.

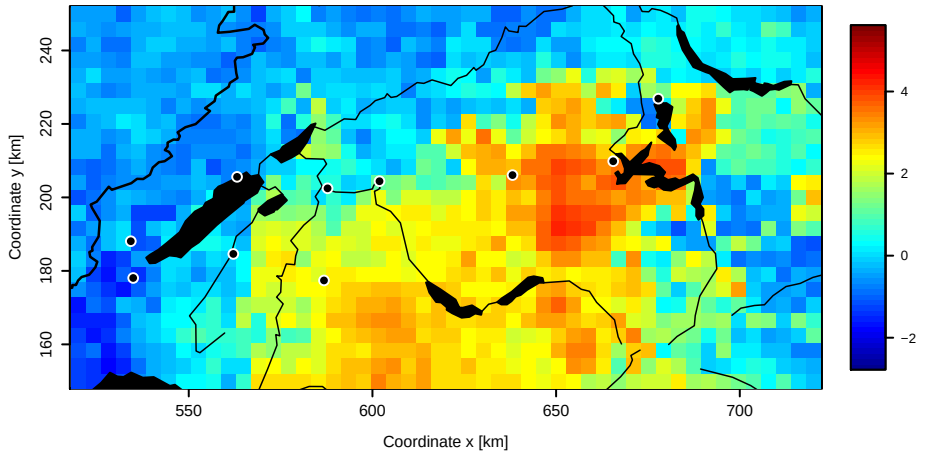

Figure H.8: Simulated spatial field, time $t = 8$.

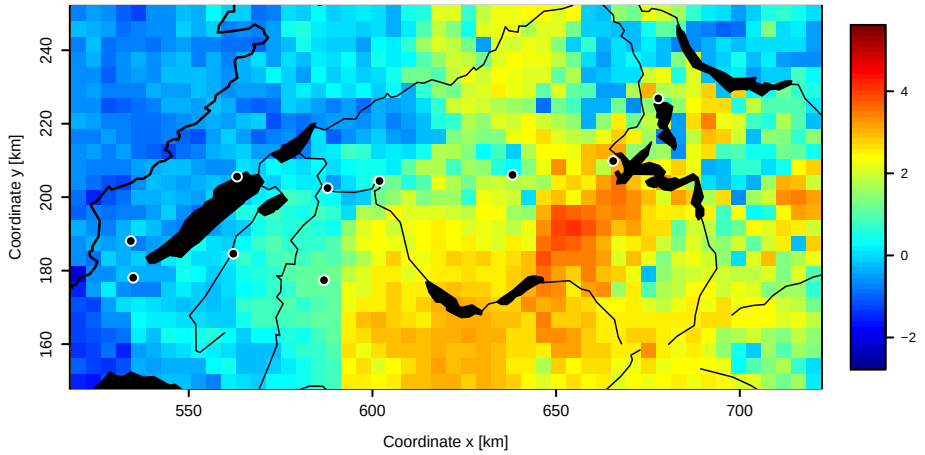

Figure H.9: Simulated spatial field, time $t = 9$.

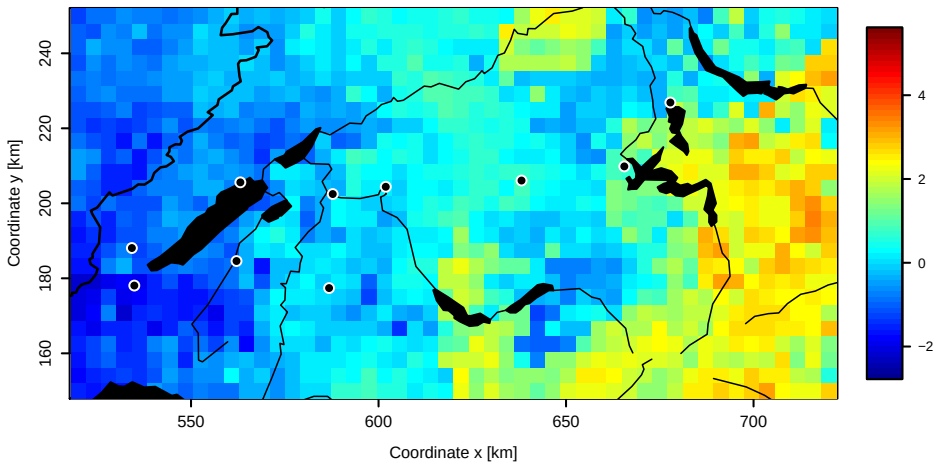

Figure H.10: Simulated spatial field, time $t = 10$.

Bibliography

Aaronson, J. (1997) *An Introduction to Infinite Ergodic Theory*. Providence: American Mathematical Society. ISBN 0821804944.

Abrahamsen, P. (1997) A Review of Gaussian Random Fields and Correlation Functions. Technical Report 917, Norwegian Computing Center, Blindern, Norway.

Ancona-Navarrete, M. A. & Tawn, J. A. (2000) A Comparison of Methods for Estimating the Extremal Index. *Extremes* **3**, 5–38. doi:10.1023/A:1009993419559.

Anderes, E. B., Huser, R., Nychka, D. & Coram, M. A. (2012) Nonstationary positive definite tapering on the plane. *Journal of Computational and Graphical Statistics*. doi:10.1080/10618600.2012.729982.

Anderes, E. B. & Stein, M. L. (2008) Estimating deformations of isotropic Gaussian random fields on the plane. *Annals of Statistics* **36**, 719–741. doi:10.1214/009053607000000893.

Anderson, C., Coles, S. G. & Hüsler, J. (1997) Maxima of Poisson-like variables and related triangular arrays. *Annals of Applied Probability* **7**, 953–971. doi:10.1214/aoap/1043862420.

Aryal, S. K., Bates, B. C., Campbell, E. P., Li, Y., Palmer, M. J. & Viney, N. R. (2009) Characterizing and Modeling Temporal and Spatial Trends in Rainfall Extremes. *Journal of Hydrometeorology* **10**, 241–253. doi:10.1175/2008JHM1007.1.

Balkema, A. A. & de Haan, L. (1974) Residual Life Time at Great Age. *Annals of Probability* **2**, 792–804. doi:10.2307/2959306.

Balkema, A. A. & de Haan, L. (1990) A Convergence Rate in Extreme-Value Theory. *Journal of Applied Probability* **27**, 577–585. doi:10.2307/3214542.

Balkema, A. A. & Resnick, S. I. (1977) Max-Infinite Divisibility. *Journal of Applied Probability* **14**, 309–319.

Ballani, F. & Schlather, M. (2011) A construction principle for multivariate extreme value distributions. *Biometrika* **98**, 633–645. doi:10.1093/biomet/asr034.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2003) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall. ISBN 9781584884101.

Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. (2004) *Statistics of Extremes: Theory and Applications*. Chichester: Wiley. ISBN 9780471976479.

Berman, S. M. (1964) Limit Theorems for the Maximum Term in Stationary Sequences. *Annals of Mathematical Statistics* **35**, 502–516. doi:10.1214/aoms/1177703551.

Bevilacqua, M., Gaetan, C., Mateu, J. & Porcu, E. (2012) Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach. *Journal of American Statistical Association* **107**, 268–280. doi:10.1080/01621459.2011.646928.

Blanchet, J. & Davison, A. C. (2011) Spatial modelling of extreme snow depth. *Annals of Applied Statistics* **5**, 1699–1725. doi:10.1214/11-AOAS464SUPP.

Blanchet, J. & Davison, A. C. (2012) Statistical modelling of ground temperature in mountain permafrost. *Proceedings of the Royal Society A: Mathematical Physical And Engineering Sciences* **468**, 1472–1495. doi:10.1098/rspa.2011.0615.

Bochner, S. (1955) *Harmonic Analysis and the Theory of Probability*. Berkeley: University of California Press. ISBN 9780486446202.

Bortot, P., Coles, S. G. & Tawn, J. A. (2000) The multivariate Gaussian tail model: an application to oceanographic data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49**, 31–49. doi:10.1111/1467-9876.00177.

Bortot, P. & Tawn, J. A. (1998) Models for the extremes of Markov chains. *Biometrika* **85**, 851–867. doi:10.1093/biomet/85.4.851.

Bradley, R. C. (2007a) *Introduction to Strong Mixing Conditions*, volume 1. Heber City: Kendrick Press. ISBN 9780974042763.

Bradley, R. C. (2007b) *Introduction to Strong Mixing Conditions*, volume 2. Heber City: Kendrick Press. ISBN 9780974042770.

Bradley, R. C. (2007c) *Introduction to Strong Mixing Conditions*, volume 3. Heber City: Kendrick Press. ISBN 9780974042787.

Brown, B. M. & Resnick, S. I. (1977) Extreme Values of Independent Stochastic Processes. *Journal of Applied Probability* **14**, 732–739. doi:10.2307/3213346.

Brown, P. E., Diggle, P. J., Lord, M. E. & Young, P. C. (2001) Space-Time Calibration of Radar Rainfall Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**, 221–241. doi:10.1111/1467-9876.00230.

Brown, P. E., Karesen, K. F., Roberts, G. O. & Tonellato, S. (2000) Blur-Generated Non-Separable Space-Time Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 847–860. doi:10.1111/1467-9868.00269.

Buishand, T. A., de Haan, L. & Zhou, C. (2008) On spatial extremes: With application to a rainfall problem. *Annals of Applied Statistics* **2**, 624–642. doi:10.1214/08-AOAS159.

Capéraà, P., Fougères, A.-L. & Genest, C. (1997) A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika* **84**, 567–577. doi:10.1093/biomet/84.3.567.

de Carvalho, M. & Ramos, A. (2012) Bivariate extreme statistics, II. *REVSTAT* **10**, 83–107.

Casson, E. & Coles, S. G. (1999) Spatial Regression Models for Extremes. *Extremes* **1**, 449–468. doi:10.1023/A:1009931222386.

Cenedese, A., Romano, G. P. & di Felice, F. (1991) Experimental testing of Taylor's hypothesis by L.D.A. in highly turbulent flow. *Experiments in Fluids* **11**, 351–358. doi:10.1007/BF00211789.

de Cesare, L., Myers, D. E. & Posa, D. (2001) Estimating and modeling space–time correlation structures. *Statistics & Probability Letters* **51**, 9–14. doi:10.1016/S0167-7152(00)00131-0.

Chan, K. S. & Ledolter, J. (1995) Monte Carlo EM Estimation for Time Series Models Involving Counts. *Journal of the American Statistical Association* **90**, 242–252. doi:10.1080/01621459.1995.10476508.

Chavez-Demoulin, V. & Davison, A. C. (2005) Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 207–222. doi:10.1111/j.1467-9876.2005.00479.x.

Chavez-Demoulin, V. & Davison, A. C. (2012) Modelling Time Series Extremes. *REVSTAT* **10**, 109–133.

Chavez-Demoulin, V., Davison, A. C. & Frossard, L. (2011) Discussion of "Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights" by P. J. Northrop and P. Jonathan. *Environmetrics* **22**, 810–812. doi:10.1002/env.1125.

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Extreme Values*. London: Springer. ISBN 9781852334598.

Coles, S. G. & Casson, E. (1998) Extreme value modelling of hurricane wind speeds. *Structural Safety* **20**, 283–296. doi:10.1016/S0167-4730(98)00015-0.

Coles, S. G., Heffernan, J. & Tawn, J. A. (1999) Dependence Measures for Extreme Value Analyses. *Extremes* **2**, 339–365. doi:10.1023/A:1009963131610.

Coles, S. G. & Pauli, F. (2002) Models and inference for uncertainty in extremal dependence. *Biometrika* **89**, 183–196. doi:10.1093/biomet/89.1.183.

Coles, S. G. & Tawn, J. A. (1991) Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **53**, 377–392.

Coles, S. G. & Tawn, J. A. (1994) Statistical Methods for Multivariate Extremes: An Application to Structural Design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **43**, 1–48. doi:10.2307/2986112.

Cooley, D., Davis, R. A. & Naveau, P. (2010) The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis* **101**, 2103–2117. doi:10.1016/j.jmva.2010.04.007.

Cooley, D. S. (2005) *Statistical Analysis of Extremes Motivated by Weather and Climate Studies: Applied and Theoretical Advances*. Ph.D. thesis, University of Colorado.

Cooley, D. S., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt-Shamseldin, E. C., Omolo, B. O. & Sun, Y. (2012) A Survey Of Spatial Extremes: Measuring Spatial Dependence And Modeling Spatial Effects. *REVSTAT* **10**, 135–165.

Cooley, D. S., Naveau, P., Jomelli, V., Rabatel, A. & Grancher, D. (2006a) A Bayesian hierarchical extreme value model for lichenometry. *Environmetrics* **17**, 555–574. doi:10.1002/env.764.

Cooley, D. S., Naveau, P. & Nychka, D. (2007) Bayesian Spatial Modeling of Extreme Precipitation Return Levels. *Journal of American Statistical Association* **102**, 824–840. doi:10.1198/016214506000000780.

Cooley, D. S., Naveau, P. & Poncet, P. (2006b) Variograms for spatial max-stable random fields. In *Dependence in Probability and Statistics*, eds. P. Bertail, P. Doukhan & P. Soulier, volume 187 of *Lecture Notes in Statistics*, pp. 373–390. New York: Springer. ISBN 9780387360621.

Cooley, D. S. & Sain, S. R. (2010) Spatial Hierarchical Modeling of Precipitation Extremes From a Regional Climate Model. *Journal of Agricultural, Biological, and Environmental Statistics* **15**, 381–402. doi:10.1007/s13253-010-0023-9.

Cornfeld, I. P., Fomin, S. V. & Sinai, Y. G. (1982) *Ergodic Theory*. Springer. ISBN 3540905804.

Cox, D. R. & Isham, V. (1980) *Point Processes*. London: Chapman & Hall. ISBN 9780412219108.

Cox, D. R. & Isham, V. (1988) A Simple Spatial-Temporal Model of Rainfall. *Proceedings of the Royal Society A: Mathematical Physical And Engineering Sciences* **415**, 317–328. doi:10.1098/rspa.1988.0016.

Cox, D. R. & Reid, N. (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737. doi:10.1093/biomet/91.3.729.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*. New York: Wiley.

Cressie, N. A. C. & Huang, H.-C. (1999) Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions. *Journal of American Statistical Association* **94**, 1330–1340. doi:10.1080/01621459.1999.10473885.

Cressie, N. A. C. & Wikle, C. K. (2011) *Statistics for Spatio-Temporal Data*. Hoboken: Wiley. ISBN 9780471692744.

Daley, D. J. & Vere-Jones, D. (2002) *An Introduction to the Theory of Point Processes*, volume 1. Springer, 2nd edition. ISBN 0387955410.

Davis, R. A., Klüppelberg, C. & Steinkohl, C. (2013a) Max-stable processes for modeling extremes observed in space and time. *Journal of the Korean Statistical Society* **42**, 399–414. doi:10.1016/j.jkss.2013.01.002.

Davis, R. A., Klüppelberg, C. & Steinkohl, C. (2013b) Statistical inference for max-stable processes in space and time. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**. doi:10.1111/rssb.12012. To appear.

Davis, R. A. & Mikosch, T. (2008) Extreme value theory for space–time processes with heavy-tailed distributions. *Stochastic Processes and their Applications* **118**, 560–584. doi:10.1016/j.spa.2007.06.001.

Davis, R. A. & Mikosch, T. (2009) The extremogram: A correlogram for extreme events. *Bernoulli* **15**, 977–1009. doi:10.3150/09-BEJ213.

Davis, R. A. & Yau, C. Y. (2011) Comments on pairwise likelihood in time series models. *Statistica Sinica* **21**, 255–277.

Davison, A. C. (2003) *Statistical Models*. New York: Cambridge University Press. ISBN 9780521773393.

Davison, A. C. & Gholamrezaee, M. M. (2012) Geostatistics of extremes. *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* **468**, 581–608. doi:10.1098/rspa.2011.0412.

Davison, A. C. & Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press. ISBN 9780521574716.

Davison, A. C., Huser, R. & Thibaud, E. (2013) Geostatistics of Dependent and Asymptotically Independent Extremes. *Mathematical Geosciences* **45**, 511–529. doi:10.1007/s11004-013-9469-y.

Davison, A. C., Padoan, S. & Ribatet, M. (2012) Statistical Modelling of Spatial Extremes (with Discussion). *Statistical Science* **27**, 161–186. doi:10.1214/11-STS376.

Davison, A. C. & Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **52**, 393–442.

Davoust, S. & Jacquin, L. (2011) Taylor's hypothesis convection velocities from mass conservation equation. *Physics of Fluids* **23**. doi:10.1063/1.3584004.

Deheuvels, P. (1991) On the limiting behavior of the Pickands estimator for bivariate extreme-value distributions. *Statistics & Probability Letters* **12**, 429–439. doi:10.1016/0167-7152(91)90032-M.

Deheuvels, P. & Tiago de Oliveira, J. (1989) On the non-parametric estimation of the bivariate extreme-value distributions. *Statistics & Probability Letters* **8**, 315–323. doi:10.1016/0167-7152(89)90038-2.

Dekkers, A. L. M., Einmahl, J. H. J. & de Haan, L. (1989) A Moment Estimator for the Index of an Extreme-Value Distribution. *Annals of Statistics* **17**, 1833–1855. doi:10.1214/aos/1176347397.

Demarta, S. & McNeil, A. J. (2005) The *t* Copula and Related Copulas. *International Statistical Review* **73**, 111–129. doi:10.1111/j.1751-5823.2005.tb00254.x.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**, 1–38.

Diggle, P. J. & Ribeiro, P. J. (2007) *Model-based Geostatistics*. New York: Springer. ISBN 9780387329079.

Dombry, C., Éyi-Minko, F. & Ribatet, M. (2013) Conditional simulation of max-stable processes. *Biometrika* **100**, 111–124. doi:10.1093/biomet/ass067.

Durbin, J. & Koopman, S. J. (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**, 669–684. doi:10.1093/biomet/84.3.669.

Durbin, J. & Koopman, S. J. (2002) A Simple and Efficient Simulation Smoother for State Space Time Series Analysis. *Biometrika* **89**, 603–615. doi:10.1093/biomet/89.3.603.

Einicke, G. A. & White, L. B. (1999) Robust Extended Kalman Filtering. *IEEE Transactions on Signal Processing* **47**, 2596–2599.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer. ISBN 9783540609315.

Engelke, S., Malinowski, A., Kabluchko, Z. & Schlather, M. (2012) Estimation of Huesler–Reiss distributions and Brown–Resnick processes. arXiv:1207.6886v2.

Falk, M. & Reiss, R. D. (2001) Estimation of canonical dependence parameters in a class of bivariate peaks-over-threshold models. *Statistics & Probability Letters* **52**, 233–242. doi:10.1016/S0167-7152(00)00194-2.

Falk, M. & Reiss, R. D. (2002) A characterization of the rate of convergence in bivariate extreme value models. *Statistics & Probability Letters* **59**, 341–351. doi:10.1016/S0167-7152(02)00209-2.

Falk, M. & Reiss, R. D. (2003a) Efficient Estimation of the Canonical Dependence Function. *Extremes* **6**, 61–82. doi:10.1023/A:1026229314063.

Falk, M. & Reiss, R. D. (2003b) Efficient estimators and LAN in canonical bivariate POT models. *Journal of Multivariate Analysis* **84**, 190–207. doi:10.1016/S0047-259X(02)00010-6.

Falk, M. & Reiss, R. D. (2005) On the distribution of Pickands coordinates in bivariate EV and GP models. *Journal of Multivariate Analysis* **93**, 267–295. doi:10.1016/j.jmva.2004.02.017.

Fawcett, L. & Walshaw, D. (2007) Improved estimation for temporally clustered extremes. *Environmetrics* **18**, 173–188. doi:10.1002/env.810.

Fearnhead, P. (2011) MCMC for State-Space Models. In *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, G. L. Jones & X.-L. Meng, pp. 513–529. Boca Raton: Chapman & Hall.

Ferrez, J., Davison, A. C. & Rebetez, M. (2011) Extreme temperature analysis under forest cover compared to an open field. *Agricultural and Forest Meteorology* **151**, 992–1001. doi:10.1016/j.agrformet.2011.03.005.

Ferro, C. A. T. & Segers, J. (2003) Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 545–556. doi:10.1111/1467-9868.00401.

Fisher, R. A. & Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* **24**, 180–190. doi:10.1017/S0305004100015681.

Fougères, A.-L. (2004) Multivariate Extremes. In *Extreme Values in Finance, Telecommunications, and the Environment*, eds. B. Finkenstädt & H. Rootzén. Chapman & Hall. ISBN 9781584884118. doi:10.1201/9780203483350.ch7.

Fougères, A.-L., Nolan, J. P. & Rootzén, H. (2009) Models for Dependent Extremes Using Stable Mixtures. *Scandinavian Journal of Statistics* **36**, 42–59. doi:10.1111/j.1467-9469.2008.00613.x.

Fuentes, M. (2006) Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference* **136**, 447–466. doi:10.1016/j.jspi.2004.07.004.

Fuentes, M., Henry, J. & Reich, B. (2011) Nonparametric spatial models for extremes: application to extreme temperature data. *Extremes* **16**, 75–101. doi:10.1007/s10687-012-0154-1.

Galambos, J. (1975) Order Statistics of Samples from Multivariate Distributions. *Journal of the American Statistical Association* **70**, 674–680. doi:10.1080/01621459.1975.10482493.

Galambos, J. (1987) *The Asymptotic Theory of Extreme Order Statistics*. Malabar: Krieger Publishing Compagny, 2nd edition. ISBN 9780898749571.

Genton, M. G. (2007) Separable approximations of space-time covariance matrices. *Environmetrics* **18**, 681–695. doi:10.1002/env.854.

Genton, M. G., Ma, Y. & Sang, H. (2011) On the likelihood function of Gaussian max-stable processes. *Biometrika* **98**, 481–488. doi:10.1093/biomet/asr020.

Genz, A. (1992) Numerical Computation of Multivariate Normal Probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–150. doi:10.1080/10618600.1992.10477010.

Genz, A. (1993) Comparison of Methods for Numerical Computation of Multivariate Normal Probabilities. *Computing Sciences and Statistics* **25**, 400–405.

Genz, A. & Bretz, F. (2002) Comparison of Methods for the Computation of Multivariate t Probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971. doi:10.1198/106186002394.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1995) *Markov Chain Monte Carlo in practice*. London: Chapman & Hall.

Gnedenko, B. (1943) Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire. *Annals of Mathematics* **44**, 423–453.

Gneiting, T. (2002) Nonseparable, Stationary Covariance Functions for Space–Time Data. *Journal of the American Statistical Association* **97**, 590–600. doi:10.1198/016214502760047113.

Gneiting, T., Genton, M. G., Guttorp, P., Finkenstädt, B., Held, L. & Isham, V. (2007) Geostatistical Space-Time Models, Stationarity, Separability, and Full Symmetry. In *Statistical Methods for Spatio-Temporal Systems*, eds. B. Finkenstädt, L. Held & V. Isham, pp. 151–175. Hoboken: Chapman & Hall. ISBN 9781584885931.

Gneiting, T. & Sasvári, Z. (1999) The Characterization Problem for Isotropic Covariance Functions. *Mathematical Geology* **31**, 105–111. doi:10.1023/A:1007597415185.

Gneiting, T., Sasvári, Z. & Schlather, M. (2001) Analogies and correspondences between variograms and covariance functions. *Advances in Applied Probability* **33**, 617–630. doi:10.1239/aap/1005091356.

Gouriéroux, C. & Monfort, A. (1991) Simulation Based Inference in Models with Heterogeneity. *Annales d'Économie et de Statistique* **20–21**, 69–107.

Gumbel, E. J. (1961) Bivariate Logistic Distributions. *Journal of the American Statistical Association* **56**, 335–349.

de Haan, L. (1970) *On Regular Variation and its Application to the Weak Convergence of Sample Extremes*. Amsterdam: Mathematisch Centrum.

de Haan, L. (1984) A Spectral Representation for Max-stable Processes. *Annals of Probability* **12**, 1194–1204. doi:10.1214/aop/1176993148.

de Haan, L. & Ferreira, A. (2006) *Extreme Value Theory: An Introduction*. New York: Springer. ISBN 9780387239460.

de Haan, L. & Pereira, T. T. (2006) Spatial extremes: Models for the stationary case. *Annals of Statistics* **34**, 146–168. doi:10.1214/009053605000000886.

Hall, P. & Tajvidi, N. (2000) Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli* **6**, 835–844.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109. doi:10.1093/biomet/57.1.97.

Heagerty, P. J. & Lele, S. R. (1998) A Composite Likelihood Approach to Binary Spatial Data. *Journal of the American Statistical Association* **93**, 1099–1111. doi:10.1080/01621459.1998.10473771.

Heffernan, J. E. (2000) A Directory of Coefficients of Tail Dependence. *Extremes* **3**, 279–290. doi:10.1023/A:1011459127975.

Heffernan, J. E. & Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 497–546. doi:10.1111/j.1467-9868.2004.02050.x.

Heffernan, J. E., Tawn, J. A. & Zhang, Z. (2007) Asymptotically (in)dependent multivariate maxima of moving maxima processes. *Extremes* **10**, 57–82. doi:10.1007/s10687-007-0035-1.

Hill, B. M. (1975) A Simple General Approach to Inference About the Tail of a Distribution. *Annals of Statistics* **3**, 1163–1174. doi:10.1214/aos/1176343247.

Hjort, N. L. & Varin, C. (2008) ML, PL, QL in Markov Chain Models. *Scandinavian Journal of Statistics* **35**, 64–82. doi:10.1111/j.1467-9469.2007.00559.x.

Hofert, M. (2010) *Sampling Nested Archimedean Copulas with Applications*. Ph.D. thesis, University of Ulm.

Hosking, J. R. M., Wallis, J. R. & Wood, E. F. (1985) Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics* **27**, 251–261. doi:10.1080/00401706.1985.10488049.

Hsing, T., Hüsler, J. & Leadbetter, M. R. (1988) On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields* **78**, 97–112. doi:10.1007/BF00718038.

Huerta, G. & Sansó, B. (2007) Time-varying models for extreme values. *Environmental and Ecological Statistics* **14**, 285–299. doi:10.1007/s10651-007-0014-3.

Huser, R. & Davison, A. C. (2013a) Composite likelihood estimation for the Brown–Resnick process. *Biometrika* **100**, 511–518. doi:10.1093/biomet/ass089.

Huser, R. & Davison, A. C. (2013b) Space-time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* To appear.

Hüsler, J. & Reiss, R.-D. (1989) Maxima of normal random vectors: Between independence and complete dependence. *Statistics & Probability Letters* **7**, 283–286. doi:10.1016/0167-7152(89)90106-5.

de Iaco, S., Myers, D. & Posa, D. (2002) Nonseparable Space-Time Covariance Models: Some Parametric Families. *Mathematical Geology* **34**, 23–42. doi:10.1023/A:1014075310344.

Jacobsen, M. (2005) *Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes*. Basel: Birkhäuser, 1 edition. ISBN 9780817642150.

Jeon, S. (2012) *Max-stable Processes for Threshold Exceedances in Spatial Extremes*. Ph.D. thesis, University of North Carolina at Chapel Hill.

Jeon, S. & Smith, R. L. (2012) Dependence Structure of Spatial Extremes Using Threshold Approach. arXiv:1209.6344v1.

Joe, H. (1990) Families of min-stable multivariate exponential and multivariate extreme value distributions. *Statistics & Probability Letters* **9**, 75–81. doi:10.1016/0167-7152(90)90098-R.

Joe, H. (1997) *Multivariate Models and Dependence Concepts*. London: Chapman & Hall. ISBN 9780412073311.

Jungbacker, B. & Koopman, S. J. (2007) Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika* **94**, 827–839. doi:10.1093/biomet/asm074.

Kabluchko, Z. & Schlather, M. (2010) Ergodic properties of max-infinitely divisible processes. *Stochastic Processes and their Applications* **120**, 281–295. doi:10.1016/j.spa.2009.12.002.

Kabluchko, Z., Schlather, M. & de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. *Annals of Probability* **37**, 2042–2065. doi:10.1214/09-AOP455.

Katz, R. W., Parlange, M. & Naveau, P. (2002) Statistics of extremes in hydrology. *Advances in Water Resources* **25**, 1287–1304. doi:10.1016/S0309-1708(02)00056-8.

Kent, J. T. (1982) Robust Properties of Likelihood Ratio Test. *Biometrika* **69**, 19–27. doi:10.1093/biomet/69.1.19.

Kim, S., Shephard, N. & Chib, S. (1998) Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies* **65**, 361–393. doi:10.1111/1467-937X.00050.

Kitagawa, G. (1996) Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics* **5**, 1–25. doi:10.2307/1390750.

Klein, A. & Mélard, G. (1995) Computation of the Fisher information matrix for time series models. *Journal of Computational and Applied Mathematics* **64**, 57–68. doi:10.1016/0377-0427(95)00006-2.

Krengel, U. (1985) *Ergodic Theorems*. Berlin: Walter de Gruyter. ISBN 9783110844641.

Lantuéjoul, C., Bacro, J.-N. & Bel, L. (2011) Storm processes and stochastic geometry. *Extremes* **14**, 413–428. doi:10.1007/s10687-010-0121-7.

Laurini, F. & Tawn, J. A. (2003) New Estimators for the Extremal Index and Other Cluster Characteristics. *Extremes* **6**, 189–211. doi:10.1023/B:EXTR.0000031179.49454.90.

Leadbetter, M. R. (1983) Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65**, 291–306. doi:10.1007/BF00532484.

Leadbetter, M. R., Lindgren, G. & Rootzén, H. (1983) *Extreme and Related Properties of Random Sequences and Processes*. New York: Springer. ISBN 9780387907314.

Ledford, A. W. & Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika* **83**, 169–187. doi:10.1093/biomet/83.1.169.

Ledford, A. W. & Tawn, J. A. (1997) Modelling Dependence within Joint Tail Regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 475–499. doi:10.1111/1467-9868.00080.

Ledford, A. W. & Tawn, J. A. (2003) Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 521–543. doi:10.1111/1467-9868.00400.

Li, B., Genton, M. G. & Sherman, M. (2007) A Nonparametric Assessment of Properties of Space–Time Covariance Functions. *Journal of the American Statistical Association* **102**, 736–744. doi:10.1198/016214507000000202.

Li, B., Murthi, A., Bowman, K. P., North, G. R., Genton, M. G. & Sherman, M. (2009) Statistical Tests of Taylor's Hypothesis: An Application to Precipitation Fields. *Journal of Hydrometeorology* **10**, 254–265. doi:10.1175/2008JHM1009.1.

Lindsay, B. G. (1988) Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

Lu, N. & Zimmerman, D. L. (2005) Testing for directional symmetry in spatial dependence using the periodogram. *Journal of Statistical Planning and Inference* **129**, 369–385. doi:10.1016/j.jspi.2004.06.058.

Ma, C. (2002) Spatio-Temporal Covariance Functions Generated by Mixtures. *Mathematical Geology* **34**, 965–975. doi:10.1023/A:1021368723926.

Ma, C. (2003a) Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference* **116**, 489–501. doi:10.1016/S0378-3758(02)00353-1.

Ma, C. (2003b) Spatio-temporal stationary covariance models. *Journal of Multivariate Analysis* **86**, 97–107. doi:10.1016/S0047-259X(02)00014-3.

Matérn, B. (1986) *Spatial Variation*. Berlin: Springer. ISBN 9780387963655.

Matheron, G. (1973) The Intrinsic Random Functions and Their Applications. *Advances in Applied Probability* **5**, 439–468. doi:10.2307/1425829.

Maulik, K. & Resnick, S. I. (2004) Characterizations and Examples of Hidden Regular Variation. *Extremes* **7**, 31–67. doi:10.1007/s10687-004-4728-4.

Maulik, K., Resnick, S. I. & Rootzén, H. (2002) Asymptotic independence and a network traffic model. *Journal of Applied Probability* **39**, 671–699. doi:10.1239/jap/1037816012.

McNeil, A. J. & Frey, R. (1998) Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach. *Journal of Empirical Finance* **7**, 271–300. doi:10.1016/S0927-5398(00)00012-8.

Mikosch, T. (2006) Copulas: Tales and facts. *Extremes* **9**, 3–20. doi:10.1007/s10687-006-0015-x.

von Mises, R. (1964) La distribution de la plus grande de *n* valeurs. In *Selected papers of Richard von Mises*, eds. P. Frank, S. Goldstein, M. Kac, W. Prager, G. Szegö & G. Birkhoff, volume 2 of *Probability and Statistics, General*, pp. 271–294. Providence: American Mathematical Society.

Mitchell, M. W., Genton, M. G. & Gumpertz, M. L. (2005) Testing for separability of space–time covariances. *Environmetrics* **16**, 819–831. doi:10.1002/env.737.

Mitchell, M. W., Genton, M. G. & Gumpertz, M. L. (2006) A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis* **97**, 1025–1043. doi:10.1016/j.jmva.2005.07.005.

Naveau, P., Guillou, A., Cooley, D. & Diebolt, J. (2009) Modelling pairwise dependence of maxima in space. *Biometrika* **96**, 1–17. doi:10.1093/biomet/asp001.

Nelsen, R. B. (2006) *An Introduction to Copulas*. New York: Springer, 2nd edition. ISBN 9780387986234.

Nikoloulopoulos, A. K., Joe, H. & Li, H. (2009) Extreme value properties of multivariate *t* copulas. *Extremes* **12**, 129–148. doi:10.1007/s10687-008-0072-4.

Northrop, P. J. & Jonathan, P. (2011) Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights (with discussion). *Environmetrics* **22**, 799–809. doi:10.1002/env.1106.

O'Brien, G. L. (1987) Extreme Values for Stationary and Markov Sequences. *Annals of Probability* **15**, 281–291. doi:10.1214/aop/1176992270.

Oesting, M., Kabluchko, Z. & Schlather, M. (2012) Simulation of Brown–Resnick processes. *Extremes* **15**, 89–107. doi:10.1007/s10687-011-0128-8.

Omey, E. & Rachev, S. T. (1991) Rates of convergence in multivariate extreme value theory. *Journal of Multivariate Analysis* **38**, 36–50. doi:10.1016/0047-259X(91)90030-6.

Opitz, T. (2013) Extremal *t* processes: Elliptical domain of attraction and a spectral representation. arXiv:1207.2296v6.

Padoan, S. A. (2011) Multivariate extreme models based on underlying skew-*t* and skew-normal distributions. *Journal of Multivariate Analysis* **102**, 977–991. doi:10.1016/j.jmva.2011.01.014.

Padoan, S. A., Ribatet, M. & Sisson, S. A. (2010) Likelihood-Based Inference for Max-Stable Processes. *Journal of the American Statistical Association* **105**, 263–277. doi:10.1198/jasa.2009.tm08577.

Pickands, J. (1975) Statistical Inference Using Extreme Order Statistics. *Annals of Statistics* **3**, 119–131.

Pickands, J. (1981) Multivariate extreme value distributions (with discussion). In *Bulletin de l'Institut International de Statistique*, volume 49, pp. 859–878, 894–902.

Politis, D. N., Romano, J. P. & Wolf, M. (1999) *Subsampling*. New York: Springer. ISBN 9780387988542.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramos, A. & Ledford, A. (2009) A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 219–241. doi: 10.1111/j.1467-9868.2008.00684.x.

Ramos, A. & Ledford, A. W. (2011) An alternative point process framework for modeling multivariate extreme values. *Communications in Statistics — Theory and Methods* **40**, 2205–2224. doi:10.1080/03610921003764233.

Reich, B. J. & Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. *Annals of Applied Statistics* **6**, 1430–1451. doi:10.1214/12-AOAS591.

Renard, D., Molenberghs, G. & Geys, H. (2004) A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis* **44**, 649–667. doi:10.1016/S0167-9473(02)00263-3.

Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*. New York: Springer.

Resnick, S. I. (2002) Hidden Regular Variation, Second Order Regular Variation and Asymptotic Independence. *Extremes* **5**, 303–336. doi:10.1023/A:1025148622954.

Resnick, S. I. (2008) Multivariate regular variation on cones: application to extreme values, hidden regular variation and conditioned limit laws. *Stochastics An International Journal of Probability and Stochastic Processes: formerly Stochastics and Stochastics Reports* **80**, 269–298. doi:10.1080/17442500701830423.

Ribatet, M. (2011) *SpatialExtremes: Modelling Spatial Extremes*. R package version 1.8-5.

Ribatet, M. (2013) An introduction to max-stable processes. Submitted to Journal de la Société Française de Statistique.

Ribatet, M., Cooley, D. S. & Davison, A. C. (2012) Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica* **22**, 813–845.

Ribatet, M. & Sedki, M. (2013) Extreme values copulas and max-stable processes. Submitted to Journal de la Société Française de Statistique.

Robert, C. Y. (2013a) Automatic declustering of rare events. *Biometrika* doi:10.1093/biomet/ast013. To appear.

Robert, C. Y. (2013b) Some new classes of stationary max-stable random fields. *Statistics & Probability Letters* **83**, 1496–1503. doi:10.1016/j.spl.2013.02.017.

Robert, C. Y., Segers, J. & Ferro, C. A. T. (2009) A sliding blocks estimator for the extremal index. *Electronic Journal of Statistics* **3**, 993–1020. doi:10.1214/08-EJS345.

Rootzén, H. & Tajvidi, N. (2006) Multivariate generalized Pareto distributions. *Bernoulli* **12**, 917–930. doi:10.3150/bj/1161614952.

Salmon, F. (2009) Recipe for disaster: The formula that killed Wall Street. `http://www.wired.com/techbiz/it/magazine/17-03/wp_quant?currentPage=al`, accessed on 12 March 2013.

Samorodnitsky, G. (2006) Long Range Dependence. *Foundations and Trends in Stochastic Systems* **1**, 163–257. doi:10.1561/0900000004.

Sang, H. & Gelfand, A. (2009) Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics* **16**, 407–426. doi:10.1007/s10651-007-0078-0.

Sang, H. & Gelfand, A. (2010) Continuous Spatial Process Models for Spatial Extreme Values. *Journal of Agricultural, Biological and Environmental Statistics* **15**, 49–65. doi:10.1007/s13253-009-0010-1.

Scarrott, C. & MacDonald, A. (2012) A Review of Extreme Value Threshold Estimation And Uncertainty Quantification. *REVSTAT* **10**, 33–60.

Schlather, M. (1999) Introduction to Positive Definite Functions and to Unconditional Simulation of Random Fields. Technical Report ST–99–10, Landcaster University.

Schlather, M. (2002) Models for Stationary Max-Stable Random Fields. *Extremes* **5**, 33–44. doi:10.1023/A:1020977924878.

Schlather, M. (2010) Some covariance models based on normal scale mixtures. *Bernoulli* **16**, 780–797. doi:10.3150/09-BEJ226.

Schlather, M. & Gneiting, T. (2006) Local approximation of variograms by covariance functions. *Statistics & Probability Letters* **76**, 1303–1304. doi:10.1016/j.spl.2006.02.002.

Schlather, M. & Tawn, J. A. (2003) A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* **90**, 139–156. doi:10.1093/biomet/90.1.139.

Segers, J. (2012) Max-stable models for multivariate extremes. *REVSTAT* **10**, 61–82.

Shaby, B. A. & Reich, B. J. (2012) Bayesian spatial extreme value analysis to assess the changing risk of concurrent high temperatures across large portions of European cropland. *Environmetrics* **23**, 638–648. doi:10.1002/env.2178.

Shao, J. & Tu, D. (1995) *The Jackknife and Bootstrap*. New York: Springer. ISBN 9780387945156.

Shi, D. (1995) Fisher information for a multivariate extreme value distribution. *Biometrika* **82**, 644–649. doi:10.1093/biomet/82.3.644.

Shumway, R. H. & Stoffer, D. S. (2004) *Time Series Analysis and Its Applications*. New York: Springer, 2 edition. ISBN 9781441978646.

Sklar, A. (1959) Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* **8**, 229–231.

Smith, R. L. (1982) Uniform Rates of Convergence in Extreme-Value Theory. *Advances in Applied Probability* **14**, 600–622. doi:10.2307/1426676.

Smith, R. L. (1985) Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90. doi:10.1093/biomet/72.1.67.

Smith, R. L. (1989) Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science* **4**, 367–377. doi:10.1214/ss/1177012400.

Smith, R. L. (1990a) Extreme Value Theory. In *Handbook of Applicable Mathematics, Supplement*, eds. W. Ledermann, E. Lloyd, S. Vajda & A. C. Chichester: Wiley. ISBN 9780471918257.

Smith, R. L. (1990b) Max-stable processes and spatial extremes. Unpublished.

Smith, R. L. (1993) Multivariate threshold methods. Technical Report 7, University of North Carolina.

Smith, R. L., Tawn, J. A. & Coles, S. G. (1997) Markov chain models for threshold exceedances. *Biometrika* **84**, 249–268. doi:10.1093/biomet/84.2.249.

Smith, R. L., Tawn, J. A. & Yuen, H. K. (1990) Statistics of Multivariate Extremes. *International Statistical Review* **58**, 47–58.

Smith, R. L. & Weissman, I. (1994) Estimating the Extremal Index. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **56**, 515–528.

Stein, M. L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer, 1 edition. ISBN 9780387986296.

Stephenson, A. (2009) High-Dimensional Parametric Modelling Of Multivariate Extreme Events. *Australian & New Zealand Journal of Statistics* **51**, 77–88. doi:10.1111/j.1467-842X.2008.00528.x.

Stephenson, A. & Tawn, J. A. (2005) Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika* **92**, 213–227. doi:10.1093/biomet/92.1.213.

Süveges, M. (2007) Likelihood estimation of the extremal index. *Extremes* **10**, 41–55. doi:10.1007/s10687-007-0034-2.

Süveges, M. (2009) *Statistical Analysis of Clusters of Extreme Events*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne.

Tajvidi, N. (1996) *Characterisation and Some Statistical Aspects of Univariate and Multivariate Generalised Pareto Distributions*. Ph.D. thesis, University of Göteborg.

Tanizaki, H. (2001) Nonlinear and Non-Gaussian State Space Modeling Using Sampling Techniques. *Annals of the Institute of Statistical Mathematics* **53**, 63–81. doi:10.1023/A:1017916420893.

Tanizaki, H. & Mariano, R. (1998) Nonlinear and non-Gaussian state-space modeling with Monte Carlo simulations. *Journal of Econometrics* **83**, 263–290. doi:10.1016/S0304-4076(97)80226-6.

Tawn, J. A. (1988a) An extreme-value theory model for dependent observations. *Journal of Hydrology* **101**, 227–250. doi:10.1016/0022-1694(88)90037-6.

Tawn, J. A. (1988b) Bivariate extreme value theory: Models and estimation. *Biometrika* **75**, 397–415. doi:10.1093/biomet/75.3.397.

Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika* **77**, 245–253. doi:10.1093/biomet/77.2.245.

Thibaud, E., Mutzner, R. & Davison, A. C. (2013) Threshold modeling of extreme spatial rainfall. *Water Resources Research* doi:10.1002/wrcr.20329. To appear.

Varin, C. (2008) On composite marginal likelihoods. *Advances in Statistical Analysis* **92**, 1–28. doi:10.1007/s10182-008-0060-7.

Varin, C. & Czado, C. (2010) A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics* **11**, 127–138. doi:10.1093/biostatistics/kxp042.

Varin, C., Reid, N. & Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.

Varin, C. & Vidoni, P. (2005) A note on composite likelihood inference and model selection. *Biometrika* **92**, 519 –528. doi:10.1093/biomet/92.3.519.

Varin, C. & Vidoni, P. (2008) Pairwise Likelihood Inference for General State Space Models. *Econometric Reviews* **28**, 170–185. doi:10.1080/07474930802388009.

Wackernagel, H. (2003) *Multivariate Geostatistics*. New York: Springer, 3 edition.

Wadsworth, J. L. & Tawn, J. A. (2012) Dependence modelling for spatial extremes. *Biometrika* **99**, 253–272. doi:10.1093/biomet/asr080.

Wadsworth, J. L. & Tawn, J. A. (2013) Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. Submitted to Biometrika.

Westra, S. & Sisson, S. A. (2011) Detection of non-stationarity in precipitation extremes using a max-stable process model. *Journal of Hydrology* **406**, 119–128. doi:10.1016/j.jhydrol.2011.06.014.

Index

RaphaëlHuser

About me
Swiss and French
27 years old
Married

Address
Croix-Rouges 6
1007 Lausanne
Switzerland

Phone
+41 (0) 21 693 5596 (*prof.*)
+41 (0) 78 730 1146 (*mobile*)

Contact
raphael.huser@gmail.com
http://mathaa.epfl.ch/~rhuser

EDUCATION

2009-2013 Swiss Federal Institute of Technology (EPFL), **PhD in Statistics**
DISSERTATION TITLE: Statistical Modeling and Inference for Spatio-Temporal Extremes
JURY MEMBERS: Prof Anthony C. Davison (*thesis director*), Prof Thomas Mountford (*jury president*), Prof Stephan Morgenthaler (*internal examiner*), Prof Holger Rootzén (*external examiner*), Prof Jonathan A. Tawn (*external examiner*).

2007-2009 EPFL, **MSc in Applied Mathematics**

2004-2007 EPFL, **BSc in Mathematics**

2001-2004 Yverdon-les-Bains high school, **Maturité fédérale and Baccalaureate**

HONORS / AWARDS

2010 **Winner of the PhD poster competition**
Workshop on Environmetrics, Boulder CO (USA)

2009 **2nd prize for the best poster summarizing the master project**
EPFL, Mathematics institute

2004 **1st prize for the best marks in advanced mathematics**
Yverdon-les-Bains high school

2004 **1st prize for the best marks in physics and applied mathematics**
Yverdon-les-Bains high school

RESEARCH INTERESTS

- Statistical modelling of extreme events in space and time
- Max-stable models
- Asymptotic independence models
- Inference for max-stable processes
- Efficiency of composite likelihoods
- Spatial interpolation for large datasets (tapering)

TEACHING

Monte Carlo inference	Principal assistant, EPFL, Fall 2012 Teaching assistant, EPFL, Fall 2010
Mathematics projects	Teaching assistant, EPFL, Spring 2012
Statistics of extremes	Principal assistant, EPFL, Fall 2011 Principal assistant, EPFL, Fall 2009
Calculus	Teaching assistant, EPFL, Fall 2011
Time series	Teaching assistant, EPFL, Spring 2011 Teaching assistant, EPFL, Spring 2010
Probability & Statistics	Teaching assistant, EPFL, Spring 2011
Statistics	Principal assistant, EPFL, Spring 2010

PUBLICATIONS

4) **Huser, R.** and Davison, A.C. (2013) Space-time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, To appear.

3) Davison, A.C., **Huser, R.** and Thibaud, E. (2013) Geostatistics of Dependent and Asymptotically Independent Extremes. *Mathematical Geosciences* 45(5), pp.511–529, DOI: 10.1007/s11004-013-9469-y.

2) **Huser, R.** and Davison, A.C. (2013) Composite likelihood estimation for the Brown–Resnick process. *Biometrika* 100(2), pp.511–518, DOI:10.1093/biomet/ass089.

1) Anderes, E., **Huser, R.**, Nychka, D. and Coram, M. (2012) Nonstationary positive definite tapering on the plane. *Journal of Computational and Graphical Statistics*, DOI:10.1080/10618600.2012.729982.

PRESENTATIONS / POSTERS

2013 **INVITED SPEAKER**, King Abdullah University of Science and Technology (KAUST), Thuwal (Saudi Arabia).
"Statistical modeling of extreme events in space and time"

TALK, *10^{th} Graduate Colloquium in Mathematics*, Bern (Switzerland).
"Statistical modeling of extreme rainfall in space and time"

2012 **INVITED TALK**, *Zürich Extremes meeting*, Zürich Development Center, Zürich (Switzerland).
"Statistical models for rainfall extremes"

INVITED SPEAKER, *Composite Likelihood Methods*, Banff International Research Station (BIRS), Banff (Canada).
"Modelling of extreme rainfall in space and time'

INVITED SPEAKER, *XII Latin American Congress of Probability and Mathematical Statistics (CLAPEM)*, Viña del Mar (Chile).
"Modelling of extreme rainfall in space and time'

2011 **INVITED TALK** (shared with A. C. Davison), *7^{th} Conference on Extreme Value Analysis, Probabilistic and Statistical Models and their Applications (EVA)*, Lyon (France).
"Modelling of extreme rainfall in space and time'

TALK, Annual meeting of the EXTREME group, Davos (Switzerland).
"Space-time modelling of extreme events and inference for max-stable processes"

2010 **INVITED SPEAKER**, *Transition Workshop on Space-time Analysis for environmental mapping, epidemiology and climate change*, Statistical and Applied Mathematical Sciences Institute (SAMSI), Raleigh NC (USA).

"Space-time modelling of extreme events and inference for max-stable processes"

AWARDED POSTER, *Workshop on Environmetrics*, National Center for Atmospheric Research (NCAR), Boulder CO (USA).
"Space-time modelling and inference for extreme events"

POSTER, *11th International Meeting on Statistical Climatology*, Edinburgh (UK).
"Space-time modelling of extreme events and inference for max-stable processes"

TALK, Annual meeting of the EXTREME group, Davos (Switzerland).
"Threshold-based inference for max-stable processes"

2009 **POSTER**, *Graybill VIII - 6th International Conference on Extreme Value Analysis*, Fort Collins (USA).
"On kriging of extreme precipitation return levels and tapering"

PROFESSIONAL EXPERIENCE

2007-2012 **SUBSTITUTE TEACHER** at the middle school level
• **Courses taught**: *Mathematics, Applied maths/physics, Informatics, Biology, Geography, Music, Sport, Professional orientation.*

Sum. 2007-8 **ASSOCIATION SUISSE ROMANDE ET ITALIENNE CONTRE LES MYOPATHIES (ASRIM)**
• Accompanying adult for young people suffering from muscular dystrophy during summer camps.
• 24h/24 support for all daily practical tasks.

Sum. 2004-6 **WESSER UND PARTNER**
• Public relations for the Swiss regional Red Crosses.
• Spokesman for the recruitment of new members and the exposition of the services delivered.

Sum. 2003 **PROTECTION ONE**
• Computer work and folder classification.
• Field work with the installation of security devices at customers' homes.

OTHER SKILLS

Languages: **FRENCH** (Mother tongue), **ENGLISH** (Good), **GERMAN** (Intermediate)

Informatics: R, Matlab, Mathematica, C++, notions of Java, LaTeX.

INTERESTS

- I'm passionate about digital **photography**. My main interests are macro, landscape and wedding photography.
- I enjoy **sports**, especially volleyball, badminton and running.
- I like **hiking** in the Alps, and have been involved in several *Jeunesse et Sport* summer camps —section *Sport de camp/Trekking*— as volunteer counsellor for adolescents.