Contributions to Likelihood-Based Modelling of Extreme Values

Thèse N° 9685

Présentée le 26 juillet 2019

à la Faculté des sciences de base
Chaire de statistique
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Léo RAYMOND-BELZILE

Acceptée sur proposition du jury

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

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acknowledgements

Completing a PhD is not unlike running a marathon: getting past the finish line is the most important aspect, yet this would be impossible without constant support throughout the race.

I would like first and foremost to thank my supervisor, Anthony Davison, for giving me the opportunity to do this PhD and for his constant encouragement. Anthony's kindness and patience have been instrumental in my success and I learned a great deal from his incredibly broad expertise and editorial skills.

I would like also to thank my jury members, Thomas Mountford for presiding the jury and Stephan Morgenthaler, Holger Rootzén and Paul Northrop for accepting to examine my thesis. The latter two did sterling work proofreading the manuscript and made several helpful suggestions.

EPFL offered many great opportunities for my professional development. I wish to express my gratitude to Nadia Kaiser and Anna Dietler, whose professionalism and hard work alleviated my administrative concerns. Thanks to Siara Isaac and Roland Tormey for the training offered by CAPE and for the logistical support for the voluntary teaching evaluation pilot-project. I had numerous teaching assignments over the year with numerous instructors: Jérôme Scherer, Jean-Marie Helbling, Anthony Davison, Emeric Thibaud and Darlene Goldstein. Jean-Marie and Emeric gave me a lot of freedom to choose the content of the exercise sessions and Darlene and Jérôme handled everything, making my teaching task much lighter.

I had the pleasure and privilege of being PhD student representative and of representing my EDMA colleagues at the section and the teaching commission throughout my studies. Thanks to Sarah, Arnaud, André, Adamandia, Kaitlin, Sameer, Vivek, Soham and all the rest for their devotion and our nice discussions about the role of PhD students. Special thanks to Margaux Voumard, with whom I worked on the voluntary teaching evaluations, for being an awesome teammate. I would also like to thank my co-representatives in mathematics, Samuel, Ondine and Émile, and the former members Rosalie, Alix and Maxime, who formed part of the QED team and help keep the program alive.

I would also like to thank my coauthors and my former supervisor, Johanna Nešlehová, for her guidance during my master and her encouragements to pursue a PhD in Switzerland.

L'environnement de recherche compte pour beaucoup dans l'expérience, et j'ai eu la chance d'être entouré de gens aimables, attentionnés et joyeux. Merci aux autres doctorants de la chaire de statistique, Alix, Emeric, Linda, Claudio, Raphaël, Thomas, Yousra, Hélène, Jonathan

et Soumaya, ainsi que les autres collègues, notamment Sebastian, Peiman, Sophie, Stefano et Erwan, d'avoir contribué à créer cette belle atmosphère de groupe et pour votre camaraderie. Je me dois de remercier mon collègue de bureau, Claudio, dont les conseils et la sagesse m'ont été d'un grand secours, ainsi que nos discussions. Merci à Hélène qui a pris la relève et avec qui je partage les dessins de Sempé. Merci à Linda et Raphaël pour les encouragements et votre aide précieuse lorsque je buttais sur des problèmes. Un merci également aux membres présents et passés des chaires de STAP et SMAT, notamment Yoav, Mikael, Marie-Hélène, Tomas[2], Rémy et Guillaume. Merci aux doctorant(e)s des autres universités pour les sorties lors des écoles de la CUSO.

J'aimerais remercier les expatriés québécois de service qui m'ont permis d'entretenir ma québécitude, Audrey-Anne, Julien et Marie-Hélène. Si la tendance se maintient, on se retrouvera tous de l'autre côté de l'Atlantique.

Merci aux équipes du théâtre de Vidy, en particulier Fanny et Claire pour leur confiance et pour m'avoir permis de faire partie des ambassadeurs culturels du théâtre. J'ai fait de belles découvertes que je garderai précieusement en mémoire. Merci aux courageux cobayes qui m'ont accompagnés lors de certaines de ces sorties, dont Béryl et Matthieu.

Merci au vigneron vaudois Alain Chollet et ses vendanges à la carte, qui m'ont permis de découvrir la culture de la vigne et les paysages spectaculaires du Lavaux.

What would be a stay in Switzerland without some exploration of the Alps? Thanks to a devoted group of mountain aficionados, notably Martin and Eva, Alex, Giuseppe, Simon, Kristina, Mario, Victor, Dmitry, Volodymyr. Thanks to you, my hiking buddies, I learned to step off the beaten path and cultivated great friendship on the way to the summits. While I stayed five years at the same place, the constant flow of flatmates has helped me forge new friendships: thanks to Florence, Roland, Kévin, Flavio, Davide, Christopher, Florian, Teodoro, Lusine, Boris, Nicolas, Angélique and Giovanni for the shared adventures and dinners.

J'aimerais remercier ma famille, notamment mes oncles et tantes Robert et Hélène, Geneviève et Guy, Jean-Guy et Lisane, Danielle, François et Linda pour leur support durant toutes mes études. Merci à Corinne et Amaryllis pour les concerts à Bâle et les visites dans la cité rhénane. J'aimerais remercier Lisanne et Simon et mes vieux amis Yolande, Carl, Eva, Yvon, Michèle, Michel, Louise, Roger, René et tous les gaspésiens et gaspésiennes qui me permettent de retrouver mes racines lorsque je rentre à la maison.

Finalement, merci à mes parents Yvan et Nicole, dont le soutien sans faille et l'amour inconditionnel m'ont porté durant toutes ces années. Merci d'avoir toujours cru en moi et de m'avoir aidé à accomplir cette dernière étape; je vous dédie cette thèse.

This thesis was financially supported by a doctoral postgraduate scholarship from National Science and Research Council of Canada (CGSD3–459751–2014). The data used in Chapter 3 was provided by MeteoSwiss.

Lausanne, le 8 juillet 2019

Abstract

Extreme value analysis is concerned with the modelling of extreme events such as floods and heatwaves, which can have large impacts. Statistical modelling can be useful to better assess risks even if, due to scarcity of measurements, there is inherently very large residual uncertainty in any analysis. Driven by the increase in environmental databases, spatial modelling of extremes has expanded rapidly in the last decade. This thesis presents contributions to such analysis.

The first chapter is about likelihood-based inference in the univariate setting and investigates the use of bias-correction and higher-order asymptotic methods for extremes, highlighting through examples and illustrations the unique challenge posed by data scarcity. We focus on parametric modelling of extreme values, which relies on limiting distributional results and for which, as a result, uncertainty quantification is complicated. We find that, in certain cases, small-sample asymptotic methods can give improved inference by reducing the error rate of confidence intervals. Two data illustrations, linked to assessment of the frequency of extreme rainfall episodes in Venezuela and the analysis of survival of supercentenarians, illustrate the methods developed.

In the second chapter, we review the major methods for the analysis of spatial extremes models. We highlight the similarities and provide a thorough literature review along with novel simulation algorithms. The methods described therein are made available through a statistical software package.

The last chapter focuses on estimation for a Bayesian hierarchical model derived from a multivariate generalized Pareto process. We review approaches for the estimation of censored components in models derived from (log)-elliptical distributions, paying particular attention to the estimation of a high-dimensional Gaussian distribution function via Monte Carlo methods. The impacts of model misspecification and of censoring are explored through extensive simulations and we conclude with a case study of rainfall extremes in Eastern Switzerland.

Key words: extreme values, ℓ-Pareto process, max-stable process, higher-order asymptotics, likelihood, uncertainty quantification, Bayesian hierarchical model.

Résumé

L'analyse des valeurs extrêmes sert à la modélisation d'événements rares tels que les inondations ou les vagues de chaleurs, dont l'impact peut être considérable. Beaucoup d'incertitude résiduelle demeure lors de l'inférence statistique en raison de la faible taille échantillonnale, or quantifier adéquatement cette incertitude est primordiale afin de mitiger les risques posés par les catastrophes naturelles. La modélisation spatiale des extrêmes a évolué rapidement au cours de la dernière décennie, portée par la disponibilité croissante de nombreux jeux de données dont la taille est conséquente. Les contributions de cette thèse servent à enrichir de telles analyses.

Le premier chapitre de cette thèse porte sur l'utilisation de la vraisemblance dans le cadre unidimensionnel. Les méthodes asymptotiques d'ordre supérieur et la correction de biais sont dérivées et validées par le biais de simulations et d'exemples, illustrant les défis uniques posés par la rareté des données. Une attention particulière est portée à la modélisation paramétrique, qui découle de l'utilisation de lois limites ; ces derniers sont approximatives à des niveaux finis, ce qui complique la quantification de l'incertitude. Dans certains cas, l'étude de simulation indique que les méthodes pour petits échantillons réduisent le taux d'erreur des intervalles de confiance. Deux analyses de données plus poussées, sur des précipitations vénézuéliennes et sur la survie de supercentenaires italiens, servent à illustrer les méthodes développées.

Le deuxième chapitre présente un survol des méthodes pour l'analyse des extrêmes fonctionnels et spatiaux, qui sert à souligner les similitudes entre les différents modèles pour les méthodes de seuillage et les maxima de blocs. Une revue de litérature exhaustive est fournie, accompagnée des quelques propositions pour des algorithmes de simulation. Les méthodes décrites dans les chapitre sont implémentés dans un paquetage statistique.

Le dernier chapitre porte sur l'estimation de modèles Bayésiens hiérarchiques pour une classe de modèles généralisés de Pareto. Différentes approches numériques pour l'estimation de la censure sont explorées et une attention particulière est portée à l'estimation des fonctions de répartitions multidimensionnelles de distributions elliptiques par le biais de méthodes de Monte Carlo. L'impact de la mauvaise spécification du modèle et la perte d'information due à la censure sont quantifiées par le biais de simulations et une étude de cas sur des précipitations en Suisse orientale conclut l'analyse.

Mots clefs : valeurs extrêmes, processus ℓ-Pareto, processus max-stable, asymptotiques d'ordre supérieur, vraisemblance, inférence statistique, quantification de l'incertitude, modèle hiérarchique bayésien.

Contents

Acknowledgements i

Abstract (English/Français) iii

Introduction 1

1 Likelihood estimation for univariate extremes 3
1.1 Introduction . . . 3
1.1.1 Extreme value distributions . . . 4
1.1.2 Return levels and extreme quantiles . . . 8
1.1.3 Penultimate approximations . . . 10
1.1.4 Asymptotic bias of maximum likelihood estimators . . . 13
1.1.5 Threshold selection methods . . . 20
1.2 Asymptotics for likelihood-based inference . . . 26
1.2.1 Likelihood and cumulants . . . 26
1.2.2 Nuisance parameters and the profile likelihood . . . 29
1.2.3 Cox–Snell bias correction . . . 31
1.2.4 Firth's score correction . . . 33
1.2.5 Test statistics based on first-order asymptotics and Bartlett adjustment . 35
1.3 Higher-order asymptotics . . . 38
1.3.1 Modified profile likelihood and the p^* approximation . . . 39
1.3.2 Orthogonal parametrization . . . 41
1.3.3 Tangent exponential model . . . 44
1.3.4 Other penalized profile likelihoods . . . 45
1.4 Simulation studies . . . 49
1.4.1 Bias corrections for extreme value distributions . . . 49
1.4.2 Higher-order asymptotic methods for confidence intervals . . . 51
1.4.3 Practical guidelines . . . 55
1.5 Data illustrations . . . 61
1.5.1 Vargas tragedy . . . 61
1.5.2 Super centenarians . . . 63
1.6 Conclusion . . . 71

2 A panorama of spatial extremes 73

2.1 Preliminary notions . . . 73
2.1.1 Point processes . . . 73
2.1.2 Properties of spatial processes . . . 76
2.1.3 Convergence of measures . . . 78
2.1.4 Regular variation . . . 80
2.2 Max-stable processes . . . 81
2.2.1 Spectral representation of max-stable processes . . . 83
2.2.2 Parametric models for max-stable processes . . . 87
2.2.3 Parametric models for multivariate extreme value distributions . . . 92
2.3 ℓ-Pareto processes and generalizations . . . 97
2.3.1 ℓ-Pareto processes . . . 99
2.3.2 Generalized ℓ-Pareto processes . . . 101
2.4 Conditional extremes . . . 102
2.4.1 Conditional spatial extremes . . . 105
2.5 Simulation . . . 106
2.5.1 Extremal functions, sub-extremal functions and hitting scenario . . . 106
2.5.2 Unconditional simulations of max-stable processes . . . 108
2.5.3 Conditional simulation of max-stable processes . . . 113
2.5.4 Unconditional simulation of ℓ-Pareto processes . . . 116
2.5.5 Unconditional simulation of generalized ℓ-Pareto processes . . . 120
2.5.6 Unconditional simulation from the conditional extremes model . . . 122
2.6 Likelihoods and estimating functions . . . 124
2.6.1 Max-stable process likelihood . . . 125
2.6.2 Likelihood for ℓ-Pareto processes and generalizations . . . 126
2.6.3 Censored likelihoods . . . 128
2.6.4 Estimating functions . . . 130
2.6.5 Nonparametric estimation of the angular measure . . . 136
2.7 Dependence measures and diagnostics . . . 139
2.7.1 Coefficients of tail dependence . . . 139
2.7.2 Extremogram . . . 142
2.7.3 Extremal coefficient . . . 144
2.7.4 Extremal concurrence probability . . . 148
2.8 Model assessment and diagnostics . . . 150
2.8.1 Proper scoring rules . . . 151
2.9 Summary . . . 152

3 Bayesian hierarchical modelling of generalized ℓ-Pareto processes 155
3.1 Basics of Markov chain Monte Carlo methods . . . 158
3.1.1 Data augmentation and pseudo-marginal methods . . . 160
3.1.2 Hamiltonian Monte Carlo . . . 161
3.1.3 Monitoring convergence and measures of efficiency . . . 162
3.1.4 Adaptive Markov chain Monte Carlo . . . 164

3.2 Numerical evaluation of elliptical distribution functions . . . 164
3.2.1 Separation of variables . . . 165
3.2.2 Minimax exponential tilting . . . 167
3.2.3 Variable reordering . . . 168
3.3 Penultimate models for extremes based on scale mixtures . . . 171
3.4 Bayesian linear model . . . 175
3.5 Bayesian inference for generalized ℓ-Pareto processes . . . 178
3.5.1 Data augmentation . . . 178
3.5.2 Simulation study: logistic max-Pareto process . . . 180
3.6 Simulation study: Bayesian hierarchical model . . . 187
3.7 Data applications . . . 197
3.7.1 Zürich rainfall . . . 197
3.7.2 Swiss rainfall . . . 202
3.7.3 Concluding remarks . . . 206

Conclusion and future work **209**

A Supplementary material for Chapter 1 **211**
A.1 Cumulants of the generalized extreme value distribution . . . 211
A.2 Cumulants of the generalized Pareto distribution . . . 213
A.3 Simulation results for higher order methods . . . 216

B Variogram and covariance models **227**

C Properties and simulation of elliptical distributions **231**
C.1 Bayesian linear model and properties of elliptical distributions . . . 231
C.2 Simulation of (truncated) Gaussian processes . . . 232

D Functionalities of the R package mev **235**

Bibliography **251**

Index **253**

Introduction

On June 11th, 2018, a storm caused havoc in downtown Lausanne. The event was unprecedented: an estimate of 41mm rain fell within 10 minutes, causing an estimated damage of more than 27 millions Swiss francs due to flooding. Pictures in the newspaper showed surreal scenes of cars sliding downhill and staircases at the train station turning into waterfalls. Extreme events are rare, but they are often destructive, fatal and costly.

Extreme events often become ingrained in the collective psyche. In Quebec, the 1996 Saguenay floods are captured by the image of a white house of the Chicoutimi borough standing alone in the middle of the flooded river. Pictures of the electric pylons crumbling under the weight of the freezing rain during the 1998 ice storm and radar picture of the clouds generated by the massive wildfires in Alberta in 2015, which forced the evacuation of more than 80 000 people, illustrate the impacts and the often large spatial extent of natural catastrophes. Extreme value analysis deals with the study of such events, which are often characterized by their complex dependence structure. The field has been booming in the last decade, in part driven by the climate crisis and the availability of new sources of data. In extreme value, the big data revolution has most often lead to larger dimensions, in spatio-temporal settings, as opposed to large sample sizes. Extreme value analysis is inherently difficult because there is little data available and even no guarantees of the quality of measurements. Long records are needed to make any inference, because the theory prescribes that only data which are large in some sense be used for inference. Nevertheless, limiting results can guide us in providing estimates, bearing in mind that the events of interest may be larger than those on record.

Outline of the thesis

The main objective of this thesis is to explore uncertainty quantification for extremes, both in the univariate setting and for spatial processes. Chapter 1 surveys classical and well-established likelihood-based modelling for univariate extremes. We consider the different modelling strategies for block maxima and threshold exceedances, review penultimate models in order to better assess the impact of working under domain of attraction conditions. The novel contribution is derivation of bias correction and higher-order asymptotic methods to refine inference for high quantiles of interest, typically those of the distribution of N-year maxima. We provide several examples of derivations of higher-order terms and implement them in generality. The relative performance of modified profile likelihood for confidence

intervals is assessed by simulation. The chapter ends with two data applications. The first, modelling of rainfall extremes in Venezuela, highlights the difficulty in assessing risk when an observation that is much larger than all previous historical records is observed. Based on the records, we show that this event was extremely rare, but not impossible. The second application, which consists of modelling survival of Italian supercentenarians, tries to answer the question of the existence of a finite limit to lifetime from a purely statistical perspective.

Chapter 2 is mostly a literature review. We focus on the functional extreme values and the theory underlying them, showing how threshold exceedances and max-stable models can be unified through a point process approach. We address likelihood inference and simulation and propose novel simulation algorithms as well as implementations of methods described therein in statistical software.

Chapter 3 deals with the joint estimation of marginal and dependence parameters in an exceedances model. We revisit recent proposals in the literature and consider the added value of a new approach for numerical estimation of Gaussian distribution functions. We review latent Gaussian models for extrapolation of marginal parameters in space and combine the latent Gaussian model with the multivariate generalized Pareto distribution, highlighting the inherent difficulties with Bayesian estimation of the model parameters. The novelty of our approach is in joint modelling of the parameters. We also provide a comparison of two-step methods and assess the loss of information loss due to censoring through simulation studies.

Notation

Throughout this thesis, we adopt the following notational conventions: $\mathbb{N}^+$ denotes the natural numbers $1, 2, \ldots$, $\mathbb{R}$ denotes the real numbers and $\mathbb{R}_+^D$ is the positive orthant $[\mathbf{0}^D, \infty^D)$. Vectors in $\mathbb{R}^D$ are denoted by boldface letters, $\boldsymbol{x} = (x_1, \ldots, x_D)$, and $\mathbf{0}_D, \mathbf{1}_D, \infty^D$ are D-vectors with identical components. We denote norms by $\|\cdot\|$, and use $\|\cdot\|_p$ for the ℓ_p-norm or L_p norm, depending on the argument, $\mathbf{1}\{\cdot\}$ is the Dirac δ function, which equals 1 if the statement is true and zero otherwise. For any $x \in \mathbb{R}$, x_+ denotes the positive part of x, $\max\{0, x\}$. All binary operations such as $\boldsymbol{x} + \boldsymbol{y}$ or $a \cdot \boldsymbol{x}$, $\boldsymbol{x}^a$ are understood as component-wise. The symbol $\overset{\mathrm{d}}{=}$ indicates equality in the sense of finite-dimensional distributions.

We use sans-serif fonts to denote stochastic processes or random functions, so X is a random function whose pointwise realization at $\boldsymbol{s}$ is $X(\boldsymbol{s})$, and $\mathsf{X}(\mathcal{S}) := \{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}}$ is a stochastic process defined on a metric space $(\mathcal{S}, d)$ for a distance function d. We assume $(\mathcal{S}, d)$ is compact throughout and that $\mathsf{X}(\mathcal{S})$ has continuous sample paths, denoted by $\mathsf{X}(\mathcal{S}) \in \mathcal{C}(\mathcal{S}, \mathbb{R})$. The sup-norm of $f \in \mathcal{C}(\mathcal{S}, \mathbb{R})$ is $\|f\|_\infty = \sup_{\boldsymbol{s}\in\mathcal{S}} |f(\boldsymbol{s})|$ and the sphere with respect to a norm $\|\cdot\|_{\mathrm{ang}}$ is $\mathbb{S}_{\mathrm{ang}} = \{f \in \mathcal{C}(\mathcal{S}) : \|f\|_{\mathrm{ang}} = 1\}$. We denote the compact sets of a complete separable metric space $\mathcal{O}$ by $\mathscr{K}(\mathcal{O})$ and the boundary of a set B by ∂B. A measure Λ on $\mathcal{O}$ is totally finite if $\Lambda(\mathcal{O}) < \infty$, Radon if $\Lambda(K) < \infty$ for any compact set $K \in \mathscr{K}(\mathcal{O})$ and boundedly finite if $\Lambda(B) < \infty$ for any bounded Borel set $B \in \mathbb{B}(\mathcal{O})$. The Λ-continuity sets of $\mathcal{O}$ are Borel sets whose boundaries have zero measure, $\{B \in \mathbb{B}(\mathcal{O}) : \Lambda(\partial B) = 0\}$.

1 Likelihood estimation for univariate extremes

1.1 Introduction

Estimating worst-case scenarios is important for risk management and policy making, but the hypothetical events that decision makers are usually interested in can lie far beyond the range of the available data and empirical estimates cannot be used. Asymptotic theory is therefore used to infer the tail behaviour of the phenomenon of interest based on the largest observations. Estimation is often based on small samples, which raises the question of the accuracy and hence practical usefulness of the asymptotic results, especially those that rely on asymptotic normality of the maximum likelihood estimator. The focus of this work is on likelihood-based methods, as the latter are necessary for Bayesian inference and can be easily extended to include covariates, to pool information, to accommodate censoring and truncation, etc.

The goal of this chapter is to investigate the use of bias correction and higher-order approximations for univariate extremes. We first introduce the extreme value distributions, and explain how they arise as limiting models for peaks over threshold and block maxima. Since the results only hold in the limit, we consider penultimate approximations and asymptotic bias due to model misspecification, providing various examples. We investigate the use of the so-called metastatistical model for inference. We then consider the estimation bias of maximum likelihood estimators and outline the use of higher-order asymptotics for better uncertainty quantification, focusing on variants of the profile likelihood for inference on a scalar of interest. The use of the tangent exponential model for the generalized extreme value distribution, as well as for functions of the parameters, is novel and we assess by simulation whether these lead to improved inference. We conclude with two examples, the first using higher order asymptotics to perform inference on the centennial maximum distribution of rainfall extremes in Venezuela and the second assessing the existence of an upper limit to human life based on survival records for Italian supercentenarians.

1.1.1 Extreme value distributions

Definition 1.1 (Generalized extreme value distribution)
The distribution function of the generalized extreme value (GEV) distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma \in \mathbb{R}_+$ and shape parameter $\xi \in \mathbb{R}$ is

$$G(x) = \begin{cases} \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}, & \xi \neq 0, \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, & \xi = 0, \end{cases}$$

defined on $\{x \in \mathbb{R} : \xi(x-\mu)/\sigma > -1\}$ where $x_+ = \max\{0, x\}$. The case $\xi = 0$ is commonly known as the Gumbel distribution. We denote the distribution by $\mathsf{GEV}(\mu, \sigma, \xi)$.

The generalized extreme value distribution is max-stable: if $F \sim \mathsf{GEV}$, then there exist location and scale constants $b \in \mathbb{R}$ and $a > 0$ such that $F^T(ax+b) = F(x)$ for any $T \in \mathbb{N}^+$. The parameters of the new distribution are easily derived: if $X_i \overset{\text{iid}}{\sim} \mathsf{GEV}(\mu, \sigma, \xi)$, then $\max\{X_1, \ldots, X_T\} \sim \mathsf{GEV}(\mu_T, \sigma_T, \xi)$ with $\mu_T = \mu - \sigma(1-T^\xi)/\xi$ and $\sigma_T = \sigma T^\xi$ for $\xi \neq 0$, or $\mu_T = \mu + \sigma\log(T)$ and $\sigma_T = \sigma$ if $\xi = 0$.

Definition 1.2 (Generalized Pareto distribution)
The distribution function of the generalized Pareto (GP) distribution with scale $\sigma \in \mathbb{R}_+$ and shape $\xi \in \mathbb{R}$ is

$$G(x) = \begin{cases} 1 - \left(1+\xi\frac{x}{\sigma}\right)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\sigma}\right)_+, & \xi = 0. \end{cases}$$

The range of the generalized Pareto distribution is $[0, -\sigma/\xi)$ if $\xi < 0$ and is $\mathbb{R}_+$ otherwise. We denote the distribution by $\mathsf{GP}(\sigma, \xi)$.

If $X \sim \mathsf{GP}(\sigma, \xi)$, straightforward calculations show that $X - u \mid X > u \sim \mathsf{GP}(\sigma + \xi u, \xi)$ for any $u \in \mathbb{R}$ such that $\sigma + \xi u > 0$, so conditional exceedances above a threshold u also follow a generalized Pareto distribution. This property is termed threshold-stability.

Under mild conditions, most univariate distribution functions are attracted to extreme value models in the sense defined next (Embrechts et al., 1997, Theorem 3.4.5).

Theorem 1.3 (Extreme value attractors)
Let $\{X_i\}_{i\in\mathbb{N}^+}$ be a sequence of independent and identically distributed random variables with distribution function F and upper endpoint $x^* := \inf\{x : F(x) = 1\}$. Assume that there exist sequences of constants $a_n = a(n) > 0$ and $b_n = b(n)$ for any $n \in \mathbb{N}^+$ such that any of the following statements hold:

1. max-stable limit: $\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$, where $G(x)$ is $\mathsf{GEV}(0, 1, \xi)$;
2. generalized Pareto limit: the conditional distribution of exceedance over a threshold

$u < x^*$ converges to a generalized Pareto distribution,

$$\lim_{u \to x^*} \frac{1 - F(xa_u + u)}{1 - F(u)} = 1 - H(x),$$

where $H(x)$ is the distribution function of $\mathsf{GP}(1, \xi)$;

3. Poisson point process: the sequence of point processes $N_n = \{i/(n+1), (X_i - b_n)/a_n\}_{i \in \mathbb{N}^+}$ converges as $n \to \infty$, for any set of the form $(0,1) \times [z_*, \infty)$, to a non-homogeneous Poisson point process with limiting measure (Coles, 2001, Theorem 7.1)

$$\Lambda\left\{[t_1, t_2] \times [x, z^*)\right\} = (t_2 - t_1)\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-1/\xi}, \qquad 0 \leq t_1 \leq t_2 \leq 1, x > z_*,$$

where z_*, z^* are the lower and upper endpoints of the corresponding generalized extreme value distribution. The case $\xi = 0$ is understood by continuity to be the limit as $\xi \to 0$.

Then, statements 1–3 are equivalent.

The equivalence between convergence to a generalized extreme value distribution and that of the point process of extremes is discussed in Section 4.4.2 of Resnick (1987).

In practice, the generalized extreme value distribution is used to model block maxima. Max-stability ensures that, if the maxima of blocks of size m form an approximate sample from a generalized extreme value distribution, then bigger blocks should follow the same distribution up to changes in location and scale parameters. The result on the convergence of maxima is due to Fisher and Tippett (1928), whose results were formalized by Gnedenko (1943).

The generalized Pareto is the asymptotic distribution for exceedances over a high threshold u. In practice, u must be lower than x^* and chosen so that there are enough exceedances above u to estimate the parameters; threshold selection is addressed in Section 1.1.5. The generalized Pareto approximation is valid only above the specified threshold u. The points that are flagged as extremes differ for the two methods, as shown by Figure 1.1.

Peaks-over-threshold analysis usually allows one to incorporate more information by discarding fewer data, but temporal dependence needs to be dealt with if observations are not independent, since extremes may cluster.

Because the scale parameter of the generalized Pareto distribution, σ_u, depends on u, it may be tempting to directly use the likelihood of the point process for exceedances falling in the region $(0,1) \times (u, \infty)$ given in Theorem 1.3 (cf. de Haan and Ferreira, 2006, § 2.1).

Proposition 1.4 (Likelihood of the Poisson point process for threshold exceedances)
Consider a sample of N observations, of which n_u exceed u and which we denote by $y_1, \ldots, y_{n_u}$.

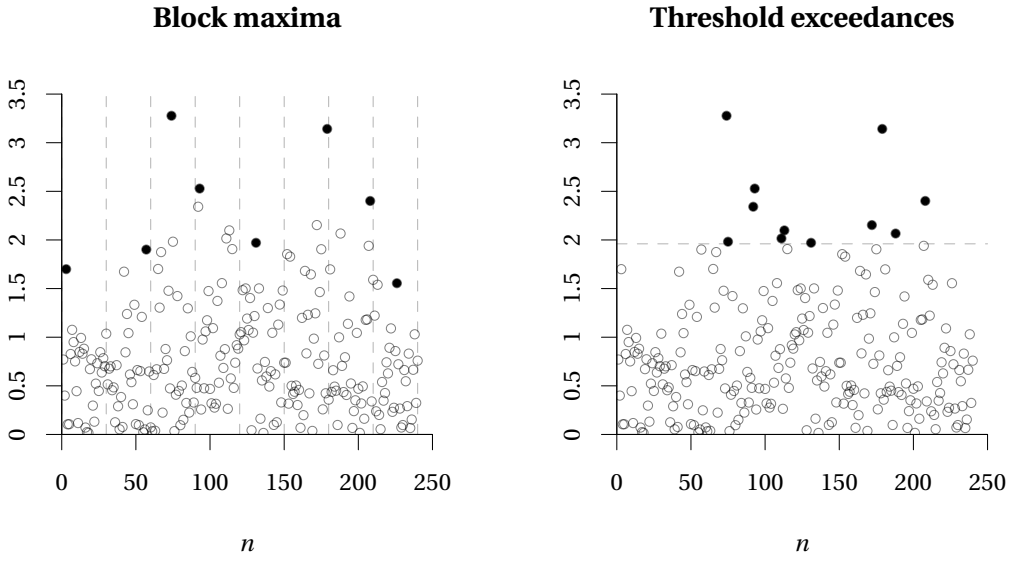

Figure 1.1 – *Extremes for block maximum (left) and peaks-over-threshold (right) methods. Only the black filled points are kept.*

The likelihood associated to the limiting distribution of threshold exceedances is

$$L(\mu,\sigma,\xi;\boldsymbol{y}) = \exp\left[-c\left\{1+\xi\left(\frac{u-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right](c\sigma)^{-n_u}\prod_{i=1}^{n_u}\left\{1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right\}_+^{-1/\xi-1}, \qquad \mu,\xi\in\mathbb{R},\sigma>0, \tag{1.1}$$

where $(\cdot)_+ = \max\{0,\cdot\}$. The quantity c is a tuning parameter whose role is described in § 7.5 of Coles (2001). If we take $c = N/m$, the parameters of the point process likelihood correspond to those of the generalized extreme value distribution fitted to blocks of size m. There is no closed form for the parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\mu},\widehat{\sigma},\widehat{\xi})$, which must be obtained through numerical optimization; we return to the maximization of the likelihood in Example 1.18.

There is often a natural block size for the block maxima method (e.g., yearly for daily data). This can lead to small samples and potentially a waste of information. A compromise is to keep the r largest observations and use a likelihood based on these order statistics.

Proposition 1.5 (Likelihood for the r-largest order statistics)
Let $Y_{(1)} \geq \cdots \geq Y_{(r)}$ denote the r largest observations from a sample. Then, under Theorem 1.3, the likelihood of the limiting distribution of the point process for the r-largest observations is

$$\ell(\mu,\sigma,\xi;\boldsymbol{y}) \equiv -r\log(\sigma) - \left(1+\frac{1}{\xi}\right)\sum_{j=1}^{r}\log\left(1+\xi\frac{y_{(j)}-\mu}{\sigma}\right)_+ - \left(1+\xi\frac{y_{(r)}-\mu}{\sigma}\right)_+^{-1/\xi}, \quad \mu,\xi\in\mathbb{R},\sigma>0.$$

The r-largest observations from the asymptotic model can be generated using the Poisson

process representation by simulating a unit rate Poisson process $0 < U_1 < U_2 < \cdots$, where $U_j = E_1 + \cdots + E_j$ and $E_j \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$, and setting $Y_{(j)} = \mu + \sigma\big(U_j^{-1/\xi} - 1\big)/\xi$.

The inverse transformation,

$$U_j = \{1 + \xi(Y_{(j)} - \mu)/\sigma\}^{-1/\xi}, \quad j = 1, \ldots, r,$$

can be used to obtain residuals by using plug-in estimates $\widehat{\mu}, \widehat{\sigma}, \widehat{\xi}$; the corresponding estimated spacings

$$\widehat{E}_1 = \widehat{U}_1, \quad \widehat{E}_j = \widehat{U}_j - \widehat{U}_{j-1}, \quad j = 2, \ldots, r,$$

are approximately independent, as the Poisson process is Markovian. By using the maximum likelihood estimator in place of the true parameter values, we can obtain approximate pivots $\widehat{E}_i$ that can be used to construct exponential quantile-quantile plots. The r-largest likelihood can be used to model block maxima, by keeping the r largest observations out of m. The choice of the number of order statistics r should typically be small, as the quality of the asymptotic approximation degrades quickly when r increases. One can also use the likelihood for peaks-over-threshold modelling, in which case the threshold is taken to be the rth order statistic.

Remark 1.1 (Notational convention)
In the context of Theorem 1.3, we say that G is the extremal attractor of F and that F is in the max-domain of attraction of G with tail index ξ, denoted by $F \in \mathsf{MDA}(G)$. Following Fisher and Tippett (1928), we say that F is in the max-domain of attraction of a Fréchet, Gumbel, or reverse Weibull distribution whenever the shape parameter ξ is positive, zero or negative, respectively. Only distribution functions with infinite upper endpoint may belong to the max-domain of attraction of the Fréchet class. Many reference textbooks consider the three possible sub-families (Fréchet, Gumbel and reverse Weibull) separately.

For the Fréchet domain of attraction, Theorem 1.3 is equivalent to regular variation of the survival function $1 - F$ with tail index $1/\xi$ (Resnick, 1987, Prop. 1.1). Roughly speaking, regular variation describes the decay rate of functions that behave like polynomials at infinity.

Definition 1.6 (Regular variation of positive measurable univariate function)
A measurable function $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is regularly varying with tail index $\alpha \in \mathbb{R}$ at ∞, denoted by RV_α if (Resnick, 2007, Definition 2.1)

$$\lim_{t\to\infty} \frac{h(tx)}{h(t)} = x^\alpha, \qquad x \in \mathbb{R}_+.$$

If $\alpha = 0$, the function is said to be slowly varying.

The concept of regular variation is central in extreme value theory since it is the mathematical basis allowing for extrapolation, and we generalize this notion in Section 2.1.4. We refer the reader to Chapters 0 and 1 of Resnick (1987) for a detailed exposition of domains of attraction.

Remark 1.2 (Normalizing constants)
If $F \in \mathsf{MDA}(G)$ for $G \sim \mathsf{GEV}(0,1,\xi)$ (cf. de Haan and Ferreira, 2006, Theorem 1.1.2), then there exists $a_t > 0$ and $b_t \in \mathbb{R}$ such that the first-order condition

$$\lim_{t\to\infty} \frac{b(tx) - b_t}{a_t} = \int_1^x s^{\xi-1}\mathrm{d}s = \begin{cases} \frac{x^\xi - 1}{\xi}, & \xi \neq 0, \\ \log(x), & \xi = 0, \end{cases}, \qquad x > 0,$$

is satisfied. Under the assumption that F is twice differentiable with density function f, the so-called von Mises conditions are sufficient to check the existence of extreme-value limits. Define the sequence $b_t \equiv b(t)$ as the solution to $F(b_t) = \exp(-t^{-1})$ and let $a_t \equiv a(t) = s(b_t) = -F(b_t)\log\{F(b_t)\}/f(b_t)$ (Smith, 1987a). Then, the condition $\lim_{x\to x^*} s'(x) = \xi \in \mathbb{R}$ is sufficient for F to belong to the maximum domain of attraction of $\mathsf{MDA}(G)$. For threshold exceedances, we consider instead a threshold $b_t = F(1-1/t)$ and scaling sequence $a_t = r(b_t)$, where $r(x) = \{1-F(x)\}/f(x)$ is the reciprocal hazard function and the extreme value limit exists if $\lim_{x\to x^*} r'(x) = \xi \in \mathbb{R}$.

Remark 1.3
Not all distributions satisfy the conditions of Theorem 1.3. Many discrete distributions cannot be suitably renormalized to extreme value distributions, including the geometric, Poisson and negative binomial distributions (Examples 3.1.4, 3.1.5 and 3.1.6 Embrechts et al., 1997). Asymptotically, these families can exhibit oscillatory behaviour due to the discreteness of the support. The log-Pareto distribution, which in its simplest form has survival function $1 - F(x) = 1 - 1/\log(x)$ for $x > e$, is an example of distribution that is too heavy-tailed to be stabilized by affine rescaling.

The support of the generalized extreme value and generalized Pareto distributions varies with their parameters, making numerical estimation of the parameters difficult. Maximum likelihood estimators for both families are asymptotically Gaussian and the models are regular whenever $\xi > -1/2$ (Smith, 1985; de Haan et al., 2004; Bücher and Segers, 2017; de Haan and Ferreira, 2006, § 3.4).

1.1.2 Return levels and extreme quantiles

Two typical questions in extreme value analysis are: given the intensity of an extreme event, what is its recurrence period? and what is a typical worst-case scenario over a given period of time? For the latter, suppose for simplicity that the daily observations are blocked into years, so that inference is based on N points for the N years during which the data were recorded. The return level is a quantile of the underlying distribution corresponding to an event of probability $p = 1 - 1/T$ for an annual maximum, which is interpreted as "the level exceeded by an annual maximum on average once every T years". If observations are independent and identically distributed, then we can approximate the probability that a return level is exceeded l times over a T year period using a binomial distribution with probability of success $1 - 1/T$ and T trials. For T large, the return level is exceeded $l = 0, 1, 2, 3, 4$ times within any T-years

period with approximate probabilities 36.8%, 36.8%, 18.4%, 6.1% and 1.5%. The probability that the maximum observation over T years is exceeded with a given probability is readily obtained from the distribution of the T-year maximum, leading Cox et al. (2002, § 3(b)) to advocate its use over return levels, among other quantities of interest such as the number of times a threshold u will be exceeded in T years or the average number of years before a threshold u is exceeded. Rootzén and Katz (2013) advocate an alternative design life measure of risk for nonstationary sequences.

Proposition 1.7 (Quantiles, mean and return levels of T-maxima)
Consider the distribution $H(x) = G^T(x)$ of the maximum of T independent and identically distributed generalized extreme value variates with parameters (μ, σ, ξ) and distribution function G. By max-stability, the parameters of $H(x)$ are $\mu_T = \mu - \sigma(1 - T^\xi)/\xi$ and $\sigma_T = \sigma T^\xi$ when $\xi \neq 0$. We denote the expectation of the T-observation maximum by $\mathfrak{e}_T$, the p quantile of the T-observation maximum by $\mathfrak{q}_p = H^{-1}(p)$ and the associated return level by $z_{1/T} := G^{-1}(1 - 1/T)$. Then, any of these three quantities can be written as

$$\begin{cases} \mu - \frac{\sigma}{\xi}\left\{1 - \kappa_\xi\right\}, & \xi < 1, \xi \neq 0, \\ \mu + \sigma\kappa_0, & \xi = 0, \end{cases}$$

where $\kappa_\xi = T^\xi \Gamma(1-\xi)$ for $\mathfrak{e}_T$, $\kappa_\xi = T^\xi \log(1/p)^{-\xi}$ for $\mathfrak{q}_p$ and $\kappa_\xi = \left\{-\log(1-1/T)\right\}^{-\xi}$ for $z_{1/T}$. In the Gumbel case, we have $\kappa_0 = \log(T) + \gamma_e$ for $\mathfrak{e}_T$, $\kappa_0 = \log(T) - \log\{-\log(p)\}$ for $\mathfrak{q}_p$ and $\kappa_0 = -\log\{-\log(1-1/T)\}$ for $z_{1/T}$. The extrapolation of the distribution based on max-stability is illustrated in Figure 1.2.

We can relate threshold exceedances to the distribution of maxima as follows: suppose a generalized Pareto distribution F is fitted to threshold exceedances above a threshold u, with parameters (ζ_u, σ_u, ξ), where ζ_u denotes the unknown proportion of points above the threshold u. If there are n_y observations per year on average, then the T-year return level is $z_{1/T} = u + \sigma_u/\xi\{(Tn_y\zeta_u)^\xi - 1\}$. We take $F^{\zeta_u T n_y}$ as the approximation to the T-year maxima distribution, conditional on exceeding u.

Example 1.1 (Maiquetía rainfall)
We illustrate graphically max-stability using a time series used in Section 1.5.1 for the purpose of illustration. The Maiquetía rainfall series consists of yearly maxima of daily cumulated rainfall measurements (in mm) for the period 1951 to 1999 recorded at the Simón Bolívar International Airport. We fit a generalized extreme value distribution G to the first 48 annual maxima and plot the corresponding density in Figure 1.2. The distribution of the centennial maximum of daily precipitation is obtained by max-stability as G^{100}. The 100-year return level based on G, corresponding to roughly the 0.37-quantile of G^{100}, and the mean of G^{100} are displayed under the density of the latter; due to the skewness, the mean of the centennial maximum is located to the right of the mode. The largest annual maximum, recorded in December 1999, is far into the tail of the estimated centennial maximum distribution.

□

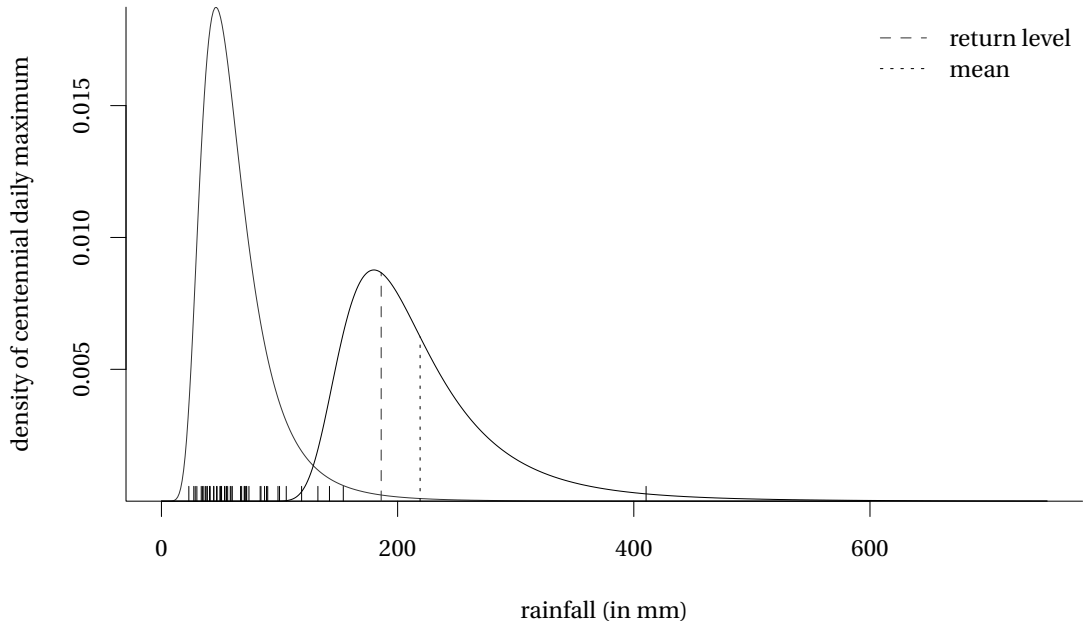

Figure 1.2 – *Extrapolated density of centennial maximum of daily rainfall (full) for the Maiquetía yearly maxima (rug) and fitted generalized extreme value density for the annual maximum distribution (grey), based on data for 1951–1998. The 100-year return level (dashed) and mean of the centennial distribution (dotted) are also displayed. A rug indicates the 410.4mm backcasted extreme record of December 1999.*

1.1.3 Penultimate approximations

The generalized extreme value (GEV) distribution arises as the non-degenerate limiting distribution of maxima. In practice, it is typically fitted to data that are partitioned into k blocks of (potentially random) sizes m_i. We shall assume that the block sizes are all the same, so that $m_i = m$ for $i = 1, \ldots, k$ and $n = km$. The quality of the generalized extreme value approximation increases with the block size m. For fixed block sizes, the estimated shape parameter generally differs from its asymptotic counterpart and this discrepancy usually introduces bias if one extrapolates far beyond the data. This section defines penultimate approximations and second-order regular variation. These are mainly of mathematical interest, since the underlying distribution of the data is unknown in practice. Still, calculations for well-known families of parametric models can be useful in assessing the performance of estimators in Monte Carlo studies.

Let $F(x)$ denote a thrice-differentiable distribution function with upper endpoint x^* and density $f(x)$. Define $s(x) = -F(x)\log\{F(x)\}/f(x)$. The existence of the limit $\xi_\infty = \lim_{n\to\infty} s'\{b_n\}$ is necessary and sufficient for existence of $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left\{-(1+\xi_\infty x)^{-1/\xi_\infty}\right\} = G(x),$$

but also for "twice-differentiable convergence", i.e., convergence of both the corresponding

density function,

$$\lim_{n\to\infty} n a_n f(a_n x + b_n) F^{n-1}(a_n x + b_n) = (1+\xi_\infty x)^{-1/\xi_\infty - 1} \exp\left\{-(1+\xi_\infty x)^{-1/\xi_\infty}\right\} = g(x),$$

and of its derivative uniformly in x on all finite intervals (Pickands, 1986, Theorem 5.2).

Smith (1987a) shows that, if $\xi_\infty = \lim_{n\to\infty} s'\{b_n\}$ holds, for any $x \in \{y : 1+\xi_\infty y > 0\}$, there exists z such that

$$\frac{-\log[F\{v + xs(v)\}]}{-\log\{F(v)\}} = \left\{1 + xs'(z)\right\}^{-1/s'(z)}, \qquad v < z < v + xs(v).$$

For each $n \geq 1$, setting $v = b_n$ and $a_n = s(b_n)$ yields

$$F^n(a_n x + b_n) = \exp\left[-\left\{1 + s'(z)x\right\}^{-1/s'(z)}\right] + \mathrm{O}(n^{-1})$$

for $z \in [\min(a_n x + b_n, b_n), \max(a_n x + b_n, b_n)]$, depending on the support of F. The ultimate approximation replaces $s'(z)$ by $s'(x^*) = \xi_\infty$, whereas Smith suggests instead suggests taking $s'(b_n)$, which is closer to $s'(a_n x + b_n)$ than is $s'(x^*)$.

The convergence rate of the generalized extreme value distribution $F^n(a_n x + b_n)$ is (Wadsworth et al., 2010)

$$\max\left\{\mathrm{O}(s'(b_m) - \xi_\infty), \mathrm{O}(s'(a_m x + b_m) - s'(b_m)), \mathrm{O}(n^{-1})\right\},$$

whereas the convergence rate for the density, pointwise, includes an additional error term of $\mathrm{O}\left(s(a_m x + b_m)/s(b_m) - (1+\xi_\infty x)\right)$. When the term $\mathrm{O}(s'(b_m) - \xi_\infty)$ dominates, Wadsworth et al. (2010) illustrate how a Box–Cox transformation can improve the rate of convergence.

For threshold exceedances, we start with the reciprocal hazard $r(x) = \{1 - F(x)\}/f(x)$ in place of $s(x)$. Smith (1987a) shows that there exists y such that

$$\frac{1 - F\{u + xr(u)\}}{1 - F(u)} = \left\{1 + xr'(y)\right\}_+^{-1/r'(y)}, \qquad u < y < u + xr(u),$$

unless $r'(x)$ is constant. The penultimate shape parameter for the generalized Pareto distribution is $r'(u)$, but the true shape parameter lies between $r'(u)$ and ξ_∞.

When we fit the limiting parametric models to finite samples, maximum likelihood estimates of the shape parameter will tend to be closer to their penultimate counterparts than to the limiting value.

Penultimate approximations show that the estimated shape will change as the threshold or the block size increases. Consider for simplicity yearly maxima arising from blocking $m = n_y$ observations per year. We are interested in the distribution of the maximum of N years and thus in ξ_{Nn_y}, but our estimate will instead target ξ_{n_y}; an extrapolation error arises from this mismatch between the shape values, which would be constant if the observations were truly

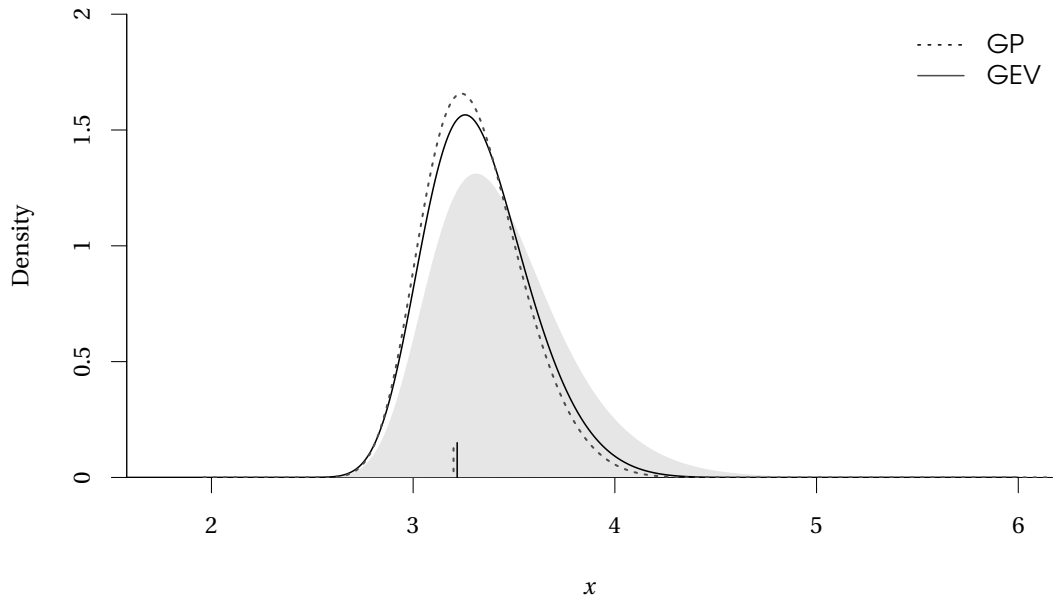

Figure 1.3 – *Estimate of the distribution of the maximum of $N = 2000$ standard Gaussian variates, based on block maxima with $m = 40$ (full) and peaks-over-threshold approximations with $u = z_{0.975}$ (dashed). The grey shaded region shows the true theoretical density function $N\phi(x)\Phi^{N-1}(x)$. The line segments near 3.2 represent the return levels $z_{1/T}$.*

max-stable. The curvature of r' determines how stable the estimates of ξ_t are when the extrapolation window increases.

Example 1.2 (Estimate of the distribution of the maxima of *N* observations)
Figure 1.3 shows the distribution of the maxima of $N = 2000$ standard Gaussian variates (shaded region). The dashed line shows the density of the block maximum penultimate approximation $G \sim \mathsf{GEV}(b_{40}, a_{40}, \xi_{40})$ with blocks of size $m = 40$. The full lines show the penultimate approximation for threshold exceedances above the 97.5% percentile of the standard Gaussian distribution, i.e., $F \sim \mathsf{GP}\{r(u), r'(u)\}$ with threshold $u = \Phi^{-1}(0.975)$ and $\zeta_u = 0.025$. The line segments show the estimated return levels $z_{1/T}$, defined for block maxima as the solution of the equation $G(z_{1/T}) = 1 - 40/2000$ and for threshold exceedances by $F(z_{1/T}) = 1 - (N\zeta_u)^{-1}$. Both correspond to the 36.4% percentile of the distribution of the maximum of N observations.

□

Example 1.3 (Penultimate approximation for lognormal variates)
We illustrate the previous discussion to provide some insight into the relevance of penultimate approximations. The left panel of Figure 1.4 illustrates how maximum likelihood estimates of the parameters for repeated samples from the model are closer, on average, to the penultimate approximation than to the limiting value $\xi_\infty = 0$. By extrapolating the model using max-stability, differences between the approximations are magnified; see the right-hand panel of Figure 1.4. In this case, the penultimate shape for $m = 1000$ is approximately $\xi_{1000} = 0.215$, compared to $\xi_{30} \approx 0.284$, which is far from the limiting value $\xi_\infty = 0$.

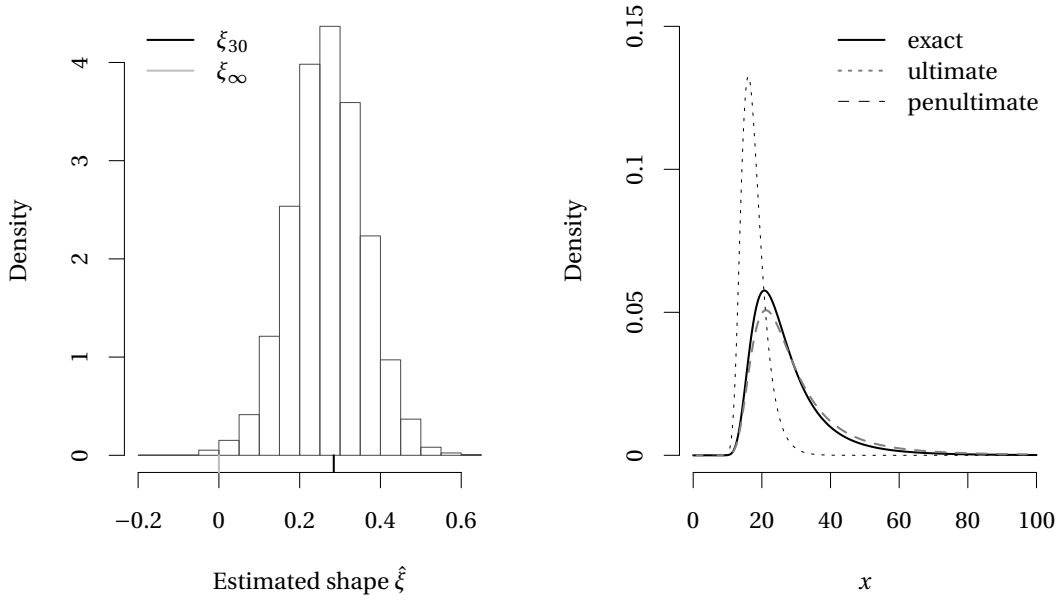

Figure 1.4 – *Penultimate approximation for standard lognormal variates. Left panel: histogram of maximum likelihood estimates of the shape of the generalized extreme value distribution fitted to* $k = 100$ *block maxima of sizes* $m = 30$*, with penultimate approximation (black line at* $\xi_{30} \approx 0.284$*) and limiting parameter (grey line at* $\xi_\infty = 0$*). Right panel: true density of the maxima of* $n = 1000$ *standard lognormal variates (heavy), density of the Smith penultimate approximation* $F_1^{1000/30}$ *for* $F_1 \sim \mathsf{GEV}(a_{30}, b_{30}, \xi_{30})$ *for the generalized extreme value approximation to block maxima of size* $m = 30$ *(dashed) and ultimate approximation* $F_2^{1000/30}$ *for* $F_2 \sim \mathsf{GEV}(a_{30}, b_{30}, \xi_\infty)$ *(short dashes).*

The extrapolated density estimates are based on $F_1^{1000/30}$, the generalized extreme value density associated to the distribution function of the penultimate approximation $F_1^{1000/30}$, with $F_1 \sim \mathsf{GEV}(a_{30}, b_{30}, \xi_{30})$. The penultimate approximation F_2 is more accurate than the ultimate approximation $\mathsf{GEV}(a_{30}, b_{30}, \xi_\infty)$, which is too short tailed.

□

1.1.4 Asymptotic bias of maximum likelihood estimators

Since we use finite block sizes and threshold, maximum likelihood estimators will be biased, more so if the fraction of sample points used for inference is too large. Moreover, this bias may not vanish asymptotically unless the fraction of observations kept for inference grows at a certain rate.

To derive the asymptotic bias of maximum likelihood estimators, one requires generalized regular variation of second order (de Haan and Stadtmüller, 1996). A distribution is generalized

regularly varying of second order if there exists a positive function $A(t)$ such that

$$H_{\xi_\infty,\rho}(x) := \lim_{t\to\infty} \frac{b(tx)-b(t)-a(t)\int_1^x s^{\xi_\infty-1}\mathrm{d}s}{a(t)A(t)} = \int_1^x s^{\xi_\infty-1}\int_1^s u^{\rho-1}\mathrm{d}u\mathrm{d}s, \qquad x>0, \rho\leq 0. \tag{1.2}$$

To avoid degeneracy, we exclude functions $A(t)$ that lead to $H_{\xi_\infty,\rho}(x) \propto \int_1^x s^{-\xi_\infty-1}\mathrm{d}s$. The second-order auxiliary function $A(t)$ satisfies the following constraints: $\lim_{t\to\infty} A(t) = 0$, $\lim_{t\to\infty} |A(tx)/A(t)| = x^\rho$ for $\rho \leq 0$, and $\text{sign}\{A(x_s)\} = \text{sign}\{A(t+x_s)\}$ for some $x_s > 0$ and for all $t > 0$.

de Haan and Resnick (1996) provide a constructive definition of the second-order auxiliary function $A(t)$ in their Theorem 2.1, assuming that $b(x)$ is twice differentiable and $b'(x)$ is eventually positive. Specifically,

$$A(t) := \frac{tb''(t)}{b'(t)} - \xi_\infty + 1 = r'(b_t) - \xi_\infty \tag{1.3}$$

is the difference between the penultimate and ultimate values of the shape parameter.

The asymptotic bias of maximum likelihood estimators derived for fixed threshold or block sizes depends on the second order regular variation parameter, ρ, which can be extracted from the relation $\lim_{t\to\infty} A(tx)/A(t) = x^\rho$ for $\rho \leq 0$, when generalized regular variation of second order holds.

Suppose $\lambda \in \mathbb{R}$ is the asymptotic rate at which the number of extreme observations grow relative to the total sample size, i.e., $k^{1/2}A(n/k) \to \lambda$ as $n, k \to \infty$. If $\lambda = 0$, the asymptotic bias of maximum likelihood estimators due to model misspecification vanishes. The optimal rate at which k must grow relative to n is thus family-dependent. Let $\boldsymbol{\theta}_k = (\widehat{\sigma}_k, \widehat{\xi}_k)$ denote maximum likelihood estimators for the generalized Pareto distribution based on the k largest order statistics from a distribution F for which eq. (1.2) holds. Following Theorem 3.4.2 of de Haan and Ferreira (2006), the rescaled sequence $\widehat{\boldsymbol{\theta}}_k$ converges as $k \to \infty$, $k/n \to 0$, to a Gaussian variate with distribution

$$k^{1/2}\begin{pmatrix} \widehat{\sigma}_k/a(n/k) - 1 \\ \widehat{\xi}_k - \xi \end{pmatrix} \xrightarrow{d} \mathsf{No}\left\{\frac{\lambda}{(1-\rho)(1+\xi-\rho)}\begin{pmatrix}-\rho \\ 1+\xi\end{pmatrix}, \begin{pmatrix} 1+(1+\xi)^2 & -(1+\xi) \\ -(1+\xi) & (1+\xi)^2 \end{pmatrix}\right\}.$$

This is proved in de Haan et al. (2004), who report a different expression in their Corollary 2.1 for the asymptotic bias to that in de Haan and Ferreira (2006); see also Smith (1987b) for the case where u is a fixed threshold.

A similar result holds for block maxima. Let k be the number of maxima of blocks of size $m = \lfloor n/k \rfloor$ and let $\ell(\boldsymbol{\theta}; x)$ and $i(\boldsymbol{\theta})$ respectively denote the log-likelihood and the Fisher information of the generalized extreme value distribution for a single observation and let $Q_{\boldsymbol{\theta}}(s)$ denote the

quantile function of a generalized extreme value distribution. Further define $\theta_0 = (0, 1, \xi)$, and

$$B(\xi, \rho) := \int_0^1 \left. \frac{\partial^2 \ell(\boldsymbol{\theta}, x)}{\partial x \partial \boldsymbol{\theta}} \right|_{x=Q_{\boldsymbol{\theta}}(s), \boldsymbol{\theta}=\boldsymbol{\theta}_0} H_{\xi,\rho}\left\{-\frac{1}{\log(s)}\right\} \mathrm{d}s,$$

where $H_{\xi_\infty,\rho}$ is defined in eq. (1.2). Then the sequence of maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_n$ based on $k = \lfloor n/m \rfloor$ maxima of blocks of length m converges in law to a Gaussian distribution as $k \to \infty$, $k/n \to 0$, viz. (Dombry and Ferreira, 2019, Theorem 2.1)

$$k^{1/2} \begin{pmatrix} (\widehat{\mu}_n - b_m)/a_m \\ \widehat{\sigma}_n/a_m - 1 \\ \widehat{\xi}_n - \xi_0 \end{pmatrix} \xrightarrow{d} \mathsf{No}\left\{\lambda i^{-1}(\boldsymbol{\theta}_0) B(\xi, \rho), i^{-1}(\boldsymbol{\theta}_0)\right\}.$$

An explicit expression for $B(\xi, \rho)$ appears in Appendix A of Dombry and Ferreira (2019), while the Fisher information of the generalized extreme value distribution is given in Prescott and Walden (1980) and in Appendix A.

Since we always work at penultimate levels, one could instead compare the bias committed by adopting an ultimate approximation, e.g., $\mathsf{GEV}(b_n, a_n, \xi_\infty)$ with a higher-order approximation such as $\mathsf{GEV}(b_n, a_n, \xi_n)$ that is closer to the limiting model (Smith, 1987b, 1988). These derivations give an avenue for selecting the optimal block size or the quantile at which to threshold the data in order to minimize the mean squared error of an estimator. Such results are inapplicable in practice, unfortunately, as the function $A(\cdot)$ is family-dependent and evaluation of the bias terms would require an estimate of ρ.

The following examples provide derivations of $a_t, b_t, \xi_t, \xi_\infty, \rho$ and $A(t)$ for various parametric families in the Gumbel max-domain of attraction, so $\xi_\infty = 0$ in all cases. It does not matter which definition we adopt, either $s(x) = -F(x)\log\{F(x)\}/f(x)$ or the reciprocal hazard $r(x) = \{1-F(x)\}/f(x)$ to derive the penultimate shape; we use the latter since it is easier to manipulate.

Example 1.4 (Penultimate approximation for Weibull variates)
The Weibull distribution has survival function $1 - F(x) = \exp\{-(x/\lambda)^\alpha\}$ for $x \geq 0$, with parameters $\alpha, \lambda > 0$. One can thus take as location and scale constants $b(t) = \lambda\{\log(t)\}^{1/\alpha}$ and $a(t) = (\lambda/\alpha)\{\log(t)\}^{1/\alpha - 1}$, respectively. The second-order auxiliary function can be derived from eq. (1.3) and is

$$r'(b_t) = A(t) = \left(\frac{1}{\alpha} - 1\right) \frac{1}{\log(t)}, \qquad t > 1,$$

which is slowly varying at ∞, so $\xi = \rho = 0$. The penultimate value ξ_t lies in the Weibull domain when $\alpha > 1$ and in the Fréchet domain when $\alpha < 1$. The case $\alpha = 1$ is the exponential distribution, which is threshold-stable.

□

Example 1.5 (Penultimate approximation for Gaussian variates)
For Gaussian variates, $b(t) \sim \{2\log(t)\}^{1/2}$ as $t \to \infty$ since (de Haan and Stadtmüller, 1996)

$$\lim_{t\to\infty} b(t)^3 \left\{ b(tx) - b(t) - \frac{\log(x)}{b(t)} \right\} = -\frac{\log^2(x)}{2} - \log(x).$$

We can choose the auxiliary functions $a(t) = 1/b(t)$ and $A(t) = \{b(t)\}^{-2}$. Expanding the reciprocal hazard using Mill's ratio, it can be seen that $r'(x) = -x^{-2} + \mathrm{O}(x^{-4})$, so $\xi = \rho = 0$ and $\xi_t = r'\{b(t)\}$ is negative. The rate of convergence of the ultimate approximation is $\mathrm{O}\{\log(t)^{-1}\}$ while that of the penultimate approximation is $\mathrm{O}\{\log(t)^{-2}\}$ (Smith, 1987a).

□

Example 1.6 (Penultimate approximation for Burr variates)
The Burr type XII distribution has survival function $1 - F(x) = \{1 + (x/\phi)^c\}^{-k}$ for $\phi, c, k > 0$ and $x \in \mathbb{R}_+$. The norming constants are $b(t) = \phi(t^{1/k} - 1)^{1/c}$ and

$$a(t) = \frac{\phi}{ck}(t^{1/k} - 1)^{1/c} \left(1 + \frac{1}{t^{1/k} - 1} \right).$$

The penultimate shape is

$$r'\{b(t)\} = \frac{1-c}{ck}(t^{1/k} - 1)^{-1} + \frac{1}{ck},$$

with $\xi = \lim_{t\to\infty} r'\{b(t)\} = (ck)^{-1} > 0$, and $\rho = -1/k$.

□

Example 1.7 (Penultimate approximation for generalized gamma variates)
The generalized gamma distribution is used by Papalexiou and Koutsoyiannis (2013) to model non-zero rainfall. Its distribution function is

$$F(x) = 1 - \Gamma\left\{ \frac{\gamma_1}{\gamma_2}, \left(\frac{x}{\beta}\right)^{\gamma_2} \right\} \Gamma^{-1}\left(\frac{\gamma_1}{\gamma_2}\right), \qquad x > 0, \gamma_1, \gamma_2, \beta > 0.$$

where $\Gamma(a, z) := \int_z^\infty x^{a-1}\exp(-x)\mathrm{d}x$ is the upper incomplete Gamma function. This can be expressed as a power series at infinity as

$$\Gamma(a, z) \sim z^{a-1}e^{-z} \sum_{k=0}^{\infty} \frac{\Gamma(a)}{\Gamma(a-k)} z^{-k}, \qquad z \to \infty.$$

We can therefore write the derivative of the reciprocal hazard function for x large as

$$\begin{aligned} r'(x) &= -1 + \left\{ \frac{1-\gamma_1}{\gamma_2} + \left(\frac{x}{\beta}\right)^{\gamma_2} \right\} \Gamma\left\{ \frac{\gamma_1}{\gamma_2}, \left(\frac{x}{\beta}\right)^{\gamma_2} \right\} \left(\frac{x}{\beta}\right)^{-\gamma_1} \exp\left\{ \left(\frac{x}{\beta}\right)^{\gamma_2} \right\} \\ &\sim -1 + \left\{ 1 + \frac{1-\gamma_1}{\gamma_2}\left(\frac{x}{\beta}\right)^{-\gamma_2} \right\} \left\{ 1 + \sum_{k=1}^{\infty} \frac{\Gamma(\gamma_1/\gamma_2)}{\Gamma(\gamma_1/\gamma_2 - k)} \left(\frac{x}{\beta}\right)^{-\gamma_2 k} \right\} \\ &\sim \frac{1}{\gamma_2}\left(\frac{x}{\beta}\right)^{-\gamma_2} + \mathrm{O}\big(x^{-2\gamma_2}\big). \end{aligned}$$

Since the upper endpoint of the generalized gamma is $x^* = \infty$ and $b_t \to \infty$, it suffices to consider $\lim_{x\to\infty} r'(x) = \xi_\infty = 0$. Following Fung and Seneta (2018, §3.2), the quantile function evaluated at $1-1/t$ admits the asymptotic expansion

$$b(t) = F^{-1}(1-1/t) \sim \left[\beta^{\gamma_2}\log\left\{\frac{\beta^{\gamma_2-\gamma_1}t}{\Gamma(\gamma_1/\gamma_2)}\right\} - k_1 + \frac{\beta^{\gamma_2}(\gamma_1-\gamma_2)}{\gamma_2}\log\left\{\left|k_2 - \frac{\gamma_2\log(t)}{\gamma_1-\gamma_2}\right|\right\}\right]^{1/\gamma_2} \times \left(1+\mathrm{O}\left[\frac{\log\{|-\log(t)|\}}{\log^2(t)}\right]\right), \qquad t\to\infty,$$

for some constants $k_1, k_2 \in \mathbb{R}$. The leading term in the expansion of $r'(x)$ is $\mathrm{O}(x^{-\gamma_2})$; it follows that the second-order auxiliary function $A(t)$ is slowly varying, so $\rho = 0$.

□

Two other distributions will appear in the simulation studies. The first is the Student distribution with ν degrees of freedom, which has shape and second order parameter $\xi = 1/\nu$ and $\rho = -2/\nu$, respectively (de Haan and Ferreira, 2006, ex. 2.15). The second is the lognormal distribution, whose auxiliary function is $a(x) \sim x/\log(x)$ as $x\to\infty$, with limits $\xi = \rho = 0$. The scaling constants a_n and b_n can be found in Embrechts et al. (1997, p. 156).

Example 1.8 (Metastatistical approach for rainfall extremes)
Marani and Ignaccolo (2015) suggest fitting a Weibull distribution to the whole distribution of N non-zero rainfall observation and use the penultimate approximation to the latter to estimate return levels. Write the distribution of the maximum of N Weibull variates as

$$H^N(x;\lambda,\alpha) = \left[1-\exp\left\{-\left(\frac{x}{\lambda}\right)^\alpha\right\}\right]^N \dot{\sim} \exp\left[-\exp\left\{-\left(\frac{x}{\lambda}\right)^\alpha + \log(N)\right\}\right] =: H_N,$$

say, using the notation of Example 1.4, and let $\boldsymbol{\theta} = (\lambda,\alpha)^\top > \mathbf{0}_2$ denote the vector of scale and shape parameters. The predictive distribution for the yearly maxima, assuming a joint distribution for $p(N,\lambda,\alpha)$, is

$$\sum_{n\in\mathrm{supp}(N)}\int_\alpha\int_\lambda H^n(\boldsymbol{y};\lambda,\alpha)p(n,\lambda,\alpha)\mathrm{d}\lambda\mathrm{d}\alpha,$$

which can be approximated by using the empirical distribution of N and by fitting a separate Weibull distribution to each of the T years of data, where

$$\widehat{H}_{\mathrm{META}}(y) = \frac{1}{T}\sum_{t=1}^{T} H_{n_t}(y,\widehat{\boldsymbol{\theta}}_t). \tag{1.4}$$

and $\widehat{\boldsymbol{\theta}}_t = (\widehat{\alpha}_t, \widehat{\lambda}_t)^\top$ are the probability weighted moments (PWM) estimates of the parameters of the Weibull distribution for the non-zero data in year t. For the Weibull distribution, these moments are

$$A_r = \mathsf{E}[X\{1-F(x)\}^r] = \frac{1}{r+1}\left[\lambda(r+1)^{-1/\alpha}\Gamma\left(1+\frac{1}{\alpha}\right)\right], \qquad r\in\mathbb{N}$$

and their empirical counterparts are

$$\widehat{A}_r = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}\frac{n-i}{n+1},$$

where $x_{(1)} \leq \cdots \leq x_{(n)}$ are the ranked observations. The first two probability weighted moment estimators are

$$\widehat{\alpha} = \frac{\log(2)}{\log(\widehat{A}_0) - \log(2\widehat{A}_1)}, \qquad \widehat{\lambda} = \frac{\widehat{A}_0}{\Gamma(1+1/\widehat{\alpha})}.$$

Marani and Ignaccolo claim that their method works better than estimation using peaks-over-thresholds or block maxima if the number of wet days and the rainfall intensity change over years. They justify this by generating pseudo-random variates from the Weibull distribution, hinting that the use of the metastatistical model leads to lower mean squared prediction error than using the GEV or the GP distributions.

We consider two distributions employed in Papalexiou and Koutsoyiannis (2013) to model rainfall around the globe, namely the generalized Gamma distribution with parameters $\beta = 1.83, \gamma_1 = 1.16$ and $\gamma_2 = 0.54$, and the Burr distribution with parameters $c = 0.91$, $k = 6.105$ and $b = 55.75$. Supposing the "true distribution" of rainfall $G(x)$ is either of these two parametric models, fitting a Weibull distribution to rainfall will lead to model misspecification and the estimated parameters $\widehat{\boldsymbol{\theta}}$ will be those for which the Kullback–Leibler divergence $\mathsf{KL}(h;g) = \int g(x)\log\{g(x)/h(x;\alpha,\lambda)\}\mathrm{d}x$ is minimized, i.e., $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathsf{KL}(h;g)$. The maximum likelihood estimator will be asymptotically Gaussian, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathsf{No}_2(\mathbf{0}_2, \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1})$ where $\mathbf{H} = \mathsf{E}_g\left(\partial^2 \log\{h(\boldsymbol{\theta})\}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top\right)$, and $\mathbf{J} = \mathsf{E}_g\left([\partial\log\{h(\boldsymbol{\theta})/\partial\boldsymbol{\theta}][\partial\log\{h(\boldsymbol{\theta})\}/\partial\boldsymbol{\theta}^\top]\right)$ (White, 1982).

Minimizing the Kullback–Leibler divergence amounts to maximizing

$$\mathsf{E}_g\left(\log\{f(x)\}\right) = \log(\alpha) - \alpha\log(\lambda) + (\alpha-1)\mathsf{E}_g\left(\log(X)\right) - \lambda^{-\alpha}\mathsf{E}_g\left(X^\alpha\right)$$

and these expectations can be computed explicitly for the generalized Gamma family,

$$\mathsf{E}_g\left(\log(X)\right) = \log(\beta) + \frac{\psi^{(0)}(\gamma_1/\gamma_2)}{\gamma_2}, \qquad \mathsf{E}_g\left(X^\alpha\right) = \beta^\alpha\frac{\Gamma\{(\gamma_1+\alpha)/\gamma_2\}}{\Gamma(\gamma_1/\gamma_2)},$$

where $\psi^{(0)}(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. We could proceed likewise to compute the Godambe information matrix $\mathbf{H}\mathbf{J}^{-1}\mathbf{H}$ analytically, but resort instead to Monte Carlo methods to evaluate these integrals numerically.

The Weibull distribution lies in the max-domain of attraction of a Gumbel distribution with $\xi = 0$, yet Gumbel random variables have lower variance than Fréchet ones. Since the Weibull model is fitted to the whole data set, the estimates of the Weibull model $\widehat{\boldsymbol{\theta}}$ are also less variable than their extreme value counterparts. This twofold reduction in variance comes

at the expense of potential bias, as there is no guarantee that the true ultimate value ξ_∞ will be zero, or that the penultimate shape $r'(b_n)$ from the fitted Weibull will match that of the true distribution. For example, the Burr distribution described has $\xi_\infty \approx 0.164$. In principle, any distribution could be fitted to the rainfall data, leading to different penultimate approximations.

An alternative approach proposed by Marani and Ignaccolo, assuming that the parameters $\boldsymbol{\theta}_t$ are constant over the T years, is to fit a Weibull distribution to the whole sample and use $H_{\bar{n}}(y, \widehat{\boldsymbol{\theta}})$ as metastatistical distribution, where $\bar{n}$ is the average number of observations per year. For the purpose of the simulation, $H_{\bar{n}}$ can replaced by its penultimate approximation; this has no practical impact on the results. The estimated metastatistical distribution $\widehat{H}_{\text{META}}(y)$ corresponds to yearly maxima. Should one wish to extrapolate the model to obtain estimates for N_y years, we would replace n_t by $n_t N_y$ in eq. (1.4).

We simulate data from the generalized Gamma, Burr and Weibull distributions, specifying the parameters $\lambda = 7.3$mm and $\alpha = 0.82$ for the latter. To reproduce the setup of Marani and Ignaccolo (2015), we suppose that daily cumulative rainfall totals follow one of the three parametric models and simulate the number of wet days in year t as $\{\lfloor N_t \rfloor\}_{t=1}^{T}$ with $N_t \sim \mathsf{No}(105, 22^2)$ and $T = 30$ as in Marani and Ignaccolo (2015) to reflect nonstationarity. We fit the Weibull distribution to each 'yearly' block of data using L-moments, but also to the whole data set, as suggested in Zorzetto et al. (2016). The parameters of the generalized extreme value distribution were estimated using maximum likelihood based on the yearly maxima. The target for inference is the median of the distribution of the maximum of 50 years of daily rainfall, which is obtained through max-stability for the generalized extreme value distribution and by using $m = Nn_y$ in the penultimate approximation for the metastatistical approach. We report the median of the estimated distribution of the maximum of Nn_y rainfall events. We can obtain the median of the 50-year maximum distribution and evaluate the estimated distribution functions at this value; if the distributions are well-calibrated, the estimate should be around 0.5. The simulations in Figure 1.5 show that adoption of the penultimate approximation leads to a bias-variance trade-off and potentially to underestimation of high quantiles. Calibration is better assessed on the percentile scale; the right panel of Figure 1.5 shows to which percentile of the fitted distribution the true value corresponds. The generalized extreme value distribution estimates are calibrated, but not sharp. In contrast, the metastatistical model estimates are not centred around the true value, hence uncalibrated.

The block maximum method delivers extremely variable estimates but, over 2500 replications, the point estimates are centred around the true value. In contrast, the estimated median of the 50-year maximum provided by the metastatistical approach is too small even when $\xi_\infty = 0$. The approach that consists in fitting a single Weibull distribution is particularly problematic, because the bulk of the data dictates the fit. Likewise, misspecification can lead to severe underestimation of risk, due to the mismatch between the penultimate shape from the data generating mechanism and that of the Weibull fit, leading to an extrapolation error. While the use of block maxima is not a panacea, it is a theoretically-justified inferential

approach. For spatial data, one could reduce the variability of the generalized extreme value parameter estimates by pooling data from multiple neighbouring sites using local likelihood, if for example the spatial dependence is strong.

□

1.1.5 Threshold selection methods

The limiting distribution of threshold exceedances is generalized Pareto as $u \to x^*$, but practitioners must choose a finite threshold in order to draw inference. In principle, the number of points above the threshold, k, should satisfy $k/n \to 0$ as $n \to \infty$. Threshold selection is subtle and it is common to select a high percentile of the data, say the 95% value, as the threshold, even if this is asymptotically incorrect, as in this case $k/n \not\to 0$ as $n \to \infty$. Most approaches for threshold selection rely on properties of the generalized Pareto distribution (moments, threshold-stability) to determine a region within which the asymptotic distribution fits the data well and the parameter estimates are stable. Comparison of model fit for different thresholds is complicated: a fraction of the data enter in both fit, yet for the generalized Pareto distribution, the model is only specified above the threshold. The latter acts as both a location parameter and a left-censoring indicator: it determines which data enter the likelihood, yet cannot be considered a parameter of the model *per se*. The threshold choice leads to a bias-variance trade-off that cannot be easily quantified because of the penultimate approximation, since the shape parameter behaves like $r'(u)$ as $u \to \infty$. Uncertainty due to threshold selection is often ignored in the subsequent analyses.

Automating threshold selection is almost necessary in multivariate or spatial data sets with large dimensions. Bader et al. (2018) propose automated selection based on goodness-of-fit tests, while Northrop et al. (2016) propose using the predictive cross-validated distribution to assess the performance for predicting holdout observations for a given threshold. Below, we focus on recent graphical selection tools, which are still standard; mixture models are reviewed in Scarrott and MacDonald (2012).

Method 1.9 (Robust selection)
The extreme value distributions have unbounded influence functions and outliers can strongly affect the estimate of the shape. Dupuis (1999) proposes an optimal B-robust estimator of the generalized Pareto parameters. Points that are outlying or for which the fit is poor are downweighted; if the weights for the largest observations are very low, this suggests that the threshold is too low. While there is no guarantee that observations that were simulated from a generalized Pareto distributed would not be downweighted, systematic downweighting of the largest exceedances may be indicative of poor fit.

Method 1.10 (Threshold stability plots)
Consider a sequence of ordered candidate thresholds $u_1 < \cdots < u_k$; one of the most widely used tools for threshold selection is the threshold-stability plots of Davison and Smith (1990). These show the point estimates of the shape ξ and the modified scale $\sigma_{u_i} - \xi u_i$, which should be constant for any threshold $u_j > u_i$ assuming that the generalized Pareto above u_i holds

exactly. In addition to the point estimates, the asymptotic pointwise 95% Wald confidence intervals are displayed; the standard errors are obtained from the observed information matrix. These plots can be difficult to interpret because no joint statement can be derived and they ignore changes in the estimated parameters due to the penultimate approximation. In practice, one could replace these confidence intervals by those obtained from the profile likelihood to better reflect the asymmetry of the estimators.

Method 1.11 (White noise process and simultaneous threshold stability plots)
The problem with the threshold stability plots lies in the point-wise nature of the estimate. Assuming a superposition of k Poisson processes, Wadsworth (2016) derives the limiting distribution of the maximum likelihood estimators from the Poisson process for overlapping windows as the number of windows $k \to \infty$ and $n_k \to \infty$. The joint asymptotic Gaussian distribution allows Wadsworth to propose two additional diagnostics: a white noise sequence of differences in estimates of the shape, standardized to have unit variance. The variables $\xi_i^* = (\widehat{\xi}_{u_{i+1}} - \widehat{\xi}_{u_i})/\{(I_{u_{i+1}}^{-1} - I_{u_i}^{-1})_{\xi,\xi}^{1/2}\}$, where I_{u_i} is the Fisher information of the Poisson process likelihood for exceedances above u_i, should form a white-noise sequence of independent variables centered around the origin; systematic deviations are indicative of inadequacy. To formally test the hypothesis, a likelihood ratio test can be used assuming a simple alternative, namely a single change point at threshold u_j. The null hypothesis is $\mathscr{H}_0 : \xi_i^* \overset{\text{iid}}{\sim} \mathsf{No}(0,1)$ for $i = 1,\ldots,k-1$ against the alternative $\mathscr{H}_a : \xi_i^* \sim \mathsf{No}(\beta,\sigma)(i = 1,\ldots,j-1)$ and $\xi_i^* \sim \mathsf{No}(0,1)$ for $j,\ldots,k-1$. This alternative is motivated by results on model misspecification (White, 1982), which suggest that the asymptotic distribution may still be Gaussian, but with a different mean and variance. This can be used to automate threshold selection, by picking the smallest threshold for which the P-value is above the level α.

For the asymptotic result to be approximately valid, the number of thresholds must be large, which implicitly requires large samples for each superposed point process. Practical experience using Wadsworth (2016) diagnostic over a range of data sets and varying thresholds point to a lack of robustness: the estimated difference in Fisher information matrices often fails to be positive definite in practice. Changing the set of thresholds $\boldsymbol{u}$ under consideration leads to potentially completely different parameter estimates being chosen by the automated procedure, so the diagnostic is also highly sensitive to the choice of k.

Method 1.12 (Changepoint tests based on penultimate approximations)
Based on the penultimate approximation, one might expect the shape to vary slowly with the threshold. Wadsworth and Tawn (2012) proposes to fit a Poisson process model to exceedances, assuming that the shape is piecewise constant. They assume there is a nonhomogeneous Poisson process with an intensity function on (v,u) and another on (u,∞). Specifically, the integrated intensities are

$$\Lambda\{(0,1)\times(x,\infty)\} = \begin{cases} \left[1+\xi_{vu}(x-\mu_{vu})/\sigma_{vu}\right]^{-1/\xi_{vu}}, & v < x < u, \\ \left[1+\xi_u(x-\mu_u)/\sigma_u\right]^{-1/\xi_u}, & u < x, \end{cases} \tag{1.5}$$

and imposing continuity constraints so that the integrated intensity Λ and the intensity λ

agree at the threshold u yields

$$\mu_{vu} = u - \frac{\sigma_{vu}}{\xi_{vu}}\left(\left[1+\xi_u\left(\frac{u-\mu_u}{\sigma_u}\right)\right]^{\frac{\xi_{vu}}{\xi_u}} - 1\right),$$
$$\sigma_{vu} = \sigma_u\left[1+\xi_u\left(\frac{u-\mu_u}{\sigma_u}\right)\right]^{1/\xi_u+1} \times \left[1+\xi_{vu}\left(\frac{u-\mu_{vu}}{\sigma_{vu}}\right)\right]^{-1/\xi_{vu}-1}.$$

Let $m = \#\{x_i > v\}$ and $m_u = \#\{x_i > u\}$; the likelihood is, up to proportionality constants,

$$\begin{aligned}\sigma_u^{-m} &\prod_{i:x_i\in(v,u)}\left\{1+\frac{\xi_{vu}(x_i-u)}{\sigma_u+\xi_u(u-\mu_u)}\right\}^{-1/\xi_{vu}-1} \prod_{i:x_i\in(u,\infty)}\left\{1+\xi_u\left(\frac{x_i-\mu_u}{\sigma_u}\right)\right\}_+^{-1/\xi_u-1}\\ &\times\left\{1+\xi_u\left(\frac{u-\mu_u}{\sigma_u}\right)\right\}^{(m-m_u)(-1/\xi_u-1)}\\ &\times\exp\left(-N\left[\left\{1+\frac{\xi_{vu}(v-u)}{\sigma_u+\xi_u(u-\mu_u)}\right\}^{-1/\xi_{vu}}\right]\left\{1+\xi_u\left(\frac{u-\mu_u}{\sigma_u}\right)\right\}^{-1/\xi_u}\right),\end{aligned}$$

which corrects the formula on page 552 of Wadsworth and Tawn (2012). We can extend this result by continuity for the cases where $\xi_{vu} = 0$ or $\xi_u = 0$. The null hypothesis is $\mathscr{H}_0 : \xi_{vu} = \xi_u$ against the alternative $\xi_{vu} \neq \xi_u$ and is tested using a likelihood ratio test statistic.

Attempts to reproduce the simulation study results in Wadsworth and Tawn (2012) by fitting the Poisson process likelihood via Markov chain Monte Carlo methods were largely unsuccessful, in parts due to the non-orthogonality of the components; this could be improved by tuning m (Sharkey and Tawn, 2017). As for threshold stability plots, the user must provide a sequence of thresholds but unlike the latter, comparisons are done pairwise. This leads to multiple testing issues and a computationally intensive procedure.

Northrop and Coleman (2014) adapt the idea of Wadsworth and Tawn (2012) and fit a generalized Pareto model with piecewise constant shape to k different thresholds; continuity constraints at the thresholds impose $k-1$ restrictions on scale parameters, so the model only has $k+1$ parameters. A score test can be used to test the hypothesis of equal shape and it only requires evaluation of the model under the null hypothesis that a generalized Pareto distribution is valid for all thresholds. A diagnostic plot is obtained by plotting P-values against threshold. One can then choose to take, e.g., (a) the lowest threshold at which the P-value is non-significant, or (b) the lowest threshold at which the P-values for all higher thresholds are non-significant: under the null hypothesis, there is an α% probability of rejection at any given threshold.

Method 1.13 (Extended generalized Pareto)
Papastathopoulos and Tawn (2013) propose three extended generalized Pareto distributions: for example, the third extended generalized Pareto model has distribution function $\{1-(1+\xi x/\sigma)_+^{-1/\xi}\}^\kappa$ for $x > 0$ and $\kappa > 0$. Each family reduces to the generalized Pareto when the additional parameter $\kappa = 1$ and share the same tail index ξ, the extended generalized Pareto provide more flexibility for modelling departures from the limiting form. Standard parameter

stability plots can be used to find a region in which $\kappa \approx 1$ and the shape parameter stabilizes.

Example 1.14 (Threshold selection for Padova rainfall)
We look at a time series of daily cumulated rainfall analyzed in Marani and Zanetti (2015) spanning 268 years between 1725–2006. The exceptional length of the series gives enough data to apply threshold selection methods; graphical diagnostics are reported in Figure 1.6. The series include 52 missing daily records, as well as 14 non-consecutive years, including change of instruments and five relocations. The white noise diagnostic (top left) suggests use of a threshold of 18.8mm, corresponding to the 87.5% of the non-zero rainfall episodes. On the contrary, the parameter stability plot (top right) indicates an increase in the shape estimates beyond the confidence intervals at this point, leading to choice of threshold above 32mm. This is corroborated by the Northrop and Coleman (2014) P-value path for the score test statistic (bottom left), which would hint that a threshold above 34mm is suitable. The parameter stability plot for the extended generalized Pareto model also suggests a threshold above 34mm.

□

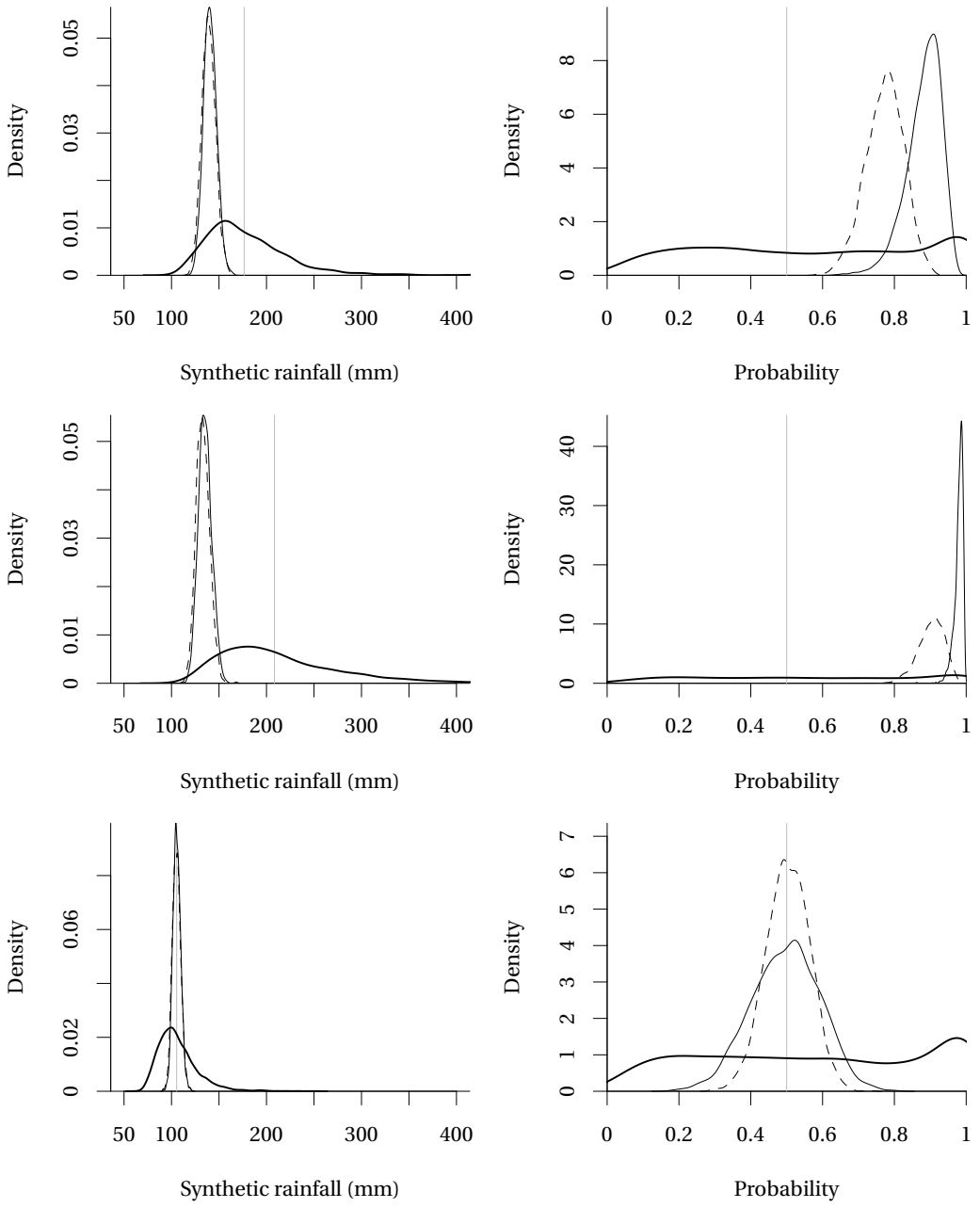

Figure 1.5 – *Comparison of the fitted metastatistical approach with 30 years of synthetic rainfall data from various parametric models. The data are generated from the generalized Gamma (top), Burr (middle) and Weibull distributions (bottom). Left panels: density of the estimated median of the maximum of 50 years of daily rainfall over the 2500 replicate datasets. Right panels: penultimate distribution functions evaluated at the true 50-year maximum median. The solid lines show kernel density estimates from 2500 replications for the metastatistical model $\widehat{H}_{\text{META}}$ (dashes), the penultimate approximation obtained by fitting a Weibull distribution to all the data (solid) and the block maximum approach (heavy). The true value is indicated by grey vertical lines.*

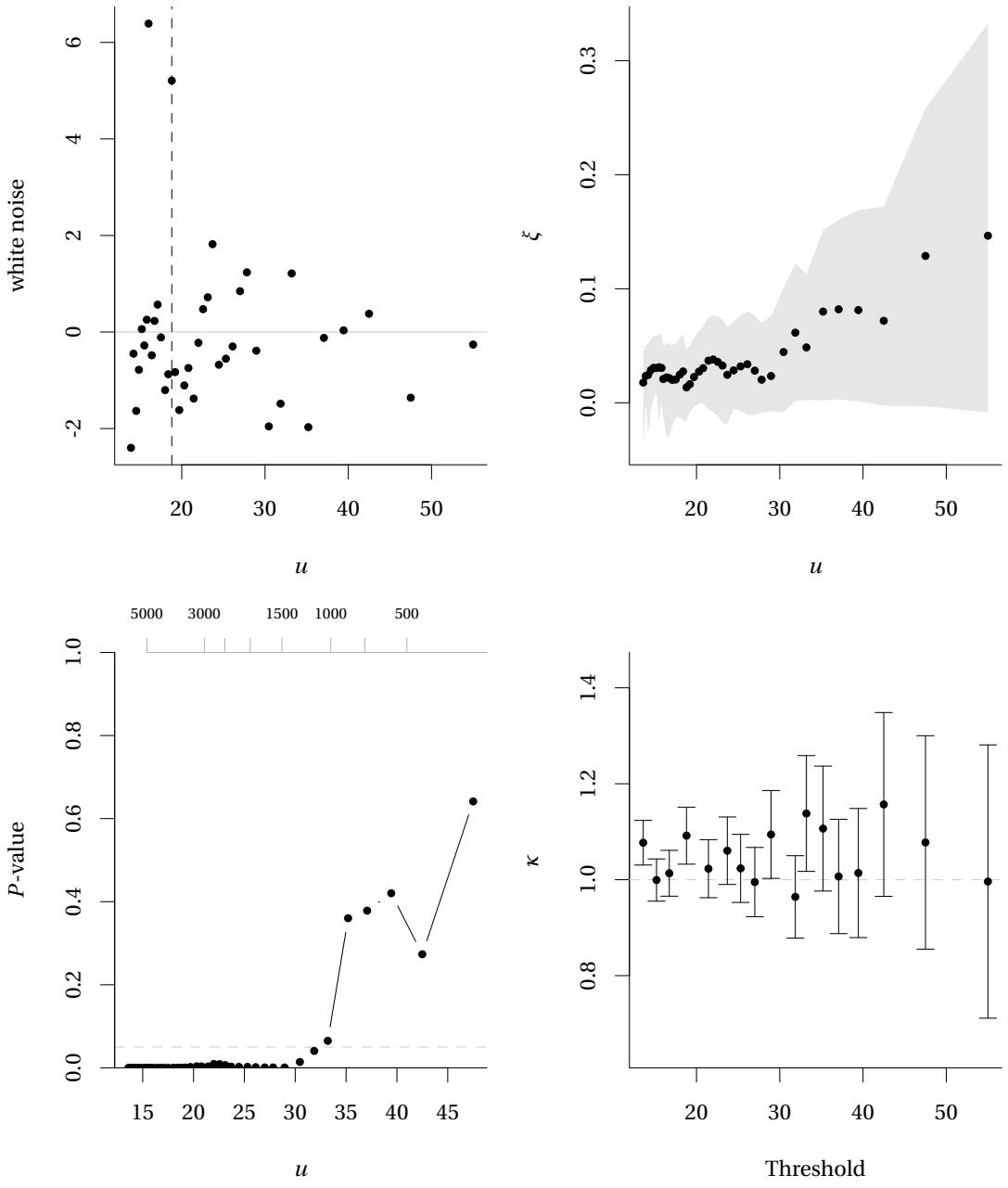

Figure 1.6 – *Threshold selection diagnostics for the Padova rainfall series. Top left: white noise sequence diagnostic of Wadsworth (2016); rescaled differences of shape parameter estimates, $\{\xi_i^*\}$. The dashed vertical line indicates the lowest threshold at which we fail to reject the likelihood ratio test that the sequence behaves like a standard Gaussian white noise sequence. Top right: parameter stability plot of Wadsworth (2015) for the transformed shape estimates with simultaneous confidence intervals. Bottom left: P-values of the score test of Northrop and Coleman (2014) against threshold. The horizontal line at $\alpha = 5\%$ gives the cutoff for individual tests, while the upper axis indicates selected quantiles and the associated number of threshold exceedances. Bottom right: parameter stability plot of Papastathopoulos and Tawn (2013) for κ with pointwise confidence intervals.*

1.2 Asymptotics for likelihood-based inference

This section focuses on parametric modelling of univariate extremes from a likelihood-based perspective, starting with definitions that enable us to derive analytical expressions for the finite-sample bias of maximum likelihood estimators. Since the quantity of interest for extreme value distributions is mostly quantiles, special attention is devoted to development and tests for scalar parameters of interest in presence of nuisance parameters, along with illustrative examples. Much of this material follows the exposition in Cordeiro and Cribari-Neto (2014), Brazzale et al. (2007), Severini (2000) and Barndorff-Nielsen and Cox (1994).

1.2.1 Likelihood and cumulants

Let $\boldsymbol{\theta}$ be a p-dimensional parameter vector and let

$$\ell(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{p(x_i;\boldsymbol{\theta})\}$$

be the log-likelihood function for an independent sample of size n with density $p(x_i;\boldsymbol{\theta})$.

We denote the derivatives of the log-likelihood by

$$\ell_r = \frac{\partial \ell}{\partial \theta_r}, \quad \ell_{rs} = \frac{\partial^2 \ell}{\partial \theta_r \partial \theta_s}, \quad \ell_{r,s} = \left(\frac{\partial \ell}{\partial \theta_r}\right)\left(\frac{\partial \ell}{\partial \theta_s}\right), \quad \ldots,$$

and their moments by $\mathsf{E}(\ell_r) = \nu_r, \mathsf{E}(\ell_{rs}) = \nu_{rs}, \mathsf{E}\left(\ell_{r,s}\right) = \nu_{r,s}$, etc. The score function $U(\boldsymbol{\theta})$ is a p-vector with rth entry ℓ_r. The Fisher (or expected) information is $i(\boldsymbol{\theta})$ with (r,s) entry $-\nu_{r,s}$, while the observed information is the function $j(\boldsymbol{\theta}) = -\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$. We use superscripts to denote entries of the inverse information matrix $i^{-1}(\boldsymbol{\theta})$ or of the inverse observed information matrix. Large sample properties of the maximum likelihood estimator depend on a Taylor expansion of the score vector but, since our focus is on higher-order asymptotics, we will also need cumulants of the log-likelihood derivatives, which are obtained from the cumulant generating function $K(\boldsymbol{\tau}) = \log\left\{\mathsf{E}\left(\exp(\boldsymbol{\tau}^\top \boldsymbol{Y})\right)\right\}$ as coefficients of the corresponding series expansion: the cumulants of the log-likelihood derivatives are

$$\begin{aligned}
&\kappa_r = \nu_r, && \kappa_{rs} = \nu_{rs}, && \kappa_{r,s} = \nu_{r,s},\\
&\kappa_{r,s,t} = \nu_{r,s,t}, && \kappa_{rs,t} = \nu_{rs,t}, && \kappa_{rst} = -\nu_{r,s,t} - [3]\nu_{rs,t},\\
&\kappa_{r,s,t,u} = \nu_{r,s,t,u} - [3]\nu_{r,s}\nu_{t,u}, && \kappa_{rs,t,u} = \nu_{rs,t,u} - \nu_{rs}\nu_{t,u}, && \kappa_{rs,tu} = \nu_{rs,tu} - \nu_{rs}\nu_{tu}, \quad \ldots,
\end{aligned}$$

for $r,s,t,u = 1,\ldots,p$, where $[\cdot]$ denotes the sums over permutations of the indices, the number of which is displayed in brackets (McCullagh, 2018, §2.3). The first few Bartlett (1953) identities for cumulants are

$$\kappa_r = 0, \qquad \kappa_{r,s} + \kappa_{rs} = 0, \qquad \kappa_{rst} + [3]\kappa_{rs,t} + \kappa_{r,s,t} = 0, \qquad \kappa_{rs,t} + \kappa_{rst} = \kappa_{rs}^{(t)}.$$

The Bartlett identities hold whenever the cumulants exist and are finite. Let p_r denote partial derivatives of the density function with respect to θ_r, using the same convention as above. The third Bartlett identity is derived by considering

$$\begin{aligned}\ell_{rst} &= \frac{\partial}{\partial\theta_t}\frac{\partial}{\partial\theta_s}\left(\frac{\partial p/\partial\theta_r}{p}\right)\\ &= \frac{\partial}{\partial\theta_t}\left(\frac{p_{rs}}{p}-\frac{p_r p_s}{p^2}\right)\\ &= \frac{p_{rst}}{p}-\frac{p_{rs}p_t}{p^2}-\frac{p_s p_{rt}}{p^2}+2\frac{p_r p_s p_t}{p^3}-\frac{p_{st}p_r}{p^2}\\ &= \frac{p_{rst}}{p}-\ell_{rs,t}-\ell_{rt,s}-\ell_{st,r}-\ell_{r,s,t},\end{aligned}$$

using the chain rule and the product rule. The former allows one to write, e.g., $\partial/\partial\theta_t \ell_r\ell_s = \ell_{r,st}+\ell_{s,rt}$, where we use the relations $p_r/p=\ell_r$ and $\ell_{rs}=p_{rs}/p-\ell_{r,s}$. Terms such as p_{rst}/p have null expectation, as can be seen by interchanging the integral and differential operators. Variants of the Bartlett identities, due to Lawley (1956), involve partial derivatives of cumulants and are most helpful. By noting that $\partial\mathsf{E}\left(g\right)/\partial\theta_r = \mathsf{E}\left(\partial g/\partial\theta_r\right)+\mathsf{E}\left(g\ell_r\right)$ for g differentiable and taking $g=\ell_{rs}$, Lawley showed that

$$\kappa_{st}^{(r)} = \frac{\partial}{\partial\theta_r}\mathsf{E}\left(\ell_{st}\right) = -\frac{\partial}{\partial\theta_r}\mathsf{E}\left(\ell_s\ell_t\right) = -\mathsf{E}\left(\ell_{rs}\ell_t\right)-\mathsf{E}\left(\ell_{rt}\ell_s\right)-\mathsf{E}\left(\ell_r\ell_s\ell_t\right) = \kappa_{rst}+\kappa_{st,r},$$

using the second and third Bartlett identities.

Example 1.15 (Fisher information for the Poisson process likelihood)
The likelihood for the non-homogeneous Poisson process with intensity measure $\Lambda\{(0,1)\times(x,\infty)\} = \{1+\xi(x-\mu)/\sigma\}^{-1/\xi}$ for $x>u$ given in eq. (1.1) includes a contribution for the random number of exceedances N_u. Calculation of the Fisher information thus require calculating the expectation with respect to both Y_i and N_u. If we write $\Lambda(R)\equiv\Lambda(\{(0,1)\times(u,\infty)\}$ and $\lambda(y_i) = \left\{1+\xi(y_i-\mu)/\sigma\right\}_+^{-1/\xi-1}$, the log-likelihood is

$$\ell(\mu,\sigma,\xi;u,\boldsymbol{y},n_u) = -\Lambda(R)+\sum_{i=1}^{n_u}\log\{\lambda(y_i)\}$$

and the Fisher information is (Wadsworth, 2016)

$$\begin{aligned}I(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{Y},N_u}\left(\nabla^2\ell(\boldsymbol{Y},N_u)\right) &= \nabla^2\Lambda(R)-\mathsf{E}_{N_u}\left(N_u\mathsf{E}_Y\left(\nabla^2\log\{\lambda(Y)\}\mathbf{1}\{Y>u\}\right)\right)\\ &= \nabla^2\Lambda(R)-\Lambda(R)\mathsf{E}_Y\left(\nabla^2\log\{\lambda(Y)\}\mathbf{1}\{Y>u\}\right),\end{aligned}$$

where the rightmost expectation is taken with respect to the conditional density of $Y\mid Y>u$, $f(y)=\lambda(y)/\Lambda(R)$. The entries of the Fisher information matrix are easily calculated (cf. Sharkey and Tawn, 2017, Appendix C).

□

Example 1.16 (Fisher information for the r-largest observations)
Let $Y_{(1)}\geq\cdots\geq Y_{(r)}$ be a r vector of $\mathsf{GEV}(\mu,\sigma,\xi)$ observations. A direct derivation of the

information matrix for the r-largest likelihood given in Proposition 1.5 is painful. Instead, note that the marginal density of $Y_{(r)}$ is

$$f_{Y_{(r)}}(y_{(r)};\mu,\sigma,\xi) = \frac{1}{(r-1)!\sigma}\left(1+\xi\frac{y_{(r)}-\mu}{\sigma}\right)_+^{-r/\xi-1}\exp\left\{-\left(1+\xi\frac{y_{(r)}-\mu}{\sigma}\right)_+^{-1/\xi}\right\},$$

so the joint density of $Y_{(1)},\ldots,Y_{(r)}$ may be written as

$$\frac{1}{(r-1)!\sigma}\left(1+\xi\frac{y_{(r)}-\mu}{\sigma}\right)_+^{-r/\xi-1}\exp\left\{-\left(1+\xi\frac{y_{(r)}-\mu}{\sigma}\right)_+^{-1/\xi}\right\}\times(r-1)!\prod_{j=1}^{r-1}\frac{1}{\sigma}\frac{\left(1+\xi\frac{y_{(j)}-\mu}{\sigma}\right)_+^{-1/\xi-1}}{\left(1+\xi\frac{y_{(r)}-\mu}{\sigma}\right)_+^{-1/\xi}};$$

that is, we write the joint density as the product of the density of $Y_{(r)}$ and the joint conditional density of $Y_{(1)},\ldots,Y_{(r-1)}$ conditional on $Y_{(r)}=y_{(r)}$. This conditional density equals that of the order statistics of $r-1$ independent variables with generalized Pareto density

$$h(\boldsymbol{y}_{-(r)}-y_{(r)};\tau,\xi) = \prod_{j=1}^{r-1}\frac{1}{\tau}\left(1+\xi\frac{y_{(j)}-y_{(r)}}{\tau}\right)_+^{-1/\xi-1},$$

where $\tau=\sigma+\xi(y_{(r)}-\mu)$. Thus the overall log likelihood is

$$\ell(\mu,\sigma,\xi;y_{(j)},y_{(r)}) \equiv \log\left\{f_{Y_{(r)}}(y_{(r)};\mu,\sigma,\xi)\right\}+\sum_{j=1}^{r-1}\log\left\{h(y_{(j)}-y_{(r)};\tau,\xi)\right\},$$

where $y_{(1)},\ldots,y_{(r-1)}$ represent the observed values of a random sample of generalized Pareto variables.

To obtain the observed information we first calculate the Hessian matrix of $-\ell$, then condition on $Y_{(r)}=y_{(r)}$ and take expectations over $X_j=Y_{(j)}-y_{(r)}$. It remains to write $\tau=\sigma+\xi(y_{(r)}-\mu)$ and integrate over $Y_{(r)}$. We see that the expression for the 3×3 Fisher information matrix based on the density of $Y_{(1)},\ldots,Y_{(r)}$ is of the form $I_r(\mu,\sigma,\xi)+(r-1)I(\mu,\sigma,\xi)$, where $I_r(\mu,\sigma,\xi)$ is the information matrix based on the density of $Y_{(r)}$, and $I(\mu,\sigma,\xi)$ is that for a single generalized Pareto variable. The formulae for these matrices are unenlightening, but they can be used to compute the information gain due to basing inference on $Y_{(1)},\ldots,Y_{(r)}$ rather than only on the sample maximum, $Y_{(1)}$. To do so, we calculate the ratios of the diagonal elements of $I_1^{-1}(\mu,\sigma,\xi)$ to those of $\{I_r(\mu,\sigma,\xi)+(r-1)I(\mu,\sigma,\xi)\}^{-1}$; an overall variance reduction for a given r is obtained by considering lth root of the ratio of determinants, where l is the number of parameters:

$$\left\{\frac{|I_1(\mu,\sigma,\xi)|}{|I_r(\mu,\sigma,\xi)+(r-1)I(\mu,\sigma,\xi)|}\right\}^{1/3}.$$

Figure 1.7 shows the variance reduction factors for μ, σ, ξ and the overall efficiency. The variance reduction factors for μ and σ suggest that, for estimation of these parameters, there is little to be gained by taking $r>5$, while for ξ the decline is closer to that of independent generalized extreme value distributed data. This stems from the structure of the matrix $I_r(\mu,\sigma,\xi)+(r-1)I(\mu,\sigma,\xi)$; the parameters μ and σ cannot be estimated based only on $I(\mu,\sigma,\xi)$,

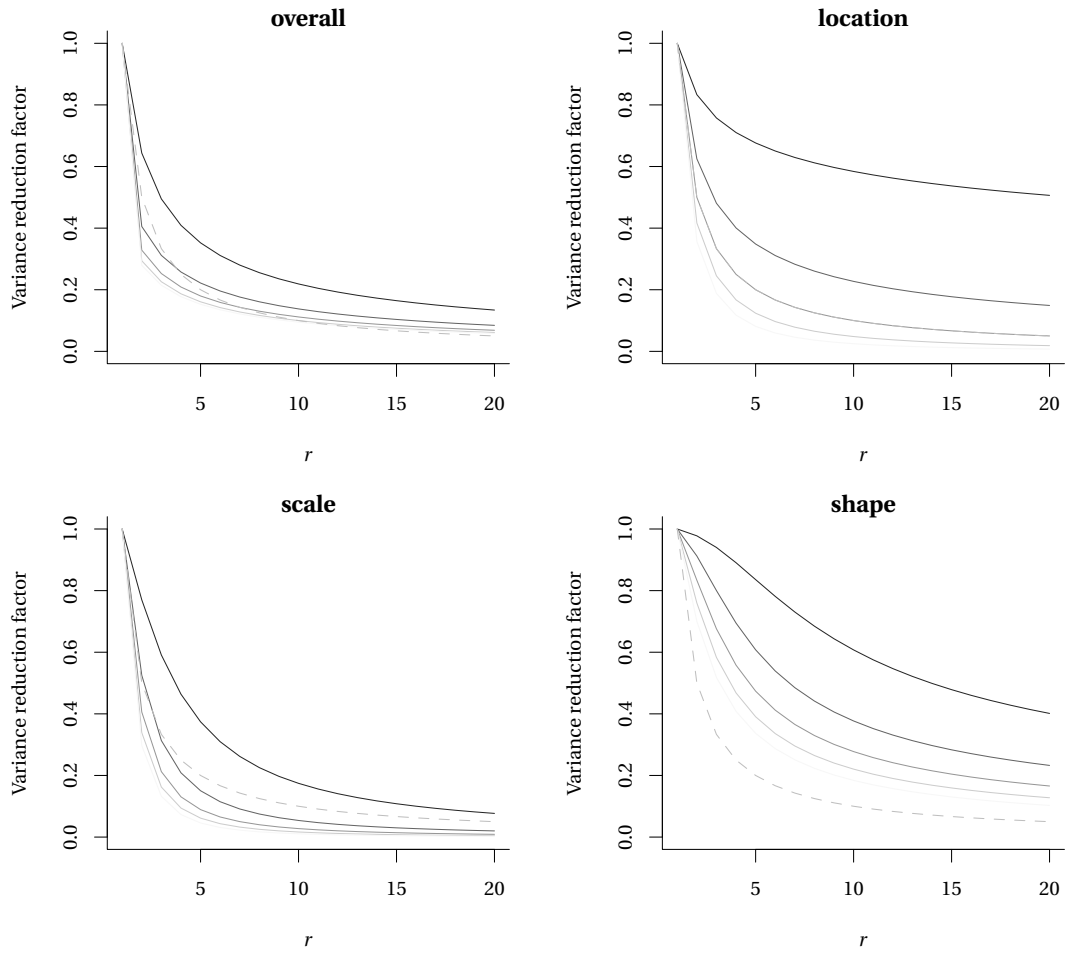

Figure 1.7 – *Variance reduction factors for inference based on the r largest order statistic for the location, scale, shape parameters and overall efficiency (clockwise from top right). The dashed grey line shows the ideal efficiency gain for independent observations. The value of the shape parameter ranges from $\xi = -0.4$ (full black) to $\xi = 0.4$ (full pale grey) in increments of* 0.2.

which has rank two, whereas both likelihood components contain information on ξ. Hence as r increases the information gain for the location and scale parameters becomes more limited. □

1.2.2 Nuisance parameters and the profile likelihood

We explained in Section 1.1.2 that the target of interest is often a quantile of the distribution of maximum over a given period. One way to derive inference is to reparametrize the likelihood in terms of the functional of interest.

Consider a parametric model with log-likelihood function $\ell(\boldsymbol{\theta})$ whose p-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ that can be decomposed into a q-dimensional parameter of interest $\boldsymbol{\psi}$

and a $(p-q)$-dimensional nuisance vector $\boldsymbol{\lambda}$. The score vector, the information matrix and its inverse are partitioned accordingly as

$$U(\boldsymbol{\theta}) = \ell_{\boldsymbol{\theta}} = \begin{pmatrix} \ell_{\boldsymbol{\psi}} \\ \ell_{\boldsymbol{\lambda}} \end{pmatrix}, \qquad i(\boldsymbol{\theta}) = \begin{pmatrix} i_{\boldsymbol{\psi}\boldsymbol{\psi}} & i_{\boldsymbol{\psi}\boldsymbol{\lambda}} \\ i_{\boldsymbol{\lambda}\boldsymbol{\psi}} & i_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{pmatrix}, \qquad i^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} i^{\boldsymbol{\psi}\boldsymbol{\psi}} & i^{\boldsymbol{\psi}\boldsymbol{\lambda}} \\ i^{\boldsymbol{\lambda}\boldsymbol{\psi}} & i^{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{pmatrix}.$$

The profile likelihood ℓ_{p}, a function of $\boldsymbol{\psi}$ alone, is obtained by maximizing the likelihood pointwise at each fixed value $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ over the nuisance vector $\boldsymbol{\lambda}_{\psi_0}$,

$$\ell_{\mathrm{p}}(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda}} \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \ell(\boldsymbol{\psi}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}).$$

The observed profile information function is

$$j_{\mathrm{p}}(\boldsymbol{\psi}) = -\frac{\partial \ell_{\mathrm{p}}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top} = \left\{ j^{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}) \right\}^{-1}.$$

The profile likelihood is not a genuine likelihood in the sense that it is not based on the density of a random variable, but one can show that (Severini, 1999, p. 128)

$$\begin{aligned} \frac{\partial \ell_{\mathrm{p}}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} &= \left. \frac{\partial \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}_{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} \right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \mathrm{O}_p(1), \\ -j_{\mathrm{p}}(\boldsymbol{\psi}) &= \left. \frac{\partial^2 \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}_{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top} \right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + \mathrm{O}_p(n^{-1/2}). \end{aligned}$$

The properties of the profile likelihood function can thus be derived using the full likelihood, and first-order asymptotic results presented in Section 1.2.5 hold also for profile likelihoods.

Example 1.17 (Profile likelihood of generalized Pareto distribution)
There is no closed-form expression for the maximum likelihood estimators of the parameters of the generalized Pareto distribution, which must be obtained by numerical optimization. Grimshaw (1993) uses a profile likelihood to reduce the problem to a one-dimensional optimization in a transformed parametrization, following Davison (1984).

Write the log-likelihood for a random sample $\boldsymbol{x}$ of size n from $\mathsf{GP}(\sigma, \xi)$.

$$\ell(\sigma, \xi; \boldsymbol{x}) = -n\log(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log\left(1 + \frac{\xi x_i}{\sigma}\right), \qquad \xi \neq 0, \sigma > -\xi x_{(n)}.$$

The upper endpoint of the distribution depends on the parameters if $\xi < 0$ and the model is non-regular. For $-1 < \xi \leq -0.5$, the maximum likelihood estimator is super-efficient, i.e., the convergence rate depends on the true parameter values, but is faster than $n^{1/2}$ and the joint limiting distribution of (σ, ξ) is non-standard because the upper endpoint reduces the effective number of parameters by one (Smith, 1985, Theorem 3). When $\xi < -1$, the log-likelihood becomes unbounded and the maximum likelihood estimates do not solve the score equation. When $\xi = -1$, the distribution is uniform with $\sigma = \max(\boldsymbol{x})$ and $\ell(\max(\boldsymbol{x}), -1; \boldsymbol{x}) = -n\log(x_{(n)})$.

If $\xi < -1$, the likelihood is unbounded since $\lim_{-\sigma/\xi \to \max(\boldsymbol{x})} \ell(\sigma, \xi; \boldsymbol{x}) = \infty$; optimization is therefore performed on the constrained parameter space $\{\xi > 0, \sigma > 0\} \sqcup \{-1 \leu \xi < 0, -\sigma/\xi > \max(\boldsymbol{x})\}$ (Grimshaw, 1993).

For numerical optimization, it is best to reparametrize the model from $\boldsymbol{\theta} = (\sigma, \xi)$ to $\boldsymbol{\vartheta} = (\xi, \eta)$, where $\eta = -\xi/\sigma$. One can then find an explicit solution for the maximizing value of ξ given $\eta \neq 0$, i.e.,

$$\widehat{\xi}_\eta = \frac{1}{n}\sum_{i=1}^{n} \log(1 - \eta x_i).$$

The profile log-likelihood for η is

$$\ell_{\mathrm{p}}(\eta; \boldsymbol{x}) = -n - \sum_{i=1}^{n} \log(1 - \eta x_i) - n \log\left\{ -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\eta} \log(1 - \eta x_i) \right\},$$

which attains a maximum for $\eta < 1/x_{(n)}$; the profile log-likelihood is unbounded at $\eta \approx 0$, so one excludes a region near the origin from the line search

□

Example 1.18 (Numerical optimization for the Poisson point process likelihood)
If the estimated probability of exceedance is small and the Poisson approximation holds,

$$c\left\{1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right\}^{-1/\xi} \approx n_u,$$

We can take advantage of Grimshaw's routine and fit a generalized Pareto distribution as discussed in Example 1.17 to obtain parameter estimates $(\widehat{\sigma}_u, \widehat{\xi})$. Then, the maximum likelihood estimates for the Poisson process likelihood (1.1) are approximately $\widehat{\sigma} = \widehat{\sigma}_u \times (n_u/c)^{\widehat{\xi}}$ and $\widehat{\mu} = u - \widehat{\sigma}\{(n_u/c)^{-\widehat{\xi}} - 1\}/\widehat{\xi}$. These points are very close to the maximum likelihood estimates and can be used as starting values for a Newton–Raphson algorithm.

□

1.2.3 Cox–Snell bias correction

The following is taken from Barndorff-Nielsen and Cox (1994, §5.3). Cox and Snell (1968) derive an expression for the first-order bias of maximum likelihood estimates, starting from a second-order Taylor series expansion of the log-likelihood,

$$\frac{\partial \ell}{\partial \theta_u} + \sum_{r,s,t=1}^{p} (\widehat{\theta}_u - \theta_u)\frac{\partial^2 \ell}{\partial \theta_u \partial \theta_r} + \frac{1}{2}(\widehat{\theta}_s - \theta_s)(\widehat{\theta}_t - \theta_t)\frac{\partial^3 \ell}{\partial \theta_r \partial \theta_s \partial \theta_t} \doteq 0. \tag{1.6}$$

From eq. (1.6), we can obtain an asymptotic expansion for $\widehat{\theta}_r - \theta_r$ in terms of log-likelihood derivatives,

$$\widehat{\theta}_r - \theta_r = j^{rs}\ell_s + \frac{1}{2} j^{rs}\ell_{stu}(\widehat{\theta}_t - \theta_t)(\widehat{\theta}_u - \theta_u) + \cdots$$

$$= j^{rs}\ell_s + \frac{1}{2} j^{rs} j^{tu} j^{vw} \ell_{stu} \ell_u \ell_w + \cdots$$

by iterating. Replacing this in eq. (1.6) and taking expectations term by term yields the first order bias

$$\begin{aligned} b_u = \mathsf{E}\left(\widehat{\theta}_u - \theta_u\right) &= \sum_{r,s,t=1}^{p} i^{ur} i^{st} \left(\frac{1}{2}\kappa_{rst} + \kappa_{rs,t}\right) + \mathrm{O}_p(n^{-2}) \\ &= \sum_{r,s,t=1}^{p} i^{ur} i^{st} \left(\kappa_{rs}^{(t)} - \frac{1}{2}\kappa_{rst}\right) + \mathrm{O}_p(n^{-2}) \\ &= -\sum_{r,s,t=1}^{p} i^{u,r} i^{s,t} \frac{1}{2}(\kappa_{r,s,t} + \kappa_{r,st}) + \mathrm{O}_p(n^{-2}). \end{aligned} \tag{1.7}$$

Cordeiro and Klein (1994) proposed the second expression, which it is advantageous in that it does not involve mixed joint cumulants $\kappa_{rs,t}$, but rather the derivatives $\kappa_{rs}^{(t)} = \partial\kappa_{rs}/\partial\theta_t$. The last expression appears in Firth (1993) and McCullagh (2018, p. 210) and is based on Bartlett's third identity, coupled with the identity

$$\sum_{r,s,t} i^{u,r} i^{s,t} \kappa_{t,rs} = \sum_{r,s,t} i^{u,r} i^{t,s} \kappa_{s,rt} = \sum_{r,s,t} i^{u,r} i^{s,t} \kappa_{s,rt},$$

which is obtained by a change of indices and symmetry of the Fisher information.

Plug-in estimates of the cumulants can be used to obtain consistent estimates of the parameters. It follows that a bias-corrected estimate can be obtained by evaluating the vector $\boldsymbol{b}$ at the maximum likelihood estimate $\boldsymbol{b}(\widehat{\boldsymbol{\theta}})$ and subtracting it from the maximum likelihood estimates, giving $\boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}} - \boldsymbol{b}(\widehat{\boldsymbol{\theta}})$, with bias that is $\mathrm{O}(n^{-2})$ (Efron 1975, remark 11). An alternative is to rely on a parametric bootstrap to estimate the bias by simulating samples from the null model with parameters $\widehat{\boldsymbol{\theta}}$, cf. § 10.6 of Efron and Tibshirani (1993). For general bias corrections, Godwin and Giles (2019) propose solving the implicit equation

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{b}(\widetilde{\boldsymbol{\theta}}). \tag{1.8}$$

The point estimate $\widetilde{\boldsymbol{\theta}}$ is obtained numerically as the root of this nonlinear system of equations.

Example 1.19 (Bias correction for the generalized Pareto distribution)
Giles et al. (2016) derive the Cox–Snell bias for the generalized Pareto distribution; the log-likelihood for a sample of size one is

$$\ell(x;\boldsymbol{\theta}) = -\log(\sigma) - (1+1/\xi)\log(1+\xi x/\sigma), \qquad 1+\xi x/\sigma > 0, x, \sigma > 0, \xi \in \mathbb{R}. \tag{1.9}$$

By making the change of variables $x \mapsto \xi(y^{-\xi} - 1)/\sigma$, it is straightforward to calculate the

cumulants of the log-likelihood. The Cox–Snell correction is (Giles et al., 2016)

$$\boldsymbol{b}(\sigma,\xi) = \left(\frac{\sigma(4\xi^2+5\xi+3)}{1+3\xi}, -\frac{(1+\xi)(3+\xi)}{1+3\xi}\right).$$

The cumulant $k_{\sigma\sigma\xi}$ is incorrectly reported in Giles et al. (2016). We can numerically compare the first-order bias with that of the maximum likelihood estimator: we generated 5000 datasets consisting each of $n = 30$ observations following a $\mathsf{GP}(1,\xi)$ for $\xi \in \{-1, -0.99, \ldots, 1\}$. Figure 1.8 shows selected quantiles of the empirical distributions of $\widehat{\sigma}/\sigma$ and $\widehat{\xi} - \xi$. For the shape, the quantiles are linear close to -1; recall that the maximum likelihood estimator is super-efficient near the lower bound -1, cf. Example 1.17. The Cox–Snell correction (dot-dashed thick line) is unbounded in a neighbourhood of $\xi = -1/3$ and should not be used unless $\xi > -1/5$, but agrees well with the mean bias (thick full line) beyond this point. For the scale σ, the bias is negligible near $\xi = -1$, when the estimator is super-efficient and the variance of the estimator increases with ξ. The bias of σ is systematically positive and that for ξ is systematically negative. This bias is well-known in the hydrology literature; Hosking and Wallis (1987) compare the bias and variance of method of moments and probability weighted moments estimators relative to those of the maximum likelihood estimators and note that moment-based estimators can display smaller root mean squared errors in small samples. While maximum likelihood estimators minimize mean squared error asymptotically, Hosking and Wallis write that very large samples are necessary for this to be visible in practice.

□

Example 1.20 (Bias correction for the generalized extreme value distribution)
The supplemental material of Roodman (2018) provide a general method to derive the cumulants of the r-largest likelihood derivatives and the generalized extreme value distribution. We derived those independently and report them in Appendix A.1.

□

1.2.4 Firth's score correction

Let $\boldsymbol{b}(\boldsymbol{\theta})$ denote the Cox–Snell first-order bias vector, which is $\mathsf{O}(n^{-1})$. In order to perform bias correction, Firth (1993) defines a modified score function,

$$U^*(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) + A(\boldsymbol{\theta}), \tag{1.10}$$

where $\mathsf{E}\,(A(\boldsymbol{\theta})) = -i(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})$. This adjusts for the fact that the score $U(\boldsymbol{\theta})$ is not usually linear in $\boldsymbol{\theta}$ and solving for $U(\boldsymbol{\theta}) = 0$ leads to bias. Natural candidates for $A(\boldsymbol{\theta})$ are the observed information, $A^{(\mathrm{o})}(\boldsymbol{\theta}) = -j(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})$ and the Fisher information $A^{(\mathrm{e})}(\boldsymbol{\theta}) = -i(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})$, but the former is typically more efficient. Firth expands the score equation U^* in a Taylor series around $\boldsymbol{\theta}$, inverts the expansion term by term and take expectations to get

$$(\theta_r^* - \theta_r) = -b_r(\boldsymbol{\theta}) + n^{-1}\sum_{s=1}^{p} i^{rs}\mathsf{E}\,(A_s) + \mathsf{O}_p(n^{-3/2}),$$

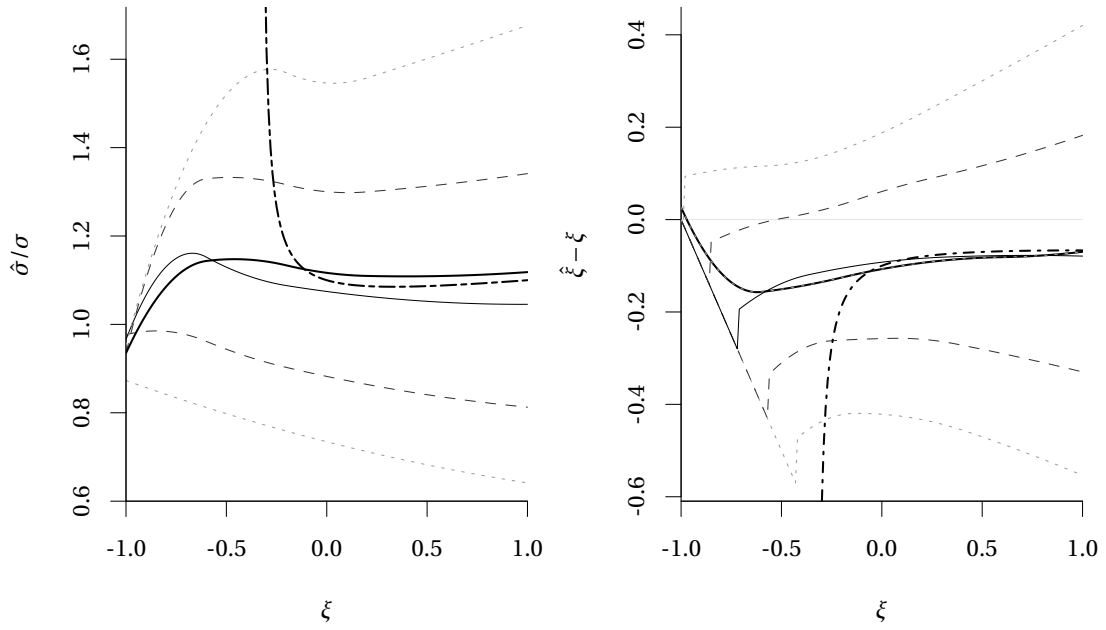

Figure 1.8 – *Distribution of the maximum likelihood estimator as a function of ξ: $\widehat{\sigma}/\sigma$ (left) and $\widehat{\xi}-\xi$ (right) based on 5000 datasets of size $n = 30$ from $\mathsf{GP}(1,\xi)$. The lines show the 0.1 and 0.9 percentiles (grey dotted), quartiles (dashed) and median (full line) of the distribution of $\widehat{\sigma}/\sigma$ and $\widehat{\xi}-\xi$ computed at each 0.01 increment and smoothed using a cubic spline. The pointwise mean (thick full) and the Cox–Snell bias correction (thick dot-dashed) $\boldsymbol{b}(1,\xi)$ of Equation* (1.9) *are overlaid.*

which entails that $\mathsf{E}(A_s) = i_{rs}b_s(\boldsymbol{\theta}) + \mathsf{O}_p(n^{-1/2})$ in order for the root of the score function U^* in eq. (1.10) to yield a bias-corrected estimator. The difference between $i(\widehat{\boldsymbol{\theta}})$ and observed information matrix $j(\widehat{\boldsymbol{\theta}})$ is usually $\mathsf{O}_p(n^{-1/2})$ and the latter is preferred for bias correction, though for full exponential families $i(\boldsymbol{\theta}) = j(\boldsymbol{\theta})$. Firth (1993) notes that for exponential family models, this is equivalent to penalizing by Jeffreys' prior. Alternatively, one could consider median bias reduction of maximum likelihood estimates (Kenne Pagui et al., 2017).

The estimators of the first-order bias (1.7) rely on the assumption that the data arise from the specified distribution. Furthermore, the cumulants appearing in the first-order bias are defined provided $-1/3 < \xi < 1$ for the generalized Pareto and generalized extreme value distributions. Bartlett's identities hold for both families despite the fact that the support depends on the parameters. Derivation of the Cox–Snell bias is in principle straightforward, but typically the vector $\boldsymbol{\theta}$ is not of interest in its own right; one is rather interested in some functional $g(\boldsymbol{\theta})$, whose first-order bias must be derived separately; this has yet to be done for extreme value distributions.

1.2.5 Test statistics based on first-order asymptotics and Bartlett adjustment

Frequentist testing and confidence interval procedures often rely on the asymptotic distribution of test statistics. The starting point for the distributional theory surrounding these statistics is the asymptotic normality of the score $U(\boldsymbol{\theta})$ by means of a central limit theorem, $U(\boldsymbol{\theta}) \dot{\sim} \mathsf{No}(0, i(\boldsymbol{\theta}))$. The variance of $U(\boldsymbol{\theta}_0)$ is exactly $i(\boldsymbol{\theta}_0)$, while that of $\widehat{\boldsymbol{\theta}}$ is approximately $i(\boldsymbol{\theta}_0)^{-1}$ under the null hypothesis $\mathscr{H}_0$.

We can expand the score function $U(\boldsymbol{\theta})$ in a neighbourhood of the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$. The latter solves $U(\widehat{\boldsymbol{\theta}}) = 0$ by definition, as the score is an unbiased estimator of zero, $\mathsf{E}\{U(\boldsymbol{\theta})\} = 0$. Following the local linearization step, one inverts the expansion and replaces the observed information term-wise by the expected information using the expansion (cf. Cordeiro and Cribari-Neto, 2014, eq 2.2)

$$j^{-1} = -j(\widehat{\boldsymbol{\theta}})^{rs} = i(\boldsymbol{\theta})^{rs} + \sum_{t,u=1}^{p} i(\boldsymbol{\theta})^{rt} i(\boldsymbol{\theta})^{su} \{i(\widehat{\boldsymbol{\theta}}_{tu}) - j(\widehat{\boldsymbol{\theta}}_{tu})\} + \cdots,$$

and thus $j(\widehat{\boldsymbol{\theta}})^{rs}/i(\boldsymbol{\theta})^{rs} \xrightarrow{\mathsf{p}} 1$ for $r, s = 1, \ldots, p$. A more rigorous derivation can be found in Barndorff-Nielsen and Cox (1994, section 3.3).

The three main classes of statistics for testing a simple null hypothesis $\mathscr{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative $\mathscr{H}_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ are the likelihood ratio, the score and the Wald statistics, defined respectively as

$$w := 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\}, \qquad w_{\mathsf{score}} := U^\top(\boldsymbol{\theta}_0) i^{-1}(\boldsymbol{\theta}_0) U(\boldsymbol{\theta}_0), \qquad w_{\mathsf{wald}} := (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top i(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate under the alternative and $\boldsymbol{\theta}_0$ is the null value of the parameter vector. A Taylor expansion of w about $\widehat{\boldsymbol{\theta}}$ gives

$$w(\boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top j(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathrm{o}_p(1) = U(\boldsymbol{\theta})^\top i^{-1}(\boldsymbol{\theta}) U(\boldsymbol{\theta}) + \mathrm{o}_p(1) \xrightarrow{d} \chi^2_p, \qquad n \to \infty.$$

The statistics $w, w_{\mathsf{score}}, w_{\mathsf{wald}}$ are all equivalent to $\mathrm{O}_p(n^{-1/2})$, but the score statistic w_{score} only requires calculation of the score and information under $\mathscr{H}_0$, which can be useful in problems where calculations under the alternative are costly. The Wald statistic w_{wald} is not parametrization-invariant.

For scalar θ, signed versions of these statistics exist, and we will mostly look at variants of

$$r(\theta) := \mathrm{sign}(\widehat{\theta} - \theta)\left[2\{\ell(\widehat{\theta}) - \ell(\theta)\}\right]^{1/2} \dot{\sim} \mathsf{No}(0, 1). \tag{1.11}$$

Example 1.21 (Regularity of the score test for the r-largest observation likelihood) In a recent paper, Bader et al. (2017) claimed that the asymptotic distribution of the score statistic for the order statistic likelihood given in Proposition 1.5 is not the usual χ^2 distribution and that the quality of the approximation deteriorates when r increases. They state that this non-regularity comes from the fact that the support of the distribution depends on the

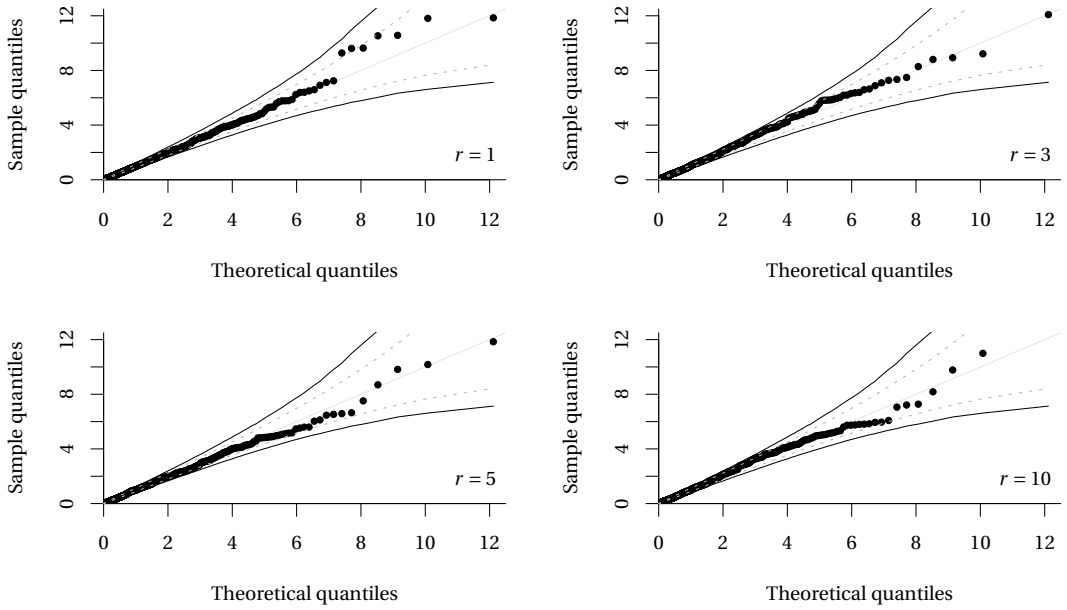

Figure 1.9 – χ^2_1 *quantile-quantile plots of the score test statistic* w_{score} *based on* 1000 *independent replicates of the score test statistic computed on a sample size* $n = 5000$ *with* $r = 1,3,5,10$ *under* $\mathscr{H}_0 : \xi = 0.1$*, with bootstrap pointwise (points) and overall (lines) 95% confidence intervals.*

parameter. Indeed, for extreme value distributions, some of the kth order cumulants ($k \in \mathbb{N}^+$) are undefined when the shape parameter $\xi \leq -1/k$. This has practical implications: the Fisher information matrix is undefined if $\xi \leq -0.5$ and cannot be evaluated at the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ if $\widehat{\xi} \leq -0.5$, which precludes the use of the score test. The first Bartlett identity holds if $\xi > -1$ and the score has expectation zero at the true parameter value, meaning that $\mathrm{E}(U(\boldsymbol{X};\boldsymbol{\theta});\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \Theta$. The maximum likelihood estimate solves the score equation, $U(\mathbf{x};\widehat{\boldsymbol{\theta}}) = 0$ provided $\widehat{\xi} > -1$. This property can be used to verify that the optimization routine has converged.

The claim of Bader et al. is erroneous and seems to be due to an incorrect implementation of the score for the r-largest likelihood. The number of order statistics, r, is not a parameter of the model and we cannot make direct comparisons between different values of r without changing the observations that enter the likelihood. The score test of Bader et al. (2017) uses an ill-posed null hypothesis $\mathscr{H}_0 : \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$: the score (and thus the score test) would be identically zero if the numerical implementation was correct.

To see this, we simulate 1000 data sets of size $n = 5000$ under the null hypothesis $\mathscr{H}_0 : \xi = \xi_0$ from the r-largest order statistics for $r \in \{1,3,5,10\}$ and compute the score statistic $w_{\mathsf{score}} \equiv U(\mathbf{x};\widehat{\boldsymbol{\theta}}_{\xi_0})^\top i^{-1}(\mathbf{x};\widehat{\boldsymbol{\theta}}_{\xi_0})U(\mathbf{x};\widehat{\boldsymbol{\theta}}_{\xi_0})$. The empirical distribution of the test statistic in Figure 1.9 matches the asymptotic χ^2_1 distribution, independent of the value of r.

□

The χ^2 approximation can be improved by Taylor series expansion, showing that

$$\mathsf{E}\left(W(\boldsymbol{\theta}_0)\right) = p\{1+\beta(\boldsymbol{\theta}_0)/n\} + \mathrm{O}(n^{-2}),$$

and this suggests using the Bartlett-adjusted statistic $w'(\boldsymbol{\theta}) = w(\boldsymbol{\theta})/\{1+\beta(\boldsymbol{\theta})/n\}$.

Sometimes only a subset $\boldsymbol{\psi}$ of $\boldsymbol{\theta}$ is of interest. The parameter vector $\boldsymbol{\theta}$ is thus partitioned into nuisance parameters $\boldsymbol{\lambda}$ and interest parameters $\boldsymbol{\psi}$. The likelihood ratio test for $\boldsymbol{\psi}$ is

$$w(\boldsymbol{\psi}) = 2\left\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}_0})\right\},$$

and the Bartlett adjustment term $\beta(\boldsymbol{\psi})$ is (Barndorff-Nielsen and Cox, 1994, p. 152)

$$\beta(\boldsymbol{\psi}) = \sum_{r,s,t,u=1}^{p} i^{rs} i^{tu} \delta_{rstu} + \sum_{r,s,t,u,v,w=1}^{p} i^{rs} i^{tu} i^{vw} \delta_{stuvw} + \mathrm{O}(n^{-2}),$$

where

$$\delta_{rstu} = \frac{1}{4}\nu_{rstu} + \nu_{rt,su} + \nu_{su,r,t} + \nu_{stu,r}\frac{1}{4}\nu_{rstu} - \nu^{(u)}_{rst} + \nu^{(su)}_{rt}$$

and

$$\begin{aligned}\delta_{rstuvw} &= \frac{1}{4}\nu_{rst}\nu_{uvw} + \nu_{st,r}\nu_{uvw} + \nu_{st,r}\nu_{uv,w} + \nu_{suw}\left(\frac{1}{2}\nu_{rtv} + \frac{1}{3}\nu_{r,t,v} + 2\nu_{tv,r}\right) + \nu_{tv,r}\nu_{su,w} \\ &= \frac{1}{6}\nu_{rtv}\nu_{suw} + \frac{1}{4}\nu_{rtu}\nu_{svw} - \nu_{rtv}\nu^{(u)}_{sw} - \nu_{rtu}\nu^{(v)}_{sw} + \nu^{(v)}_{rt}\nu^{(u)}_{sw} + \nu^{(u)}_{rt}\nu^{(v)}_{sw}.\end{aligned}$$

The second parametrization, due to Lawley (1956), avoids calculation of mixed cumulants (cf. Cordeiro and Cribari-Neto, 2014, p. 18). The Bartlett correction for the parameter of interest is $\beta(\boldsymbol{\psi}) = \beta(\boldsymbol{\theta}) - \beta(\boldsymbol{\lambda})$.

In general, if we test a restriction $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, the expectation of the log-likelihood ratio becomes $p - q + \beta(\boldsymbol{\theta}) - \beta(\boldsymbol{\lambda})$. The terms are calculated as before, but summation only takes place over the parameters in $\boldsymbol{\lambda}$, since $\boldsymbol{\psi}_0$ is fixed.

An alternative to analytical derivation of the Bartlett correction is the use of bootstrap (Rocke, 1989), whereby one estimates $\mathsf{E}\{W(\boldsymbol{\theta}_0)\}$ by a Monte Carlo average $\bar{w}(\boldsymbol{\theta}_0)$, obtained by computing repeatedly the likelihood ratio test statistic under B samples from the null model, i.e., under $f(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}_0})$. The empirical Bartlett correction gives $w(\boldsymbol{\theta})p/\bar{w}(\boldsymbol{\theta}_0) \sim \chi^2_p$. The benefit of the Bartlett correction is that one only needs to compute the expectation pointwise at $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}_0}$ rather than approximate the distribution of the test statistic. The Bartlett corrected statistics are distributed as χ^2_p to order $\mathrm{O}(n^{-2})$.

Example 1.22 (Bartlett correction for the generalized Pareto distribution)
Tedious calculations give the overall Bartlett correction

$$\beta(\boldsymbol{\theta}) = -\frac{2\left(118\xi^5 - 285\xi^4 - 1216\xi^3 - 944\xi^2 - 210\xi - 7\right)}{3\left(4\xi+1\right)\left(3\xi+1\right)^2\left(2\xi+1\right)}, \qquad \xi > -1/4.$$

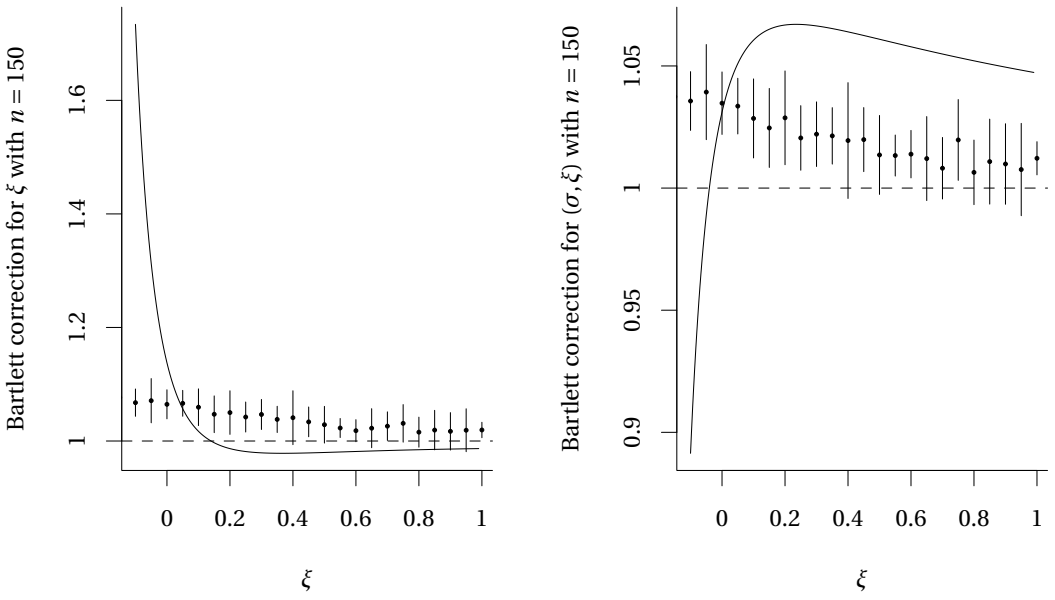

Figure 1.10 – *Analytic and bootstrap Bartlett correction (with pointwise 95% confidence intervals) for the generalized Pareto distribution for $n = 150$. Left panel: adjustment for the shape alone. Right panel: adjustment correction for joint tests of (σ, ξ). In both cases, the bootstrap corrections, $p/\bar{w}(\boldsymbol{\theta}_0)$, were computed by sampling $B = 10^5$ samples from $\mathsf{GP}(1, \xi_0)$.*

The Bartlett correction for the scale parameter is

$$\beta(\sigma) = -\frac{2\left(944\xi^8 + 1584\xi^7 + 880\xi^6 + 3540\xi^5 + 6753\xi^4 + 4626\xi^3 + 1090\xi^2 - 42\xi - 31\right)}{3(4\xi+1)(3\xi+1)^2(2\xi+1)^4},$$

for $\xi > -1/4$, while that for the shape is

$$\beta(\xi) = \frac{4\left(108\xi^5 + 298\xi^4 + 228\xi^3 + 43\xi^2 - 10\xi - 3\right)}{(4\xi+1)(3\xi+1)^2(2\xi+1)^2}, \qquad \xi > -1/4.$$

The Bartlett corrections are unbounded in a neighbourhood of $-1/4$ and can be negative; this is contrast to bootstrap-based corrections, as shown in Figure 1.10; the variability of the bootstrap estimates of the Bartlett correction for ξ is about 0.015 in the left panel of the figure. □

1.3 Higher-order asymptotics

The use of the profile likelihood and first-order methods is the starting point for statistical inference for parametric models. These methods may fare poorly if the dimension of the nuisance parameter $\boldsymbol{\lambda}$ is large. Additionally, testing procedures that rely on first-order theory may perform poorly in small samples. The distribution of such procedures is typically obtained

by keeping the first term of the asymptotic expansion of the test statistics, usually to an order of $\mathrm{O}(n^{-1})$. By keeping more terms from the expansion of the distribution, one may hope to achieve a more accurate approximation. Such so-called higher-order methods are intended to provide improved inferences. If the numerical discrepancy between first order and higher-order methods is negligible, this provides reassurance that the initial assessment was correct; otherwise higher-order development should in principle improve over first-order methods.

Much of the development of higher-order asymptotics relies on dimension reduction through conditioning. We recall some key notions related to data compression. The statistic $\boldsymbol{S}$ is sufficient if the conditional distribution of the sample $\boldsymbol{X}$ given $\boldsymbol{S}$ does not depend on the parameter $\boldsymbol{\theta}$. By the sufficiency principle, the maximum likelihood estimate, if it is unique, is a function of the sufficient statistic. Ancillary statistics are quantities whose distribution does not depend on $\boldsymbol{\theta}$. By Basu's theorem, if $(\boldsymbol{S}, \boldsymbol{A})$ is a minimal sufficient statistic and $\boldsymbol{S}$ is complete sufficient with $\boldsymbol{A}$ ancillary, then $\boldsymbol{S}$ and $\boldsymbol{A}$ are independent.

In the presence of nuisance parameters, we require a different notion of ancillarity. Consider a decomposition of the parameter vector into features of interest and nuisance, $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, assuming the sample space factorizes as $\Omega_{\boldsymbol{\Theta}} = \Omega_{\boldsymbol{\Psi}} \times \Omega_{\boldsymbol{\Lambda}}$. Then $\boldsymbol{A}$ is a cut, or ancillary for the minimal sufficient statistic $(\boldsymbol{S}, \boldsymbol{A})$, if the factorization

$$p(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}) = p(\boldsymbol{a}; \boldsymbol{\lambda}) p(\boldsymbol{s} \mid \boldsymbol{a}; \boldsymbol{\psi})$$

holds (Barndorff-Nielsen and Cox 1994, p. 38). In such cases, we may be interested in performing conditional inference on $\boldsymbol{\psi}$ by considering only the term $p(\boldsymbol{s} \mid \boldsymbol{a}; \boldsymbol{\psi})$. An alternative, if we can factorize $\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2)$ into

$$p(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{a}; \boldsymbol{\psi}, \boldsymbol{\lambda}) = p(\boldsymbol{a}; \boldsymbol{\lambda}) p(\boldsymbol{s}_1 \mid \boldsymbol{a}; \boldsymbol{\psi}) p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{a}; \boldsymbol{\psi}, \boldsymbol{\lambda}), \tag{1.12}$$

is to base inference for $\boldsymbol{\psi}$ on the marginal likelihood $p(\boldsymbol{s}_1 \mid \boldsymbol{a}; \boldsymbol{\psi})$ by neglecting the rightmost term of eq. (1.12) if the latter is complicated. Chapter 12 of Davison (2003) gives examples of using conditional or marginal likelihood for inference.

1.3.1 Modified profile likelihood and the p^* approximation

For a given value of the q-dimensional vector $\boldsymbol{\psi}$, we denote the partial maximum likelihood estimator of $\boldsymbol{\lambda}$ by $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ and the maximum likelihood estimator by $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}}_{\widehat{\boldsymbol{\psi}}})$. The profile likelihood is $\ell_{\mathrm{p}}(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})$ and the score of the profile is asymptotically $U_{\mathrm{p}}(\widehat{\boldsymbol{\psi}}) \sim \mathsf{No}(\boldsymbol{0}, j_{\mathrm{p}}(\widehat{\boldsymbol{\psi}})) + \mathrm{O}(n^{-1/2})$ (Barndorff-Nielsen and Cox, 1994, p. 181). First order inference is based on the signed profile likelihood root,

$$r(\psi) = \mathrm{sign}(\widehat{\psi} - \psi)[2\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi)\}]^{1/2} \sim \mathsf{No}(0, 1). \tag{1.13}$$

Higher order asymptotic developments can be used to obtain more accurate confidence intervals. Barndorff-Nielsen derives a modified likelihood root statistic r^* for scalar ψ (Barndorff-

Nielsen and Cox, 1994, §6.6.1)

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\left\{\frac{q(\psi)}{r(\psi)}\right\}. \tag{1.14}$$

The asymptotic distribution of $q(\psi)$ in eq. (1.14) is parameter-free and $q(\psi)$ provides an approximate pivot. The distribution of the modified signed likelihood root $r^*(\psi)$ is closer to a standard Gaussian distribution than its counterpart $r(\psi)$, since the relative error of $r^*(\psi)$ is $\mathrm{O}(n^{-3/2})$, or $\mathrm{O}(n^{-1})$ for discrete likelihoods, compared to $\mathrm{O}(n^{-1/2})$ for $r(\psi)$. A scaled profile function can be obtained by taking $\exp(-r^2/2)$.

Barndorff-Nielsen (1980, 1983, 1988) proposed the p^* approximation for the conditional distribution of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ given an ancillary statistic $\boldsymbol{a}$,

$$p^*(\widehat{\boldsymbol{\theta}}; \boldsymbol{\theta} \mid \boldsymbol{a}) \doteq c(\boldsymbol{\theta}, \boldsymbol{a}) |j(\widehat{\boldsymbol{\theta}})|^{1/2} \exp\{\ell(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \boldsymbol{a}) - \ell(\widehat{\boldsymbol{\theta}}; \widehat{\boldsymbol{\theta}}, \boldsymbol{a})\},$$

where $c(\boldsymbol{\theta}, \boldsymbol{a})$ is a normalizing constant. A main benefit of the p^* formula is that it is invariant to one-to-one transformations of the data while simultaneously being parametrization-invariant. The p^* approximation compresses the information contained in the n-sample $\boldsymbol{x}$ into the statistic $\widehat{\boldsymbol{\theta}}$, achieving dimension reduction from n to p in such a way that the information lost is minimized. In transformation models, the pair $(\widehat{\boldsymbol{\theta}}, \boldsymbol{a})$ is minimal sufficient and the result holds exactly (Barndorff-Nielsen and Cox, 1994). Define the sample space and mixed derivatives

$$\ell_{;\widehat{\boldsymbol{\theta}}}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \boldsymbol{a}) = \frac{\partial \ell(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \boldsymbol{a})}{\partial \widehat{\boldsymbol{\theta}}}, \quad \ell_{\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \boldsymbol{a}) = \frac{\partial^2 \ell(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \boldsymbol{a})}{\partial \boldsymbol{\theta} \partial \widehat{\boldsymbol{\theta}}^\top};$$

the correction term $q(\psi)$ that appears in eq. (1.14) for scalar ψ is

$$q(\psi) = \frac{\left|\ell_{;\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}}) - \ell_{;\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}}_\psi)\, \ell_{\boldsymbol{\lambda};\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}}_\psi)\right|}{\left|\ell_{\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}})\right|} \frac{\left|j_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\right|^{1/2}}{\left|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_\psi)\right|^{1/2}}.$$

Modifications of the profile log-likelihood function to achieve higher-order accuracy are possible if we have an explicit bijection between the data $\boldsymbol{y}$ and $(\widehat{\boldsymbol{\theta}}, \boldsymbol{a})$; the Barndorff-Nielsen approximation takes the form

$$\ell_{\mathrm{b}}(\psi) = \ell_{\mathrm{p}}(\psi) + \frac{1}{2}\log\left\{\left|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_\psi)\right|\right\} - \log\left\{\left|\frac{\partial^2 \ell(\psi, \boldsymbol{\lambda}; \widehat{\psi}, \widehat{\boldsymbol{\lambda}}, \boldsymbol{a})}{\partial \widehat{\boldsymbol{\lambda}} \partial \boldsymbol{\lambda}^\top}\right|\right\}. \tag{1.15}$$

In general, exact ancillary statistics need not exist and this precludes the use of the p^* formula. The sample space derivative with respect to $\widehat{\boldsymbol{\lambda}}$ that appears in the Jacobian term of eq. (1.15) can be approximated if one resorts to the use of an approximate ancillary statistic, i.e., a statistic $\boldsymbol{a}$ such that the pair $(\widehat{\boldsymbol{\theta}}, \boldsymbol{a})$ is sufficient up to $\mathrm{O}_p(1/n)$, i.e., $p(\boldsymbol{a}; \boldsymbol{\theta}_0 + \boldsymbol{\delta} n^{-1/2}) = p(\boldsymbol{a}; \boldsymbol{\theta}_0)\{1 + O(n^{-1})\}$; see eq. (8.3) of McCullagh (2018). In the case of multiple parameters, it is customary

to do the inference for each variable in turn; alternatives include the Skovgaard (2001) and Fraser et al. (2016) analogs of r^* in the case of vector $\boldsymbol{\psi}$.

1.3.2 Orthogonal parametrization

Orthogonality between parameters helps reduce variability in estimators. The parameters of interest $\boldsymbol{\psi}$ are said to be globally orthogonal to the nuisance parameters $\boldsymbol{\lambda}$, written $\boldsymbol{\psi} \perp \boldsymbol{\lambda}$, if the information matrix is block diagonal, i.e., $i_{\boldsymbol{\psi\lambda}}(\boldsymbol{\theta}) = \boldsymbol{O}$ for any value of $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\Theta}}$. Cox and Reid (1987) show that, for values of $\boldsymbol{\psi}$ in a neighbourhood of the maximum likelihood estimate with $\boldsymbol{\psi} \perp \boldsymbol{\lambda}$ having $\boldsymbol{\psi} - \widehat{\boldsymbol{\psi}} = \mathrm{O}_p(n^{-1/2})$, then $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}} - \widehat{\boldsymbol{\lambda}} = \mathrm{O}_p(n^{-1})$ instead of the usual $\mathrm{O}_p(n^{-1/2})$. Indeed, when both ψ, λ are scalar (Barndorff-Nielsen and Cox 1994, § 3.6)

$$\widehat{\lambda}_{\psi} - \widehat{\lambda} = -j^{-1}_{\lambda\lambda}(\widehat{\boldsymbol{\theta}}) j_{\psi\lambda}(\widehat{\boldsymbol{\theta}})(\psi - \widehat{\psi}) + \mathrm{O}_p(n^{-1}) = -i^{-1}_{\lambda\lambda} i_{\psi\lambda}(\psi - \widehat{\psi}) + \mathrm{O}_p(n^{-1}),$$

showing that in the general case $\widehat{\lambda}_{\psi} - \widehat{\lambda} = \mathrm{O}_p(n^{-1/2})$ unless the expected information block $i_{\psi\lambda}$ or its observed counterpart $j_{\psi,\lambda}(\widehat{\boldsymbol{\theta}})$ are zero matrices. One could thus omit the Jacobian term from the modified profile likelihood eq. (1.15) in the hope that the $\mathrm{O}_p(n^{-1})$ term is negligible, giving the Cox–Reid log-likelihood,

$$\ell_{\mathsf{cr}_1}(\psi) \coloneqq \ell(\psi, \widehat{\boldsymbol{\lambda}}_{\psi}) - \frac{1}{2}\log\{|j_{\boldsymbol{\lambda\lambda}}(\psi, \widehat{\boldsymbol{\lambda}}_{\psi})|\}.$$

The bias of the score is $\mathrm{O}(n^{-1})$, but the Cox–Reid likelihood does not correct the bias of the information and is not invariant to reparametrization. An alternative to global orthogonality is to require only local orthogonality, i.e., $i_{\boldsymbol{\psi\lambda}}(\boldsymbol{\theta}_0) = \mathbf{0}$ for a given value of $\boldsymbol{\theta}_0$.

Transformation to orthogonality is always possible if ψ is scalar, but compatibility conditions appear when $\boldsymbol{\psi}$ is a vector and may not be satisfied. The choice of orthogonal parametrization is not unique: if $\boldsymbol{\lambda} \perp \psi$, then any smooth one-to-one transformation $g(\boldsymbol{\lambda})$ is also orthogonal to ψ, and this raises the question as to which parametrization is best suited for the analysis.

We now provide details on how to find an orthogonal parametrization for scalar ψ. Consider a transformation from $(\psi, \boldsymbol{\eta})$ to $(\psi, \boldsymbol{\lambda})$ such that $\psi \perp \boldsymbol{\lambda}$; then one can consider the mixed partial derivatives $\{\partial^2\ell/(\partial\psi\partial\lambda_r)\}$ expressed as a function of $(\psi, \boldsymbol{\eta})$ using the chain rule. On taking expectations, the system of partial differential equations to be solved reduces to

$$\left(\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\lambda}}\right)^{\top}\left(i_{\boldsymbol{\eta}\psi} + i_{\boldsymbol{\eta\eta}}\frac{\partial\boldsymbol{\eta}}{\partial\psi}\right) = 0,$$

so $\boldsymbol{\lambda}(\psi, \boldsymbol{\eta})$ is a solution of

$$\frac{\partial\boldsymbol{\eta}}{\partial\psi} = -i^{-1}_{\boldsymbol{\eta\eta}} i_{\boldsymbol{\eta}\psi}.$$

If the Jacobian of the transformation from $\boldsymbol{\eta} \mapsto \boldsymbol{\lambda}$ is non-zero, we can find $g(\boldsymbol{\lambda})$ as the constant of integration in the solution of this system of first-order partial differential equations.

It is also possible to express the likelihood in terms of the original parameters $(\psi, \boldsymbol{\eta})$, without deriving $\boldsymbol{\lambda}$ explicitly in terms of the latter. The modified profile log-likelihood is then (Davison 2003, § 12.4.2)

$$\ell_{\mathsf{cr}_1} \coloneqq \ell(\psi, \widehat{\boldsymbol{\eta}}_\psi) - \frac{1}{2}\log\left\{\left|j_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\psi}, \widehat{\boldsymbol{\eta}}_\psi)\right|\right\} + \log\left\{\left|\frac{\partial \lambda(\psi, \widehat{\boldsymbol{\eta}}_\psi)}{\partial \boldsymbol{\eta}^\top}\right|\right\}. \tag{1.16}$$

Orthogonality of the parameters has many important properties that are summarized in Cox and Reid (1987). A key property is asymptotic independence of the maximum likelihood estimators $\widehat{\psi}$ and $\widehat{\boldsymbol{\lambda}}$. Another is that the conditional estimates $\widehat{\psi}_{\boldsymbol{\lambda}}$ vary less than their non-orthogonal counterparts $\widehat{\psi}_{\boldsymbol{\eta}}$ for given $\boldsymbol{\lambda}, \psi$.

Cox and Reid (1993) propose an alternative profile likelihood modification that does not require formal derivation of an orthogonal parametrization for a scalar parameter of interest ψ, viz.

$$\ell_{\mathsf{cr}_2}(\psi, \widehat{\boldsymbol{\eta}}_\psi) - \frac{1}{2}\log\left\{|j_{\boldsymbol{\eta}\boldsymbol{\eta}}(\psi, \widehat{\boldsymbol{\eta}}_\psi)|\right\} - (\psi - \widehat{\psi})\mathrm{tr}\left\{\frac{\partial}{\partial \boldsymbol{\eta}}\left(i^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}} i_{\boldsymbol{\eta}\psi}\right)\right\}\Big|_{\psi=\widehat{\psi}, \boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}}.$$

The formula is not invariant to interest-respecting transformations, as it involves an approximate linearization step.

Example 1.23 (Quantiles of the generalized Pareto distribution)
The following is Example 12.25 from Davison (2003). Suppose one is interested in the $(1-p)$ quantile of the generalized Pareto distribution,

$$\psi = \begin{cases} \sigma(p^{-\xi} - 1)/\xi, & \xi \neq 0, \\ -\sigma\log(p), & \xi = 0. \end{cases}$$

The information matrix can be written as

$$i(\psi, \xi) = \frac{1}{(1+\xi)(1+2\xi)}\begin{pmatrix} i_{\psi\psi} & i_{\psi\xi} \\ i_{\xi\psi} & i_{\xi\xi} \end{pmatrix},$$

where

$$\begin{aligned} i_{\psi\psi} &= \frac{a_{\psi\psi}(\xi)}{\psi^2} = \frac{1+\xi}{\psi^2}, \\ i_{\psi\xi} &= \frac{a_{\psi\xi}(\xi)}{\psi} = \frac{(1+\xi)\{1+\xi\log(p) - p^\xi\}}{\psi\xi(1-p^\xi)} + \frac{1}{\psi}, \\ i_{\xi\xi} &= a_{\xi\xi}(\xi) = \frac{(1+\xi)\{1+\xi\log(p) - p^\xi\}^2}{\xi^2(1-p^\xi)^2} + \frac{2\{1+\xi\log(p) - p^\xi\}}{\xi(1-p^\xi)} + 2. \end{aligned}$$

The information matrix is incorrectly reported in Davison (2003). The partial differential

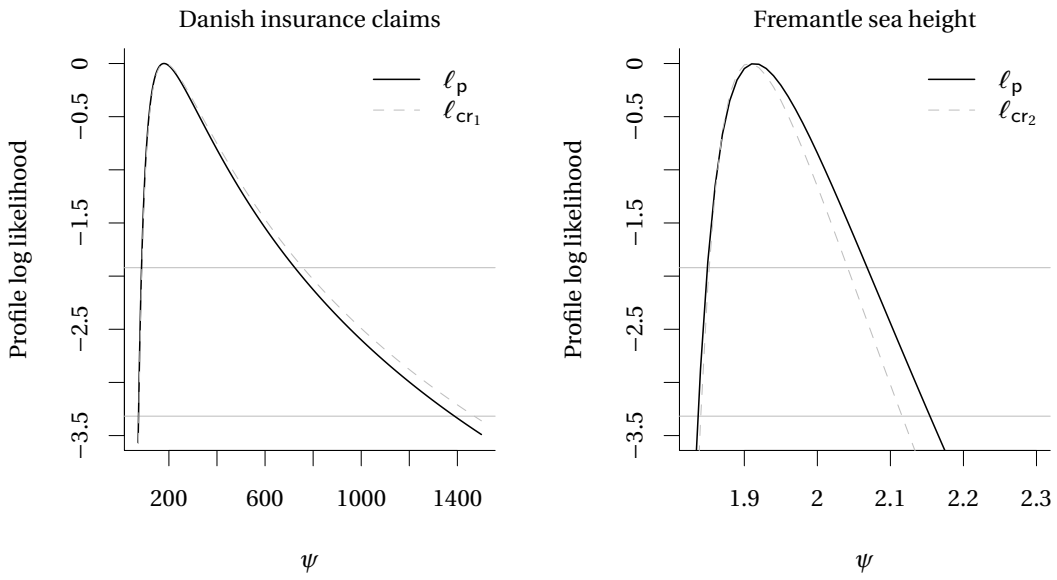

Figure 1.11 – *Left: profile log-likelihood for the 99 percentile of the generalized Pareto distribution fitted to the Danish insurance claim data. Estimation is based on the $k = 60$ exceedances over the threshold $u = 15$. The solid line shows the profile for the parametrization (ψ, ξ), while the dashed line shows orthogonal parametrization (ψ, λ). Right: profile log-likelihood for the average of the 100-year maximum sea level data, with original parametrization (solid line) and approximate orthogonal parametrization (dashed grey). The horizontal grey lines indicate cutoff values for 95% and 99% percentiles based on the asymptotic χ^2_1 distribution.*

equation

$$\psi \frac{\partial \xi}{\partial \psi} = -\frac{a_{\psi\xi}(\xi)}{a_{\xi\xi}(\xi)}$$

can be solved by the method of separation of variables and we can use eq. (1.16) with $\partial\lambda/\partial\psi = a_{\xi\xi}(\xi)/a_{\psi\xi}(\xi)$.

The left panel of Figure 1.11 shows the profile log-likelihood for the 99 percentile of the generalized Pareto distribution fitted to exceedances over $u = 15$ for the Danish insurance claim data (Embrechts et al., 1997), which yields $k = 60$ observations. The profile log-likelihood in the orthogonal parametrization has a slightly wider right tail. Both have been rescaled so that the maximum is attained at zero.

□

Example 1.24 (Orthogonal parametrization for Fremantle sea level dataset)
The Fremantle sea level data consists of $n = 86$ measurements of maximum annual sea height levels measured between 1897 to 1989 (Coles, 2001). The profile log-likelihood for the Cox–Reid approximate orthogonal parametrization is easily computed numerically after fitting a generalized extreme value distribution to the data, and the right panel of Figure 1.11 displays

the profile log-likelihoods for the mean of the 100-years maximum. The estimated average $\widehat{z_q} = 1.91$ lies between the first and second order statistics.

□

1.3.3 Tangent exponential model

Calculation of Barndorff-Nielsen's p^* approximation requires a decomposition of the parameter $\boldsymbol{\theta}$ in terms of the maximum likelihood and an explicit ancillary statistic. Fraser et al. (1999) propose an approximation to the correction factor q, built in three steps. First, by conditioning on an approximately ancillary statistic with relative error $\mathrm{O}(n^{-1})$, the dimension of the data is reduced from the sample size n to the dimension of the parameter p. The model is then approximated by an exponential family with canonical parameter $\boldsymbol{\psi}$. Lastly, marginalization over the nuisance parameter $\boldsymbol{\lambda}$ of the exponential family yields a pivot that is a function of $\boldsymbol{\psi}$ and whose distribution does not depend on $\boldsymbol{\lambda}$.

Rather than approximate the density of the maximum likelihood estimator for every point $\boldsymbol{y} \in \mathbb{R}^d$, one can restrict attention to the sample points $\boldsymbol{y}^0$. The Fraser–Reid approximation, termed the tangent exponential model, does precisely this. Its rationale arises from consideration of the full exponential family model, where the derivative of the log-likelihood with respect to the minimal sufficient statistic gives $\boldsymbol{\varphi}$, the canonical parameter. Following the lines of Fraser (2003) and Brazzale et al. (2007), we sketch the arguments leading to its construction.

Let Y be a random variable with distribution function F, density function f and p-dimensional vector of parameters is $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda})$; the observed data is $\boldsymbol{y}^0$. Define the $n \times d$ matrix of approximate ancillary statistics $\mathbf{V}$. The ith row of $\mathbf{V}$ for continuous parameters could be, e.g.,

$$\mathbf{V}_i = -\frac{\partial F\big(y_i^0;\boldsymbol{\theta}\big)}{\partial \boldsymbol{\theta}} \frac{1}{f\big(y_i^0;\boldsymbol{\theta}\big)}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}},$$

whereas for discrete models, $\mathbf{V}_i = \mathrm{d}\mathsf{E}\left(Y_i;\boldsymbol{\theta}\right)/\mathrm{d}\boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$.

Further define the gradient $\partial\ell/\partial\boldsymbol{y}$ in the direction $\mathbf{V}$,

$$\varphi(\boldsymbol{\theta})^\top = \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\mathbf{V}_i}\ell(\boldsymbol{\theta};\boldsymbol{y})\bigg|_{\boldsymbol{y}_i=\boldsymbol{y}_i^0} = \mathbf{V}^\top \frac{\partial\ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{y}}\bigg|_{\boldsymbol{y}=\boldsymbol{y}^0},$$

where the directional derivative $(\mathrm{d}/\mathrm{d}\mathbf{v})h(\mathbf{y}) = (\mathrm{d}/\mathrm{d}\mathbf{x})h(\mathbf{y}+\mathbf{x}\mathbf{v})|_{\mathbf{x}=\mathbf{0}}$. The parameter φ can be interpreted as the projection of the sample space derivative onto the span of $\mathbf{V}$. The approximate pivot q of eq. (1.14) for the tangent exponential model approximation is

$$q(\psi) = \frac{\Big|\varphi(\widehat{\boldsymbol{\theta}}) - \varphi(\widehat{\boldsymbol{\theta}}_\psi) \quad \partial\varphi/\partial\boldsymbol{\lambda}^\top(\widehat{\boldsymbol{\theta}}_\psi)\Big|}{\big|\partial\varphi/\partial\boldsymbol{\theta}^\top(\widehat{\boldsymbol{\theta}})\big|} \frac{\big|j(\widehat{\boldsymbol{\theta}})\big|^{1/2}}{\big|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_\psi)\big|^{1/2}}. \tag{1.17}$$

The first matrix in the numerator of eq. (1.17) is formed by binding the $d \times 1$ column vector

$\varphi(\widehat{\boldsymbol{\theta}}) - \varphi(\widehat{\boldsymbol{\theta}}_\psi)$ with the $d \times (d-1)$ matrix $\partial\varphi/\partial\boldsymbol{\lambda}^\top$.

Instead of the p^* approximation, one could use the tangent exponential model approximation directly in place of the rightmost term of eq. (1.15), giving an analogous modified profile likelihood (Fraser, 2003, eq. 1.3)

$$\ell_{\mathrm{fr}}(\boldsymbol{\psi}) = \ell_{\mathrm{p}}(\boldsymbol{\psi}) + \frac{1}{2}\log\left\{\left|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\right|\right\} - \frac{1}{2}\log\left\{\left|\boldsymbol{\varphi}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})^\top j_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}})\boldsymbol{\varphi}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\right|\right\}.$$

where $\boldsymbol{\varphi}_{\boldsymbol{\lambda}}$ is $p \times 1$ Jacobian of $\boldsymbol{\varphi}$ with respect to the nuisance parameter $\boldsymbol{\lambda}$ and $j_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\theta}}) = -(\partial^2/\partial\varphi^2)\ell(\boldsymbol{\theta}; \boldsymbol{y}^0)$ is the observed information in the $\boldsymbol{\varphi}$ parametrization (Fraser, 2004, pp. 338–339). Since a closed-form expression for the likelihood in terms of $\boldsymbol{\psi}$ is rarely available, we report instead $\ell_{\mathrm{fr}}(\boldsymbol{\psi}) = \log\{\phi(r^*)\}$, where r^* is computed using the tangent exponent model approximation in eq. (1.17).

1.3.4 Other penalized profile likelihoods

The profile likelihood is not based on the density of a random variable and replacing the nuisance parameter by its maximum likelihood estimate $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ is not harmless, particularly if the dimension of $\boldsymbol{\lambda}$ is large. Severini (2000) lists many modifications of the profile likelihood that can be viewed as penalized estimators that approximate true marginal or conditional likelihood in the sense of the Barndorff-Nielsen p^* approximation. The general form of these modified profile log-likelihoods is

$$\ell_{\mathrm{m}}(\boldsymbol{\psi}) = \ell_{\mathrm{p}}(\boldsymbol{\psi}) + \frac{1}{2}\log\left\{\left|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\right|\right\} - \log\left\{\left|\ell_{\boldsymbol{\lambda};\widehat{\boldsymbol{\lambda}}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}};\widehat{\boldsymbol{\theta}}, \boldsymbol{a})\right|\right\}.$$

We list two further analytical approximations to ℓ_{m}, which consist in replacing $\ell_{\boldsymbol{\lambda};\widehat{\boldsymbol{\lambda}}}$ by an approximation. The first replaces the tangent space derivative by that obtained from the tangent exponential model construction,

$$\ell_{\mathrm{m}}^{\mathrm{tem}}(\boldsymbol{\psi}) = \ell_{\mathrm{p}}(\boldsymbol{\psi}) + \frac{1}{2}\log\left\{\left|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\right|\right\} - \log\left\{\left|\ell_{\boldsymbol{\lambda};\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\widehat{\mathbf{V}}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})\right|\right\}, \tag{1.18}$$

where $\ell_{\boldsymbol{\lambda};\boldsymbol{y}} = \partial^2\ell/\partial\boldsymbol{\lambda}\partial\boldsymbol{y}^\top$ is a mixed partial derivative. The second class of approximations, due to Severini and similar in spirit to ideas in Skovgaard (1996), uses covariances to approximate the sample space derivative, with an error of $\mathrm{O}(n^{-1})$ in the moderate deviation sense, i.e., for values of $\boldsymbol{\psi}$ with $\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi} = \mathrm{O}(n^{-1/2})$ and $\mathrm{O}(n^{-1/2})$, in the large deviation sense, i.e., for values $\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi} = \mathrm{O}(1)$; cf. Severini (2000). The covariance-based modified profile likelihood is

$$\ell_{\mathrm{m}}^{\mathrm{cov}}(\boldsymbol{\psi}) = \ell_{\mathrm{p}}(\boldsymbol{\psi}) + \frac{1}{2}\log\left\{\left|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\right|\right\} - \log\left\{\left|i_{\boldsymbol{\lambda};\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}};\widehat{\boldsymbol{\theta}})\right|\right\}, \tag{1.19}$$

where

$$i_{\boldsymbol{\lambda};\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}};\widehat{\boldsymbol{\theta}}) = \mathsf{E}\left(\ell_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\ell_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})^\top\right)$$

is unknown but can be replaced by its empirical counterpart,

$$\widehat{i}_{\boldsymbol{\lambda};\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}};\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell^{(i)}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\ell^{(i)}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})^{\top},$$

where $\ell^{(i)}_{\boldsymbol{\lambda}}$ denotes the score associated to the ith observation. This version has the same order of approximation error.

We look at two novel examples to illustrate the derivations required to derive the tangent exponential model approximation.

Example 1.25 (Profile likelihood for the scale of max-stable variates)
Suppose $Y_i \overset{\text{iid}}{\sim} \mathsf{GEV}(\mu,\sigma,\xi)$ and we have an n sample $\{y_i\}_{i=1}^n$; we consider $\psi=\sigma$ and $\boldsymbol{\lambda}=(\mu,\xi)$. The observed information $j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$ is easily obtained as the negative Hessian matrix of the log-likelihood. Since the distribution is a location-scale family, we have $\mathrm{V}_{i,\mu}=1$. Letting $s_i(\boldsymbol{\theta}) = \{1+\xi(y_i^0-\mu)/\sigma\}$, the second column of the $n\times 2$ matrix $\mathbf{V}$ has elements

$$\mathrm{V}_{i,\widehat{\xi}} = \frac{\widehat{\sigma}\, s_i(\widehat{\boldsymbol{\theta}})\log\{s_i(\widehat{\boldsymbol{\theta}})\}}{\widehat{\xi}^2} + \frac{y_i^0-\widehat{\mu}}{\widehat{\xi}},$$

while

$$\begin{aligned}
\ell_{\widehat{\mu}_\sigma,y_i}(\widehat{\boldsymbol{\theta}}_\sigma) &= \frac{1+\widehat{\xi}_\sigma}{\sigma^2}s_i(\widehat{\boldsymbol{\theta}}_\sigma)^{-1/\widehat{\xi}_\sigma-2} - \frac{\widehat{\xi}_\sigma(1+\widehat{\xi}_\sigma)}{\sigma^2 s_i(\widehat{\boldsymbol{\theta}}_\sigma)^2},\\
\ell_{\widehat{\xi}_\sigma,y_i}(\widehat{\boldsymbol{\theta}}_\sigma) &= \frac{s_i(\widehat{\boldsymbol{\theta}}_\sigma)^{-1/\widehat{\xi}_\sigma-1}}{\sigma}\left(\frac{(y_i^0-\widehat{\mu}_\sigma)(1+\widehat{\xi}_\sigma)}{\sigma\widehat{\xi}_\sigma s_i(\widehat{\boldsymbol{\theta}}_\sigma)} + \frac{\log\{s_i(\widehat{\boldsymbol{\theta}}_\sigma)\}}{\widehat{\xi}_\sigma^2}\right) + \frac{1}{\sigma s_i(\widehat{\boldsymbol{\theta}}_\sigma)} + \frac{(1+\widehat{\xi}_\sigma)(y_i^0-\widehat{\mu}_\sigma)}{\sigma^2 s_i(\widehat{\boldsymbol{\theta}}_\sigma)^2}.
\end{aligned}$$

These expressions can be substituted into $\ell^{\mathrm{tem}}_{\mathsf{m}}$ or $\ell^{\mathrm{cov}}_{\mathsf{m}}$ to obtain a modified profile log-likelihood function.

□

Example 1.26 (Expectation of N threshold exceedances)
Suppose data $\{x_i\}_{i=1}^{n_u}$ above a high threshold u are modelled with a generalized Pareto distribution, meaning

$$\mathsf{P}(X\leq z \mid X>u) = F(z) = 1-\left\{1+\xi\left(\frac{z-u}{\sigma_u}\right)\right\}_+^{-1/\xi}, \qquad z>u, \sigma_u>0, \xi\in\mathbb{R}.$$

Then, the maximum of N exceedances has expectation

$$\mathfrak{z}_N := \int_u^\infty zNF(z)^{N-1}f(z)\mathrm{d}z = u+\frac{\sigma}{\xi}\left\{\frac{\Gamma(N+1)\Gamma(1-\xi)}{\Gamma(N+1-\xi)}-1\right\}, \qquad \xi<1.$$

The pth percentile of the maxima of N observations is simply the $p^{1/N}$th percentile of the generalized Pareto distribution. We can profile over $\mathfrak{z}_N$ by parametrizing the likelihood in

terms of $(\mathfrak{z}_N, \xi)$. The n-vector $\mathbf{V}$ appearing in the tangent exponential model has elements

$$\mathrm{V}_i = \frac{\{(\mathfrak{z}_N - u) + (x_i - u)\}\log(\mathfrak{q}) - N(x_i - u)\mathrm{Be}(N, 1-\xi)\left\{\log(\mathfrak{q}) + \psi^{(0)}(1-\xi) - \xi\psi^{(0)}(N+1-\xi)\right\}}{\xi\{N\mathrm{Be}(N, 1-\xi) - 1\}},$$

where $\psi^{(0)}(\cdot) = \Gamma'(x)/\Gamma(x)$ is the digamma function and

$$\mathfrak{q} = \left[1 + \left(\frac{y_i - u}{\mathfrak{z}_N - u}\right)\{1 - N\mathrm{Be}(N, 1-\xi)\}\right].$$

The elements of the canonical parameter are

$$\varphi_i = \mathrm{V}_i \times \left(1 + \frac{1}{\xi}\right)\frac{(N!\Gamma(1-\xi) - \Gamma(N+1-\xi))}{-(x-u)N!\Gamma(1-\xi) + \{(\mathfrak{z}_N - u) + (y_i - u)\}\Gamma(N+1-\xi)}, \qquad i = 1, \ldots, n_u.$$

These expressions are used to calculate the modified profile log-likelihood appearing in Figure 1.12.

We illustrate the different approximations of the profile log-likelihood on simulated standard Gaussian data and compare inferences drawn using block maximum and peaks-over-threshold methods. We sample $n = 2000$ observations and target the mean of the maximum of $N = 2000$ variates, using $u = 1.96$ as threshold and with $m = 50$ for block maxima. The profile log-likelihood ℓ_{p} (heavy black), the tangent exponential model ℓ_{fr} (solid grey) and Severini's corrections with a tangent exponential model penalty $\ell_{\mathsf{m}}^{\mathrm{tem}}$ (dashed grey) and an empirical covariance approximation $\ell_{\mathsf{m}}^{\mathrm{cov}}$ (dashed black) are displayed in Figure 1.12. Most of the higher-order methods yield wider confidence intervals in the right tails. The true value, 3.43, is well within the confidence intervals in all cases.

□

Other approximations for the profile likelihood exist in the literature: Barndorff-Nielsen (1986) proposes a correction to handle non-orthogonal parameters, as do Cox and Reid (1993). However, the conditional profile likelihood is not parametrization-invariant and the choice of orthogonal parameter is not unique. Another approximation is proposed in McCullagh and Tibshirani (1990), who modify the profile likelihood to restore the Bartlett identities up to second order, so that it is information-unbiased. To do so, they generate bootstrap samples at points $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}}$ and evaluate a location-scale transformation based on the simulations. While their approach is invariant under interest-preserving transformations, such transformations may not exist if $\boldsymbol{\psi}$ is multidimensional. If the profile likelihood is not an explicit function of ψ, it is necessary to linearize $\ell_{\mathrm{p}}(\psi)$ in terms of $\ell(\psi, \boldsymbol{\lambda})$ to get a first-order approximation. We shall not pursue these approaches further.

Another approach worth mentioning is bootstrap-based estimation of the distribution of the likelihood ratio statistic $w(\psi) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\psi)\}$ (cf. Lee and Young, 2005). The procedure, described in Algorithm 1.1, is computationally intensive, as it requires running a parametric bootstrap for each value of ψ over a grid $\{\psi_i\}_{i=1}^N$. The significance function obtained from

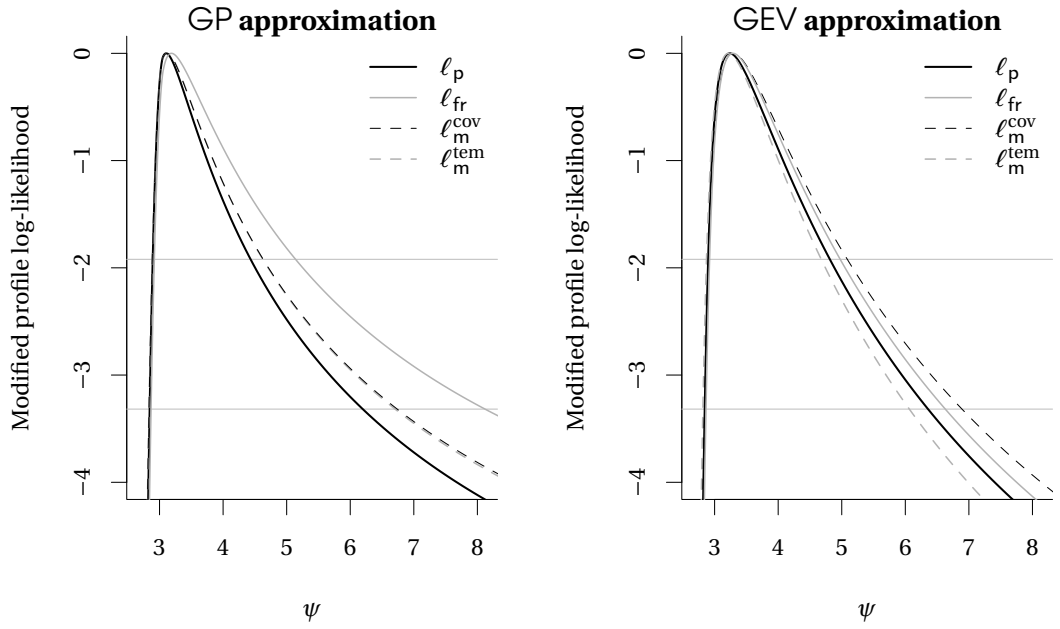

Figure 1.12 – *Profile log-likelihood ℓ_{p} (heavy black), profile log-likelihood of thetangent exponential model approximation ℓ_{fr} (solid grey) and Severini's modified profile log-likelihood with a tangent exponential model penalty $\ell_{\mathrm{m}}^{\mathrm{tem}}$ (dashed grey) and an empirical covariance approximation $\ell_{\mathrm{m}}^{\mathrm{cov}}$ (dashed black) for maxima of $N = 2000$ variates using peaks-over-threshold (left) and block maximum (right) methods. Horizontal lines show the cut-off for 95% and 99% confidence intervals based on the quantiles of the χ_1^2 distribution.*

Algorithm 1.1 Bootstrap estimation of the distribution of the likelihood ratio statistic w

Require: n sample $\boldsymbol{x}$ from $F(\boldsymbol{x};\boldsymbol{\theta})$.
1: compute the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$;
2: **for all** $i = 1,\ldots,N$ **do**
3: estimate $\widehat{\boldsymbol{\theta}}_{\psi_i}$ and $w(\psi_i)$ based on $\boldsymbol{x}$;
4: **for all** $b = 1,\ldots,B$ **do**
5: simulate n observations $\boldsymbol{x}^{(b,i)}$ from $F(\cdot;\widehat{\boldsymbol{\theta}}_{\psi_i})$;
6: find the constrained MLE $\widehat{\boldsymbol{\theta}}_{\psi_i}^{(b,i)}$ based on $\boldsymbol{x}^{(b,i)}$;
7: find the MLE $\widehat{\boldsymbol{\theta}}^{(b,i)}$ based on $\boldsymbol{x}^{(b,i)}$;
8: **return** $w^{(b,i)} = 2\{\ell(\widehat{\boldsymbol{\theta}}^{(b,i)}) - \ell(\widehat{\boldsymbol{\theta}}_{\psi_i}^{(b)})\}$;
9: **return** the bootstrap P-value at ψ_i, $p_i = \#\{b : w^{(b,i)} > w(\psi_i)\}/B$.

the procedure is not smooth because of Monte Carlo variability, but the approximation to the distribution is valid up to second order. One can obtain a $(1-\alpha)$ confidence interval by computing $\min_i\{p_i < \alpha/2, \psi_i > \widehat{\psi}\}$ and $\min_i\{p_i < \alpha/2, \psi_i < \widehat{\psi}\}$, where p_i is the boostrap P-value at ψ_i.

1.4 Simulation studies

This section explores two small-sample corrections for extremes. The first Monte Carlo simulation looks at the estimation bias of maximum likelihood estimates when fitting generalized Pareto and generalized extreme value distributions to data simulated from these models. The second study looks at small-sample corrections for confidence intervals, including modifications of the profile log-likelihood, the tangent exponential model approximation and Severini's corrections.

It is important to recall that extreme value models are valid only asymptotically and the use of limiting models in finite samples leads to biased estimates.

1.4.1 Bias corrections for extreme value distributions

Sample sizes for extremes, whether block maxima or threshold exceedances, tend to be small due to the need to compromise between closeness of approximation (and asymptotic bias) and estimation variability. The estimation bias is often neglected in asymptotic studies (Dombry and Ferreira, 2019) because of the consistency of maximum likelihood estimators, but it is important when dealing with small datasets.

The derivation of high order cumulants appearing in the first-order bias term $\boldsymbol{b}(\boldsymbol{\theta})$ is tedious, particularly for the generalized extreme value distribution. They can be obtained by automatic integration routines after a change of variables so that the supports of the integrals do not depend on the model parameters. We performed most of the symbolic calculations in Sage and resorted to Mathematica whenever the former could not provide a closed-form expression. These quantities may be of interest in their own right and are provided in Appendix A.

The first-order term in the expansion of the score equation about the maximum likelihood estimator yields the dominant bias term, which is $\mathrm{O}(n^{-1})$. The third-order (mixed) cumulants are defined only when $-1/3 < \xi < 1$, yet even values close to $\xi \approx -1/5$ give unbounded bias terms. For the generalized extreme value distribution, all cumulants are defined by continuity for $\xi \to 0$, but numerical instability at the origin implies that care must be taken when coding the cumulants, and the implementation in the mev R package uses linear interpolation between the limiting case $\xi = 0$ and small values of ξ to deal with the discontinuities when r and q are zero, as advocated in Brazzale et al. (2007, p. 149).

Both generalized extreme value distribution and generalized Pareto variates are simulated by applying the quantile transform to simulated uniform variates. Maximum likelihood estimates are obtained using a constrained optimizer for the generalized extreme value distribution (augmented Lagrange with boundary constraints) to ensure that $\widehat{\xi} > -1$ and that $\widehat{\sigma} + \widehat{\xi}(x_i - \widehat{\mu}) > 0$ $(i = 1, \ldots, n)$, so that the likelihood is finite and non-zero. The generalized Pareto distribution is fitted using the algorithm of Grimshaw (1993). Since the bias correction for $\widehat{\xi} \leq -1/3$ doesn't exist, calculation of $\boldsymbol{b}(\widehat{\boldsymbol{\theta}})$ is impossible when $\widehat{\xi} < -1/3$.

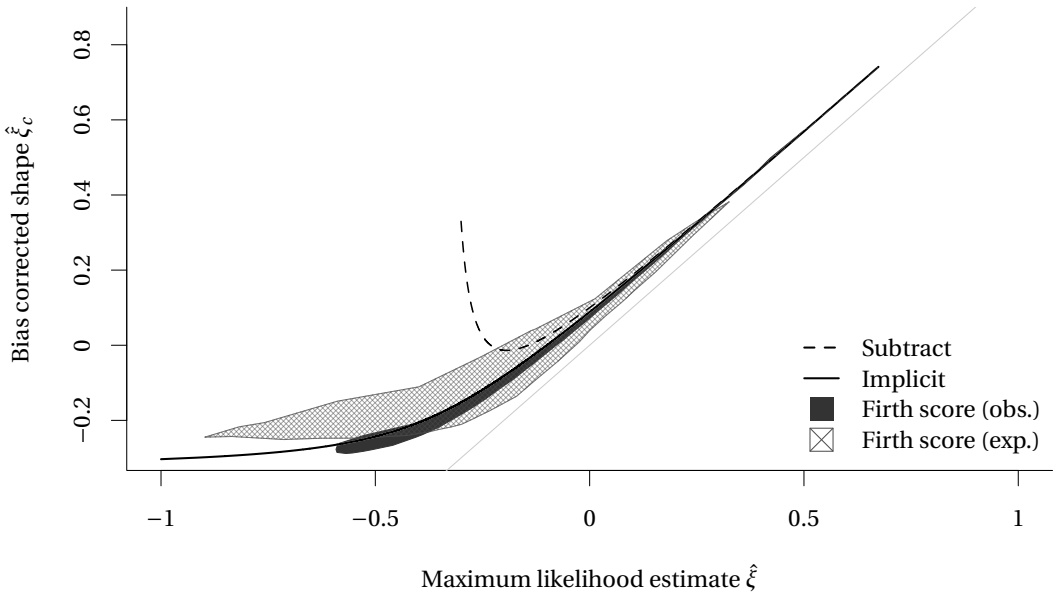

Figure 1.13 – *Bias correction for the shape parameter ξ of the generalized Pareto distribution* $\mathsf{GP}(\sigma = 1, \xi = -0.2)$*, with $n = 30$, based on $B = 13000$ replications. The grey line shows the line $y = x$; as the bias-corrected shapes are necessarily larger than $-1/3$. The shaded and dashed polygons for Firth's score corrections are the α-convex hull with parameter $\alpha = 0.5$, which show the region in which bias-corrected estimates fall relative to the maximum likelihood estimates. Both bias subtraction and implicit corrections are monotonic functions of ξ.*

For each replication, we calculate the maximum likelihood estimates and the bias-corrected estimates. The resulting point estimates for more than 10 000 replications are summarized in Tables 1.1 and 1.2. While the mean squared error of the estimate would normally be reported, its calculation is upset by the fact that bias-corrected estimates, wherever computable, yield estimates $\widehat{\xi} > -1/3$. Moreover, the estimate need not be available when $\widehat{\xi}$ is strongly negative or for the bias subtraction with the correction $b(\widehat{\boldsymbol{\theta}})$ when $\widehat{\xi} < -1/3$. This happens most often for the generalized Pareto distribution, for which the negative bias of $\widehat{\xi}$ can be substantial. For the generalized extreme value distribution, the so-called bias subtraction $\boldsymbol{b}(\widehat{\boldsymbol{\theta}})$ cannot be calculated in 34% of the scenarios when $\xi_0 = -0.2, n = 20$, decreasing to 7% when $n = 80$. The second method that fails most often is Firth's observed score, which cannot be calculated in 3% of the cases for the same setting $\xi_0 = -0.2, n = 20$. The distribution of the estimator $\widehat{\xi}_c$ is thus highly skewed to the right and truncated at $-1/3$. As n increases, the bias correction goes to zero at rate n^{-1} and the estimates all approach the maximum likelihood estimates. The variability of the estimate obtained using Firth's score is smaller if we use the observed information rather than the Fisher information, as predicted. Figure 1.13 illustrates the correction, with the implicit bias correction always yielding estimates. The α-convex hull represents the bias corrected point estimates for a given value of $\widehat{\xi}$, since the correction is random and depend on the data through the score and the observed information.

θ_i	Correction | ξ_0	−0.2	−0.1	0	0.1	0.2	0.3	0.4	0.5
μ	Theoretical (1st order)	5.0	3.1	2.2	1.7	1.3	0.9	0.6	0.3
	none	2.3	1.8	1.3	0.8	0.3	−0.2	−0.7	−1.1
	Subtract	−3.2	−2.8	−1.7	−1.3	−1.1	−1.2	−1.2	−1.2
	Implicit	−3.2	−2.0	−1.4	−1.1	−1.1	−1.2	−1.2	−1.1
	Firth score (obs.)	−3.0	−2.0	−1.4	−1.1	−1.1	−1.2	−1.3	−1.4
	Firth score (exp.)	−5.1	−3.1	−1.9	−1.3	−0.7	−0.4	−0.0	0.4
σ	Theoretical (1st order)	−1.6	−2.9	−3.5	−3.7	−3.9	−4.0	−4.0	−3.9
	none	−3.0	−3.4	−3.6	−4.0	−4.4	−4.8	−5.2	−5.4
	Subtract	−3.0	−1.8	−1.0	−0.8	−0.9	−1.2	−1.5	−1.6
	Implicit	−1.6	−0.9	−0.6	−0.5	−0.7	−0.9	−1.2	−1.3
	Firth score (obs.)	−2.6	−1.4	−1.0	−1.0	−1.2	−1.6	−2.0	−2.4
	Firth score (exp.)	−1.9	−1.1	−0.6	−0.2	−0.0	0.3	0.5	0.8
ξ	Theoretical (1st order)	−7.2	−3.5	−1.8	−0.6	0.3	1.1	1.9	2.6
	none	−2.8	−2.0	−1.5	−0.8	−0.3	0.3	0.9	1.4
	Subtract	8.4	3.2	0.9	−0.0	−0.4	−0.8	−1.1	−1.6
	Implicit	2.8	1.2	0.3	−0.1	−0.5	−0.8	−1.0	−1.6
	Firth score (obs.)	1.8	0.7	0.2	−0.0	−0.2	−0.3	−0.4	−0.5
	Firth score (exp.)	6.7	2.8	0.6	−0.3	−1.4	−2.4	−3.3	−4.3

Table 1.1 – *Median absolute bias (*×100*) for the generalized extreme value distribution,* $n = 30$.

The overall conclusion of the simulation study, reported in Tables 1.1 and 1.2, is that bias correction is useful only for the generalized Pareto distribution, for which the first-order bias of the shape is consistently negative (eigth line of Table 1.2) and of similar magnitude regardless of the value of ξ_0. The Monte Carlo average bias of the maximum likelihood estimator, $\boldsymbol{b} = \boldsymbol{\theta}_0 - B^{-1}\sum_{b=1}^{B}\widehat{\boldsymbol{\theta}}_b$, is much smaller in magnitude than the first-order correction would imply. The implicit bias correction (eq. (1.8)) performs best and maintains an ordering relative to the maximum likelihood estimates. Firth's bias correction, particularly the observed information, provide good bias reduction on average, but the variance for a given value of $\widehat{\xi}$ and the potential lack of convergence for the root-finding algorithm makes implicit bias correction a more interesting avenue.

1.4.2 Higher-order asymptotic methods for confidence intervals

Pires et al. (2018) performed a simulation study to assess the performance of higher order asymptotic methods for the shape parameter of the generalized Pareto distribution. Their study concluded that the profile likelihood was oversized, meaning that the coverage probability was smaller than the nominal coverage probability; the authors also noted that the point estimates provided by modified profile likelihood were less biased. Although the sign of the shape parameter may help to understand the behaviour of extrapolated extremes, ξ is not usually of interest by itself. Use of higher-order methods should in principle provide more accurate coverage than first order methods. Since the use of profile likelihood-based confidence intervals is standard for extreme value analysis, it is interesting to see how they fare relative to penalized methods.

θ_i	Correction | ξ_0	−0.2	−0.1	0	0.1	0.2	0.3	0.4	0.5
σ	Theoretical (1st order)	18.0	12.1	10.0	9.1	8.7	8.5	8.6	8.7
	none	8.8	7.9	7.7	6.9	6.3	5.8	5.5	5.2
	Subtract	−7.6	−10.1	−8.4	−6.6	−5.3	−4.5	−3.9	−3.0
	Implicit	−6.3	−4.4	−3.1	−2.8	−2.7	−2.8	−2.8	−2.5
	Firth score (obs.)	−6.2	−4.3	−3.2	−3.0	−3.0	−3.2	−3.7	−4.1
	Firth score (exp.)	−12.8	−9.7	−7.7	−6.2	−5.4	−5.0	−4.8	−4.7
ξ	Theoretical (1st order)	−18.7	−12.4	−10.0	−8.7	−8.0	−7.5	−7.2	−7.0
	none	−10.3	−9.6	−9.2	−8.6	−8.5	−8.2	−8.2	−7.9
	Subtract	18.7	10.7	5.3	2.3	0.6	−0.2	−1.0	−2.0
	Implicit	4.6	2.2	0.7	0.1	−0.5	−0.8	−1.2	−1.7
	Firth score (obs.)	2.2	0.5	−0.6	−0.6	−1.0	−1.1	−1.2	−1.1
	Firth score (exp.)	12.5	8.0	4.8	2.9	1.7	1.0	0.4	0.0

Table 1.2 – *Median absolute bias (×100) for the generalized Pareto distribution, $n = 30$.*

We conduct a Monte Carlo simulation to study the coverage, bias and variability of the higher order asymptotic methods for high quantiles. Specifically, we generate samples of size $n = 1800$ and target the quantile corresponding to an exceedance every 9000 observations. This setting mimics 100-year return levels estimated from observations over a season of roughly 90 days, collected over a period of 20 years. The alternative is to look directly at the distribution of the maximum of N observations, derived using max-stability arguments. We selected the median and the mean as representative point estimates; the latter is larger than the former whenever $\xi > 0$. The three parameters are therefore ordered.

For fixed sample size n, we contrast two methods. We fit the generalized extreme value distribution to maxima of blocks of sizes $m = 30, 45, 90$. We also use peaks-over-threshold modelling, discarding all but the largest $m \in \{20, 40, 60\}$ observations and fitting a generalized Pareto distribution to the exceedances.

Six parametric families of distributions (Burr, Weibull, generalized gamma, Gaussian and lognormal, Student) were selected as data generating processes to exemplify the use of extreme value approximations when the data are only in the max-domain of attraction of the latter. The distributions mimic standard choices appearing in the literature. The Burr and Student distributions are heavy-tailed, while the others are in the Gumbel max-domain of attraction. Parameter values for the generalized Gamma distribution were selected to reflect values found in hydrology in estimating of the shape index for rainfall (Papalexiou and Koutsoyiannis, 2013).

Five methods were used to derive estimates and confidence intervals. The first is the normal confidence interval derived from the Wald statistic, which is computed on the log scale and back-transformed, i.e., $\exp\{\log(\widehat{\psi}) \pm \mathfrak{z}_{1-\alpha/2}\, \mathrm{se}(\widehat{\psi})/\widehat{\psi}\}$. The second is the signed profile likelihood root r of eq. (1.13), while the third is the r^* approximation presented in Section 1.3.3. The last two are the modified profile likelihoods based on the tangent exponential model (eq. 1.18) or the modified empirical covariance of Severini (eq. 1.19). The last three should provide bias-corrected point estimates as well as more accurate confidence intervals. For each scenario, we

calculated the relative bias, the coverage and the average width of the confidence intervals. The profile log-likelihood estimates are obtained using constrained optimization methods at selected values of ψ, using dedicated algorithms such as sequential quadratic programming to obtain profile log-likelihood values for each of the values of ψ on a grid. Every model is parametrized in terms of the parameter of interest, along with tangent space derivatives, canonical parameters, score and information matrices. Once the profile log-likelihood values have been obtained, we compute the signed likelihood root $r = \text{sign}(\widehat{\psi} - \psi)\{2\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\psi)\}^{1/2}$. The correction term q for the tangent exponential model is calculated as per eq. (1.17) and used to get a value of $r^* = r + r^{-1}\log(q/r)$. The maximum of the tangent exponential model is found as the root of $r^*(\psi_0)r(\psi_0)$ by a line search starting from the maximum likelihood estimate, for which $r(\psi_0) = 0$. For the penalized likelihood of Severini, we compute penalty terms that are added to the profile log-likelihood values and the optimum is obtained as the maximum of the modified profile.

The values of r^* obtained from this routine can be numerically unstable, often due to numerical instability near $r = 0$ or to failure of the optimization routines. To palliate this, we remove values of r^* that are not in line with their neighbours and interpolate the rest using splines. Specifically, we fit constrained quantile regression B splines to $r^* - r$ with a Bayesian information criterion to select the number of knots, using r as regressor. The default quantile selected is the median. Since the values most susceptible to cause numerical overflow are located near $r = 0$, we down-weight those observations using as weights in the regression $q(r^2)$, where q is the quantile function of a χ^2_1 variable. The departure of the values of $r^* - r$ from the fitted curve is calculated, and the estimates of the squared standardized estimates that exceed a 70% confidence interval from the χ^2_1 are excluded. A second spline regression is fitted to the remaining values and the values of r^* are interpolated using the median. This ad-hoc scheme usually does not affect estimates by more than an order of 10^{-6}, but is effective at removing the outliers and was found to work well in practice.

For the confidence intervals, we fit constrained B-splines with response ψ and values r (respectively r^*) and predict the quantile corresponding to $r = \pm \mathfrak{q}^{1/2}$, where $\mathfrak{q}$ is the quantile of the χ^2_1 distribution with one degree of freedom at level $1 - \alpha$. For the penalized methods, we compute the equivalent of the signed likelihood root statistic and proceed analogously.

The following paragraphs summarize the findings of the simulation study, which was conducted using the routines from the mev package using the infrastructure provided by the R package simsalapar (Hofert and Mächler, 2016a).

Table 1.3 gives the penultimate shape parameters for the various levels under consideration for both block maximum and peaks-over-threshold methods. The limiting value is given alongside the penultimate value for maxima of $m = 9000$ observations. The difference between the estimated shape and ξ_{9000} is less than 0.1 in magnitude.

Distribution |	$q = 0.967$	$q = 0.978$	$q = 0.989$	$m = 30$	$m = 45$	$m = 90$	$m = 9000$	ξ_∞
Burr	0.01	0.03	0.05	0.03	0.04	0.06	0.10	0.1
Weibull	0.15	0.13	0.11	0.16	0.14	0.12	0.05	0
Gen. gamma	0.15	0.14	0.12	0.17	0.15	0.13	0.07	0
Gaussian	−0.18	−0.16	−0.13	−0.16	−0.14	−0.12	−0.06	0
Lognormal	0.27	0.26	0.25	0.28	0.27	0.25	0.19	0
Student	−0.05	−0.03	0.00	−0.03	−0.02	0.01	0.07	0.1

Table 1.3 – *Penultimate shape parameters for six distributions, based on threshold exceedances with threshold at q percentile (first three columns), block maxima with maximum of m observations (fourth to sixth column). The penultimate shape parameter for the maximum of* 9000 *observations is the reference, still far from the tail index ξ_∞ in the last column.*

Block maxima

The general coverage properties are comparable for all three parameters under consideration. The mean of the N-observation maximum is usually stochastically larger than the quantile or N-observation median, and has the widest confidence intervals and the smallest coverage; these are due to the extrapolation error. The penultimate effects are not really visible. The coverages of all but the Wald method are within 2% of the nominal level, so nominal errors are reported instead; see Table 1.4. The transformed (Wald) normal confidence intervals, $\exp\{\log(\widehat{\psi}) \pm \mathfrak{z}_{1-\alpha/2}\, \mathrm{se}(\widehat{\psi})/\widehat{\psi}\}$, severely undercover, especially when the sample size gets smaller; with 20 observations (Table 1.5), the 99% confidence intervals capture the true value roughly 82% of the time, but the 5% nominal error rate for the lower interval is 0%, indicating that the interval is too large on the left and strikingly too short for the upper limit. Despite the transformation, the Wald confidence intervals fail to capture the positive skewness of the distribution of the estimated values $z_{1/N}, \mathfrak{q}_{1/2}$ (not shown) and $\mathfrak{e}_N$, defined in Proposition 1.7. If the data are generated from the generalized extreme value distribution (F_7–F_9), the empirical error rates are closer to nominal for the TEM but no method is universally better.

Further results are reported in Appendix A.3. When $k = 60$ (Table A.7), the tangent exponential model-based intervals are closest to the nominal level, but the higher-order methods, and in particular intervals based on r^*, overcover slightly if the sample size is smaller and the blocks are larger, e.g., when $m = 90$ and $k = 20$ (Table 1.5). The average widths of the confidence intervals are comparable between methods for the block maximum method with $m = 30, k = 60$ (Table A.9). For this setting, the intervals based on $\ell_{\mathsf{m}}^{\mathsf{tem}}$ are the shortest among the five methods implemented. As the sample size decreases, the confidence intervals get wider and their coverage increases beyond the nominal level. The largest standard error of the average relative width of the confidence intervals is 0.2; it was estimated using a nonparametric bootstrap with $B = 1000$ replications. The point estimates are all comparable, but the average unbiased estimator of the TEM is roughly 2% more positively biased than the MLE (Table A.2). All the methods are comparable, so it is reassuring that the higher-order methods and the intervals derived using the profile likelihood largely agree. For block maxima, all the profile or

higher-order methods seem impervious to the effects of extrapolation and their coverage is excellent.

Peaks-over-threshold methods

The results for peaks-over-thresholds are more contrasted, though the performance of Wald-type confidence intervals remains calamitous even when these are computed on the log-scale to better capture the skewness; the empirical error rate for the upper 5% of the untransformed Wald confidence intervals is close to 30% in all scenarios and between 20% and 30% for the transformed intervals. With $k = 20$ observations (Table 1.7), most higher-order methods overcover even when the model is not misspecified and the TEM breaks down and suffers from overcoverage, while the profile likelihood appears to be more robust. The TEM overcoverage is mostly in the upper tail, whereas Severini's corrections display higher empirical error rate in the lower tail. This break-down of the TEM could be attributed to penultimate effects and small sample bias; it vanishes as the sample size grows and the TEM behaves as expected when $k = 60$ (Table 1.6). Whereas maximum likelihood estimators are negatively biased for any sample size $k \leq 60$, the other point estimators correct for the bias. For $k = 20$, the TEM point estimates are positively biased, but have the lowest bias of all the methods when $k > 40$. Both of Severini's modified profiles have lower bias than the MLE. Higher-order methods give wider confidence intervals, which in turn lead to higher coverage: for example, compared to the confidence intervals based on the profile likelihood, the TEM intervals are between 1.75 and 2 times larger when $k = 20$ and about 1.25 times larger than when $k = 60$. Confidence intervals based on the profile likelihood typically have good coverage, but their empirical error rates in the upper tail are typically more than double the nominal values.

1.4.3 Practical guidelines

Poor finite sample properties of maximum likelihood estimators have been documented before (Hosking and Wallis, 1987, Table 5). However, extension of moment-based estimators to generalized additive modelling of the covariates and handling of censoring and truncation is not as straightforward as in a likelihood-based setting. For the generalized Pareto distribution, maximimization of the profile likelihood nearly guarantees a solution if the sample is large enough to yield a maximum. While there is a scale/shape trade-off, this typically translate into large differences only in the tails of the distribution. An alternative to bias correction based on the Cox–Snell bias is use of the bootstrap, but it is computationally intensive. Rather, one could consider computing bias curves such as those given in Figure 1.8 for different values of ξ and for different n, even if the first-order bias term is inversely proportional to the sample size. By interpolating these curve slices, one could compute an implicit correction. The advantage is two-fold; the correction is deterministic, so the variance is not affected, and we need not run a new simulation for every point estimator. By virtue of the bootstrap, one can also correct any functional directly without explicit derivations for all cases of interest. Of course, the penultimate approximation is also likely to result in discrepancies because

of model misspecification (Dombry and Ferreira, 2019), so how well bias correction works depends on the underlying data-generating process.

Profile likelihood-based confidence intervals have good coverage properties overall and we can recommend them for estimating quantiles of the distribution of N-observation maxima; the discrepancy between the nominal error rates in the lower and upper tails seems to be due to the bias of the quantile estimates themselves. Wald-based confidence intervals should never be used; their performance is dismal and their coverage is constantly too low, even after transformation. Tangent exponential model-based confidence intervals have very good coverage and smaller bias when the sample size is larger than 50, say, for the approximation to be reliable; while no higher-order method is always better, the TEM correction usually improves over the penalized profile likelihood of Severini.

	Parameter	Quantile						N-obs. mean					
F	Method | Error rate	0.5	2.5	5	5	2.5	0.5	0.5	2.5	5	5	2.5	0.5
F_1	Wald	0.0	0.0	0.0	19.5	15.5	10.5	0.0	0.0	0.0	20.5	17.0	12.0
	profile	0.5	2.0	4.0	7.0	3.0	0.5	0.5	2.0	4.0	7.5	3.5	0.5
	TEM	0.5	2.5	5.0	5.0	2.0	0.5	0.5	2.5	4.5	5.5	2.5	0.5
	Severini (TEM)	0.5	1.5	3.5	8.0	3.5	0.5	0.5	1.5	3.5	8.5	4.0	0.5
	Severini (cov.)	0.5	2.0	4.5	5.0	2.0	0.5	0.5	2.5	4.5	5.5	2.0	0.5
F_2	Wald	0.0	0.0	1.5	13.0	10.5	7.0	0.0	0.0	1.0	13.0	10.5	7.0
	profile	0.5	2.5	5.0	4.5	2.5	0.5	0.5	3.0	6.0	4.0	2.0	0.5
	TEM	0.5	3.0	6.0	3.5	1.5	0.5	1.0	3.5	7.0	3.0	1.5	0.5
	Severini (TEM)	0.5	2.0	4.5	5.0	2.5	0.5	0.5	2.5	5.0	4.5	2.5	0.5
	Severini (cov.)	0.5	2.5	5.0	4.5	2.5	0.5	0.5	3.0	6.0	4.0	2.0	0.5
F_3	Wald	0.0	0.0	1.0	15.0	12.0	8.0	0.0	0.0	0.5	14.5	12.0	8.0
	profile	0.0	1.5	3.5	6.0	3.5	0.5	0.5	2.0	4.0	5.5	3.0	0.5
	TEM	0.0	1.5	4.0	5.0	2.5	0.5	0.5	2.0	5.0	4.5	2.0	0.5
	Severini (TEM)	0.0	1.0	3.0	6.5	3.5	1.0	0.0	1.5	3.5	6.0	3.0	0.5
	Severini (cov.)	0.0	1.5	3.0	6.0	3.0	0.5	0.5	2.0	4.0	5.5	3.0	0.5
F_4	Wald	0.0	0.0	0.0	26.0	22.5	17.0	0.0	0.0	0.0	27.5	24.0	18.0
	profile	0.5	1.5	3.5	10.0	5.5	1.0	0.5	1.5	3.0	10.5	5.5	1.0
	TEM	0.5	2.5	4.5	6.5	3.0	0.5	0.5	2.5	4.5	7.0	3.5	0.5
	Severini (TEM)	0.5	1.5	2.5	10.5	6.0	1.5	0.0	1.0	2.5	11.0	6.0	1.5
	Severini (cov.)	0.5	1.5	3.0	7.5	4.0	0.5	0.5	1.5	3.0	8.0	4.0	0.5
F_5	Wald	0.0	0.5	2.0	11.5	8.5	5.5	0.0	0.0	1.0	11.0	8.5	5.5
	profile	0.5	2.5	5.5	4.5	2.5	0.5	1.0	3.0	7.0	4.0	2.0	0.5
	TEM	0.5	2.5	5.5	3.5	2.0	0.5	1.0	3.5	7.0	3.5	1.5	0.5
	Severini (TEM)	0.5	2.0	4.5	5.0	2.5	0.5	0.5	2.5	5.5	4.5	2.5	0.5
	Severini (cov.)	0.5	2.5	5.0	4.5	2.5	0.5	0.5	3.0	6.5	4.0	2.0	0.5
F_6	Wald	0.0	0.0	0.5	22.5	19.0	13.5	0.0	0.0	0.0	24.5	20.5	15.0
	profile	0.5	2.0	4.0	10.0	5.5	1.5	0.5	2.0	3.5	11.0	6.0	1.5
	TEM	0.5	2.5	5.0	7.5	4.0	1.0	0.5	2.5	4.5	8.0	4.5	1.0
	Severini (TEM)	0.5	1.5	3.0	11.0	6.5	1.5	0.5	1.5	3.0	11.5	7.0	2.0
	Severini (cov.)	0.5	2.0	3.5	9.5	5.0	1.0	0.5	2.0	3.5	10.0	5.5	1.5
F_7	Wald	0.0	0.0	1.0	17.0	13.5	9.5	0.0	0.0	0.5	17.5	14.5	10.0
	profile	0.5	2.0	4.5	7.0	3.5	1.0	0.5	2.5	4.5	7.0	3.5	1.0
	TEM	0.5	2.5	5.5	5.5	2.5	0.5	0.5	3.0	5.5	5.5	3.0	0.5
	Severini (TEM)	0.5	2.0	4.0	8.0	4.0	1.0	0.5	2.0	4.0	8.0	4.5	1.0
	Severini (cov.)	0.5	2.0	4.0	7.0	3.5	1.0	0.5	2.5	4.5	7.0	3.5	1.0
F_8	Wald	0.0	0.0	0.5	19.0	15.5	11.0	0.0	0.0	0.5	19.5	16.5	12.0
	profile	0.5	2.0	4.5	7.0	3.5	1.0	0.5	2.0	4.5	7.0	3.5	1.0
	TEM	0.5	3.0	5.5	5.0	2.5	0.5	0.5	3.0	5.5	5.0	2.5	0.5
	Severini (TEM)	0.5	1.5	3.5	7.5	4.0	1.0	0.5	1.5	3.5	7.5	4.0	1.0
	Severini (cov.)	0.5	2.0	4.0	6.5	3.5	0.5	0.5	2.0	4.5	6.5	3.0	0.5
F_9	Wald	0.0	0.0	0.0	22.5	19.0	14.5	0.0	0.0	0.0	23.0	19.5	15.0
	profile	0.5	1.5	3.5	8.0	4.0	1.0	0.5	1.5	3.5	7.5	4.0	0.5
	TEM	0.5	2.5	5.0	5.0	2.5	0.5	0.5	2.5	5.0	5.0	2.0	0.0
	Severini (TEM)	0.0	1.0	2.5	8.5	4.5	1.0	0.0	1.0	2.5	8.0	4.0	0.5
	Severini (cov.)	0.0	1.5	3.0	6.5	3.0	0.5	0.5	1.5	3.5	6.5	3.0	0.5

Table 1.4 – *One-sided nominal error rate (in %) for lower (first to third columns) and upper (fourth to sixth columns) confidence intervals, block maximum method with* $m = 45, k = 40$. *The distributions (from top to bottom) are Burr* (F_1)*, Weibull* (F_2)*, generalized gamma* (F_3)*, normal* (F_4)*, lognormal* (F_5)*, Student t* (F_6)*,* $\mathsf{GEV}(\xi = 0.1)$ (F_7)*, Gumbel* (F_8) *and* $\mathsf{GEV}(\xi = -0.1)$ (F_9).

	Parameter	Quantile						N-obs. mean					
F	Method | Error rate	0.5	2.5	5	5	2.5	0.5	0.5	2.5	5	5	2.5	0.5
F_1	Wald	0.0	0.0	0.0	23.0	20.0	15.5	0.0	0.0	0.0	24.0	21.0	17.0
	profile	0.5	2.0	4.0	6.5	3.0	0.5	0.5	2.0	4.0	6.5	3.0	0.5
	TEM	0.5	2.5	5.0	3.5	1.5	0.0	0.5	2.5	5.0	3.5	1.5	0.0
	Severini (TEM)	0.0	1.0	3.0	7.5	3.5	0.5	0.0	1.0	3.0	7.5	3.5	0.5
	Severini (cov.)	0.5	2.0	4.5	3.0	1.0	0.0	0.5	3.0	6.0	3.0	1.0	0.0
F_2	Wald	0.0	0.0	0.5	20.0	17.5	13.5	0.0	0.0	0.0	20.0	17.5	13.5
	profile	0.5	2.0	4.0	6.0	3.0	0.5	0.5	2.5	4.5	5.5	2.5	0.5
	TEM	0.5	2.5	5.0	3.5	1.5	0.0	0.5	2.5	5.5	3.0	1.5	0.0
	Severini (TEM)	0.0	1.0	2.5	7.0	3.5	0.5	0.0	1.5	3.0	6.5	3.0	0.5
	Severini (cov.)	0.0	1.5	3.5	5.0	2.5	0.5	0.5	2.0	5.0	4.5	2.0	0.5
F_3	Wald	0.0	0.0	0.5	21.5	19.0	15.0	0.0	0.0	0.0	21.5	19.0	15.5
	profile	0.0	1.0	3.0	7.5	4.0	1.0	0.0	1.0	3.5	7.0	3.5	0.5
	TEM	0.0	1.5	3.5	5.0	2.5	0.0	0.5	1.5	4.0	4.5	2.0	0.0
	Severini (TEM)	0.0	0.5	1.5	9.0	5.0	1.0	0.0	0.5	2.0	8.5	4.5	1.0
	Severini (cov.)	0.0	1.0	2.5	6.5	3.5	0.5	0.0	1.5	3.5	6.0	3.0	0.5
F_4	Wald	0.0	0.0	0.0	28.5	25.5	20.5	0.0	0.0	0.0	29.0	26.0	21.5
	profile	0.5	1.5	3.5	7.0	3.5	0.5	0.5	1.5	3.5	6.5	3.0	0.5
	TEM	0.5	2.5	5.0	3.0	1.0	0.0	0.5	2.5	5.0	3.0	1.0	0.0
	Severini (TEM)	0.0	1.0	2.0	8.0	3.5	0.5	0.0	1.0	2.0	8.0	3.5	0.5
	Severini (cov.)	0.0	1.5	3.5	3.5	1.5	0.0	0.5	1.5	4.0	3.0	1.0	0.0
F_5	Wald	0.0	0.0	1.0	18.0	15.0	11.5	0.0	0.0	0.0	18.5	15.5	12.0
	profile	0.5	2.5	4.5	6.0	3.0	0.5	0.5	2.5	5.5	5.5	3.0	0.5
	TEM	0.5	2.5	5.0	4.5	2.0	0.5	1.0	3.0	6.0	4.0	2.0	0.5
	Severini (TEM)	0.0	1.5	3.0	7.0	4.0	1.0	0.5	1.5	4.0	7.0	3.5	1.0
	Severini (cov.)	0.5	2.0	4.0	5.5	3.0	0.5	0.5	3.0	6.0	5.0	2.5	0.5
F_6	Wald	0.0	0.0	0.0	25.5	22.0	17.5	0.0	0.0	0.0	27.0	23.5	19.0
	profile	0.5	2.0	4.0	8.5	4.5	1.0	0.5	2.0	4.0	8.0	4.0	1.0
	TEM	0.5	3.0	5.5	5.0	2.5	0.5	0.5	3.0	5.5	5.0	2.0	0.0
	Severini (TEM)	0.0	1.5	3.0	9.5	5.0	1.0	0.0	1.5	3.0	9.5	5.0	1.0
	Severini (cov.)	0.5	2.0	4.0	6.0	3.0	0.5	0.5	2.5	4.5	5.5	2.5	0.5
F_7	Wald	0.0	0.0	0.5	22.0	19.0	15.0	0.0	0.0	0.0	22.5	20.0	15.5
	profile	0.5	2.0	4.5	7.5	4.0	0.5	0.5	2.0	4.5	7.0	3.5	0.5
	TEM	0.5	2.5	5.5	5.0	2.0	0.5	0.5	2.5	5.5	4.5	2.0	0.5
	Severini (TEM)	0.0	1.5	3.0	8.5	4.5	1.0	0.0	1.5	3.0	8.5	4.5	1.0
	Severini (cov.)	0.5	1.5	4.0	6.5	3.0	0.5	0.5	2.5	5.0	6.0	3.0	0.5
F_8	Wald	0.0	0.0	0.0	23.5	20.5	16.5	0.0	0.0	0.0	24.0	21.0	17.0
	profile	0.5	2.0	4.0	7.0	3.5	0.5	0.5	2.0	4.0	6.5	3.0	0.5
	TEM	0.5	2.5	5.5	4.0	1.5	0.0	0.5	2.5	5.5	3.5	1.5	0.0
	Severini (TEM)	0.0	1.0	2.5	8.0	4.0	0.5	0.0	1.0	2.5	7.5	3.5	0.5
	Severini (cov.)	0.0	1.5	3.5	5.0	2.5	0.5	0.5	2.0	4.5	4.5	2.0	0.0
F_9	Wald	0.0	0.0	0.0	27.0	24.0	19.0	0.0	0.0	0.0	27.0	24.5	20.0
	profile	0.0	1.5	3.0	6.5	3.0	0.5	0.0	1.5	3.5	6.0	2.5	0.0
	TEM	0.5	2.5	5.0	2.5	1.0	0.0	0.5	2.5	5.0	2.5	0.5	0.0
	Severini (TEM)	0.0	0.5	1.5	7.5	3.5	0.5	0.0	0.5	2.0	6.5	3.0	0.0
	Severini (cov.)	0.0	1.0	3.0	4.0	1.5	0.0	0.0	1.5	3.5	3.0	1.0	0.0

Table 1.5 – *One-sided nominal error rate (in %) for lower (first to third columns) and upper (fourth to sixth columns) confidence intervals, block maximum method with $m = 90, k = 20$. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile						N-obs. mean					
F	Method | Error rate	0.5	2.5	5	5	2.5	0.5	0.5	2.5	5	5	2.5	0.5
F_1	Wald	0.0	0.0	0.0	29.0	26.0	20.5	0.0	0.0	0.0	31.5	28.5	23.5
	profile	0.5	1.5	3.0	5.5	1.5	0.0	0.5	1.5	3.5	5.0	1.0	0.0
	TEM	1.0	3.5	6.5	0.5	0.0	0.0	0.5	3.5	7.5	0.5	0.0	0.0
	Severini (TEM)	0.0	1.0	2.5	3.0	0.5	0.0	0.0	1.0	2.5	3.0	0.5	0.0
	Severini (cov.)	0.0	1.0	2.5	4.0	1.0	0.0	0.0	1.5	3.0	3.0	1.0	0.5
F_2	Wald	0.0	0.0	0.0	28.0	24.5	19.5	0.0	0.0	0.0	29.5	26.0	21.5
	profile	0.5	1.5	3.0	6.0	2.0	0.0	0.5	1.5	3.0	4.5	1.0	0.0
	TEM	0.5	3.0	6.0	0.5	0.0	0.0	0.5	3.5	7.0	0.0	0.0	0.0
	Severini (TEM)	0.0	1.0	2.0	3.0	1.0	0.0	0.0	1.0	2.5	2.5	0.5	0.0
	Severini (cov.)	0.0	1.0	2.0	3.5	1.0	0.0	0.0	1.0	3.0	2.5	1.0	0.5
F_3	Wald	0.0	0.0	0.0	28.0	25.5	21.0	0.0	0.0	0.0	29.5	26.5	22.5
	profile	0.0	0.5	2.0	7.0	2.5	0.0	0.0	1.0	2.5	5.5	2.0	0.0
	TEM	0.5	2.0	4.5	1.0	0.0	0.0	0.5	2.5	5.5	0.5	0.0	0.0
	Severini (TEM)	0.0	0.5	1.5	4.0	1.0	0.0	0.0	0.5	2.0	3.5	1.0	0.0
	Severini (cov.)	0.0	0.5	1.5	4.5	1.5	0.0	0.0	1.0	2.5	3.5	1.5	0.5
F_4	Wald	0.0	0.0	0.0	29.0	25.5	20.0	0.0	0.0	0.0	31.0	28.0	23.0
	profile	0.5	1.5	3.0	3.0	0.5	0.0	0.5	1.5	3.0	2.0	0.5	0.0
	TEM	1.0	3.5	7.0	0.0	0.0	0.0	1.0	3.5	6.5	0.0	0.0	0.0
	Severini (TEM)	0.0	1.0	2.0	1.5	0.0	0.0	0.0	1.0	2.0	1.0	0.0	0.0
	Severini (cov.)	0.0	0.5	2.0	2.5	1.0	0.0	0.0	1.0	2.0	2.0	1.0	0.5
F_5	Wald	0.0	0.0	0.0	28.5	25.0	20.5	0.0	0.0	0.0	29.5	26.5	22.0
	profile	0.0	1.5	3.0	8.0	3.5	0.0	0.5	2.0	4.0	7.0	3.0	0.0
	TEM	0.5	3.0	6.0	1.5	0.0	0.0	1.0	4.0	7.5	1.0	0.0	0.0
	Severini (TEM)	0.0	1.5	3.0	5.5	2.0	0.0	0.0	1.5	3.5	4.5	1.5	0.0
	Severini (cov.)	0.0	1.5	3.0	5.5	2.0	0.0	0.5	2.0	4.0	4.5	1.5	0.5
F_6	Wald	0.0	0.0	0.0	29.5	26.5	21.0	0.0	0.0	0.0	32.0	28.5	24.0
	profile	0.5	1.5	3.5	5.5	1.5	0.0	0.5	2.0	3.5	5.0	1.0	0.0
	TEM	1.0	4.0	7.5	0.5	0.0	0.0	1.0	4.0	7.5	0.5	0.0	0.0
	Severini (TEM)	0.0	1.5	3.0	3.0	0.5	0.0	0.0	1.0	3.0	3.0	0.5	0.0
	Severini (cov.)	0.0	1.0	2.5	4.0	1.5	0.0	0.5	1.5	3.0	3.5	1.5	0.5
F_7	Wald	0.0	0.0	0.0	29.0	25.5	20.5	0.0	0.0	0.0	30.5	27.5	22.5
	profile	0.5	1.5	3.0	6.0	2.0	0.0	0.5	1.5	3.5	5.5	1.5	0.0
	TEM	1.0	3.5	6.5	0.5	0.0	0.0	1.0	3.5	7.5	0.5	0.0	0.0
	Severini (TEM)	0.0	1.0	2.5	4.0	1.0	0.0	0.0	1.0	3.0	3.5	1.0	0.0
	Severini (cov.)	0.0	1.0	2.5	4.5	1.5	0.0	0.0	1.5	3.0	3.5	1.5	0.5
F_8	Wald	0.0	0.0	0.0	27.5	24.0	19.5	0.0	0.0	0.0	29.0	26.0	21.5
	profile	0.5	1.5	3.5	4.0	1.0	0.0	0.5	1.5	3.5	3.0	0.5	0.0
	TEM	0.5	3.5	6.5	0.0	0.0	0.0	0.5	3.5	7.0	0.0	0.0	0.0
	Severini (TEM)	0.0	1.0	2.0	2.0	0.5	0.0	0.0	1.0	2.5	2.0	0.5	0.0
	Severini (cov.)	0.0	0.5	2.0	3.0	1.0	0.0	0.0	1.0	3.0	2.5	1.0	0.5
F_9	Wald	0.0	0.0	0.0	28.0	24.0	18.5	0.0	0.0	0.0	29.5	26.5	21.0
	profile	0.5	1.0	2.5	2.5	0.5	0.0	0.5	1.5	2.5	1.5	0.0	0.0
	TEM	1.0	3.5	6.5	0.0	0.0	0.0	0.5	3.0	6.5	0.0	0.0	0.0
	Severini (TEM)	0.0	0.5	1.5	1.0	0.0	0.0	0.0	0.5	1.0	0.5	0.0	0.0
	Severini (cov.)	0.0	0.5	1.5	2.0	0.5	0.0	0.0	1.0	2.0	1.5	1.0	0.5

Table 1.6 – *One-sided nominal error rate (in %) for lower (first to third columns) and upper (fourth to sixth columns) confidence intervals, peaks-over-threshold method with $k = 20$. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile						N-obs. mean					
F	Method | Error rate	0.5	2.5	5	5	2.5	0.5	0.5	2.5	5	5	2.5	0.5
F_1	Wald	0.0	0.0	0.0	24.0	20.0	14.5	0.0	0.0	0.0	25.5	22.0	16.5
	profile	0.5	1.5	3.5	9.5	4.5	0.5	0.5	1.5	3.0	10.0	4.5	0.5
	TEM	0.5	3.0	5.5	4.5	1.5	0.0	0.5	2.5	5.5	5.0	2.0	0.0
	Severini (TEM)	0.5	1.5	3.5	8.0	3.5	0.5	0.5	1.5	3.5	8.5	4.0	0.5
	Severini (cov.)	0.5	1.5	3.5	8.0	3.5	0.5	0.5	1.5	3.5	8.5	4.0	0.5
F_2	Wald	0.0	0.0	0.0	19.5	16.5	11.5	0.0	0.0	0.0	19.5	16.5	12.0
	profile	0.5	1.5	3.5	7.5	4.0	1.0	0.5	2.0	3.5	7.5	4.0	1.0
	TEM	0.5	3.0	5.5	4.0	2.0	0.5	0.5	3.0	6.0	4.0	1.5	0.5
	Severini (TEM)	0.5	2.0	3.5	6.5	3.5	0.5	0.5	2.0	4.0	6.0	3.0	0.5
	Severini (cov.)	0.5	2.0	3.5	6.5	3.5	0.5	0.5	2.0	4.0	6.0	3.0	0.5
F_3	Wald	0.0	0.0	0.0	21.0	18.0	13.0	0.0	0.0	0.0	21.0	18.0	13.5
	profile	0.0	1.0	2.0	9.0	5.5	1.5	0.0	1.0	2.5	9.0	5.0	1.0
	TEM	0.0	2.0	4.0	5.0	2.5	0.5	0.5	2.0	5.0	5.0	2.5	0.5
	Severini (TEM)	0.0	1.0	2.5	8.0	4.5	1.0	0.0	1.0	3.0	8.0	4.5	1.0
	Severini (cov.)	0.0	1.0	2.5	8.0	4.5	1.0	0.0	1.5	3.0	8.0	4.5	1.0
F_4	Wald	0.0	0.0	0.0	30.0	26.5	20.5	0.0	0.0	0.0	32.0	28.0	22.5
	profile	0.5	1.5	3.0	11.0	6.0	1.0	0.5	1.0	2.5	11.5	6.0	1.0
	TEM	0.5	3.0	5.0	4.5	2.0	0.0	0.5	2.5	4.5	5.0	2.0	0.0
	Severini (TEM)	0.0	1.0	2.5	9.5	4.5	0.5	0.0	1.0	2.0	9.5	5.0	0.5
	Severini (cov.)	0.0	1.0	2.5	9.0	4.5	0.5	0.0	1.0	2.5	9.5	4.5	0.5
F_5	Wald	0.0	0.0	0.5	17.0	14.0	9.5	0.0	0.0	0.0	17.5	14.5	10.0
	profile	0.5	2.0	3.5	7.5	4.0	1.0	0.5	2.0	4.0	7.0	4.0	1.0
	TEM	0.5	3.0	5.5	4.5	2.0	0.5	1.0	3.5	7.0	4.5	2.0	0.5
	Severini (TEM)	0.5	2.0	4.0	6.5	3.5	1.0	0.5	2.5	5.0	6.5	3.5	0.5
	Severini (cov.)	0.5	2.0	4.0	6.5	3.5	1.0	0.5	2.5	5.0	6.5	3.5	0.5
F_6	Wald	0.0	0.0	0.0	27.0	23.5	18.0	0.0	0.0	0.0	29.0	25.0	19.5
	profile	0.5	2.0	3.5	12.5	7.0	2.0	0.5	1.5	3.0	13.5	7.5	2.0
	TEM	0.5	3.0	5.5	7.0	3.5	0.5	0.5	2.5	5.5	7.5	4.0	0.5
	Severini (TEM)	0.5	1.5	3.5	11.0	6.5	1.5	0.5	1.5	3.5	11.5	6.5	1.5
	Severini (cov.)	0.5	1.5	3.5	11.0	6.5	1.5	0.5	1.5	3.5	11.5	6.5	1.5
F_7	Wald	0.0	0.0	0.0	21.5	18.0	13.0	0.0	0.0	0.0	22.5	19.0	14.0
	profile	0.5	1.5	3.5	9.5	5.5	1.5	0.5	1.5	3.5	10.0	5.5	1.5
	TEM	0.5	3.0	5.5	5.5	3.0	0.5	0.5	3.0	6.0	5.5	3.0	0.5
	Severini (TEM)	0.5	2.0	4.0	8.5	5.0	1.0	0.5	2.0	4.0	8.5	5.0	1.0
	Severini (cov.)	0.5	2.0	4.0	8.5	5.0	1.0	0.5	2.0	4.0	8.5	4.5	1.0
F_8	Wald	0.0	0.0	0.0	23.0	19.5	14.5	0.0	0.0	0.0	23.5	20.5	15.5
	profile	0.5	1.5	3.5	9.0	5.0	1.0	0.5	1.5	3.5	9.0	5.0	1.0
	TEM	0.5	3.0	5.5	4.5	2.0	0.5	0.5	3.0	5.5	4.5	2.0	0.5
	Severini (TEM)	0.0	1.5	3.5	7.5	4.0	1.0	0.0	1.5	3.5	7.5	4.0	1.0
	Severini (cov.)	0.5	1.5	3.5	7.5	4.0	1.0	0.0	1.5	3.5	7.5	4.0	0.5
F_9	Wald	0.0	0.0	0.0	27.5	23.5	18.0	0.0	0.0	0.0	28.0	24.5	19.0
	profile	0.5	1.5	2.5	9.5	5.0	0.5	0.5	1.0	2.5	9.5	5.0	0.5
	TEM	0.5	2.5	5.0	4.0	1.5	0.0	0.5	2.5	5.0	3.5	1.5	0.0
	Severini (TEM)	0.0	1.0	2.5	8.0	4.0	0.5	0.0	1.0	2.5	8.0	3.5	0.5
	Severini (cov.)	0.0	1.0	2.5	8.0	3.5	0.5	0.0	1.0	2.5	8.0	3.5	0.5

Table 1.7 – *One-sided nominal error rate (in %) for lower (first to third columns) and upper (fourth to sixth columns) confidence intervals, peaks-over-threshold method with* $k = 60$. *The distributions (from top to bottom) are Burr* (F_1), *Weibull* (F_2), *generalized gamma* (F_3), *normal* (F_4), *lognormal* (F_5), *Student t* (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), *exponential* (F_8) *and* $\mathsf{GP}(\xi = -0.1)$ (F_9).

1.5 Data illustrations

1.5.1 Vargas tragedy

The heavy rainfall of December 1999 caused an estimated 30,000 deaths in the state of Vargas in Venezuela, due to debris flow that hit coastal installations. Daily cumulated rainfall data recorded at the Maiquetía *Simón Bolívar International Airport* is available for the period 1961–1999, in addition to yearly maximum daily rainfall for the period 1951–1960. Anecdotal records are also available: for example, during the floods of February 1951, a reported 282 mm of rain fell in Maiquetía, while the neighbouring station of El Infiernito, located in the Cordillera de la Costa, between Caracas and Maiquetía, recorded 529 mm for the same day (Wieczorek et al., 2001). Other series include monthly total precipitation, which mostly agree with the daily records (except for 1999). The December 1999 storm led to estimated cumulated precipitation of 911mm (backcasted) over a three-day period at Maiquetía airport. These data were analyzed in Coles and Pericchi (2003) and Coles et al. (2003), who argued that an event of this magnitude would not have been considered impossible had uncertainty been properly accounted for. December is typically dry month, but also more prone to very intense rainfall episodes; the records combine different types of meteorological events.

While one could combine information about the yearly maxima for the period 1951-1960 using the Poisson point process, we model the tail of daily cumulated rainfall for 1961–1998 using a generalized Pareto distribution, selecting the empirical 99.8% value, $u = 57.5$mm, as threshold. This leaves $n = 24$ exceedances for inference. Threshold stability plots for the shape (not shown) indicate this is a reasonable choice; moreover, the changepoint score test of Northrop and Coleman (2014) presented in Method 1.12 above the 98.5% value, 20.9mm, yields P-values that exceed 0.2. It would be tempting to select a lower threshold, but the conclusions (unlike the shape parameter estimates) change drastically for a threshold between 45mm and 55mm, due to a large number of high, but not extreme events. Based on the tangent exponential model approximation for the distribution of the maximum of the centennial cumulated daily rainfall, the P-value for the 410.4mm event recorded on December 16th, 1999 is 4% with $u = 38.7$mm ($n = 67$, 99.5%), whereas it is 25% with $u = 57.5$mm ($n = 24$, 99.8%). This additional uncertainty results from a combination of lower sample size and heavier tail. Figure 1.14 shows the profile likelihood for the median of the centennial maximum.

The bias-corrected estimate of the shape for the generalized Pareto distribution with threshold $u = 57.5$mm is 0.04, while the maximum likelihood estimate is $\widehat{\xi} = -0.07$; since smaller shape estimates also have smaller standard errors, this can magnify the underestimation of risk at the extrapolation stage. The quantile-quantile plot of Figure 1.15 for the fitted model also hints that we better capture the tail of the distribution with the bias-corrected estimates.

Although the extrapolation allows us to draw statements about quantities of interest, a Bayesian approach is more natural for assessing the likelihood of the 1999 tragedy. We thus consider the posterior predictive distribution, which the metastatistical distribution is meant

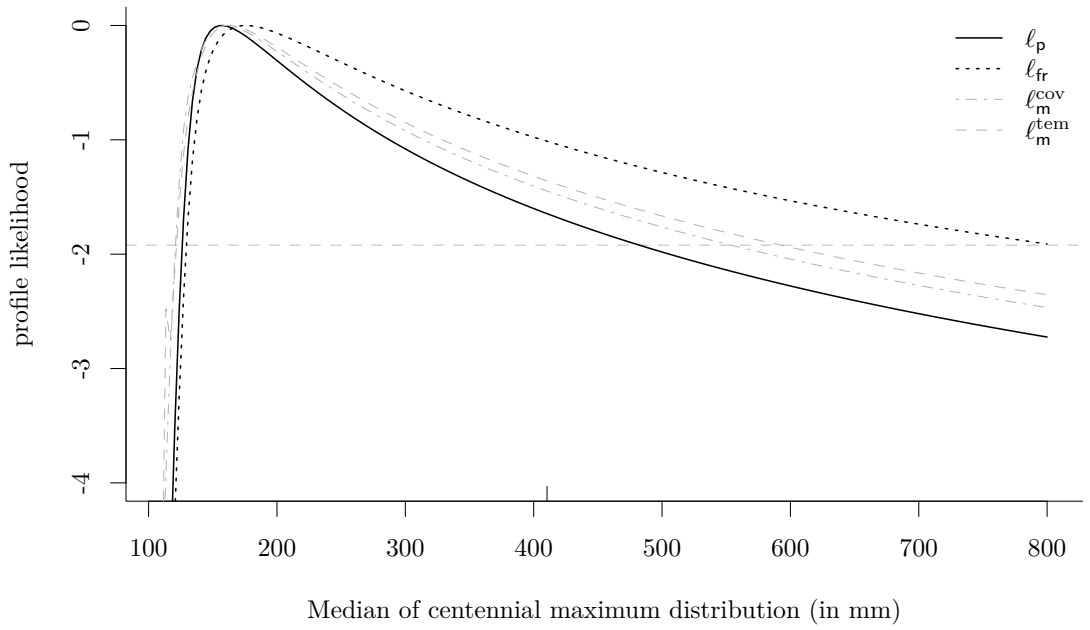

Figure 1.14 – *Profile likelihood and higher order approximations for the median of the centennial maximum daily rainfall at Maiquetía based on* 24 *threshold exceedances above* $u = 57.5$*mm, using daily data from 1961–1998. Profile likelihood (full black line), tangent exponential model approximation (dotted) and Severini's modified profiles (grey dashed and dot-dashed). The grey horizontal line at* -1.92 *indicates cutoff values for 95% confidence intervals based on the asymptotic* χ^2_1 *distribution. The rug at 410.4mm indicates the maximum record of December 1999.*

to approximate. We fit the generalized Pareto distribution to the same data using the ratio-of-uniforms method (Wakefield et al., 1991) with prior $\sigma^{-1}(\xi-1/2)^3(1-\xi)^5\mathbf{1}\{-1/2\leq\xi\leq 1\}$, a modified version of the Beta prior of Martins and Stedinger (2000) whose mean is $1/10$, in agreement with rainfall studies of Serinaldi and Kilsby (2014). This prior enforces finite moments and regularizes the posterior, discarding most very heavy-tailed distributions that would lead to nonsensical predictions in the context of rainfall. Based on $N = 10\,000$ independent draws from the posterior obtained using the R package `revdbayes` (Northrop, 2019a), we then sample 73 exceedances ≈ 0.002 exceedances/day $\times 365.25$ days/year $\times 100$ years and compute the maximum of these samples. The metastatistical model is obtained by considering a mixture of generalized extreme value distributions; their parameters are obtained through a max-stability argument based on the empirical distribution of the number of wet days and probability-weighted moment estimates of the Weibull parameters fitted to non-zero rainfall for each year. The value of December 16th, 1999, corresponds to the 0.955-quantile of the posterior predictive distribution and the 0.98-quantile of the metastatistical model. The corresponding posterior predictive distributions are shown in Figure 1.16. That based on the generalized Pareto fit is more concentrated around values in the range 150–250mm and has a slightly more open tail. As before, the metastatistical model gives an estimate that seems too low in view of the records and the time series.

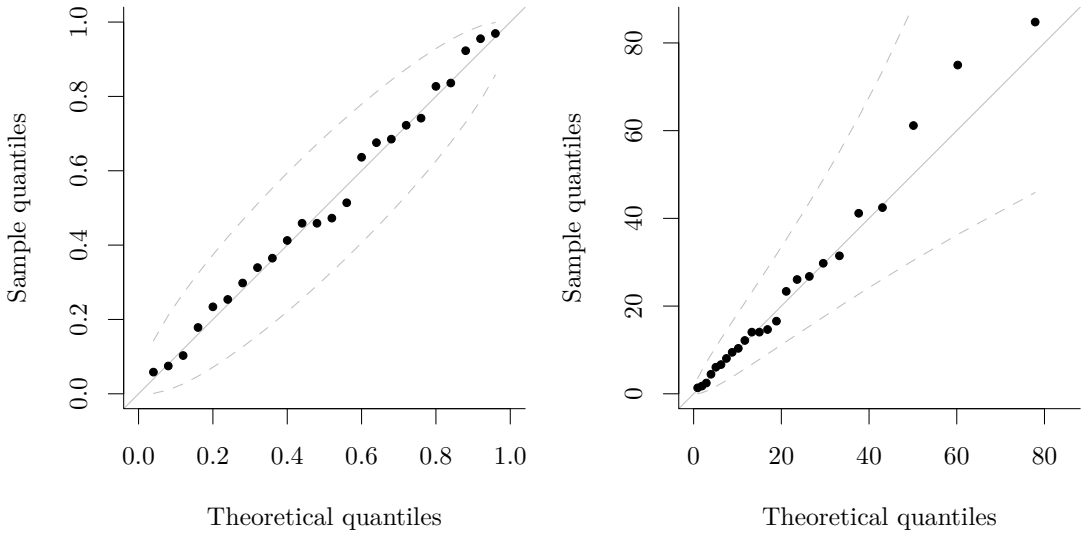

Figure 1.15 – *Probability-probability plot (left) and quantile-quantile plot (right) for daily cumulated rainfall threshold exceedances based on the generalized Pareto distribution with threshold $u = 57.5$mm, using the bias-corrected estimator of eq.* (1.8)*. The dashed grey lines show approximate 95% pointwise confidence intervals based on order statistics.*

The Maiquetía data provide a good illustration of the pitfalls associated with analysis of extremes: the choice of threshold has a huge impact on the conclusions of the analysis, and revised estimates taking into account the data for 1999 would lead to tremendous changes. Measurements are not available after the largest event was recorded; unpublished work of Barlow et al. (2019) explore effects of stopping rules on inference. Concurring with Coles and Pericchi (2003), the Vargas tragedy was much larger than expected and corresponds to an unlikely, but possible event, even if it would fall outside its 95% posterior predictive credible interval.

1.5.2 Super centenarians

The possible existence of a finite upper limit for human lifetime is a longstanding question that has recently sparked interest in the extreme value community (Hanayama and Sibuya, 2016; Einmahl et al., 2019; Rootzén and Zholud, 2017). The Italian centenarian data set, kindly provided by Holger Rootzén, contains the birth date and age of 3836 individuals from the super centenarian survey conducted by Istituto Nazionale di Statistica (Istat); the data are analyzed in Barbi et al. (2018). Individuals appear in the database if they reached age 105 years between January 1st, 2009 (c_1) and January 1st, 2016 (c_2); and the survival time is censored for individuals alive at c_2. The cohort thus comprises people born between 1896 and 1910 whose excess lifetime is measured as days of survival beyond $u = 38351$ days. Nonparametric estimates of the survival function and the cumulative hazard function are displayed in Figure 1.17.

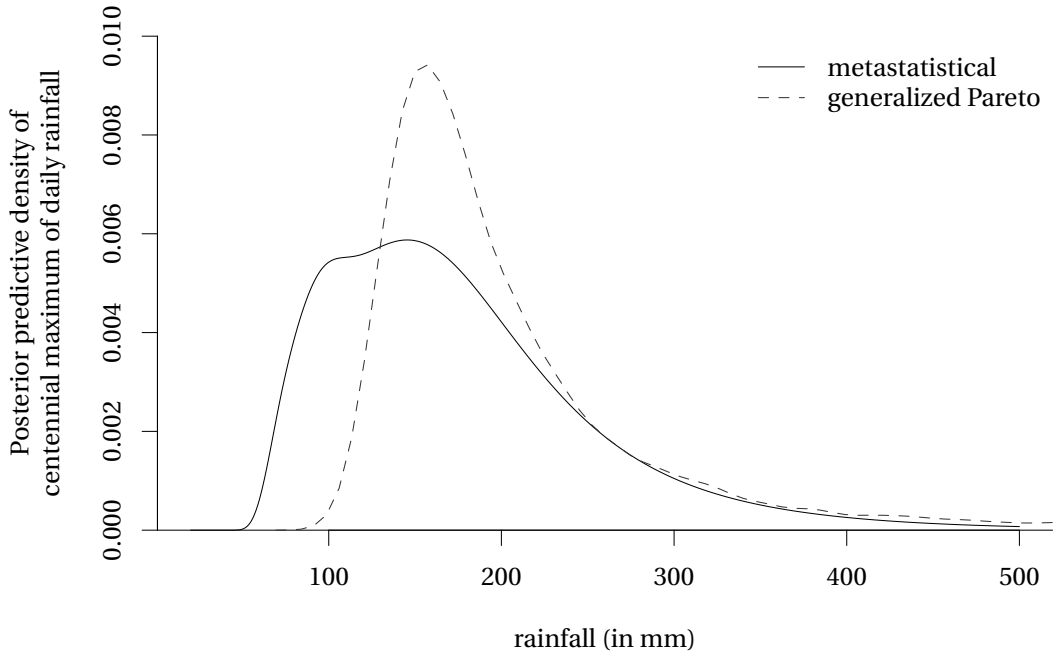

Figure 1.16 – *Posterior predictive distribution of rainfall conditional on exceeding $u = 57.5$mm based on the maximum of 73 generalized Pareto variates (dashed). Metastatistical distribution for centennial maxima obtained through penultimate approximation and max-stability argument (full).*

Figure 1.18 shows the construction of the likelihood contributions for individuals in the sample whose age exceeded u_0 between calendar times c_1 and c_2; the observations are potentially left-truncated and right-censored and taking this into account is important. Failure to account for the censoring would lead to negative bias for the shape parameter ξ since, for example, individuals born after 1910 would not be given a chance to reach 116 years. Let S and f denote the survival and the density functions, let x_i denote the calendar date at which individual i reached u_0 years, let t_i denote the excess lifetime at calendar time c_2, and let s_i denote an indicator variable taking value 1 if individual i was alive at calendar time c_2 and zero otherwise. The likelihood is

$$L(\boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{s}) = \prod_{i=1}^{n} \left[\frac{f(t_i)}{S\{(c_1 - x_i)_+\}} \right]^{1-s_i} \left[\frac{S(t_i)}{S\{(c_1 - x_i)_+\}} \right]^{s_i},$$

where $x_+ = \max\{x, 0\}$. We fit a generalized Pareto distribution to excess lifetimes over a range of thresholds starting from $u = 105$ years. We performed likelihood ratio tests (not reported) to check for differences between men and women, by fitting different generalized Pareto and exponential distributions to both subsamples; none of the improvements in fit is statistically significant at the 10% level. Maximum likelihood estimates of the generalized Pareto parameters and upper endpoints are given in Table 1.8. The largest excess lifetime is censored (Emma Moreno, who died aged 117 years in 2017) and the estimated shape $\widehat{\xi}$ tends to be close to zero, although for high u its variability is large.

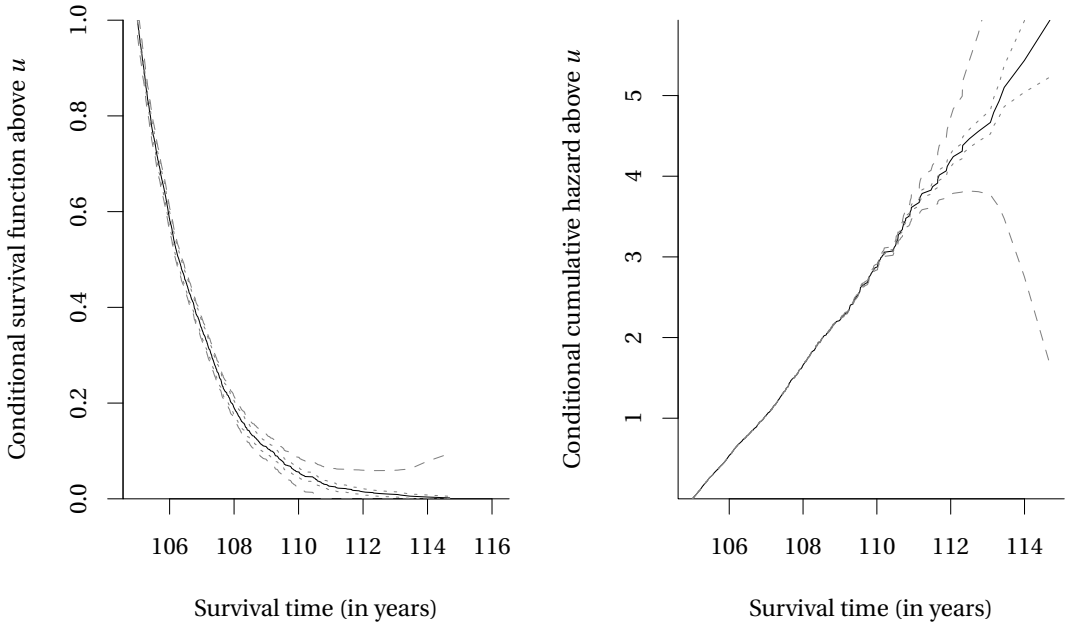

Figure 1.17 – *Estimated conditional survival function (left) and conditional cumulative hazard function (right) above u based respectively on the Kaplan–Meier and Nelson–Aalen estimators for left-truncated right-censored data (Tsai et al., 1987) for the super centenarian data. Dotted (dashed) lines give 95% pointwise confidence intervals (95% confidence bands). The intervals for the cumulative hazard functions are arcsine-transformed intervals (Bie et al., 1987) and the critical levels for the confidence bands are based on results in Wellner and Hall (1980).*

Whenever $\widehat{\xi}$ is negative, the estimated distribution has a finite upper endpoint $\iota = -\sigma/\xi$ and one could reparametrize the profile likelihood in terms of the latter; if $\widehat{\xi} \geq 0$, excess lifetime is unbounded. Figure 1.19 shows the profile likelihood and the modified version ℓ_{fr} of Section 1.3.3 for the upper endpoint for two thresholds; the 95% confidence interval based on the likelihood root r and the tangent exponential model approximation given for two scenarios in Table 1.10 have similar lower endpoints and the tangent exponential model approximation has a heavier right tail; right endpoints are nearly infinite for higher thresholds (not shown), since $\widehat{\xi}_\iota \to 0$ as $\iota \to \infty$, so the drop in deviance is negligible starting from 200 years or so. As is often the case, the tangent exponential model approximation behaves erratically in a neighbourhood of $r = 0$; we interpolate the values from the root $r^* = 0$ using a spline. To check the accuracy of the approximations, we used Algorithm 1.1 to compute the distribution of the likelihood ratio statistic w using the bootstrap by simulating from the generalized Pareto model (Lee and Young, 2005). For each ι, we compute the bootstrap P-value $\mathsf{P}\left(w_{\mathrm{p}}(\iota)\right)$ and compare it with the P-value obtained from the asymptotic χ^2_1 distribution. Figure 1.20 shows that, in this setting, the bootstrap and asymptotic χ^2_1 P-values agree up to Monte Carlo variability. The P-values decrease as the distance between ι and $\widehat{\iota}$ increases.

One drawback of the frequentist paradigm in the present case is the inability to compute

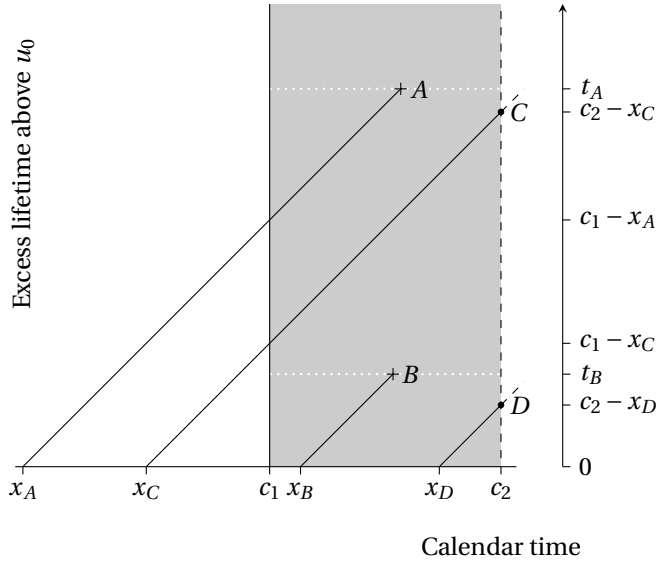

Figure 1.18 – *Lexis diagram showing the life trajectory of observed individuals; those who would attain age u_0 beyond calendar time c_2 or those who died before calendar date c_1 or below age u_0 are unobserved. Individual A passes age u_0 at date $x_A < c_1$, has excess lifetime (beyond u_0) of $c_1 - x_A$ at date c_1, and dies at age $u_0 + t_A$ (left-truncated). Individual B reaches age u_0 at calendar time x_B and dies between calendar time c_1 and c_2 at age $u_0 + t_B$. Individual C is left-truncated at excess lifetime $c_1 - x_C$ and is right-censored at excess lifetime $c_2 - x_C$. Individual D is right-censored at excess lifetime $c_2 - x_D$.*

u	n_u	$\widehat{\sigma}$	$\widehat{\xi}$	$\widehat{\iota}$	$\ell(\widehat{\theta})$
105	3836	1.67 (0.04)	−0.04 (0.02)	142	−4253.7
105.5	2665	1.72 (0.05)	−0.07 (0.02)	130	−2955.2
106	1874	1.70 (0.06)	−0.07 (0.03)	129	−2064.3
106.5	1345	1.63 (0.07)	−0.06 (0.03)	132	−1448.2
107	946	1.47 (0.08)	−0.02 (0.04)	196	−999.3
107.5	650	1.40 (0.09)	0.01 (0.05)	∞	−665.7
108	415	1.47 (0.11)	−0.01 (0.06)	210	−440.6
108.5	278	1.55 (0.14)	−0.05 (0.07)	139	−294.6
109	198	1.33 (0.15)	0.03 (0.09)	∞	−202.9
109.5	134	1.24 (0.18)	0.09 (0.13)	∞	−132.7
110	88	1.22 (0.23)	0.12 (0.17)	∞	−85.4
110.5	60	0.83 (0.23)	0.43 (0.28)	∞	−54.6
111	34	1.50 (0.47)	0.06 (0.30)	∞	−34.9
111.5	23	1.23 (0.51)	0.24 (0.45)	∞	−24.9

Table 1.8 – *Maximum likelihood estimates of the generalized Pareto for the Italian supercentenarian data. From left to right, threshold u (in years), number n_u of observations above the threshold, estimates (standard errors) of the scale σ, shape ξ and endpoint in years ι parameters, maximum log-likelihood, $\ell(\widehat{\theta})$.*

u	$\widehat{\sigma}$	$\ell(\widehat{\theta})$	$\widehat{\sigma}_e$	$\widehat{\alpha}_e$	$\ell_e(\widehat{\theta})$
105	1.61 (0.03)	−4255.8	1.68 (0.05)	0.05 (0.03)	−4253.7
105.5	1.62 (0.04)	−2959.1	1.75 (0.06)	0.10 (0.04)	−2954.9
106	1.60 (0.04)	−2067.3	1.73 (0.07)	0.11 (0.05)	−2063.8
106.5	1.55 (0.05)	−1450.0	1.67 (0.08)	0.10 (0.06)	−1447.8
107	1.45 (0.05)	−999.3	1.48 (0.08)	0.02 (0.05)	−999.2
107.5	1.42 (0.06)	−665.7	—	—	—
108	1.45 (0.08)	−440.6	1.47 (0.12)	0.02 (0.07)	−440.6
108.5	1.49 (0.10)	−294.8	1.59 (0.17)	0.09 (0.13)	−294.5
109	1.36 (0.11)	−202.9	—	—	—
109.5	1.34 (0.13)	−133.0	—	—	—
110	1.35 (0.17)	−85.7	—	—	—
110.5	1.22 (0.18)	−56.4	—	—	—
111	1.58 (0.32)	−34.9	—	—	—
111.5	1.49 (0.35)	−25.1	—	—	—

Table 1.9 – *Maximum likelihood estimates (standard errors) and maximum likelihood values for the exponential distribution (second and third columns) and extended exponential distribution $F(x) = 1 - \exp\{-x/\sigma_e - \alpha_e(x/\sigma_e)^2/2\}$ when $\widehat{\alpha}_e \neq 0$ (last three columns) for the Italian super-centenarian data, as a function of the threshold u.*

potential upper endpoints when the estimate shape is positive, since the (infinite) upper endpoint ι lies on the boundary of the parameter space. To remedy this, we consider Bayesian inference for the model and assign vague priors to the parameters (standard Gaussian and

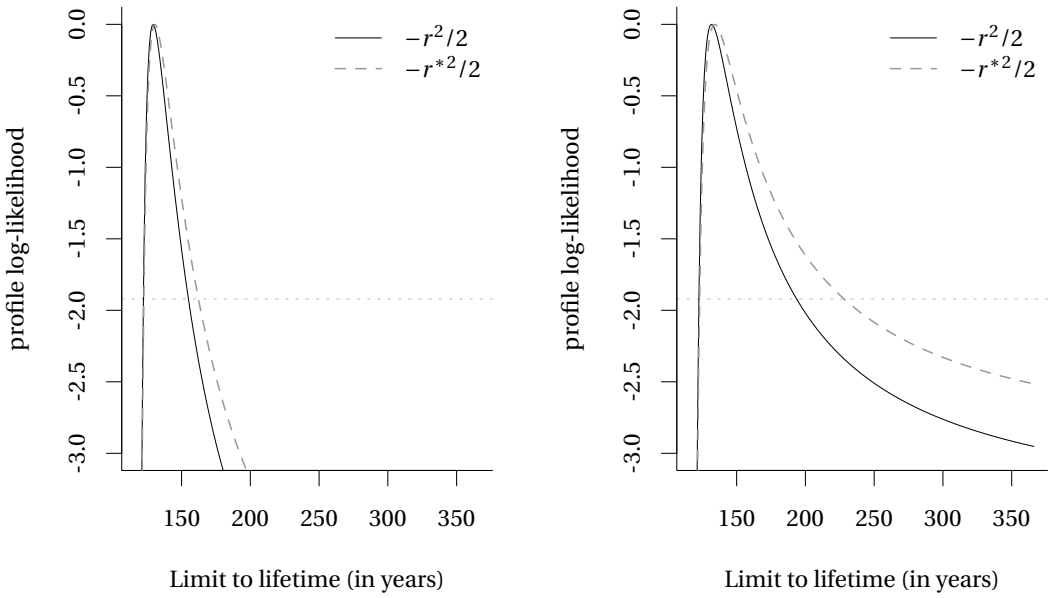

Figure 1.19 – *Profile likelihood ℓ_{p} (full black) and modified profile likelihood based on the tangent exponential model approximation ℓ_{fr} (dashed grey) as a function of the upper endpoint ι based on exceedances of lifetime beyond* 106 *years (left) and* 106.5 *years (right) for the Italian super centenarian data. The horizontal dotted grey lines indicates cutoff values for the 95% confidence interval based on the asymptotic χ^2_1 distribution.*

u	r	r^*
105	142.2 (131.7, 212.9)	143.4 (132.4, 235.8)
105.5	129.9 (120.6, 145.6)	134.2 (125.0, 151.2)
106	129.3 (119.4, 150.2)	134.0 (123.9, 160.4)
106.5	131.7 (120.9, 194.6)	137.0 (124.8, 227.9)

Table 1.10 – *Point estimates (95% confidence intervals) for the upper limit to lifetime ι (in years) based on the profile likelihood ratio statistic w (middle) and the modified likelihood ratio statistic w^* for the tangent model approximation (right) using threshold exceedances of u for the Italian semi-super centenarian data set.*

half-Cauchy for ξ and σ, respectively). We use a Hamiltonian Monte Carlo algorithm (cf. Section 3.1.2) to sample five chains of 25 000 draws from the posterior distribution, after discarding the first 1000 iterations and thinning to retain every fifth iteration for memory allocation even if unnecessary (Carpenter et al., 2017). We use the 25 000 remaining draws to compute 90% credible intervals which are reported in Table 1.11 with the posterior median. For low thresholds, the posterior medians are close to the maximum likelihood estimates given in Table 1.10. The contours of the joint posterior density for (σ, ξ) are further from bivariate Gaussian and become curved since smaller values of ξ are incompatible with the data, resulting in asymmetric intervals and larger point estimates. The mixing of the chains is

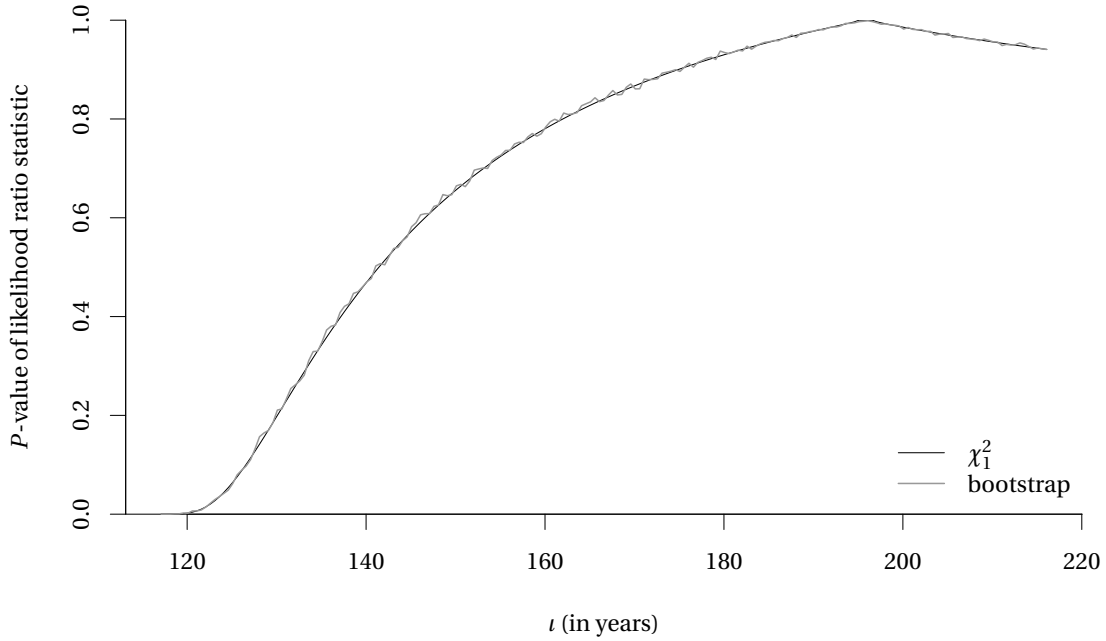

Figure 1.20 – *Italian super centenarian data: P-value function,* $\mathsf{P}\{w_{\mathrm{p}}(\iota)\}$*, for the profile likelihood ratio statistic* $w_{\mathrm{p}} = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\iota\}$ *using excess lifetime above* $u = 107$ *years based on the asymptotic* χ^2_1 *distribution (black) and the bootstrap distribution (grey).*

u	ι
105	142.9 (127.4, 305.3)
105.5	130.6 (122.5, 164.3)
106	130.1 (121.8, 178.2)
106.5	133.2 (122.1, 437.2)
107	230.6 (127.7, ∞)
107.5	∞ (131.6, ∞)
108	484.0 (124.2, ∞)
108.5	148.9 (120.7, ∞)
109	∞ (125.6, ∞)
109.5	∞ (129.6, ∞)
110	∞ (127.6, ∞)
110.5	∞ (∞, ∞)
111	∞ (118.8, ∞)
111.5	∞ (120.1, ∞)

Table 1.11 – *Posterior median (90% credible intervals) for the upper limit to lifetime* ι *(in years) using uninformative priors, as a function of the threshold* u*.*

good in all the scenarios considered.

The apparent stability of the estimates and the large standard errors for $\widehat{\xi}$ do not allow us to

u	105	105.5	106	106.5	107	107.5	108	108.5	109	109.5	110	110.5	111	111.5
$\widehat{\xi}$	98	100	99	96	63	36	56	69	32	17	16	1	31	16
n_c	15	10	14	13	21	33	62	72	57	24	8	5	6	11

Table 1.12 – *Percentage of bootstrap samples in which $\widehat{\xi}^{(b)} > \widehat{\xi}$ (first row) and in which the number of right-censored simulated data $n_c^{(b)} > n_c$ for individuals reaching age u in $[c_1, c_2]$ (second row) as a function of threshold u. The standard error for the Monte Carlo estimator is less than 0.5%.*

rule out the exponential tail for thresholds $u > 107$ years at level 10% (the P-values for the lower four thresholds are 4.1%, 0.5%, 1.4% and 6.0%). We fix $\xi = 0$ and compute maximum likelihood estimates for the exponential model; see the second column of Table 1.9.

Another way to assess the hypothesis that excess lifetimes are exponentially distributed is by simulation; we condition on the birth dates and simulate excess lifetimes from a left-truncated exponential distribution in the time frame $[c_1, c_2]$, censoring observations that fall outside the frame. For low thresholds, there is some indication of lack of fit. The simulated proportion of individuals that are censored agrees with the proportion in the data set for all thresholds. For each simulated data set, we can fit a generalized Pareto distribution and compare the shape $\widehat{\xi}^{(b)}$ in replication $b = 1, \ldots, B$ with that obtained in the original sample, $\widehat{\xi}$. We see discrepancies at low thresholds in Table 1.12, when the sample size is large enough to allow one to detect departure from exponentiality. This could be due to a decrease in the force of mortality at low thresholds, followed by an increase. Adjustment for increase in the number of individuals, as indicated by potentially different rate of exceedances of u for older ages, depends on information that is yet unavailable; see Davison (2018) for a discussion of the impacts of non-stationarity.

The limiting distribution of excess lifetime could be exponential, even if we reject this hypothesis for low thresholds; such an hypothesis would be compatible with penultimate effects. To assess this, consider a survival function of the form $S(x) = \exp\{-x/\sigma - \alpha(x/\sigma)^2/2\}$ for $x, \alpha, \sigma > 0$; this includes the exponential and is, up to second order, the penultimate approximation for Gaussian variates, for which the penultimate shape is negative. The corresponding parameter estimates and maximum log-likelihood values are reported in the last three columns of Table 1.9. If excess lifetime was ultimately exponential, we would expect to observe a decreasing α and this is the case for all thresholds beyond 108.5 years. Simulations (not reported here) indicate that improvements in fit appear only if the penultimate shape of the generalized Pareto is negative, which is consistent with the estimates $\widehat{\xi}$. The estimates are thus in some sense compatible with penultimate effects.

The fact that the upper bounds of the credible and confidence intervals are infinite reflects the fact that the shape could be zero or positive. Rootzén and Zholud (2017) argue that one could consider the exponential model, on the ground of parsimony. Under the latter, the probability of surviving one additional year conditional on survival up to u years is $\exp(-1/\sigma)$ as a consequence of the the memoryless property. Based on exceedances of $u = 110$ years,

the model would yield an estimated probability of surviving an additional year of 0.476 with 95% confidence interval (0.416, 0.537) and fewer than four in a thousand super-centenarians would be expected to live older than Emma Moreno. This justifies the claim of Rootzén and Zholud that life is infinite, but short.

1.6 Conclusion

This chapter discussed likelihood-based estimation of extreme value distributions, focusing on higher-order methods and bias correction. The extreme value distributions are non-standard models: their support is parameter-dependent and the cumulants of the log-likelihood derivatives exist only for sufficiently large values of the shape parameter. These restrictions make inference challenging, in particular because the distribution depends on the threshold choice in peaks-over-threshold analysis or the block size for block maxima. We have advocated the use of the distribution of the maximum of N observations because its interpretation is more natural than return levels; the discrepancy between the penultimate approximation and the estimated distributions can be substantial, as illustrated through examples. The penultimate approximation can be useful as a gold standard in simulation studies and in understanding that, even if asymptotically constant, the shape parameter is often bound to change as the threshold or the block size changes. Its use in practice is difficult unless one is willing to assume a parametric family for the underlying data, as the metastatistical approach does: the choice of distribution introduces arbitrariness that practitioners need to understand better. The penultimate approach requires derivation of exact normalizing constants, which is easily done by assuming a parametric model for the whole distribution of the data. Estimation of the latter directly from the limiting model is hopeless, and this explains why we only carry out frequentist and Bayesian inference using the likelihood of the limiting model. While the Poisson process formulation is the most useful for combining multiple sources of data, optimization of the likelihood is highly dependent on the tuning parameter even if guidelines for orthogonalization are available (Sharkey and Tawn, 2017).

Most of the methods described in the chapter have been programmed and implemented in the `mev` package (Belzile et al., 2019); this include likelihood, score functions and Fisher and observed information matrices for different likelihoods, profile and modified profile likelihoods for the generalized extreme value and generalized Pareto distributions in different parametrizations, optimization routines, threshold selection methods described in Section 1.1.5 and bias correction methods; see the link provided in Appendix D for a detailed list of the functionalities implemented.

The novel contributions in the chapter include derivation of several examples second order condition (Examples 1.6 and 1.7), the comparison of the metastatistical approach with the generalized extreme value distribution, the derivation of the Fisher information for the r-largest observations, a comparison of the Cox–Snell correction for the generalized extreme value distribution and the generalized Pareto distribution as reported in Giles et al. (2016)

and Roodman (2018) with the implicit bias correction and the Firth's score correction through simulation, the derivation of the Bartlett correction for the generalized Pareto distribution, comparison of profile likelihood-based confidence intervals for mean and quantiles of the distribution of the N-observations maximum, derivation of the TEM for mean and quantiles of the N-observation maximum, return levels, parameters and expected shortfall numerically, including some analytical examples in Examples 1.25 and 1.26. Several real data examples illustrate the methods.

Likelihood methodology provides a convenient framework for investigating the performance of estimators and for derivation of explicit expressions for cumulants, first-order bias, etc. The peaks-over-threshold method, while potentially allowing for more data points to be used in the analysis, has been shown to lead to significantly negatively-biased estimates of the shape. The bias for ξ is systematic over the range of ξ and bias correction may be advised, particularly if extrapolation for periods far beyond that of the data is envisioned. The study of higher-order asymptotics for extreme value distributions has shown that the first order approximation is poor, but that the use of analytic higher-order methods such as bias correction or Bartlett correction is not a panacea; in fact, it seems that a parametric bootstrap provides more sensible estimates when ξ is negative, as illustrated in Examples 1.19 and 1.22. Simulation studies show that, while the profile-likelihood confidence intervals based on r have adequate coverage, their error rate is too small. The confidence intervals for the tangent exponential model approximation are much wider, reflecting the skewness of the estimated distribution. Inclusion of covariates, non-stationary modelling and data transformations are not easily studied, despite the possibility of including these in the framework.

2 A panorama of spatial extremes

This chapter presents the three main approaches for modelling spatial extremes, namely peaks-over-threshold, max-stable and conditional extremes models. The literature review includes some novel contributions, with a particular emphasis on simulation. The reader is invited to consult page 2 for a description of the notation used in this chapter. Whenever a proof is provided, it is either novel or else serves to provide further insight in the result.

2.1 Preliminary notions

Because extreme events are, by definition, rare, it is impossible to estimate quantities of interest empirically since these lie outside the range of the data. One may be interested in predicting extremes at a site where no measurements are available, or in assessing the risk of complex extreme physical phenomena, such as floods or heatwaves, which are inherently spatio-temporal. In such cases, extrapolation to unobserved periods and sites is necessary, but it cannot be handled in the univariate or the multivariate frameworks. Thus, we focus on functional extremes and review methodological and theoretical developments in the field of spatial extremes.

2.1.1 Point processes

Point processes are a key tool for describing occurrences of extremes, so we detail key properties, following van Lieshout (2010). Let $(\mathcal{O}, d_{\mathcal{O}})$ be a complete separable metric space with associated Borel σ-algebra $\mathbb{B}(\mathcal{O})$ and let $\mathbb{B}_b(\mathcal{O})$ be the subset of bounded Borel sets. A point process Φ is a random configuration of "points" on $\mathcal{O}$, an unordered set $\{\phi_1, \ldots, \phi_n\} \in \mathcal{O}^n$ for any $n \in \mathbb{N} \cup \{\infty\}$. For any set B, the (random) number of points falling in B is the counting measure $\Phi(B) = \text{card}(B \cap \Phi)$.

A configuration of points ϕ is locally finite (or boundedly finite) if $\text{card}(\{\phi\} \cap B) < \infty$ for any bounded set $B \subseteq \mathcal{O}$. Spatial point processes are often defined on bounded subsets: it is thus customary to define point processes Φ on the space of locally finite configurations $\mathcal{N}^{\text{lf}}$,

equipped with the σ-algebra

$$\mathbb{B}(\mathcal{N}^{\mathrm{lf}}) := \sigma\left\{\phi : \mathrm{card}(\{\phi\} \cap B) = n, n \in \mathbb{N} \cup \{\infty\}, B \in \mathbb{B}(\mathcal{O}) \text{ bounded}\right\}.$$

For any bounded set $B \in \mathbb{B}(\mathcal{O})$, we identify the process via the integer-valued random variable $\Phi(B) = \sum_{i=1}^{n} \mathbf{1}\{\phi_i \in B\}$. A point process is said to be simple (or nonatomic) if all points are pairwise distinct, meaning that $\Phi(\{\phi\}) \leq 1$ almost surely for all $\phi \in \mathcal{O}$.

Moments of the point process are defined through those of the counting variable. Suppose that the expected number of points falling in any compact $K \subset \mathcal{O}$ is finite, i.e., $\mathsf{E}\{\Phi(K)\} < \infty$; the first-order moment measure, also termed the intensity measure, is $\Lambda(B) = \mathsf{E}\{\Phi(B)\}$ for any Borel set B. If Φ is regular, so the σ-finite measure Λ is dominated by the Lebesgue measure, we can write $\Lambda(\mathrm{d}\phi) = \lambda(\phi)\mathrm{d}\phi$ and $\Lambda(B) = \int_B \lambda(\phi)\mathrm{d}\phi$. The Radon–Nikodym derivative of $\Lambda(\cdot)$, denoted $\lambda(\cdot)$, is the intensity function of the nonatomic process. The process is said to be homogeneous if the intensity $\lambda(\cdot)$ is constant over $\mathcal{O}$. The expected number of points in a bounded set A is then the measure of that set, $\lambda\nu(A)$ for $\lambda \geq 0$.

Higher-order moments of a point process take into account counts in multiple sets. Consider bounded Borel sets $B_1, \ldots, B_D$ and $\Phi(B_1), \ldots, \Phi(B_D)$ for any $D \in \mathbb{N}^+$. The Dth order measure is simply

$$\Lambda_n(B_1 \times \cdots \times B_D) = \mathsf{E}\{\Phi(B_1) \cdots \Phi(B_D)\}.$$

If the sets are all identical to B, then the Dth order measure of B is simply the Dth power of its first-order moment, $\Lambda_D(B) = \Lambda(B)^D$. The Dth-order measure Λ_n can be extended to a σ-finite measure: for any measurable mapping $f : \mathcal{O} \to \mathbb{R}$ which is either nonnegative or else integrable and provided Λ_D exists,

$$\mathsf{E}\left(\sum_{\phi_1,\ldots,\phi_D \in \Phi} f(\phi_1, \ldots, \phi_D)\right) = \int_{\mathcal{O}^D} f(\phi_1, \ldots, \phi_D)\Lambda_n(\mathrm{d}\phi_1, \ldots, \mathrm{d}\phi_D). \tag{2.1}$$

This result is often referred to as Campbell's theorem.

Poisson point processes

Simple parametric families can be used to describe the counts of the point process. The class most used is the class of Poisson point processes, denoted hereafter by $\mathsf{PPP}(\Lambda)$. A point process is Poisson if

(i) for any Borel $B \in \mathbb{B}(\mathcal{O})$, $\Phi(B) \sim \mathsf{Pois}\{\Lambda(B)\}$;

(ii) for all disjoint Borel sets $B_1, \ldots, B_n$, the variables $\Phi(B_i)$ are independent (complete randomness).

The simplest Poisson point process example is the homogeneous Poisson process on the positive real line $\mathbb{R}_+$.

Definition 2.1 (Homogeneous Poisson point process on $\mathbb{R}_+$)
The homogeneous Poisson point process on $\mathbb{R}_+$ with intensity $\mathrm{d}\lambda$ can be characterized in terms of waiting time, that is, the distance between two events of the point process. Specifically, let $E_i \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$ be waiting times and define the sequence of arrival times $\{\Gamma_n\}_{n\in\mathbb{N}^+}$ with $\Gamma_n = \sum_{i=1}^{n} E_i$.

This definition can be used to generate a homogeneous Poisson point process with unit rate on $\mathbb{R}_+$. For more general domains $S \subset \mathcal{O}$ satisfying $\Lambda(S) < \infty$, simulation of a homogeneous Poisson process can be done by first simulating the number of points in S as $N_S \sim \mathsf{Pois}(\Lambda(S))$ and then by simulating n_S points uniformly from S. If S is irregular, it may be simpler to simulate on a larger spatial domain and use an accept-reject scheme.

To generate samples from inhomogeneous Poisson processes, the simplest option is thinning: if the Poisson process has a bounded intensity function $\lambda(\phi) \leq M$ for all $\phi \in \mathcal{O}$, one can generate points from a homogeneous Poisson process on $\mathcal{O}$ with intensity M and keep each simulated observation with probability $\lambda(\phi)/M$. While thinning is typically easy to implement, it is not the most efficient simulation method. We can instead use the preservation property of Poisson processes, which states that Poisson processes remain Poissonian under suitable transformations (cf. Resnick, 2007, Proposition 5.2).

Proposition 2.2 (Transformation of Poisson processes)
Let Φ be a Poisson point process with intensity measure Λ. Let $T : \mathcal{O} \to \mathcal{O}'$ be a measurable transformation such that for any compact set $K' \in \mathscr{K}(\mathcal{O}')$, the inverse image $T^{-1}K' \in \mathscr{K}(\mathcal{O})$ is also compact and that $T^{-1}(\{\varphi\}) = 0$ for all $\varphi \in \mathcal{O}'$. Then $\Phi \circ T^{-1} \sim \mathsf{PPP}(\Lambda \circ T^{-1})$ on $\mathcal{O}'$.

Example 2.1
Consider the mapping $x \mapsto 1/x$ of points of the homogeneous Poisson process on $\mathbb{R}_+$ with intensity $\mathrm{d}\zeta$; the transformation is bijective over $(0,\infty)$ and thus preserves the properties of the Poisson distribution as per Proposition 2.2. The transformed process Γ_n^{-1} is also Poisson with intensity $\zeta^{-2}\mathrm{d}\zeta$.

□

Marked point process

It is possible to enlarge the state space of a (Poisson) point process with intensity measure Λ and consider simultaneously a location and the realization of an event (a mark) attached to this coordinate. More specifically, consider elements of the product state space $\mathcal{E} = \mathcal{O} \times \mathcal{F}$, where $\mathcal{O}$ is a metric space to be understood as the location component of the process and $\mathcal{F}$ is the mark space $\mathcal{F}$, i.e., a metric space with a suitable associated σ-algebra. Suppose that random components m_i on $\mathcal{F}$ are independent and identically distributed with probability measure μ. Then, the (Poisson) point process defined on $\mathcal{E}$ with elements $\sum_{i\in\mathbb{N}^+} \mathbf{1}\{\phi_i, m_i\}$ is a marked point process and its intensity factorises as $\Lambda \times \mu$. In particular, one can generate from the marked process by sampling the points of a Poisson process with intensity Λ and attach to each realization a mark m_i drawn from μ (cf. de Haan and Ferreira, 2006, Lemma 9.4.7). The

reader is referred to § 4.1.1 of Heinrich (2013) for a more rigorous definition of marked point processes.

Slivnyak–Mecke theorem

The Palm distribution of a point process is, heuristically, the conditional distribution of the point process Φ given a point at the location ϕ. The Slivnyak–Mecke theorem provides an equivalence between this Palm distribution and that of the point process $\Phi \cup \{\phi\}$. It allows for interchange of measures, an analog of Fubini's theorem for point processes.

Theorem 2.3 (Slivnyak–Mecke)
Let Φ be a Poisson point process with finite intensity measure Λ on a complete separable metric space $\mathbb{O}$. For any nonnegative function $f : \mathbb{O} \times \mathcal{N}^{\mathrm{lf}} \to \mathbb{R}_+$,

$$\mathsf{E}\left(\sum_{\phi\in\Phi} f(\phi, \Phi \setminus \{\phi\})\right) = \int_{\mathbb{O}} \mathsf{E}\left(f(\phi, \Phi)\right) \Lambda(\mathrm{d}\phi). \tag{2.2}$$

More generally, for $D \in \mathbb{N}^+$ pairwise distinct points (denoted by the superscript $\neq$) and $f : \mathbb{O}^D \times \mathcal{N}^{\mathrm{lf}} \to \mathbb{R}_+$,

$$\mathsf{E}\left(\sum_{\phi_1,\ldots,\phi_D\in\Phi}^{\neq} f(\{\phi_1,\ldots,\phi_D\}, \Phi \setminus \{\phi_1,\ldots,\phi_D\})\right) = \int_{\mathbb{O}^D} \mathsf{E}\left(f(\phi_1,\ldots,\phi_D,\Phi)\right) \prod_{i=1}^{D} \Lambda(\mathrm{d}\phi_i).$$

A proof can be found in Møller and Waagepetersen (2003), Section 3.2.1.

2.1.2 Properties of spatial processes

Throughout, we use the notation $\mathsf{Z}(\mathcal{S}) = \{Z(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}}$ for the stochastic process on $\mathbb{R}^m$ whose pointwise realizations are $Z(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathbb{R}^m$ and denote by $\boldsymbol{Z}$ the D-vector with components $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_D)$. The translation operator $T_{\boldsymbol{h}}$ is $T_{\boldsymbol{h}}\left[\{Z(\boldsymbol{s}_i)\}_{i=1}^D\right] = \{Z(\boldsymbol{s}_i + \boldsymbol{h})\}_{i=1}^D$ for any $\boldsymbol{h} \in \mathbb{R}^m$.

Two major concepts for stochastic processes are stationarity and ergodicity.

Definition 2.4 (Strict stationarity)
Let $T_{\boldsymbol{h}}$ be the translation operator. A stochastic process $\mathsf{Z}(\mathcal{S})$ with sample paths in $\mathbb{R}^m$ is said to be strictly stationary if all its finite-dimensional distributions are invariant under translation, i.e., $\boldsymbol{Z} \overset{\mathrm{d}}{=} T_{\boldsymbol{h}}(\boldsymbol{Z})$ for any $\boldsymbol{h} \in \mathbb{R}^m$.

Loosely speaking, a strictly stationary process is ergodic if realizations sufficiently far apart are near independent. This in particular means that properties of the process can be obtained by replacing expectation of functionals by empirical averages (for example covariance). Ergodicity is described in more detail on pages 54–58 of Cressie (1993), from which the following definition is extracted.

Definition 2.5 (Ergodicity)
Let $T_{\boldsymbol{h}}$ be the translation operator and let B be a measurable set with $T_{\boldsymbol{h}}^{-1}(B) = \{b : T_{\boldsymbol{h}}(b) \in B\}$. A strictly stationary process Z is ergodic if $T_{\boldsymbol{h}}^{-1}(B) = B$ for all $\boldsymbol{h}$ implies $\mathsf{P}(\mathsf{Z} \in B) = 0$ or $\mathsf{P}(\mathsf{Z} \in B) = 1$, i.e., the only sets invariant to translation are sets of measure zero or the whole sample space.

The requirement of strict stationarity is too stringent for many applications, so two weaker notions are used.

Definition 2.6 (Weak stationarity)
The stochastic process $Z(\boldsymbol{s})$ is termed weakly (or second order) stationary if its mean process is constant and the covariance between sites depends only on the distance between them, i.e., $\mathsf{E}\{Z(\boldsymbol{s})\} = \mu$ for all $\boldsymbol{s} \in \mathcal{S}$ and $\mathsf{Cov}\{Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{h})\} = C(\boldsymbol{h})$ for all $\boldsymbol{s}, \boldsymbol{h} \in \mathcal{S}$.

For the second notion, we need first the following definition.

Definition 2.7 (Semi-variogram)
The semi-variogram $\gamma : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ is half the process incremental variance, i.e.,

$$2\gamma(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) := \mathsf{Var}\{Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})\}.$$

The following conditions are both necessary and sufficient for a nonnegative function $\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2)$ to yield a valid semi-variogram: (a) $\gamma(\boldsymbol{s}, \boldsymbol{s}) = 0$; (b) $\gamma(\cdot,\cdot)$ is even, meaning $\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2) = \gamma(\boldsymbol{s}_2, \boldsymbol{s}_1)$ for any $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{S}$; and (c) $\gamma(\cdot,\cdot)$ is conditional negative definite, i.e., $\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) \leq 0$ for any pair of sites $\boldsymbol{s}_i, \boldsymbol{s}_j$ and for any $\boldsymbol{a} \in \mathbb{R}^n$ such that $\sum_{i=1}^n a_i = 0$. These three relations must hold for any potential values of the variogram parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Definition 2.8 (Intrinsic stationarity)
A stochastic process $\mathsf{Z}(\mathbb{R}^m)$ is intrinsically stationary if its increments are second order stationary, i.e., $\mathsf{E}\{Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})\} = 0$ and $\gamma(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) = \gamma(\boldsymbol{0}, \boldsymbol{h})$ for all $\boldsymbol{s}, \boldsymbol{h} \in \mathbb{R}^m$.

We use the short-hand notation $\gamma(\boldsymbol{h}) \equiv \gamma(\boldsymbol{0}, \boldsymbol{h})$ for the semi-variogram if the process under consideration is intrinsically stationary. If we fix the value of $\mathsf{Z}(\mathbb{R}^m)$ at the origin, setting $Z(\boldsymbol{o}) = 0$ almost surely, then the covariance matrix derived for increments from $Z(\boldsymbol{o})$ is

$$\mathsf{Cov}\{Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2)\} = \gamma(\boldsymbol{s}_1) + \gamma(\boldsymbol{s}_2) - \gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2).$$

Second order stationarity implies intrinsic stationarity, so the covariance function C and semi-variogram γ are related via $\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h})$ with $\lim_{\|\boldsymbol{h}\|\to\infty} \gamma(\boldsymbol{h}) = C(\boldsymbol{0})$ as $\lim_{\|\boldsymbol{h}\|\to\infty} C(\boldsymbol{h}) = 0$. (Chilès and Delfiner, 2012, eq. 3.2). The converse is false in general but, if Z is ergodic, the semi-variogram is bounded and $C(\boldsymbol{h}) \to 0$ as $\|\boldsymbol{h}\| \to \infty$, resulting in $\lim_{\|\boldsymbol{d}\|\to\infty} \gamma(\boldsymbol{d}) = C(\boldsymbol{0})$ and

$$C(\boldsymbol{h}) = \lim_{\|\boldsymbol{d}\|\to\infty} \gamma(\boldsymbol{d}) - \gamma(\boldsymbol{h}).$$

If the variogram or the covariance does not depend on the direction of the lag vector $\boldsymbol{h}$, then the process is said to be isotropic and we write $\gamma(\boldsymbol{h}) = \gamma(h)$, where $h = \|\boldsymbol{h}\|$. Otherwise, the

process is anisotropic. It is common to incorporate a rotation matrix for processes in $\mathbb{R}^2$ to account for any geometric anisotropy, which can be described by the composition of a rotation and a dilation matrix, $\mathbf{A}$, in which case we use an isotropic variogram with distance $h = \|\mathbf{A}\boldsymbol{h}\|$. A common parametrization for the latter is (cf. Chilès and Delfiner, 2012, p.98)

$$\mathbf{A} = \begin{pmatrix} a_1\cos(\varrho) & a_1\sin(\varrho) \\ -a_2\sin(\varrho) & a_2\cos(\varrho) \end{pmatrix}, \qquad a_1, a_2 > 1, \varrho \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right). \tag{2.3}$$

To ensure identifiability, it is customary to set $a_1 = 1$ when the variogram or covariance model includes a scale parameter.

2.1.3 Convergence of measures

We will be interested in the sequel in point processes and their extremal attractor, the functional analog of Theorem 1.3. To this effect, we review a particular notion of weak convergence for stochastic process excluding some regions since we want the limiting point process to have a finite intensity measure. Let $(\mathcal{S}, d)$ be a compact metric space and consider the separable Banach space of bounded continuous functions $\mathcal{F} = \{f : f \in \mathcal{C}(\mathcal{S}, \mathbb{R})\}$ equipped with the $\|\cdot\|_\infty$ norm and the subset $\mathcal{F}^+$ of nonnegative bounded functions $\{f : f \in \mathcal{C}(\mathcal{S}, [0,\infty))\}$. We follow Lindskog et al. (2014) and consider subsets of $\mathcal{F}$ that are bounded away from a cone $\mathscr{C}$.

Definition 2.9 (Cone)
A cone $\mathscr{C}$ is a measurable subset of $\mathcal{F}$ which is closed under dilation, meaning $tx \in \mathscr{C}$ for all $x \in \mathscr{C}$ and $t \in (0,\infty)$.

Definition 2.10 (Bounded away from a cone)
Let the open ball of radius r around $\mathscr{C}$ be $\mathscr{C}_{(r)} := \{f \in \mathcal{F} : d(f, \mathscr{C}) := \inf_{g\in\mathscr{C}} d(f,g) < r\}$. A Borel set $B \in \mathbb{B}(\mathcal{F} \setminus \mathscr{C})$ is bounded away from a cone $\mathscr{C}$ if $d(B, \mathscr{C}) > 0$, meaning that $B \subset \mathcal{F} \setminus \mathscr{C}_{(r)}$ for some $r > 0$. We say that a function f is bounded away from $\mathscr{C}$ if $f(x) = 0$ for all $x \in \mathscr{C}_{(r)}$. A totally finite measure is said to be bounded away from a cone $\mathscr{C}$ if, for any positive radius r,

$$\Lambda\{f \in \mathcal{F}(\mathcal{S}) : d(f, \mathscr{C}) > r\} < \infty.$$

Before introducing the concept of functional regular variation, we must define convergence of measures on metric spaces; see Daley and Vere-Jones (2002), Appendix A.2, for a detailed exposition of the topic. Below, we always consider measures that are neither constantly null nor degenerate at a point.

We consider nonnegative bounded continuous functions $\mathcal{F}^+$: define $\mathcal{F}^+_{\mathscr{C}} = \mathcal{F}^+ \setminus \mathscr{C}$ for some closed cone $\mathscr{C}$. The open cone $\mathcal{F}^+_{\mathscr{C}}$ is a metric subspace of $\mathcal{F}^+$ with associated σ-algebra $\mathbb{B}(\mathcal{F}^+_{\mathscr{C}}) = \{B : B \subset \mathcal{F}^+_{\mathscr{C}}, B \subset \mathbb{B}(\mathcal{F}^+)\}$. We consider the set $\mathbb{M}_{\mathcal{F}^+_{\mathscr{C}}}$ of all Borel measures on $\mathcal{F}^+_{\mathscr{C}}$ that are finite on $\mathcal{F}^+ \setminus \mathscr{C}_{(r)}$ for every $r > 0$ (Lindskog et al., 2014).

The subspace $\mathcal{F}^+_{\mathscr{C}}$ is a complete separable metric space and, if we equip the space of bounded Borel measures on $\mathcal{F}^+_{\mathscr{C}}$, $\mathbb{M}(\mathcal{F}^+_{\mathscr{C}})$, with the topology of weak-hash convergence, then $\mathbb{M}(\mathcal{F}^+_{\mathscr{C}})$

is also metrizable and can be made into a complete separable metric space (Lindskog et al., 2014, Theorem 2.2).

A set in $\mathbb{M}_{\mathcal{F}_\mathscr{C}}$ is relatively compact if every sequence $\{\Lambda_n\}_{n\in\mathbb{N}^+}$ for $M \subset \mathbb{M}_{\mathcal{F}_\mathscr{C}}$ contains a convergent subsequence (Hult and Lindskog, 2006).

The notion of convergence of measures we consider in the sequel is w$^\#$ convergence (Daley and Vere-Jones 2002, Proposition A2.6.II).

Definition 2.11 (Weak-hash convergence)
Let $\{\Lambda_n\}_{n\in\mathbb{N}^+}$ and Λ_∞ be boundedly finite measures on a complete separable metric space $\mathcal{F}$. We say that Λ_n converges in the w$^\#$ topology, $\Lambda_n \xrightarrow{\text{w}^\#} \Lambda_\infty$ as $n \to \infty$, if $\int f \mathrm{d}\Lambda_n \to \int f \mathrm{d}\Lambda_\infty$ on $\mathcal{F}$ for any bounded continuous functions f vanishing outside a bounded set.

Alternatively, for $\{\Lambda_n\}_{n\in\mathbb{N}^+}$, $\Lambda \in \mathbb{M}_{\mathcal{F}_\mathscr{C}}$, the sequence of measures $\Lambda_n \xrightarrow{\text{w}^\#} \Lambda_\infty$ if $\lim_{n\to\infty} \Lambda_n(B) = \Lambda_\infty(B)$ for any Λ_∞-continuous Borel set $B \in \mathbb{B}(\mathcal{F})$ that is bounded away from $\mathscr{C}$ (Lindskog et al., 2014, Theorem 2.1).

Remark 2.1
Weak-hash convergence as defined in Definition 2.11 implies vague convergence (Lindskog et al., 2014, Lemma 2.1), but the two notions are equivalent for locally compact spaces. Similarly, weak convergence (Daley and Vere-Jones, 2002, Definition A2.3.I), which we denote by $\xrightarrow{\text{w}}$, reduces to w$^\#$ convergence whenever $\mathcal{S}$ is compact.

Theorem 2.3 in Lindskog et al. (2014) provides a continuous mapping theorem for measures on complete separable metric spaces.

Theorem 2.12 (Mapping theorem)
Let $\mathcal{F}^1$ and $\mathcal{F}^2$ be complete separable metric spaces and let $\mathscr{C}_1 \in \mathcal{F}^1$ and $\mathscr{C}_2 \in \mathcal{F}^2$ be closed cones. Consider the associated metric subspace $\mathcal{F}^1_\mathscr{C} = \mathcal{F}^1 \setminus \mathscr{C}_1$ and $\mathcal{F}^2_\mathscr{C} = \mathcal{F}^2 \setminus \mathscr{C}_2$. If the measurable transformation $T : (\mathcal{F}^1_\mathscr{C}, \mathbb{B}(\mathcal{F}^1_\mathscr{C})) \to (\mathcal{F}^2_\mathscr{C}, \mathbb{B}(\mathcal{F}^2_\mathscr{C}))$ is a homeomorphism such that $T^{-1}(A)$ is bounded away from $\mathscr{C}_1$ for any set $A \in \mathbb{B}(\mathcal{F}^2_\mathscr{C}) \cap T(\mathcal{F}^1_\mathscr{C})$ bounded away from $\mathscr{C}_2$, then the pushforward measure $\mathbb{M}_{\mathcal{F}^1_\mathscr{C}} \to \mathbb{M}_{\mathcal{F}^2_\mathscr{C}}$ defined by $\Lambda \circ T^{-1}$ for any $\Lambda \in \mathbb{M}_{\mathcal{F}^1_\mathscr{C}}$ is continuous.

Example 2.2 (Pseudo-polar transformation in $\mathbb{R}^D_+$)
For the pseudo-polar transformation $T(\boldsymbol{x}) = (\|\boldsymbol{x}\|, \boldsymbol{x}/\|\boldsymbol{x}\|)$ to be bijective, we consider the punctured space from which the origin is removed: because the cone $\{\mathbf{0}_D\}$ is compact in $\mathbb{R}^D_+$ and the image set $\{0\} \times \mathbb{S}$ is closed, the pushforward measure is continuous.

□

In spatial statistics, data are assumed to arise from a random field that is only observed at a finite number of sites. Projection is a continuous mapping, so the parameters of the spatial process can be estimated based on the observed realizations at the measurement stations.

Corollary 2.13 (Finite-dimensional projections)
Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D \in \mathcal{S}$ be D sites. The mapping $T_{\text{proj}} : \mathcal{F}^+_\mathscr{C} \mapsto \mathbb{R}^D_+ \setminus T_{\text{proj}}(\mathscr{C})$, where $T_{\text{proj}}(f) = \{f(\boldsymbol{s}_1), \ldots, f(\boldsymbol{s}_D)\}$ is continuous by Theorem 2.12.

Under stricter assumptions, the continuity of the pushforward measure is readily established.

Corollary 2.14
Suppose that either (a) $T : \mathcal{F}^1 \to \mathcal{F}^2$ is absolutely continuous with $\mathscr{C}_2 = T(\mathscr{C}_1)$ closed in $\mathcal{F}^2$ or (b) T is continuous and either $\mathcal{F}^1$ or $\mathscr{C}_1$ is compact. Then $T_{\mathbb{M}} : \mathbb{M}_{\mathcal{F}^1_{\mathscr{C}}} \mapsto \mathbb{M}_{\mathcal{F}^2_{\mathscr{C}}}$, defined by $T_{\mathbb{M}}(\Lambda) := \Lambda \circ T^{-1}$, is continuous in $\mathbb{M}_{\mathcal{F}^2_{\mathscr{C}}}$.

Convergence of a sequence of measures $\{\Lambda_n\}$ to a limiting measure Λ_∞ in $\mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$ therefore implies convergence of the pushforward measures $\Lambda_n \circ T^{-1} \xrightarrow{\text{w}^\#} \Lambda_\infty \circ T^{-1}$ in $\mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$.

Remark 2.2 (Extensions)
We focus solely on stochastic processes with continuous sample paths. This excludes processes taking values in the Skorohod space of càdlàg functions (right continuous with left limits) and the class of upper-semi continuous processes, a notable example of which is the disk model of Schlather (2002), see also Davison and Gholamrezaee (2011) and Huser and Davison (2013). The reader is referred to Hult and Lindskog (2005) for more details on regular variation on these types of spaces.

2.1.4 Regular variation

A general formulation of regular variation for metric spaces (Hult and Lindskog, 2005) allows one to relax the assumption of local compactness of the metric space. To avoid complications discussed on p. 274 of Lindskog et al. (2014), we will follow Lindskog et al. and define tail regions as subsets of $\mathcal{S}$ bounded away from a cone $\mathscr{C}$, without compactifying the space. This allows extrapolation on rays.

Definition 2.15 (Regular variation on $\mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$)
A sequence of measures $\{\Lambda_n\}_{n\in\mathbb{N}^+}$ in $\mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$ is regularly varying if there exists a positive increasing regularly varying sequence $\{a_n\}_{n\in\mathbb{N}^+}$ such that, as $n \to \infty$, $a_n\Lambda_n \xrightarrow{\text{w}^\#} \Lambda_\infty$ in $\mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$ for a nonzero measure $\Lambda_\infty \in \mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$.

Alternatively, regular variation of a measure $\Lambda \in \mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$ is equivalent to the existence of a nonzero measure $\Lambda_\infty \in \mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$ and an increasing positive function $a : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$t\Lambda\{a(t)\cdot\} \xrightarrow{\text{w}^\#} \Lambda_\infty(\cdot), \qquad t \to \infty.$$

Lemma 2.16
Regular variation of the measure Λ implies that the limit measure Λ_∞ is homogeneous of order $-\alpha$ for $\alpha > 0$, meaning

$$\Lambda_\infty(tA) = t^{-\alpha}\Lambda_\infty(A)$$

for all $A \in \mathbb{B}(\mathcal{F}^+_{\mathscr{C}})$ and $t \in \mathbb{R}_+$ (Lindskog et al., 2014, Theorem 3.1).

As a consequence of the continuous mapping theorem (Theorem 2.12), functional regular variation also implies regular variation on $\mathbb{R}^D_+$; the latter is usually defined in the multivariate

setting using vague convergence on $\mathbb{R}_+^D \setminus \{\mathbf{0}_D\}$ with the one-point uncompactification (Resnick, 2007, Theorem 6.1).

A useful result for analyzing the tail behaviour of scale mixture models is Breiman's lemma (cf. Basrak et al., 2002).

Lemma 2.17 (Breiman's lemma)
Let R be a positive random variable with survival function $\bar{F}_R$, and let $\boldsymbol{S}$ be a D-dimensional random vector independent of R. Suppose that either

1. $\bar{F}_R \in \mathsf{RV}_{-\alpha}$ for $\alpha \in (0,\infty)$ and $\mathsf{E}\left(\|\boldsymbol{S}\|_\infty^{\alpha+\varepsilon}\right) < \infty$ for some $\varepsilon > 0$ or
2. $\bar{F}_R(r) \sim Cr^{-\alpha}$ as $r \to \infty$ for some $C > 0$ and $\mathsf{E}\left(\|\boldsymbol{S}\|_\infty^{\alpha}\right) < \infty$.

Then, the product $R\boldsymbol{S}$ is regularly varying on $[-\infty_D, \infty_D] \setminus \{\mathbf{0}_D\}$ with tail index $\alpha > 0$. In the univariate case,

$$\lim_{y\to\infty} \frac{\mathsf{P}\left(RS > y\right)}{\mathsf{P}\left(R > y\right)} = \mathsf{E}\left(S^\alpha\right).$$

Furthermore, if $S \in (0,1)$ almost surely, then the conclusion also holds for $\alpha = 0$ and $\alpha = \infty$, where $\mathsf{E}\left(S^\infty\right) = 0$ by convention.

The proof of the multivariate version of Breiman's lemma follows from a pseudo-polar decomposition of $\boldsymbol{S} \mapsto (\|\boldsymbol{S}\|_\infty, \boldsymbol{S}/\|\boldsymbol{S}\|_\infty)$ and an application of the univariate Breiman lemma. Extensions of Breiman's lemma with vector radial variables $\boldsymbol{R} \in \mathbb{R}^D$ and dependent radial and angular components are given in Theorem 1 of Fougères and Mercadier (2012).

Regular variation can only describe processes that have positive shape parameters pointwise, but we can derive convergence results nevertheless by considering marginal transformations, coupling Definition 2.15 and Theorem 2.12.

2.2 Max-stable processes

Definition 2.18 (Max-stable process)
A stochastic process $\mathsf{Z}(\mathcal{S})$ on a compact metric space $\mathcal{S}$ is max-stable if, for any $k \in \mathbb{N}^+$, there exist sequences of scaling functions $\mathsf{a}_k : \mathcal{S} \to (0,\infty)$ and $\mathsf{b}_k : \mathcal{S} \to \mathbb{R}$ such that, for k independent and identical copies $\mathsf{Z}_1(\mathcal{S}), \ldots, \mathsf{Z}_k(\mathcal{S})$ of $\mathsf{Z}(\mathcal{S})$,

$$\frac{\max_{i=1}^k Z_i(\boldsymbol{s}) - b_k(\boldsymbol{s})}{a_k(\boldsymbol{s})} \overset{\mathrm{d}}{=} Z(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{S}.$$

As a consequence of max-stability, $Z(\boldsymbol{s}) \sim \mathsf{GEV}\{\mu(\boldsymbol{s}), \sigma(\boldsymbol{s}), \xi(\boldsymbol{s})\}$ for each $\boldsymbol{s} \in \mathcal{S}$.

Max-stable processes arise as non-degenerate limits of pointwise maxima of random processes: consider independent and identically distributed copies of a stochastic process $\mathsf{X}(\mathcal{S})$. If there exists sequences of normalizing functions $\mathsf{a}_n : \mathcal{S} \to (0,\infty)$ and $\mathsf{b}_n : \mathcal{S} \to \mathbb{R}$ such that

$$\frac{\max_{i=1}^n X_i(\boldsymbol{s}) - b_n(\boldsymbol{s})}{a_n(\boldsymbol{s})} \longrightarrow Z(\boldsymbol{s}), \qquad n \to \infty, \quad \boldsymbol{s} \in \mathcal{S},$$

in the sense of finite-dimensional distributions, then the process $\mathsf{Z}(\mathcal{S}) = \{Z(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}}$ is max-stable and is termed the max-stable extremal attractor of $\mathsf{X}(\mathcal{S})$.

We decouple the convergence of the marginal distribution from that of the dependence structure and require that the marginal distribution function of $X(\boldsymbol{s})$, $F_{\boldsymbol{s}}$, be absolutely continuous in order to use the probability integral transform to transform the process onto a standardized scale; further assumptions are necessary to do this for upper semi-continuous processes (Sabourin and Segers, 2017). If the pointwise marginals of Z are unit Fréchet for any $\boldsymbol{s} \in \mathcal{S}$, we say the process is simple max-stable and denote it by Z^* to indicate that it is standardized. More generally, if $Z(\boldsymbol{s}) \sim \mathsf{GEV}\{\mu(\boldsymbol{s}), \sigma(\boldsymbol{s}), \xi(\boldsymbol{s})\}$, we can obtain the standardized process $Z^*(\boldsymbol{s}) = T\{Z(\boldsymbol{s})\}$ through the transformation

$$T\{Z(\boldsymbol{s})\} = \begin{cases} \left[1 + \xi(\boldsymbol{s})/\sigma(\boldsymbol{s})\{Z(\boldsymbol{s}) - \mu(\boldsymbol{s})\}\right]^{1/\xi(\boldsymbol{s})}, & \xi(\boldsymbol{s}) \neq 0, \\ \exp\left[\{Z(\boldsymbol{s}) - \mu(\boldsymbol{s})\}/\sigma(\boldsymbol{s})\right], & \xi(\boldsymbol{s}) = 0, \end{cases} \qquad \boldsymbol{s} \in \mathcal{S}. \tag{2.4}$$

The multivariate setting is treated in Section 2.2.2.

So far, we have focused on convergence in the sense of finite-dimensional distributions. Functional convergence is, however, necessary to compute, e.g., the area of the exceedance region, $|\mathcal{S}|^{-1}\int_{\mathcal{S}} \mathbf{1}\{Z(\boldsymbol{s})>u\}\mathrm{d}\boldsymbol{s}$ or the spatial average $|\mathcal{S}|^{-1}\int_{\mathcal{S}} Z(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$. Following Theorem 2.8 of de Haan and Lin (2001), functional convergence of a sequence of continuous stochastic processes in $\mathscr{F}_0^+$ can be described as follows.

Theorem 2.19 (Convergence to a simple max-stable process in $\mathscr{F}_0^+$)
Let $\{\mathsf{X}_i(\mathcal{S})\}_{i\in\mathbb{N}^+}$ be independent and identically distributed stochastic processes with sample paths in $\mathscr{F}_0^+(\mathcal{S})$ such that their pointwise marginal distribution function $F_{\boldsymbol{s}}(x) := \mathsf{P}\{X(\boldsymbol{s}) \leq x\}$ is continuous in $\boldsymbol{s} \in \mathcal{S}$ for all x. Let $F_{\boldsymbol{s}}^{\leftarrow}(u) = \inf\{x : F_{\boldsymbol{s}}(x) \geq u\}$ for $u \in [0,1]$. The following are equivalent:

(i) There exist continuous sequences of functions $\mathsf{a}_n : \mathcal{S} \to (0,\infty)$ and $\mathsf{b}_n : \mathcal{S} \to \mathbb{R}$ such that

$$\frac{\max_{i=1}^n X_i(\boldsymbol{s}) - b_n(\boldsymbol{s})}{a_n(\boldsymbol{s})} \longrightarrow Z(\boldsymbol{s}), \qquad n \to \infty, \quad \boldsymbol{s} \in \mathcal{S},$$

in $\mathscr{F}_0^+$, where $\mathsf{Z}(\mathcal{S})$ is a max-stable process whose pointwise margins $Z(\boldsymbol{s})$ at $\boldsymbol{s}$ are Fréchet with shape $\xi(\boldsymbol{s})$ and $\xi : \mathcal{S} \to \mathbb{R}$ is continuous.

(ii) The first-order condition

$$\lim_{t\to\infty} \frac{F_{\boldsymbol{s}}^{\leftarrow}(1-1/tx) - F_{\boldsymbol{s}}^{\leftarrow}(1-1/t)}{a_t(\boldsymbol{s})} = \frac{x^{\xi(\boldsymbol{s})}-1}{\xi(\boldsymbol{s})} \tag{2.5}$$

holds uniformly in $\boldsymbol{s}\in\mathcal{S}$ and locally uniformly in $x\in(0,\infty)$ for $\xi(\boldsymbol{s})$ continuous. Furthermore,

$$\frac{1}{n}\max_{i=1}^{n}\frac{1}{1-F_{\boldsymbol{s}}(X_i(\boldsymbol{s}))} \xrightarrow{\text{w}^{\#}} [1+\xi(\boldsymbol{s})Z(\boldsymbol{s})]^{1/\xi(\boldsymbol{s})} \equiv Z^*(\boldsymbol{s}), \qquad n\to\infty, \quad \boldsymbol{s}\in\mathcal{S}, \tag{2.6}$$

in $\mathcal{F}_0^+$ for continuous $\xi:\mathcal{S}\to\mathbb{R}$. The process $\mathsf{Z}^*(\mathcal{S})$ appearing on the right-hand side of (2.6) is simple max-stable.

Theorem 2.19 separates convergence to a max-stable process into a requirement of marginal uniform convergence, eq. (2.5) and a statement on the convergence of measure of the dependence structure in standardized margins, eq. (2.6). These criteria are essentially the same as those of Deheuvels (1978) in the multivariate setting.

2.2.1 Spectral representation of max-stable processes

The spectral representation of a max-stable process, which first appeared in de Haan (1984), is a constructive characterization of the process in terms of a Poisson point process. It is the building-block for simulation and for likelihood inference. Many variants of the spectral representation exist (cf. Vatan 1985; Giné, Hahn and Vatan 1990; Schlather 2002), but we shall restrict attention to that presented in Corollary 9.4.5 of de Haan and Ferreira (2006).

Theorem 2.20 (Spectral representation of max-stable processes (1))
Any simple max-stable process $\mathsf{Z}(\mathcal{S})$ on the compact set $\mathcal{S}$ with continuous sample paths in $\mathcal{F}_0^+$ admits the representation

$$Z^*(\boldsymbol{s}) \overset{\text{d}}{=} \max_{i=1}^{\infty}\zeta_i W_i(\boldsymbol{s}), \qquad \boldsymbol{s}\in\mathcal{S}, \tag{2.7}$$

where

- $\zeta_1>\zeta_2>\cdots$ are points of a Poisson point process on $\mathbb{R}_+$ with intensity $\zeta^{-2}\mathrm{d}\zeta$;
- the spectral functions, $\mathsf{W}(\mathcal{S})$, are independent and identically distributed copies of a stochastic process in $\mathcal{F}_0^+$ satisfying $\mathsf{E}\{W(\boldsymbol{s})\}=1$ for all $\boldsymbol{s}\in\mathcal{S}$ and $\mathsf{E}\left(\|\mathsf{W}\|_\infty\right)<\infty$;
- $\{\zeta_i\}$ is independent of $W_j(\boldsymbol{s})$ for all $\boldsymbol{s}\in\mathcal{S}$ and for all $i,j\in\mathbb{N}^+$.

The condition $\mathsf{E}\left(\|\mathsf{W}(\mathcal{S})\|_\infty\right)<\infty$ holds whenever $\mathsf{W}(\mathcal{S})$ is bounded.

Theorem 2 in Schlather (2002) provides a more general spectral representation which is useful for creating new parametric classes of max-stable processes.

Theorem 2.21 (Spectral representation of max-stable processes (2))
Let $\{\zeta_i\}_{i\in\mathbb{N}^+}$ be decreasing points of a Poisson point process on $\mathbb{R}_+$ with intensity measure $\zeta^{-2}\mathrm{d}\zeta$. Independently of $\{\zeta_i\}_{i\in\mathbb{N}^+}$, let $\mathsf{W}_1, \mathsf{W}_2, \ldots$, be independent and identically distributed copies of an arbitrary stochastic process $\mathsf{W}(\mathbb{R}^m)$ and let $\mathsf{W}_+(\mathbb{R}^m)$, where $W_+(\boldsymbol{s}) := \max\{0, W(\boldsymbol{s})\}$, be such that $\mathsf{E}\{W_+(\boldsymbol{s})\} = 1$ for all $\boldsymbol{s} \in \mathbb{R}^m$. Then,

$$Z^*(\boldsymbol{s}) := \max_{i=1}^{\infty} \zeta_i W_{i+}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathbb{R}^m, \tag{2.8}$$

is a simple max-stable process on $\mathbb{R}^m$. If $\mathsf{W}(\mathbb{R}^m)$ is stationary, so is $\mathsf{Z}^*(\mathbb{R}^m)$.

The requirement $\mathsf{E}\{W_+(\boldsymbol{s})\} = 1$ ensures that pointwise marginals are unit Fréchet.

Remark 2.3
The representation in eq. (2.7) is not unique, so different spectral functions may lead to the same max-stable process $\mathsf{Z}^*(\mathcal{S})$. One could also multiply a max-stable process by a random field with unit expectation and get a max-stable process with the same distribution (see, e.g., Ribatet, 2013). For identifiability, one needs to impose additional restrictions, e.g., by fixing the value of the process at a particular site $\boldsymbol{o} \in \mathcal{S}$. This leads to no loss in generality for intrinsically stationary processes, since we can always define W', with $W'(\boldsymbol{s}) = W(\boldsymbol{s}) - W(\boldsymbol{o})$ for $\boldsymbol{s} \in \mathcal{S}$; both W' and W will have the same incremental variance.

de Haan's spectral representation of the simple max-stable process Z^* gives

$$\Lambda(B) = \int_0^{\infty} \mathsf{P}\left(\zeta W(\boldsymbol{s}) \in B\right) \zeta^{-2}\mathrm{d}\zeta,$$

and an application of Fubini's theorem shows that the intensity function, if it exists, is

$$\lambda(\boldsymbol{z}) = \int_0^{\infty} f_{\boldsymbol{W}}(\boldsymbol{z}/\zeta)\zeta^{-D-2}\mathrm{d}\zeta. \tag{2.9}$$

If spectral functions are truncated in the spectral representation (2.8), the intensity measure Λ_+, defined on $\mathscr{F}_0^+$, has positive mass on $\{f \in \mathscr{F}_0^+ : \min_{\boldsymbol{s}\in\mathcal{S}} f(\boldsymbol{s}) = 0\}$ (cf. Coles and Tawn, 1991).

Remark 2.4 (Change of measure for non-standardized processes)
Consider the pointwise map T given in eq. (2.4) for μ, σ, ξ continuous functions on $\mathcal{S}$, and suppose that

$$n\mathsf{P}\left\{T\left(\frac{\mathsf{X} - \mathsf{b}_n}{\mathsf{a}_n}\right) \in \cdot\right\} \xrightarrow{\mathrm{w}^{\#}} \Lambda(\cdot), \qquad n \to \infty,$$

for $\Lambda \in \mathbb{M}_{\mathscr{F}_{\mathscr{C}}}$ a homogeneous measure of order -1. By Theorem 2.20, the simple max-stable process with intensity measure Λ admits the spectral representation of Equation (2.7). Because T is continuous on $\mathscr{F}_{\mathscr{C}}^+$, it follows by virtue of Proposition 2.2 that the max-stable attractor of $\mathsf{X}(\mathcal{S})$ admits the representation $\max_{i=1}^{\infty} T^{-1}\{\zeta W(\boldsymbol{s})\}$ with limiting intensity measure $\Lambda \circ T(\cdot)$. Consider the projection of the random function onto D sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D \subset \mathcal{S}^D$ and assume that

$\boldsymbol{W}$ has density function h for any $B \in \mathbb{B}(\mathbb{R}^D)$,

$$\Lambda \circ T(B) = \int_0^\infty \int_{\mathbb{R}_+^D} \mathbf{1}\{T^{-1}(\zeta \boldsymbol{w}) \in B\} h(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \zeta^{-2} \mathrm{d}\zeta,$$

and making the change of variable $z_j = T^{-1}(\zeta w_j)(j = 1, \ldots, D)$ gives

$$\Lambda \circ T(B) = \int_0^\infty \int_B h\left\{T(z_j)\zeta^{-1}, j = 1, \ldots, D\right\} \left| \prod_{j=1}^D \left| \frac{\partial T(z_j)}{\partial z_j} \right| \right| \zeta^{-D-2} \mathrm{d}\boldsymbol{z}\mathrm{d}\zeta. \tag{2.10}$$

If Λ has a Radon–Nikodym derivative λ, we see from (2.9) that the intensity function associated to $\Lambda \circ T$ is $\lambda\{T(\boldsymbol{z})\} \prod_{j=1}^D \left|\partial T(z_j)/\partial z_j\right|$ by the chain rule.

Definition 2.22 (Exponent measure)
Let $\mathcal{S}$ be a compact metric space and let ρ_∞ be a totally finite measure on $\mathbb{S}_\infty$ satisfying the moment constraint $\int f \rho_\infty(\mathrm{d}f) = 1$ for all $\boldsymbol{s} \in \mathcal{S}$ and such that for any $D \in \mathbb{N}^+$, any compact subsets $K_1, \ldots, K_D$ of $\mathcal{S}$ and $\boldsymbol{z} > \mathbf{0}_D$, (Giné et al., 1990)

$$\begin{aligned} -\log\left[\mathsf{P}\left(\bigcap_{j=1}^D \left\{\sup_{\boldsymbol{s}\in K_j} Z^*(\boldsymbol{s}) \le z_j\right\}\right)\right] &= \Lambda\left\{f \in \mathcal{F}_0^+ : \bigcup_{j=1}^D \left\{\sup_{\boldsymbol{s}\in K_j} f(\boldsymbol{s}) > z_j\right\}\right\} \\ &= \int_{\mathbb{S}_\infty} \max_{j=1}^D \left(\frac{\sup_{\boldsymbol{s}\in K_j} f(\boldsymbol{s})}{z_j}\right) \rho_\infty(\mathrm{d}f). \end{aligned}$$

In pseudo-polar coordinates, the exponent measure Λ factorises into radial and angular measures, $\mathrm{d}\Lambda = \zeta^{-2}\mathrm{d}\zeta\mathrm{d}\rho_\infty$.

The intensity measure of the limiting Poisson point process Z^* gives the void probabilities of the complements of measurable sets B, viz. $-\log\left\{\mathsf{P}\left(\mathsf{Z}^*(\mathcal{S}) \in B\right)\right\} = \Lambda(B^{\mathsf{c}})$ (Giné et al., 1990).

Definition 2.23 (Distribution function of multivariate max-stable vector)
For a fixed collection of sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$ with $\boldsymbol{s} \in \mathcal{S}$, the distribution function of the random vector $\boldsymbol{Z}^* = \{Z^*(\boldsymbol{s}_1), \ldots, Z^*(\boldsymbol{s}_D)\}$ is

$$\mathsf{P}\left(\boldsymbol{Z}^* \le \boldsymbol{z}\right) = \exp\{-V(\boldsymbol{z})\} = \exp\left\{-\Lambda\left([\mathbf{0}_D, \boldsymbol{z})^{\mathsf{c}}\right)\right\}, \qquad \boldsymbol{z} > \mathbf{0}_D.$$

The function $V(\boldsymbol{z}) = \Lambda\{(\mathbf{0}_D, \boldsymbol{z}]^{\mathsf{c}}\}$ is termed the exponent measure.

Consider the transformation of the random vector $\boldsymbol{Z}$ into pseudo-polar coordinates, $\boldsymbol{Z} \mapsto (R, \boldsymbol{\Omega})$ from $\mathbb{R}_+^D \to (0,\infty) \times \mathbb{S}_1$, where $\mathbb{S}_1 = \{\boldsymbol{\omega} \in \mathbb{R}_+^D : \|\boldsymbol{\omega}\|_1 = 1\}$ is the ℓ_1-simplex. The measure Λ factorises as a product measure $\Lambda(\mathrm{d}\boldsymbol{z}) = D\zeta^{-2}\mathrm{d}\zeta\rho_1(\mathrm{d}\boldsymbol{\omega})$ with angular measure ρ_1, a probability measure satisfying the moment constraint $D\int_{\mathbb{S}_1} \omega_j \rho_1(\mathrm{d}\boldsymbol{\omega}) = 1$ for $j = 1, \ldots, D$; this moment constraint holds for any ρ_1, whereas it will be measure dependent if the radial measure is not $\|\cdot\|_1$ (Einmahl and Segers, 2009). We can re-express the event $\{\boldsymbol{Z} \le \boldsymbol{z}\}^{\mathsf{c}}$ as

$$B_{\boldsymbol{z}} = \{(R, \boldsymbol{W}) \in (0,\infty) \times \mathbb{S}_1 : R\boldsymbol{W} \not\le \boldsymbol{z}\} = \left\{(R, \boldsymbol{W}) \in (0,\infty) \times \mathbb{S}_1 : R > \min_{j=1}^D \frac{z_j}{W(\boldsymbol{s}_j)}\right\},$$

and it follows that (cf. Coles and Tawn, 1991)

$$V(\boldsymbol{z}) = D\int_{\mathbb{S}_1}\int_0^\infty \mathbf{1}\{\zeta > \min \boldsymbol{z}/\boldsymbol{w}\}\zeta^{-2}\mathrm{d}\zeta\rho_1(\mathrm{d}\boldsymbol{w}) = D\int_{\mathbb{S}_1}\max_{j=1}^{D}\frac{w(\boldsymbol{s}_j)}{z_j}\rho_1(\mathrm{d}\boldsymbol{w}). \tag{2.11}$$

The distribution function of the D-dimensional simple multivariate extreme value model is thus

$$\mathsf{P}\left(\boldsymbol{Z}^* \leq \boldsymbol{z}\right) = \exp\left(-\mathsf{E}\left[\max\left\{\frac{W(\boldsymbol{s}_1)}{z_1},\ldots,\frac{W(\boldsymbol{s}_D)}{z_D}\right\}\right]\right), \qquad \boldsymbol{z}\in\mathbb{R}_+^D\setminus\{\mathbf{0}_D\}. \tag{2.12}$$

An alternative characterization of the exponent measure uses a decomposition of the set $[\mathbf{0}_D,\boldsymbol{z})$ into disjoint regions in which the maximum is attained by component j ($j=1,\ldots,D$) in the expectation in eq. (2.12). This yields the formula (Huser and Davison, 2013)

$$V(\boldsymbol{z}) = \sum_{j=1}^{D}\frac{1}{z_j}\mathsf{E}\left(W(\boldsymbol{s}_j)\mathbf{1}\{W(\boldsymbol{s}_i)<z_iW(\boldsymbol{s}_j)/z_j, i\neq j\}\right). \tag{2.13}$$

Another formula for the exponent measure in terms of the intensity function is (cf. Ho, 2018, proof of Proposition 3.2),

$$V(\boldsymbol{u}) = \Lambda\{[\mathbf{0}_D,\boldsymbol{u})^{\mathsf{c}}\} = \int \mathbf{1}\{\boldsymbol{z}\not\leq\boldsymbol{u}\}\lambda(\boldsymbol{z})\mathrm{d}\boldsymbol{z},$$

where $\lambda(\boldsymbol{z})$ is the intensity function. Partition $\{\boldsymbol{z}\not\leq\boldsymbol{u}\}$ into $\bigsqcup_{i=1}^{D}\{z_i>u_i, z_j/u_j\leq z_i/u_i, j\neq i\}$. Using the $-(D+1)$-homogeneity of $\lambda(\cdot)$,

$$V(\boldsymbol{u}) = \sum_{i=1}^{D}\int_{\mathbb{R}^{D-1}}\int_{u_i}^\infty \mathbf{1}\{\boldsymbol{z}_{-i}/z_i\leq\boldsymbol{u}_{-i}/u_i\}z_i^{-1-D}\lambda(\boldsymbol{z}/z_i)\mathrm{d}z_i\mathrm{d}\boldsymbol{z}_{-i}$$

and making a change of variable $y_j = z_j/z_i$ for $j\in\{1,\ldots,D\}\setminus\{i\}$, one can rewrite the expression in terms of the D-vector $\boldsymbol{y}$ with jth component y_j and ith component 1, giving

$$\begin{aligned} V(\boldsymbol{u}) &= \sum_{i=1}^{D}\int_{\mathbb{R}^{D-1}}\int_{u_i}^\infty \mathbf{1}\{\boldsymbol{y}_{-i}\leq\boldsymbol{u}_{-i}/u_i\}\lambda(\boldsymbol{y})z_i^{-2}\mathrm{d}z_i\mathrm{d}\boldsymbol{y}_{-i} \\ &= \sum_{i=1}^{D}u_i^{-1}\int_{\mathbb{R}^{D-1}}\mathbf{1}\{\boldsymbol{y}_{-i}\leq\boldsymbol{u}_{-i}/u_i\}\lambda(\boldsymbol{y})\mathrm{d}\boldsymbol{y}_{-i}, \end{aligned} \tag{2.14}$$

which gives a symmetric form for the exponent measure.

Hierarchical kernel extreme value process

Oesting (2018) extends the spectral representation to a more general class of max-stable processes that encompasses the hierarchical kernel extreme value process of Reich and Shaby. Let $\boldsymbol{x}\in\mathbb{R}_+^\infty$ and let $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^\infty x_i^p\right)^{1/p}$ be the l_p norm for $p\in(1,\infty]$, the case $p=\infty$ being understood in the limiting sense as $\|\boldsymbol{x}\|_\infty = \max_{i=1}^\infty x_i$. The following result is Theorem 2.1 of

Oesting (2018).

Theorem 2.24 (ℓ_p representation of simple max-stable processes)
Let $\{\zeta_i\}_{i\in\mathbb{N}^+}$ be decreasing points of a Poisson point process on $\mathbb{R}_+$ with intensity measure $\zeta^{-2}\mathrm{d}\zeta$. Independently of $\{\zeta_i\}_{i\in\mathbb{N}^+}$, let $\mathsf{W}_1, \mathsf{W}_2, \ldots$, be independently and identically distributed copies of a stochastic process W on $\mathcal{S}$ with unit expectation for all $\boldsymbol{s} \in \mathcal{S}$. For $0 < p \leq \infty$ and any $\boldsymbol{s} \in \mathcal{S}$, let $A(\boldsymbol{s})$ be a collection of independent p-Fréchet variables with distribution function $\exp(-x^{-p})\mathbf{1}\{x \geq 1\}$, the case $p = \infty$ being understood to mean $A_i = 1$ almost surely. Let $\mathsf{Y}_i = \zeta_i \mathsf{W}_i$ for $i \in \mathbb{N}^+$ and define the corresponding sequence $\mathbf{Y} = \{\mathsf{Y}_i\}_{i=1}^{\infty}$. Then, the process

$$Z^*(\boldsymbol{s}) \overset{\mathrm{d}}{=} \frac{A(\boldsymbol{s})}{\Gamma(1-1/p)} \|\boldsymbol{Y}(\boldsymbol{s})\|_p, \qquad \boldsymbol{s} \in \mathcal{S}, \tag{2.15}$$

is simple max-stable.

We say that the max-stable process $\mathsf{Z}^*(\mathcal{S})$ admits an l_p representation if eq. (2.15) holds. A max-stable process with l_p representation also admits an l_q representation for $p < q < \infty$, but the converse requires further conditions (Oesting, 2018). The extremal coefficient $\theta(\cdot)$, defined in Section 2.7.3, is bounded below for l_p processes; in particular the pairwise extremal coefficient is bounded below by $2^{1/p}$, and is linked to a nugget in Reich and Shaby (2012). Oesting therefore terms the standard max-stable process, which corresponds to $p = \infty$, a denoised process.

2.2.2 Parametric models for max-stable processes

In the univariate case, the only limiting max-stable distribution is the generalized extreme value distribution. In the functional case, the de Haan spectral representation of max-stable processes shows that we can build max-stable processes using any spectral function W that satisfy the mean constraint of Theorem 2.21. We use log-Gaussian or truncated Gaussian processes as spectral functions to create parametric models for extremes.

Example 2.3 (Spectral representation of the Brown–Resnick process)
The Brown–Resnick process on $\mathcal{S} \subset \mathbb{R}^m$ associated to the semi-variogram γ (cf. Kabluchko et al., 2009) is formed by considering an intrinsically stationary log-Gaussian process $\mathsf{W}(\mathcal{S})$ with $\exp\{X(\boldsymbol{s}) - \gamma(\boldsymbol{s})\}$ where $X(\boldsymbol{s})$ is a sample-continuous Gaussian random field with stationary increments, mean zero and semi-variogram γ with $X(\boldsymbol{o}) = 0$ almost surely. Remarkably, although X only has stationary increments, the process

$$Z^*(\boldsymbol{s}_i) = \max_{i\in\mathbb{N}^+} \zeta_i \exp\left\{X(\boldsymbol{s}_i) - \gamma(\boldsymbol{o}, s)\right\}, \qquad \boldsymbol{s} \in \mathcal{S},$$

is simple max-stable, stationary and its law depends only on the variogram $\gamma(\cdot)$. The same is true if we start with a non-stationary Gaussian process with (necessarily finite) covariance function $\sigma^2(\boldsymbol{s})$. Kabluchko (2011) shows under conditions that the weak limit (in the sense of finite-dimensional distributions) of suitably rescaled triangular arrays of Gaussian processes is of Brown–Resnick type and max-stable, but that $\mathsf{Z}(\mathcal{S})$ is not necessarily stationary.

The Brown–Resnick process is named after Brown and Resnick (1977), who constructed a univariate process on $\mathbb{R}_+$ by taking a Brownian motion $B(s)$ and setting $W(s) = \exp\{B(s) - |s|/2\}$.

□

Example 2.4 (Spectral representation of the extremal Student process)
The max-stable attractor of elliptically contoured processes is the extremal Student process (Thibaud and Opitz, 2015), obtained by considering spectral functions of the form

$$W(\boldsymbol{s}) = c_\nu \max\{0, X(\boldsymbol{s})\}^\nu, \qquad \boldsymbol{s} \in \mathcal{S}, \tag{2.16}$$

for $\nu > 0$, where $\mathsf{X}(\mathcal{S})$ a stationary continuous Gaussian process with correlation function $\rho(\cdot)$, and c_ν a normalizing constant $\pi^{1/2}2^{1-\nu/2}[\Gamma\{(1+\nu)/2\}]^{-1}$. The extremal Student process with $\nu = 1$, $W(\boldsymbol{s}) = (2\pi)^{1/2}\max\{0, X(\boldsymbol{s})\}$ for $\boldsymbol{s} \in \mathcal{S}$, is called the Schlather model, following Schlather (2002). The power ν corresponds to the regular variation index of R. The extremal Student process includes the Brown–Resnick process as a limiting case when $\nu \to \infty$ (Nikoloulopoulos et al., 2009). In dimension D, the extremal Student distribution admits the stochastic representation $\boldsymbol{Y} = R\mathbf{A}\boldsymbol{\Omega} + \boldsymbol{\mu}$ with $\boldsymbol{\mu} = 1$, and where R is a nonnegative radial variable, i.e., $\mathsf{P}(R \geq 0) = 1$, and $\mathbf{A}$ is the Cholesky root of the correlation matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{A}\mathbf{A}^\top = \boldsymbol{\Sigma}$ (Opitz, 2013b). If R is regularly varying with positive tail index, then the process $\boldsymbol{Y}$ is also regularly varying by Breiman's lemma (Lemma 2.17).

□

Example 2.5 (Intensity and conditional intensity of Brown–Resnick processes)
Consider a log-Gaussian process $W(\boldsymbol{s}) = \exp\{X(\boldsymbol{s}) - \gamma(\boldsymbol{s})\}$, with $X(\boldsymbol{s})$ a zero-mean Gaussian process with stationary increments such that $W(\boldsymbol{o}) = 0$ almost surely. Define the column vector $\boldsymbol{\gamma} = [\gamma(\boldsymbol{o}, \boldsymbol{s}_1), \ldots, \gamma(\boldsymbol{o}, \boldsymbol{s}_D)]^\top$ and the quantities

$$\mathbf{A} = \boldsymbol{\Sigma}^{-1} - \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_D\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}, \qquad \mathbf{K}_{10} = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{O}_{D-k,k} \end{pmatrix}, \qquad \mathbf{K}_{01} = \begin{pmatrix} \mathbf{O}_{k,D-k} \\ \mathbf{I}_{D-k} \end{pmatrix},$$
$$\boldsymbol{\Gamma}_k^{-1} = \mathbf{K}_{01}^\top\mathbf{A}\mathbf{K}_{01}, \qquad \boldsymbol{\mu}_k = -\boldsymbol{\Gamma}_k\mathbf{K}_{01}^\top\mathbf{A}\mathbf{K}_{10}\log(\boldsymbol{z}_{1:k}) - \boldsymbol{\Gamma}_k\mathbf{K}_{01}^\top\left(\mathbf{A}\boldsymbol{\gamma} + \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}\right).$$

Dombry et al. (2013), § 2.2, give the intensity function of the Brown–Resnick process,

$$\lambda(\boldsymbol{z}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}(\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D)^{-1/2}}{(2\pi)^{(D-1)/2}\prod_{j=1}^D z_j}\exp\left\{-\frac{1}{2}\log(\boldsymbol{z})^\top\mathbf{A}\log(\boldsymbol{z}) - \left(\boldsymbol{\gamma}^\top\mathbf{A} + \frac{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}\right)\log(\boldsymbol{z})\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}^\top\mathbf{A}\boldsymbol{\gamma} + \frac{2\boldsymbol{\gamma}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D - 1}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}\right)\right\}, \qquad \boldsymbol{z} \in [\mathbf{0}_D, \infty_D) \setminus \{\mathbf{0}_D\}.$$

The conditional intensity of the last $D-k$ components coincides with the density of the log-Gaussian vector with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Gamma}_k$ (Wadsworth and Tawn, 2014,

§ 3.3),

$$\frac{|\boldsymbol{\Gamma}_k|^{-1/2}}{(2\pi)^{(D-k)/2}\prod_{j=k+1}^{D} z_i} \exp\left[-\frac{1}{2}\{\log(\boldsymbol{z}_{(k+1):D}) - \boldsymbol{\mu}_k\}^\top \boldsymbol{\Gamma}_k^{-1}\{\log(\boldsymbol{z}_{(k+1):D}) - \boldsymbol{\mu}_k\}\right].$$

The conditional intensity is readily obtained upon dividing the intensity for all D components by that of the first k, upon writing $\log(\boldsymbol{z}) = \mathbf{K}_{10}\log(\boldsymbol{z}_{1:k}) + \mathbf{K}_{01}\log(\boldsymbol{z}_{(k+1):D})$. Incidentally, Wadsworth and Tawn (2014) condition on the first k components, while the formulas presented in Dombry et al. (2013) condition on the last k. The matrix $\mathbf{A}$ is a singular nonnegative definite precision matrix, given that $\ker(\mathbf{A}) = \mathrm{span}\{\mathbf{1}_D\}$.

□

Example 2.6 (Intensity and conditional intensity of Brown–Resnick processes (2))
Engelke et al. (2015) condition explicitly on a site and consider the process in terms of the semivariogram $\gamma_{i,j} = \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ for $(i,j) \in \{1,\ldots,D\}^2$. We assume that the semivariogram has no nugget component, so that $\gamma(\boldsymbol{s},\boldsymbol{s}) = 0$ for all $\boldsymbol{s} \in \mathcal{S}$. Suppose without loss of generality that the D-vector of observations has an exceedance for site $\boldsymbol{s}_1$ and define the $(D-1)\times(D-1)$ covariance matrix $\boldsymbol{\Sigma}$ with entries $\Sigma_{ij} = \gamma_{i+1,1} + \gamma_{j+1,1} - \gamma_{i+1,j+1}$ $(1 \leq i \leq j \leq D-1)$ and let $\boldsymbol{x}$ be a $(D-1)$ vector with jth entry $\log(z_{j+1}/z_1) + \gamma_{1,(j+1)}$ $(j = 1,\ldots,D-1)$. We denote the inverse of the covariance matrix by $\boldsymbol{Q}$.

If the first k components exceed their marginal threshold $\boldsymbol{u}_{1:k}$, the conditional intensity of the last $D-k$ components conditional on the first site is (Asadi et al., 2015)

$$\lambda_{|1:k}(\boldsymbol{z}) = \frac{1}{z_1^2 z_2 \cdots z_k}\phi_{k-1}\left(\boldsymbol{x}_{1:(k-1)}, \boldsymbol{\Sigma}_{1:(k-1),1:(k-1)}\right)\Phi_{D-k}\left(\boldsymbol{\eta}_C, \boldsymbol{\Sigma}_C\right), \qquad 1 < k < D,$$

where $\boldsymbol{\Sigma}_C = \boldsymbol{Q}^{-1}_{(D-k):(D-1),(D-k):(D-1)}$ and

$$\boldsymbol{\eta}_C = \log\left(\frac{\boldsymbol{u}_{(D-k+1):D}}{z_1}\right) + \boldsymbol{\gamma}_{(D-k+1):D,1} - \boldsymbol{\Sigma}_{(D-k):(D-1),1:(k-1)}\boldsymbol{\Sigma}^{-1}_{1:(k-1),1:(k-1)}\boldsymbol{x}_{1:(k-1)}.$$

If $k = 1$, $\lambda_{|1}(\boldsymbol{z}) = z_1^{-2}\Phi_{D-1}\{\log(\boldsymbol{u}_{2:D}) - \log(z_1)\mathbf{1}_{D-1} + \boldsymbol{\gamma}_{2:D}, \boldsymbol{\Sigma}\}$ whereas, if $k = D$, the intensity function is $\lambda(\boldsymbol{z}) = z_1^{-1}\prod_{j=1}^{D} z_j^{-1}\phi_{D-1}(\boldsymbol{x},\boldsymbol{\Sigma})$. Since $\boldsymbol{Z}_{2:D} \mid Z_1$ is log-normal, the censored contribution is the integral of the conditional distribution of the last $D-k$ components up to $\log(\boldsymbol{u}_{(D-k+1):D})$, cf. Proposition C.2.

□

Example 2.7 (Intensity and conditional intensity of extremal Student processes)
The intensity function of the extremal Student distribution with correlation function $\boldsymbol{\Sigma}$ and ν degrees of freedom is (Ribatet, 2013, Appendix A)

$$\lambda(\boldsymbol{z}) = \frac{\left(\mathrm{sign}(\boldsymbol{z})|\boldsymbol{z}|^{1/\nu\top}\boldsymbol{\Sigma}^{-1}\mathrm{sign}(\boldsymbol{z})|\boldsymbol{z}|^{1/\nu}\right)^{-(\nu+D)/2}}{|\boldsymbol{\Sigma}|^{1/2}\pi^{(D-1)/2}\nu^{D-1}}\frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\prod_{j=1}^{D}|z_j|^{1/\nu-1}, \quad \boldsymbol{z} \in \mathbb{R}^D \setminus \{\mathbf{0}_D\}, \quad (2.17)$$

where $\mathrm{sign}(\boldsymbol{z})$ is the vector with ith entry $\mathrm{sign}(z_i) = -\mathbf{1}\{z_i<0\} + \mathbf{1}\{z_i>0\}$. The index ν is the exponent of the Gaussian process in the de Haan spectral representation of the extremal

Student process, whose spectral functions are of the form

$$\frac{\pi^{1/2}2^{1-\nu/2}}{\Gamma\{(1+\nu)/2\}}\mathsf{X}_+^\nu$$

for X a Gaussian process with correlation function $\rho(\cdot)$. The conditional intensity function of the $D-k$ last components given the first k coincides with that of $\boldsymbol{T}_+^\nu$, where $\boldsymbol{T}$ is a multivariate Student random vector with $\nu+k$ degrees of freedom. For $\boldsymbol{z} \in (\mathbf{0}_D, \infty_D)$, the conditional intensity of components $(z_{k+1},\ldots,z_D)$ given $(z_1,\ldots,z_k)$ is (Ribatet, 2013)

$$\lambda_{|1:k}(\boldsymbol{z}) = \frac{|\boldsymbol{\Omega}_k|^{-1/2}\prod_{j=k+1}^D z_j^{1/\nu-1}}{\nu^{(D-k)}\{\pi(\nu+k)\}^{(D-k)/2}}\frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{k+\nu}{2}\right)}\left\{1+\frac{\left(\boldsymbol{z}_{(k+1):D}^{1/\nu}-\boldsymbol{\eta}_k\right)^\top\boldsymbol{\Omega}_k^{-1}\left(\boldsymbol{z}_{(k+1):D}^{1/\nu}-\boldsymbol{\eta}_k\right)}{\nu+k}\right\}^{-(\nu+D)/2},$$

with conditional mean $\boldsymbol{\eta}_k = \boldsymbol{\Sigma}_{(k+1):D,1:k}\boldsymbol{\Sigma}_{1:k}^{-1}\boldsymbol{z}_{1:k}^{1/\nu}$ and conditional variance

$$\boldsymbol{\Omega}_k = \frac{\nu+\boldsymbol{z}_{1:k}^{1/\nu\top}\boldsymbol{\Sigma}_{1:k}^{-1}\boldsymbol{z}_{1:k}^{1/\nu}}{k+\nu}\left(\boldsymbol{\Sigma}_{(k+1):D}-\boldsymbol{\Sigma}_{(k+1):D,1:k}\boldsymbol{\Sigma}_{1:k,1:k}^{-1}\boldsymbol{\Sigma}_{1:k,(k+1):D}\right).$$

Suppose without loss of generality that the components of $\boldsymbol{z}$ are ordered from largest to smallest; the extremal Student process places mass on subsets of $\mathbb{R}_+^D\setminus\{\mathbf{0}_D\}$ of Lebesgue measure zero, i.e., on sets of the form $\left\{\boldsymbol{z}\in\mathbb{R}_+^D\setminus\{\mathbf{0}_D\}:\boldsymbol{z}_{1:k}>\mathbf{0}_k,\boldsymbol{z}_{(k+1):D}=\mathbf{0}_{D-k}\right\}$ for $1\leq k<D$. The restricted intensity function for vectors lying on the boundary of $\mathbb{R}_+^D\setminus\{\mathbf{0}_D\}$ is obtained by integrating λ over negative components from $-\infty$ to 0, yielding (Thibaud and Opitz, 2015, eq. 14)

$$\lambda_{|1:k}^+(\boldsymbol{z}) = \frac{\mathsf{St}_{k+\nu}\left(\mathbf{0}_{D-k-1};\boldsymbol{\eta}_k,\boldsymbol{\Omega}_k\right)}{\nu^{k-1}\pi^{(k-1)/2}|\boldsymbol{\Sigma}_{1:k}|^{1/2}}\frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\left(\prod_{j=1}^k z_j\right)^{1/\nu-1}\left(\boldsymbol{z}_{1:k}^{1/\nu\top}\boldsymbol{\Sigma}_{1:k}^{-1}\boldsymbol{z}_{1:k}^{1/\nu}\right)^{-(\nu+k)/2},\quad \boldsymbol{z}\in\mathbb{R}_+^D\setminus\{\mathbf{0}_D\}.$$

For an observation $\boldsymbol{z}$ with ordered components, if $\boldsymbol{z}_{1:k}>\mathbf{0}_k$ and $\boldsymbol{z}_{(k+1):D}=\mathbf{0}_{D-k}$, the density is $\lambda_{|1:k}^+(\boldsymbol{z}_{1:k},\mathbf{0}_{D-k})$.

□

Example 2.8 (Exponent measure of the Brown–Resnick process)
Consider sites $\{\boldsymbol{s}_j, j=1,\ldots,D\}$ and a semivariogram $\gamma(\cdot)$ such that $\gamma(\boldsymbol{o})=0$. We define $\boldsymbol{\Gamma}$ as the $D\times D$ matrix with entries $\gamma_{i,j}=\gamma(\boldsymbol{s}_i-\boldsymbol{s}_j)/2$ for $i,j=1\ldots,D$ and let $\boldsymbol{\Lambda}$ be the matrix of squared coefficients $\{\lambda_{ij}^2\}_{i,j=1}^D$, where $\lambda_{ij}^2=\gamma(\boldsymbol{s}_i,\boldsymbol{s}_j)/2$ for a semi-variogram function $\gamma(\cdot,\cdot)$ (Engelke et al., 2015). Huser and Davison (2013) showed that the exponent measure of the Brown–Resnick process can be written as

$$\begin{aligned}V(\boldsymbol{z}) &= \sum_{j=1}^D\frac{1}{z_j}\Phi_{D-1}\left(\log(\boldsymbol{z}_{-j})-\log(z_j\mathbf{1}_{D-1});-\boldsymbol{\Gamma}_{-j,j},\boldsymbol{\Gamma}_{-j,j}\mathbf{1}_{D-1}^\top+\mathbf{1}_{D-1}\boldsymbol{\Gamma}_{j,-j}-\boldsymbol{\Gamma}_{-j,-j}\right)\\ &= \sum_{j=1}^D\frac{1}{z_j}\Phi_{D-1}\left(\boldsymbol{\lambda}_{.j}+\frac{1}{2\boldsymbol{\lambda}_{.j}}\circ\log\left(\frac{\boldsymbol{z}_{-j}}{z_j}\right);\mathbf{0}_{D-1},\mathbf{R}_{-j}\right),\end{aligned}$$

where $\circ$ denotes the Hadamard product and the partial correlation matrix $\mathbf{R}_{-i}$ has elements

$$\varrho_{j,k;i} = \frac{\lambda_{ij}^2 + \lambda_{ik}^2 - \lambda_{jk}^2}{2\lambda_{ij}\lambda_{ik}}, \qquad j,k \in \{1,\dots,D\} \setminus \{i\}.$$

□

Remark 2.5 (Links between the different parametrizations of the Brown–Resnick model)
Wadsworth and Tawn (2014) express the Brown–Resnick process as a function of a covariance matrix $\boldsymbol{\Sigma}$ which, for a collection of D sites, has elements $\sigma_{ij} = \gamma(\boldsymbol{s}_i, \boldsymbol{o}) + \gamma(\boldsymbol{s}_j, \boldsymbol{o}) - \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ with $\sigma_i^2 \equiv \sigma_{i,i}$. This yields an alternative formula for the exponent measure,

$$V(\boldsymbol{z}) = \sum_{i=1}^{D} \frac{1}{z_i} \Phi_{D-1}\left(\log\left(\frac{\boldsymbol{z}_{-i}}{z_i}\right); -\frac{\boldsymbol{\sigma}_{-i}^2}{2} - \frac{\sigma_i^2}{2} + \boldsymbol{\sigma}_{i,-i}, \mathbf{T}^{(i)}\boldsymbol{\Sigma}\mathbf{T}^{(i)\top}\right),$$

where $\mathbf{T}^{(i)}$ is a $(D-1)\times D$ transformation matrix with $\mathbf{T}^{(i)}_{,-i} = \mathbf{I}_{D-1}$ and $\mathbf{T}^{(i)}_{,i} = -\mathbf{1}_{D-1}$.

Define the column vector $\boldsymbol{\gamma}_{\boldsymbol{o}} = (\gamma(\boldsymbol{o}, \boldsymbol{s}_1), \dots, \gamma(\boldsymbol{o}, \boldsymbol{s}_D))^\top$ and the quantities (Dombry et al., 2013)

$$\mathbf{A} = \boldsymbol{\Sigma}^{-1} - \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_D\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}, \qquad \mathbf{L} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_{\boldsymbol{o}} - \frac{1 - \mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_{\boldsymbol{o}}}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}\boldsymbol{\Sigma}^{-1}\mathbf{1}_D.$$

The expressions in Huser and Davison (2013), Wadsworth and Tawn (2014), Engelke et al. (2015) and Dombry et al. (2013) can be related upon noting that

$$\begin{aligned}(\mathbf{A}_{-i})^{-1}\mathbf{L}_{-i} = \boldsymbol{\Gamma}_{-i,i} &= \frac{\boldsymbol{\sigma}_{-i}^2}{2} + \frac{\sigma_i^2}{2} - \boldsymbol{\sigma}_{i,-i},\\ (\mathbf{A}_{-i})^{-1} &= \boldsymbol{\Gamma}_{-i,i}\mathbf{1}_{D-1}^\top + \mathbf{1}_{D-1}\boldsymbol{\Gamma}_{i,-i} - \boldsymbol{\Gamma}_{-i,-i} = \mathbf{T}^{(i)}\boldsymbol{\Sigma}\mathbf{T}^{(i)\top}.\end{aligned}$$

Example 2.9 (Exponent measure of the extremal Student model)
The exponent measure of the extremal Student model first appeared in Demarta and McNeil (2005) in the bivariate case and in Nikoloulopoulos et al. (2009), Theorem 2.3, in the general case. Let $\boldsymbol{\Sigma}$ be a correlation matrix with inverse $\boldsymbol{\Sigma}^{-1} = \mathbf{Q}$, say. Then, the exponent measure of the extremal Student model is (Opitz, 2013a)

$$V(\boldsymbol{z}) = \sum_{i=1}^{D} \frac{1}{z_i} \mathsf{St}_{D-1}\left(\left(\frac{\boldsymbol{z}_{-i}}{z_i}\right)^{1/\nu}; \boldsymbol{\Sigma}_{-i,i}, (\nu+1)^{-1}\left(\boldsymbol{\Sigma}_{-i,-i} - \boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{i,-i}\right), \nu+1\right).$$

□

Proof
We follow the argument leading to Equation (2.14). Write

$$V(\boldsymbol{u}) = \sum_{i=1}^{D} \int_{\mathbb{R}^{D-1}} \mathbf{1}\{\boldsymbol{z}_{-i}/z_i \le \boldsymbol{u}_{-i}/u_i\} \int_{u_i}^{\infty} z_i^{-1-D} \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{|\mathbf{Q}|^{1/2}}{\pi^{(D-1)/2}\nu^{D-1}}$$

$$\times \left\{ \left(\frac{\boldsymbol{z}}{z_i}\right)^{1/\nu^\top} \mathbf{Q} \left(\frac{\boldsymbol{z}}{z_i}\right)^{1/\nu} \right\}^{-(\nu+D)/2} \prod_{\substack{j=1 \\ j\neq i}}^{D} \left(\frac{z_j}{z_i}\right)^{1/\nu-1} \mathrm{d}z_i \mathrm{d}\boldsymbol{z}_{-i}.$$

Make the change of variable $y_j = (z_j/z_i)^{1/\nu}$ for $j \neq i$, $j = 1, \ldots, D$, and write $\boldsymbol{y}$ for the D-vector with jth coordinate y_j and ith coordinate 1; noting that $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{-i,-i} - \boldsymbol{\Sigma}_{-i,i}\Sigma_{i,i}\boldsymbol{\Sigma}_{i,-i}||\Sigma_{i,i}| = |\mathbf{Q}_{-i,-i}|^{-1}$ because $\boldsymbol{\Sigma}$ is a correlation matrix, we get

$$\begin{aligned} V(\boldsymbol{u}) = \sum_{i=1}^{D} \int_{\mathbb{R}^{D-1}} \int_{u_i}^{\infty} \mathbf{1}\{\boldsymbol{y}_{-i} \leq (\boldsymbol{u}_{-i}/u_i)^{1/\nu}\} z_i^{-2} \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{\Sigma_{i,i}^{-1/2} |\mathbf{Q}_{-i,-i}|^{1/2}}{\pi^{(D-1)/2}} c_i^{-(\nu+D)/2} \\ \times \left[1 + \frac{(\boldsymbol{y}_{-i} - \mathbf{Q}_{-i}^{-1}\mathbf{Q}_{-i,i})^\top (\nu+1)\mathbf{Q}_{-i}/c_i (\boldsymbol{y}_{-i} - \mathbf{Q}_{-i}^{-1}\mathbf{Q}_{-i,i})}{\nu+1} \right]^{-(\nu+D)/2} \mathrm{d}z_i \mathrm{d}\boldsymbol{y}_{-i}, \end{aligned}$$

where we expanded the quadratic form $\boldsymbol{y}^\top \mathbf{Q} \boldsymbol{y}$ into $\mathrm{Q}_{i,i} + 2\boldsymbol{y}_{-i}^\top \mathbf{Q}_{-i,i} + \boldsymbol{y}_{-i}^\top \mathbf{Q}_{-i,-i} \boldsymbol{y}_{-i}$ and factored $c_i = \mathrm{Q}_{i,i} - \mathbf{Q}_{i,-i}\mathbf{Q}_{-i,-i}\mathbf{Q}_{-i,i} = \Sigma_{i,i}^{-1} = 1$ out of the expression in square brackets. Finally,

$$V(\boldsymbol{u}) = \sum_{i=1}^{D} \frac{1}{u_i} \mathsf{St}_{D-1}\left(\left(\frac{\boldsymbol{u}_{-i}}{u_i}\right)^{1/\nu}; -\mathbf{Q}_{-i}^{-1}\mathbf{Q}_{-i,i}, (\nu+1)^{-1}\mathbf{Q}_{-i}^{-1}, \nu+1 \right).$$

This expression coincides with eq. 11 of Thibaud and Opitz (2015).

■

2.2.3 Parametric models for multivariate extreme value distributions

Many parametric models exist in the D-variate case: we limit the presentation to those most widely used for modelling. Multivariate extreme value distributions are often presented in unit Fréchet margins in terms of their distribution function $H(\boldsymbol{z}) = \exp\{-V(\boldsymbol{z})\}$, so we report the exponent measure $V(\boldsymbol{z})$.

Example 2.10 (Exponent measure of the logistic model)
The exponent measure of the logistic model of Gumbel (1960) with dependence parameter $\alpha \in (0, 1]$ is

$$V(\boldsymbol{z}) = \left\{ \sum_{j=1}^{D} \left(\frac{1}{z_j}\right)^{1/\alpha} \right\}^{\alpha}, \qquad \boldsymbol{z} \in \mathbb{R}_+^D,$$

and the derivative of $V(\boldsymbol{z})$ with respect to the first k components is (cf. Castruccio et al., 2016)

$$V_{1:k}(\boldsymbol{z}) = -\alpha^{-k} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha-k+1)} \prod_{j=1}^{k} z_j^{-1/\alpha-1} \left(\sum_{j=1}^{D} z_j^{-1/\alpha} \right)^{\alpha-k}.$$

Perfect dependence is obtained in the limit as $\alpha \to 0$, whereas $\alpha = 1$ yields the independence model. The logistic multivariate extreme value distribution is popular for theoretical purposes

due to its simplicity. As a consequence of exchangeability, the mean squared error of the maximum likelihood estimator $\widehat{\alpha}$ is proportional to $n^{-1}D^{-1}$ if the marginal parameters are known (Hofert et al., 2012). The score and Fisher information are given in Shi (1995).

□

Example 2.11 (Exponent measure of the asymmetric logistic model)
Let $\mathbb{P}_D$ be the collection of all nonempty subsets of $\{1,\dots,D\}$. The exponent measure of the asymmetric logistic model (Tawn, 1990) with asymmetry parameters $\theta_{i,b} \in [0,1]$ satisfying $\sum_{b\in\mathbb{P}_D}\theta_{i,b} = 1$ for $i = 1,\dots,D$ and dependence parameters $0 < \alpha_b \le 1$ is

$$V(\boldsymbol{z}) = \sum_{b\in\mathbb{P}_D}\left(\sum_{i\in b}\left(\frac{\theta_{i,b}}{z_i}\right)^{1/\alpha_b}\right)^{\alpha_b}, \qquad \boldsymbol{z}\in\mathbb{R}_+^D.$$

Stephenson (2003) gives some insight into the construction of the asymmetric logistic max-stable distribution, which is a max-mixture of logistic max-stable vectors. Consider sampling $\boldsymbol{Z}_b$ from a logistic distribution of dimension $|b|$ (or Fréchet variates if $|b| = 1$) with parameter α_b). Each marginal value corresponds to the maximum of the corresponding weighted entries, i.e., $X_j = \max_{b\in\mathbb{P}_D}\theta_{j,b}Z_{j,b}$ for all $j = 1,\dots,D$. The asymmetric logistic model is rarely used in high dimensions without further constraints, as it is overparametrized.

□

Example 2.12 (Exponent measure of the negative logistic model)
The exponent measure of the negative logistic distribution with parameter $\alpha \le 0$ is (Galambos, 1975)

$$V(\boldsymbol{z}) = \sum_{b\in\mathbb{B}}(-1)^{|b|}\left(\sum_{i\in b} z_i{}^{-\alpha}\right)^{\frac{1}{\alpha}}, \qquad \boldsymbol{z}\in\mathbb{R}_+^D.$$

□

Example 2.13 (Exponent measure of the asymmetric negative logistic distribution)
Joe (1990) mentions the asymmetric negative logistic model as a generalization of Galambos' negative logistic model. It is constructed in the same way as the asymmetric logistic distribution; see Theorem 1 in Stephenson (2003). Let $\alpha_b \le 0$ for all $b \in \mathbb{B}_D$ and $\theta_{i,b} \ge 0$ with $\sum_{b\in\mathbb{B}_D}\theta_{i,b} = 1$; the exponent measure is

$$V(\boldsymbol{z}) = \sum_{b\in\mathbb{B}}\sum_{c\in b}(-1)^{|c|}\left\{\sum_{i\in c}\left(\frac{\theta_{i,b}}{z_i}\right)^{\alpha_b}\right\}^{\frac{1}{\alpha_b}}, \qquad \boldsymbol{z}\in\mathbb{R}_+^D. \tag{2.18}$$

This does not correspond to the "negative logistic distribution" given in §4.2 of Coles and Tawn (1991) or §3.5.3 of Kotz and Nadarajah (2000), which is not a valid distribution function in dimension $D \ge 3$ as the constraints therein on the parameters $\theta_{i,b}$ are necessary but not sufficient to yield a valid distribution function. The proof that $\exp\{-V(\boldsymbol{z})\}$, with $V(\boldsymbol{z})$ as in eq. (2.18) for $\boldsymbol{z} \in [\mathbf{1}_D, \infty_D)$, is a valid distribution follows from the fact that it is a max-mixture (Stephenson, 2003, Theorem 1).

□

Some models are more easily defined in terms of the angular (or spectral) density $\rho_1(\boldsymbol{w})$ given in Equation (2.11).

Example 2.14 (Spectral density of the multilogistic model)
This multivariate extension of the logistic model, proposed by Boldi (2009), places mass on the interior of the simplex. Let $\boldsymbol{W} \in \mathbb{S}_1$ be the solution of

$$\frac{W_j}{W_D} = \frac{C_j U_j^{-\alpha_j}}{C_D U_D^{-\alpha_D}}, \quad j = 1, \ldots, D,$$

where $C_j = \Gamma(D-\alpha_j)/\Gamma(1-\alpha_j)$ for $j = 1, \ldots, D$ and $\boldsymbol{U} \in \mathbb{S}_1$ follows a D-mixture of Dirichlet with the jth component being $\mathsf{Dir}(\mathbf{1}_D - \delta_j \alpha_j)$, so that the mixture has density function

$$h_{\boldsymbol{U}}(\boldsymbol{u}) = \frac{1}{D} \sum_{j=1}^{D} \frac{\Gamma(D-\alpha_j)}{\Gamma(1-\alpha_j)} u_j^{-\alpha_j}, \qquad 0 < \alpha_j < 1, j = 1, \ldots, D.$$

The angular density, defined for $\boldsymbol{\alpha} \in [0,1]^D$, is

$$\rho_1(\boldsymbol{w}) = \frac{1}{D} \left(\sum_{j=1}^{D} \alpha_j u_j \right)^{-1} \left(\prod_{j=1}^{D} \alpha_j u_D \right) \left(\sum_{j=1}^{D} \frac{\Gamma(D-\alpha_j)}{\Gamma(1-\alpha_j)} u_j^{-\alpha_j} \right) \prod_{j=1}^{D} w_j^{-1}, \qquad \boldsymbol{w} \in \mathbb{S}_1.$$

□

The following max-stable distribution, the scaled extremal Dirichlet model, is taken from Belzile and Nešlehová (2017) and encompasses three widely used parametric families of multivariate extreme value distributions (logistic, negative logistic and Coles–Tawn extremal Dirichlet). We give its spectral representation, its intensity function λ and its angular density ρ_1.

Example 2.15 (Spectral representation of the scaled extremal Dirichlet max-stable model)
Let $a, c > 0$ and $b \neq 0$; if $G \sim \mathsf{Ga}(c,1)$ is Gamma with shape parameter c, we say $aG^{1/b} \sim \mathsf{sGa}(a,b,c)$ has a scaled Gamma distribution with density

$$f(y; a, b, c) = \frac{|b|}{\Gamma(c)} a^{-bc} y^{bc-1} \exp\left\{ -\left(\frac{y}{a}\right)^b \right\}, \qquad y > 0. \tag{2.19}$$

Consequently, $\mathsf{E}(Y) = a\Gamma(c+1/b)/\Gamma(c) < \infty$ when $b < -1/c$. The scaled Gamma family includes several well-known distributions as special cases, notably the Gamma when $b = 1$, the Weibull when $c = 1$ and $b > 0$, the inverse Gamma when $b = -1$, and the Fréchet when $c = 1$ and $b < 0$. When $b > 0$, the scaled Gamma is the generalized Gamma distribution of Stacy (1962), albeit in a different parametrization.

Consider the parameter vector $\boldsymbol{\alpha} > \mathbf{0}_D$, $\bar{\alpha} = \|\boldsymbol{\alpha}\|_1$ and $\kappa > -\min(\alpha_1, \ldots, \alpha_D)$, $\kappa \neq 0$ and independent scaled Gamma variables $V_i \sim \mathsf{sGa}\{1/c(\alpha_i, \kappa), 1/\kappa, \alpha_i\}$ with $c(\alpha, \kappa) = \Gamma(\alpha+\kappa)/\Gamma(\alpha)$ for $\alpha > 0$. If $\boldsymbol{G}$ is a random vector with independent Gamma components, $G_i \sim \mathsf{Ga}(\alpha_i, 1)$ then

for all $i = 1,\ldots,D$, $V_i \stackrel{\mathrm{d}}{=} G_i^\kappa / c(\alpha_i,\kappa)$. Furthermore, $\|\boldsymbol{G}\| \sim \mathsf{G}\mathsf{a}(\bar{\alpha},1)$ is independent of $\boldsymbol{G}/\|\boldsymbol{G}\|$, which has the same distribution as the Dirichlet vector $\boldsymbol{D}_{\boldsymbol{\alpha}} = (D_1,\ldots,D_D)$. The requirement that $\kappa > -\min(\alpha_1,,\ldots,,\alpha_D)$ ensures that the expectation of G_i^κ is finite for all $i \in \{1,\ldots,D\}$ and thus that $\mathsf{E}(\boldsymbol{V}_i) = \mathbf{1}_D$. For $\boldsymbol{z} \in \mathbb{R}_+^D$, the exponent measure is

$$V(\boldsymbol{z};\kappa,\boldsymbol{\alpha}) = \mathsf{E}\left\{\max_{i=1}^D \frac{V_i}{z_i}\right\} = \mathsf{E}\left[\max_{i=1}^D \left\{\frac{G_i^\kappa}{z_i c(\alpha_i,\kappa)}\right\}\right] = \mathsf{E}\left(\|\boldsymbol{G}\|^\kappa\right)\mathsf{E}\left[\max_{i=1}^D \left\{\frac{D_i^\kappa}{z_i c(\alpha_i,\kappa)}\right\}\right], \tag{2.20}$$

and $\mathsf{E}(\|\boldsymbol{G}\|^\kappa) = c(\bar{\alpha},\kappa)$. Equation (2.20) implies that the scaled extremal Dirichlet model admits the de Haan spectral representation (2.8) with $\boldsymbol{W}_i$ a sequence of D-vectors with independent components with distribution $W_{ij} \sim \mathsf{sG}\mathsf{a}\{1/c(\alpha_j,\kappa), 1/\kappa, \alpha_j\}$ for $j = 1,,\ldots,,D$.

□

Example 2.16 (Intensity of the scaled extremal Dirichlet distribution)
Using the spectral representation given in Example 2.15, consider independent scaled Gamma components $Z_i \sim \mathsf{sG}\mathsf{a}\{\Gamma(\alpha_i+\kappa)/\Gamma(\alpha_i), \kappa^{-1}, \alpha_i\}$, where $a_i = \Gamma(\alpha_i+\kappa)/\Gamma(\alpha_i)$, $\boldsymbol{\alpha} > \mathbf{0}_D$ and $\kappa \geq -\min_{i=1}^D \alpha_i$. The intensity function of the scaled Dirichlet model is

$$\lambda(\boldsymbol{z}) = \frac{\Gamma\left(\sum_{j=1}^D \alpha_j + \kappa\right)}{|\kappa|^{D-1}} \left[\sum_{j=1}^D \left\{\frac{z_j\Gamma(\alpha_j+\kappa)}{\Gamma(\alpha_j)}\right\}^{1/\kappa}\right]^{-\sum_{j=1}^D \alpha_j - \kappa} \prod_{j=1}^D \frac{z_j^{\alpha_j/\kappa-1}}{\Gamma(\alpha_j)} \left\{\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+\kappa)}\right\}^{-\alpha_j/\kappa-1}.$$

We retrieve the Coles and Tawn Dirichlet model if $\kappa = 1$, the logistic model if $\boldsymbol{\alpha} = \mathbf{1}_D, \kappa \in [-1,0)$ and the negative logistic model if $\boldsymbol{\alpha} = \mathbf{1}_D, \kappa > 0$ (Belzile and Nešlehová, 2017).

□

Proof
Write $a_i = \Gamma(\alpha_i+\kappa)/\Gamma(\alpha_i)$; the intensity function is

$$\lambda(\boldsymbol{z}) = \int_0^\infty |\kappa|^{-D} \prod_{j=1}^D \frac{a_j^{\alpha_j/\kappa}}{\Gamma(\alpha_j)} \left(\frac{z_j}{\zeta}\right)^{\alpha_j/\kappa-1} \exp\left\{-\zeta^{-1/\kappa}\sum_{j=1}^D \left(\frac{z_j}{a_j}\right)^{1/\kappa}\right\} \zeta^{-D-2}\mathrm{d}\zeta$$

and making the change of variable $\zeta^{-1/\kappa} = u$ yields

$$\begin{aligned}\lambda(\boldsymbol{z}) &= |\kappa|^{-D+1} \prod_{j=1}^D \frac{a_j^{\alpha_j/\kappa}}{\Gamma(\alpha_j)} z_j^{\alpha_j/\kappa-1} \int_0^\infty u^{\sum_{j=1}^D \alpha_j + \kappa - 1} \exp\left\{-u\sum_{j=1}^D \left(\frac{z_j}{a_j}\right)^{1/\kappa}\right\} \mathrm{d}u \\ &= |\kappa|^{-D+1} \prod_{j=1}^D \frac{a_j^{\alpha_j/\kappa}}{\Gamma(\alpha_j)} z_j^{\alpha_j/\kappa-1} \Gamma\left(\sum_{j=1}^D \alpha_j + \kappa\right) \left\{\sum_{j=1}^D \left(\frac{z_j}{a_j}\right)^{1/\kappa}\right\}^{\sum_{j=1}^D \alpha_j + \kappa},\end{aligned}$$

after recovering a Gamma integral.

■

Example 2.17 (Spectral density of the (scaled) extremal Dirichlet model)
The extremal Dirichlet model of Coles and Tawn (1991) with parameters $\boldsymbol{\alpha} > \mathbf{0}_D$ is similarly

defined in terms of the angular density, with

$$\rho_1(\boldsymbol{w}) = \frac{1}{D}\frac{\Gamma\left(1+\sum_{j=1}^D \alpha_j\right)}{\prod_{j=1}^D \alpha_j w_j}\left(\sum_{j=1}^D \alpha_j w_j\right)^{-(D+1)} \prod_{j=1}^D \alpha_j \prod_{j=1}^D \left(\frac{\alpha_j w_j}{\sum_{k=1}^D \alpha_k w_k}\right)^{\alpha_j - 1}, \qquad \boldsymbol{w} \in \mathbb{S}_1.$$

This is one of the few models, outside of the asymmetric logistic models, that allows asymmetry. The scaled extremal Dirichlet model (Belzile and Nešlehová, 2017) introduced in Example 2.15 and whose intensity function is given in Example 2.16 includes the logistic, negative logistic and extremal Dirichlet models as special cases. For $\kappa > -\min(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha} \in \mathbb{R}_+^D$, the angular density of the scaled extremal Dirichlet model is

$$\rho_1(\boldsymbol{w}) = \frac{\Gamma(\bar{\alpha}+\kappa)}{D\kappa^{D-1}\prod_{i=1}^D \Gamma(\alpha_i)}\left\langle \{\boldsymbol{c}(\boldsymbol{\alpha},\kappa)\}^{1/\kappa}, \boldsymbol{w}^{1/\kappa}\right\rangle^{-\kappa-\bar{\alpha}} \prod_{i=1}^D \{c(\alpha_i,\kappa)\}^{\alpha_i/\kappa} w_i^{\alpha_i/\kappa - 1}, \qquad \boldsymbol{w} \in \mathbb{S}_1,$$

where $\boldsymbol{c}(\boldsymbol{\alpha},\kappa)$ is the D-vector with entries $\Gamma(\alpha_i+\kappa)/\Gamma(\alpha_i)$ for $i = 1,\ldots,D$ and $\langle \mathbf{a},\mathbf{b}\rangle = \mathbf{a}^\top\mathbf{b}$ denotes the inner product between two D-vectors.

When $\kappa = 1$, the scaled extremal Dirichlet model becomes the Coles–Tawn Dirichlet extremal model (Segers, 2012). When $\boldsymbol{\alpha} = \mathbf{1}_D$ and $\kappa > 0$, the model reduces to the negative logistic model (Dombry et al., 2016). Similarly, when $\kappa < 0$ and $\boldsymbol{\alpha} = \mathbf{1}_D$, the scaled extremal Dirichlet model becomes the logistic model (cf. Appendix A.2.4 of Dombry et al., 2016).

□

The scaled extremal Dirichlet and multilogistic models have a dependence structure derived from Dirichlet distributions, and have fewer than one parameter per variable. Alternative multivariate models derived from elliptical distributions yield more flexibility as they include potentially one parameter per pair of sites.

Example 2.18 (Hüsler–Reiss distribution)

Provided that the pairwise correlation between variables $|\rho_{ij}| <$ for $i \neq j$, the max-stable attractor of Gaussian processes is the independence max-stable process, which has D-variate exponent measure $V(\boldsymbol{z}) = \sum_{j=1}^D z_j^{-1}$: a different limiting model is obtained if we increase the dependence between sites with n at a specified rate: the Hüsler–Reiss class of multivariate extreme value distributions arise as limiting distributions of maxima of normalized triangular arrays of standardized zero-mean unit variance D-dimensional Gaussian variables whose correlation matrix $\boldsymbol{\Sigma}_n$ satisfies

$$\boldsymbol{\Lambda} = \lim_{n\to\infty} \log(n)(\mathbf{1}_D^\top \mathbf{1}_D - \boldsymbol{\Sigma}_n) \in \mathscr{D}$$

where

$$\mathscr{D} = \left\{ \mathbf{A} \in \mathbb{R}_+^{D\times D} : \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} < 0 \ \forall\, \boldsymbol{x} \in \mathbb{R}^D \setminus \{\mathbf{0}\} \text{ with } \sum_{i=1}^D x_i = 0, a_{ij} = a_{ji}, a_{ii} = 0, i,j \in \{1,\ldots,D\} \right\}.$$

The D-dimensional realizations of Brown–Resnick max-stable processes associated to a semi-variogram γ are Hüsler–Reiss distributed with parameter matrix $\boldsymbol{\Lambda} = [\gamma_{ij}/2)^{1/2}]_{i,j=1}^D \in \mathscr{D}$ for

any $D > 2$ since the semi-variogram is a negative definite function.

□

2.3 ℓ-Pareto processes and generalizations

Max-stable processes are the functional generalization of the generalized extreme value distribution. The spectral representation in Theorem 2.20 shows that max-stable processes arise as the pointwise maximum of an infinite collection of random functions $\varphi_i = \zeta_i W_i$, where ζ_i gives the intensity of the storm and the spectral functions W_i are the profile of the storm over space (Smith, 1990). The individual events φ_i may be of interest themselves, and we turn to threshold exceedances for the analysis of φ_i that are extreme in some sense, looking at the functional equivalent of peaks-over-threshold.

While exceedances are unambiguous in the univariate framework because $\mathbb{R}_+$ is ordered, many possible definitions could be adopted in the multivariate framework due to the lack of ordering (Barnett, 1976). One could be interested in events consisting of marginal exceedances at all sites, or for at least one site, or at large cumulated values over the domain. We consider exceedances defined in terms of a functional ℓ, that may be context-dependent.

Definition 2.25 (Risk functional)
A risk functional is a nonnegative 1-homogeneous continuous functional $\ell : \mathcal{F}^+ \setminus \mathscr{C} \mapsto [0,\infty)$, i.e., $\ell(tf) = t\ell(f)$ for all $t > 0$ (Dombry and Ribatet, 2015).

By definition, regular variation means there exists a non-decreasing sequence $a_n > 0$ such that

$$\Lambda_n(\cdot) = n\mathsf{P}\left(X/a_n \in \cdot\right) \xrightarrow{\mathrm{W}^\#} \Lambda(\cdot), \qquad n \to \infty.$$

Describing regular variation in terms of weak-hash convergence of measures in $\mathbb{M}_{\mathcal{F}_\mathscr{C}}$ is mathematically accurate, but also amounts to weak convergence of conditional distributions and the latter is easier to handle (Meinguet and Segers, 2010). We consider sets bounded away from the closed cone $\mathscr{C}_\ell = \{f \in \mathcal{F}^+ : \ell(f) = 0\}$ and threshold exceedances $\{\ell(X) \geq u_n\}$, where u_n is an nondecreasing sequence of thresholds such that $\mathsf{P}\{\ell(X) \geq u_n\} \to 0$ as $n \to \infty$. We restrict the region of interest to $\{f \in \mathcal{F}^+_\mathscr{C} : \ell(f) > 0\}$: if we have convergence of Λ_n to Λ on $\mathcal{F}^+ \setminus \mathscr{C}$ for some cone $\mathscr{C}$, then applying the continuous mapping theorem with the function $T(f) = f\mathbf{1}\{\ell(f)>0\}$ gives convergence of the pushforward on $\mathcal{F}^+ \setminus \mathscr{C}_\ell$, where $\mathscr{C}_\ell = \mathscr{C} \cup \{f \in \mathcal{F}^+ : \ell(x) = 0\}$.

We consider now a generalized pseudo-polar representation with radial component $\ell(\cdot)$.

Definition 2.26 (Generalized pseudo-polar transformation)
Let ℓ be a continuous homogeneous risk functional and let $\mathcal{F}_\mathscr{C}$ be $\mathcal{F} \setminus \mathscr{C}$. The generalized pseudo-polar transformation considered in Opitz (2013b) and in de Fondeville and Davison (2018) is the mapping

$$T : \mathcal{F}_\mathscr{C} \to [0,\infty) \times \mathbb{S}_{\mathrm{ang}} \setminus T(\mathscr{C}), \qquad T(f) = (r,\omega) = (\ell(f), f/\|f\|_{\mathrm{ang}}), \tag{2.21}$$

over the sphere defined by the angular norm, $\mathbb{S}_{\text{ang},\ell} := \{f \in \mathcal{F} : \ell(f) > 0, \|f\|_{\text{ang}} = 1\}$. This mapping is an homeomorphism with inverse $T^{-1}(r,\omega) = r\ell(\omega)^{-1}\omega$ when restricted to $\{f \in \mathcal{F}_{\mathscr{C}} : \ell(f) > 0\}$ and, by Corollary 2.14, the pushforward measure is continuous if $\mathscr{C}$ is compact (this is the case, for example, for the cone consisting of the zero element). Suppose $f \in \mathsf{RV}(\mathcal{F}^+_{\mathscr{C}}, a_n, \Lambda)$ and Λ is a homogeneous measure of order $-\alpha$, with $\alpha > 0$. By the continuous mapping theorem, the associated pseudo-polar measure is

$$\Lambda \circ T^{-1}\{\mathrm{d}(r,\omega)\} = \alpha r^{-\alpha-1}\Lambda\{f \in \mathcal{F}_{\mathscr{C}} : \ell(f) > 1\}\mathrm{d}r\rho_{\ell,\text{ang}}(\mathrm{d}\omega),$$

where (Opitz, 2013b)

$$\rho_{\ell,\text{ang}}(\cdot) := \frac{\Lambda(\{f \in \mathcal{F}_{\mathscr{C}} : \ell(f) > 1, f/\|f\|_{\text{ang}} \in \cdot\})}{\Lambda\{f \in \mathcal{F}_{\mathscr{C}} : \ell(f) > 1\}} \tag{2.22}$$

on $(0,\infty) \times \{\omega \in \mathbb{S}_{\text{ang}} : \ell(\omega) > 0\}$.

Under the generalized pseudo-polar decomposition (2.21), the measure factorises into a product measure whose first component decays as a Pareto tail with scale $\Lambda\{f \in \mathcal{F}_{\mathscr{C}} : \ell(f) > 1\}^{1/\alpha}$ and shape α. We can consider regular variation as defined in Definition 2.15 for a generalized pseudo-polar transformation, conditional on exceedances of the risk functional being large. Coupling Theorem 2.12 with regular variation gives an alternative characterization of $\mathbb{M}_{\mathcal{F}_{\mathscr{C}}}$ convergence in terms of conditional distributions (cf. Hult and Lindskog, 2005, Theorem 4).

Proposition 2.27
Let $u_n = \inf\{u \geq 0 : \mathsf{P}\left(\ell(\mathsf{X}) > u\right) = 1/n\}$; if $\{f \in \mathcal{F}_{\mathscr{C}} : \ell(f) = 0\} = \{0\}$ such as when $\ell = \|\cdot\|_\infty$, $\max_{j=1}^D f(\boldsymbol{s}_j)$ or $|\mathcal{S}|^{-1}\int_{\mathcal{S}} X(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$, then the pseudo-polar transformation is an homeomorphism and regular variation (Definition 2.15) is equivalent to convergence of the conditional distribution (de Fondeville, 2018, Theorem 1.6)

$$n\mathsf{P}\left(\ell(\mathsf{X}) > r\,u_n, \mathsf{X}/\|\mathsf{X}\|_{\text{ang}} \in B \mid \ell(\mathsf{X}) > u_n\right) \xrightarrow{\text{w}^\#} r^{-\alpha}\rho_{\ell,\text{ang}}(B), \qquad n \to \infty, \quad r > 0, \tag{2.23}$$

in $\mathcal{F}_0$ for $B \in \mathbb{B}(\mathbb{S}_{\text{ang},\ell})$ such that $\rho_{\ell,\text{ang}}(\partial B) = 0$.

With the generalized pseudo-polar representation (2.21), the limiting measure Λ factorise, a consequence of the fact that Λ is homogeneous of order $-\alpha$. As in the multivariate case, the choice of the angular norm is essentially arbitrary (Wadsworth, Tawn, Davison and Elton, 2017; Einmahl and Segers, 2009) and does not affect the convergence result. The choice of risk functional does not affect the limit measure, but impacts the type of observations that are used to draw inferences.

If inference is performed with one risk functional ℓ_2, it is possible to refer back to the measure for a different risk functional ℓ_1 in cases where $\ell_1(f) > 1$ ρ_{ℓ_2,ℓ_2}-almost everywhere. Using a common angular norm allows for easier comparisons.

Proposition 2.28 (Change of measure for angular distributions)
Consider two absolutely continuous risk functionals ℓ_1, ℓ_2 with a common angular norm and

Λ a $-\alpha$-homogeneous measure. Then, the measures $\rho_{\ell_1,\text{ang}}$ and $\rho_{\ell_2,\text{ang}}$ are related via (Opitz, 2013b)

$$\rho_{\ell_2,\text{ang}}(\mathrm{d}\omega) = \frac{\Lambda(B_{\ell_1})}{\Lambda(B_{\ell_2})} \left\{ \frac{\ell_2(\omega)}{\ell_1(\omega)} \right\}^{\alpha} \rho_{\ell_1,\text{ang}}(\mathrm{d}\omega), \tag{2.24}$$

where $B_\ell = \Lambda\{f \in \mathcal{F} \setminus \mathscr{C}_\ell : \ell(f) \geq 1\}$ for any $\omega \in \{\omega \in \mathbb{S}_{\text{ang}} : \ell_2(\omega) > 0\}$, provided $\{\omega \in \mathbb{S}_{\text{ang}} : \ell_2(\omega) > 0\} \subset \{\omega \in \mathbb{S}_{\text{ang}} : \ell_1(\omega) > 0\}$.

Proof
For any $B \in \mathbb{B}(\mathbb{S}_{\text{ang}})$,

$$\begin{aligned}
\Lambda\{f \in \mathcal{F}_{\mathscr{C}} : \ell_2(f) > 1\}\rho_{\ell_2,\text{ang}}(B) &= \Lambda\{f \in \mathcal{F}_{\mathscr{C}} : \ell_2(f) > 1\} \int_{\mathbb{S}_{\text{ang}}} \mathbf{1}\{y \in B\} \rho_{\ell_2,\text{ang}}(\mathrm{d}y) \\
&= \int_{\mathcal{F}_{\mathscr{C}}} \mathbf{1}\{x : \ell_2(x) > 1\} \mathbf{1}\{x : x/\|x\|_{\text{ang}} \in B\} \Lambda(\mathrm{d}x) \\
&= \int_{\mathcal{F}_{\mathscr{C}}} \mathbf{1}\{x : \ell_1(x) > 1\} \mathbf{1}\{x : x/\|x\|_{\text{ang}} \in B\} \left\{ \frac{\ell_2(x)}{\ell_1(x)} \right\}^{\alpha} \Lambda(\mathrm{d}x) \\
&= \Lambda\{f \in \mathcal{F}_{\mathscr{C}} : \ell_1(f) > 1\} \int_{\mathbb{S}_{\text{ang}}} \mathbf{1}\{y \in B\} \left\{ \frac{\ell_2(y)}{\ell_1(y)} \right\}^{\alpha} \rho_{\ell_1,\text{ang}}(\mathrm{d}y),
\end{aligned}$$

where we make the change of variable $x \mapsto x\ell_1(x)/\ell_2(x)$ to go from the second to the third line. Since ℓ_2 is homogeneous of order 1, $\ell_2\{x\ell_1(x)/\ell_2(x)\} = \ell_1(x)$. The scaling $\{\ell_2(x)/\ell_1(x)\}^{\alpha}$ arises from the $-\alpha$-homogeneity of Λ. The last line follows from the definition of $\rho_{\ell,\text{ang}}$.

■

2.3.1 ℓ-Pareto processes

The decomposition of the limiting measure Λ in generalized pseudo-polar coordinates justifies the following definition.

Definition 2.29 (ℓ-Pareto processes)
A ℓ-Pareto process admits the stochastic representation

$$\mathsf{X} = P \frac{\mathsf{S}}{\ell(\mathsf{S})}, \tag{2.25}$$

for P a Pareto random variable with survival function $\mathsf{P}(P > x) = x^{-\alpha}\mathbf{1}\{x \geq 1\}$ and a stochastic process S, independent of P, with sample paths in $\mathbb{S}_{\text{ang}} := \{f \in \mathcal{F}_0 : \|f\|_{\text{ang}} = 1\}$.

ℓ-Pareto processes arise as the unique non-degenerate weak limit of conditional threshold exceedances of $\ell(f)$ (Dombry and Ribatet, 2015, Theorem 2).

Proposition 2.30 (Representation of ℓ-Pareto processes)
Let $\mathsf{X}(\mathcal{S}) := \{X(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}\}$ be a continuous stochastic process with sample paths in $\mathcal{F}_0$. Then

$\mathsf{X}(\mathcal{S})$ is an ℓ-Pareto process with tail index α and spectral (probability) measure

$$\rho_{\ell,\text{ang}}(\cdot) = \mathsf{P}\left(\mathsf{X}/\|\mathsf{X}\|_{\text{ang}} \in \cdot, \ell(\mathsf{X}) > 1\right)$$

on $\mathbb{S}_{\text{ang}}$ if any of the following three equivalent statements hold:

1. Constructive definition: (a) $\ell(\mathsf{X})$ is Pareto distributed on $[1,\infty)$ with shape $\alpha > 0$ and (b) $\ell(\mathsf{X})$ and $\mathsf{X}/\|\mathsf{X}\|_{\text{ang}}$ are independent.

2. Homogeneity of the exponent measure: (a) $\mathsf{P}\left(\ell(\mathsf{X}) > 1\right) = 1$ and (b) for all $u \geq 1$ and $B \subset \{f \in \mathcal{F}_0, \ell(f) \geq 1\}$ measurable, $\mathsf{P}(\mathsf{X}/u \in B) = u^{-\alpha}\mathsf{P}(\mathsf{X} \in B)$.

3. Peaks-over-threshold stability: (a) $\mathsf{P}\left(\ell(\mathsf{X}) > 1\right) > 0$ and (b) $\mathsf{P}\left(u^{-1}\mathsf{X} \in B \mid \ell(\mathsf{X}) > u\right) = \mathsf{P}(\mathsf{X} \in B)$ for all Borel $B \in \mathbb{B}(\mathcal{F})$ and $u \geq 1$ such that $\mathsf{P}\left(\ell(\mathsf{X}) > u\right) > 0$.

Pareto processes were introduced by Ferreira and de Haan (2014) for the special case $\ell(\mathsf{X}) = \sup_{\boldsymbol{s}\in\mathcal{S}} X(\boldsymbol{s})$ but, while mathematically convenient, threshold exceedances induced by risk functional cannot be determined from a realization of the random field at a finite number of sites. Pareto processes are functional analogs of the multivariate Pareto distribution (Rootzén and Tajvidi, 2006; Rootzén, Segers and Wadsworth, 2018), which is the D-variate restriction with $\ell(\mathsf{X}) = \max_{j=1}^{D} X(\boldsymbol{s}_j)$.

In general, there is no closed form for the marginal distribution of a ℓ-Pareto process at a site $\boldsymbol{s}_j$ and the latter must be calculated numerically pointwise through (Thibaud, 2014, pp.115–116)

$$\mathsf{P}\left(X(\boldsymbol{s}_j) > x\right) = \frac{\Lambda\{\mathsf{X} : X(\boldsymbol{s}_j) > x\}}{\Lambda\{\mathsf{X} : \ell(\mathsf{X}) > u\}};$$

though a notable exception is when the risk functional is the uniform norm $\|\cdot\|_\infty$. However, since the angle $\Omega(\boldsymbol{s})$ lies in $\mathbb{S}_{\text{ang}}$ and is bounded, by Breiman's lemma, $X(\boldsymbol{s}_j)$ is tail equivalent to a Pareto variable with shape α.

It is important to stress that threshold exceedances $\{f \in \mathcal{F}_\mathscr{C} : \ell(\mathsf{X}) > u_n\}$ may not yield pointwise realizations that are extreme everywhere in the domain $\mathcal{S}$. However, the conditional distribution at site $\boldsymbol{s}_j \in \mathcal{S}$ above a marginal threshold u_j within the risk region, i.e.,

$$\{\mathsf{X} : X(\boldsymbol{s}_j) > u_j\} \subseteq \{\mathsf{X} : \ell(\mathsf{X}) > u\}$$

is generalized Pareto distributed. D-dimensional marginal distributions of ℓ-Pareto processes need not have the same distribution if there is marginalization over the coordinates that create the exceedance $\ell(\mathsf{X}) > u$ (cf. Rootzén et al., 2018).

2.3.2 Generalized ℓ-Pareto processes

All the conditional exceedances of ℓ-Pareto processes are tail equivalent to a Pareto distribution with tail index $\xi > 0$, which is unrealistic in many practical settings. To circumvent this restriction, it is tempting to transform the data at each site to have, e.g., a unit Pareto distribution. This has a major disadvantage: the risk region is defined on the standardized Pareto scale, so events that correspond to exceedances need not be those of interest on the data scale. Chapter 4 of de Fondeville (2018) proposes a generalization of ℓ-Pareto processes under the assumption that the shape parameter $\xi \in \mathbb{R}$ is fixed over the domain, extending the notion of risk exceedances. While it may seem restrictive, this hypothesis excludes processes where the site with the largest tail index dominates asymptotically. A generalized form of regular variation follows from the developments described in Remark 2.4.

Proposition 2.31 (Generalized functional regular variation)
Consider the continuous map $T(x)$ equal to $(1+\xi x)^{1/\xi}$ if $\xi \neq 0$ or $\exp(x)$ if $\xi = 0$ and let X be a stochastic process with sample paths in the Banach space of continuous functions $\mathcal{F}$. We write $\mathsf{X} \in \mathsf{RV}(\mathcal{F}_{\ell}, \xi, \mathsf{a}_n, \mathsf{b}_n, \Lambda)$ if, for sequences of continuous scaling functions $\{\mathsf{a}_i\}_{i=1}^{\infty} \equiv \{a_i(\cdot)\}_{i=1}^{\infty} > 0$ and $\{\mathsf{b}_i\}_{i=1}^{\infty} = \{b_i(\cdot)\}_{i=1}^{\infty}$, a homogeneous measure Λ of order -1 and $\xi \in \mathbb{R}$,

$$n\mathsf{P}\left(T\left(\frac{\mathsf{X}-\mathsf{b}_n}{\mathsf{a}_n}\right) \in \cdot\right) \xrightarrow{\mathrm{w}^{\#}} \Lambda(\cdot), \qquad n \to \infty$$

in $\mathbb{M}_{\mathcal{F}_{\ell}}$, where the scaling functions are chosen so that, for any $\boldsymbol{s} \in \mathcal{S}$ and $x > 0$,

$$\lim_{n\to\infty} n\mathsf{P}\left\{\frac{X(\boldsymbol{s}) - b_n(\boldsymbol{s})}{a_n(\boldsymbol{s})} > x\right\} = \begin{cases} (1+\xi x)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-x), & \xi = 0. \end{cases}$$

Equipped with the hypothesis of functional regular variation as described in Proposition 2.31, one can obtain the limiting measure of ℓ-exceedances; see de Fondeville (2018, § 4.3) for detailed statements of convergence. The unique limiting distribution is that of a generalized ℓ-Pareto process and follows from the continuous mapping theorem. Rather than present the convergence statement, we focus on the stochastic representation equivalent to Proposition 2.30.

Consider the standardized ℓ-Pareto process X, with unit Pareto margins and associated -1-homogeneous measure Λ, defined on the risk region

$$\mathscr{A}_u = \left\{\mathsf{X} \in \mathcal{F}_0^+ : \ell\left(\boldsymbol{\tau}\frac{\mathsf{X}^{\xi}-1}{\xi} + \boldsymbol{\eta}\right) \geq u\right\},$$

with $(\boldsymbol{x}^{\xi}-1)/\xi$ understood as $\log(\boldsymbol{x})$ when $\xi = 0$. Because we consider processes with continuous sample path (with potentially negative components), the notion of risk functional must be extended: $\ell : \mathcal{F} \to \mathbb{R}$ must be monotone increasing and additional requirements are imposed to ensure that $\mathscr{A}_u$ is bounded away from 0 (de Fondeville, 2018, § D.2). The probability mea-

sure of Z^* over $\mathscr{A}_u$ is $\Lambda(\cdot)/\Lambda\{\mathscr{A}_u\}$. The stochastic representation of the generalized ℓ-Pareto vector is

$$\mathsf{Z} = \boldsymbol{\tau}\frac{\mathsf{X}^{\xi}-1}{\xi} + \boldsymbol{\eta}. \tag{2.26}$$

If ℓ is a linear risk functional, Z is the product of an angular distribution and a generalized Pareto distribution (de Fondeville, 2018, Definition 2).

The rescaled process $\mathsf{Z} = \boldsymbol{\tau}(\mathsf{X}^{\xi}-1)/\xi + \boldsymbol{\eta}$ is conditionally generalized Pareto distributed above some threshold: for any site $\boldsymbol{s}_0 \in \mathcal{S}$ such that $\{Z(\boldsymbol{s}_0) > u\} \subset \ell(\mathsf{Z}) > u'$,

$$\mathsf{P}\left(Z(\boldsymbol{s}_0) > x\right) \propto \left\{1+\xi\frac{x-u}{\sigma(\boldsymbol{s}_0)}\right\}_+^{-1/\xi}, \qquad x \geq u,$$

with $\sigma(\boldsymbol{s}_0) = \tau(\boldsymbol{s}_0) + \xi\{u - \eta(\boldsymbol{s}_0)\}$. Initial estimates of the pointwise marginal parameters, ξ, $\tau(\boldsymbol{s}_0)$ and $\eta(\boldsymbol{s}_0)$ can thus be obtained by maximizing the generalized Pareto likelihood. If the risk functional satisfies $\ell(\mathsf{X}+b) = \ell(\mathsf{X}) + b$ for $b \in \mathbb{R}$, then we can consider direct modelling of $\mathsf{X} - \ell(\mathbf{b}_n)$; for $\ell(\mathsf{X}) = \max_{j=1}^{D} X(\boldsymbol{s}_j)$, this amounts to modelling the threshold exceedances $X(\boldsymbol{s}_j) - u_j$ for some threshold u_j and setting $\eta(s) = 0$ for all $\boldsymbol{s} \in \mathcal{S}$.

2.4 Conditional extremes

A major drawback of the convergence theory underlying (generalized) ℓ-Pareto processes is the assumption that each variable grows at the same rate; many models are asymptotically independent despite exhibiting positive dependence at penultimate levels. One could refine the standardization by allowing different normalizing constants for each variable to avoid this. An alternative approach, termed conditional extremes, was proposed by Heffernan and Tawn (2004) by looking at the behaviour of random vectors when one component is large.

Heffernan and Tawn (2004) propose to use projections as a basis for inference in multivariate extremes. Specifically, consider a D-random vector $\boldsymbol{X}$ with absolutely continuous margins, transformed to be exponential-tailed, for example standardized Gumbel or Laplace margins (Keef et al., 2013); denote the transformed observations by $\boldsymbol{X}^{\mathrm{e}}$. For $X_k > u_k$, Heffernan and Tawn (2004) assume that there exist normalizing functions $a_{|k}(\cdot) : \mathbb{R} \to \mathbb{R}_+^{D-1}$ and $b_{|k}(\cdot) : \mathbb{R} \to \mathbb{R}^{D-1}$ and a non-degenerate distribution function H, absolutely continuous with respect to Lebesgue measure, such that

$$\mathsf{P}\left(\frac{\boldsymbol{X}_{-k} - b_{|k}(X_k)}{a_{|k}(X_k)} \leq \boldsymbol{x}_{-k}, X_k - u_k > x_k \,\middle|\, X_k > u_k\right) \to H(\boldsymbol{x}_{-k})\exp(-x_k), \qquad x_k > 0, u_k \to \infty.$$

The marginals of H must be such that $\lim_{x_j\to\infty} H_j = 1$ for $j = 1,\ldots,D$, $j \neq k$; the theory for this convergence result based on random renormalization by X_k is formalized in Heffernan and Resnick (2007, § 4); with exponential-tailed variables, the vector functions a and b must be regularly varying, of index 1 for a and of index $\rho < 1$ for b, componentwise.

Heffernan and Tawn (2004) use simple parametric form for the scaling vectors that covers many copulas, taking $a_{|k}(x) = \boldsymbol{\alpha} x$ and $b a_{|k}(x) = x^{\boldsymbol{\beta}}$ for $\boldsymbol{\alpha} \in [\mathbf{0}_{D-1}, \mathbf{1}_{D-1}]$ and $\boldsymbol{\beta} \in [\mathbf{0}_{D-1}, \mathbf{1}_{D-1})$ for Gumbel margins, restricting to positive dependence; negative dependence would correspond to $\beta_{j|k} < 0$ and $\alpha_{j|k} = 0$ and then $\alpha_{j|k}$ may be on the boundary of the parameter space. If $\alpha_{j|k} = 1$ and $\beta_{j|k} = 0$, then X_k and X_j are asymptotically dependent; otherwise they are asymptotically independent.

Theoretical assumptions

The underlying theoretical results are given in Heffernan and Resnick (2007), who focus on the bivariate setting. Suppose without loss of generality that X_k has asymptotically a unit Pareto tail; this can be achieved using the probability integral transform. Heffernan and Resnick (2007) assume the existence of a Radon measure Λ such that

$$t\mathsf{P}\left(\frac{\boldsymbol{X}_{-k} - \boldsymbol{b}_{|k}(t)}{\boldsymbol{a}_{|k}(t)} \leq \boldsymbol{x}_{-k}, \frac{X_k}{t} > x_k\right) = \Lambda\{[\infty_{D-1}, \boldsymbol{x}_{-k}] \times [x_k, \infty)\}, \qquad \boldsymbol{x}_{-k} \in \mathbb{R}^{D-1}, x_k > 0.$$

Since $\mathsf{P}(X > x) \sim x^{-1}$ as $x \to \infty$,

$$\begin{aligned}\mathsf{P}\left(\frac{\boldsymbol{X}_{-k} - \boldsymbol{b}_{|k}(t)}{\boldsymbol{a}_{|k}(t)} \leq \boldsymbol{x}_{-k} \,\middle|\, \frac{X_k}{t} > 1\right) &\sim t\mathsf{P}\left(\frac{\boldsymbol{X}_{-k} - \boldsymbol{b}_{|k}(t)}{\boldsymbol{a}_{|k}(t)} \leq \boldsymbol{x}_{-k}, \frac{X_k}{t} > 1\right) \\ &\to \Lambda\{[\infty_{D-1}, \boldsymbol{x}_{-k}] \times [1, \infty)\}, \qquad t \to \infty.\end{aligned}$$

The measure Λ factorises into a product measure, i.e.,

$$\Lambda\{[-\infty_{D-1}, \boldsymbol{x}_{-k}] \times (x, \infty]\} = x^{-1}\Lambda\{[-\infty_{D-1}, \boldsymbol{x}_{-k}] \times (1, \infty]\},$$

provided $a_{j|k} \in \mathsf{RV}_0$ and

$$\lim_{t\to\infty} \frac{b_{j|k}(ct) - b_{j|k}(t)}{a_{j|k}(t)} = 0.$$

We can contrast this convergence result with the one underlying ℓ-Pareto processes with $\ell = X(\boldsymbol{s}_k)$: if $\boldsymbol{X}$ is regularly varying, then Λ is -1-homogeneous and we recover the same convergence results, albeit in a slightly different form. The conditional extremes approach however does not however assume regular variation of $\boldsymbol{X}$, allowing for modelling of asymptotically independent processes.

Inference for the conditional extremes model

Inference for the conditional extremes model is performed in two stages. In the first, the data are transformed to Gumbel margins, so that $\mathsf{P}(X > x) \sim \exp(-x)$ as $x \to \infty$; Laplace margins can be used should one consider the modelling of negative dependence (Keef et al., 2013), in which case the support of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ changes and inference is simplified relative to

Algorithm 2.1 Unconditional simulation from the conditional extremes model given $X_j > u_j$, (Heffernan and Tawn, 2004)

Require: parameter estimates $\widehat{\boldsymbol{\alpha}}_{|j}, \widehat{\boldsymbol{\beta}}_{|j}$, threshold u_j.

1: set $\widehat{\boldsymbol{z}}_{i|j} = \left\{(\boldsymbol{x}_{i,-j} - \widehat{\boldsymbol{\alpha}} x_{i,j}) / x_{i,j}^{\widehat{\boldsymbol{\beta}}}\right\}_{i=1,\ldots,n_{u_j}}$;
2: simulate $X_j \sim \mathsf{Exp}(1) + u_j$;
3: sample $\boldsymbol{Z} \sim \mathsf{U}\big(\widehat{\boldsymbol{z}}_{1|j}, \ldots, \widehat{\boldsymbol{z}}_{n_{u_j}|j}\big)$, i.e. uniformly from (2.27);
4: set $\boldsymbol{X}_{-j} \leftarrow \widehat{\boldsymbol{\alpha}}_{|j} X_j + X_j^{\widehat{\boldsymbol{\beta}}_{|j}} \boldsymbol{Z}$;
5: **return** $\boldsymbol{X}$

the form of Heffernan and Tawn (2004). Standardized margins can be obtained through the probability integral transform, using the empirical distribution or a semi-parametric model with a generalized Pareto model for tail exceedances (2.29).

In the second stage, the risk region $\mathbb{A}$ is split into mutually disjoint components,

$$\mathbb{A}_k = \mathbb{A} \cap \{\boldsymbol{x} \in \mathbb{R}^D : F_{X_k}(x_k) > F_{X_j}(x_j), j = 1, \ldots, D, j \neq k\}, \qquad k = 1, \ldots, D,$$

and we can write

$$\mathsf{P}(\boldsymbol{X} \in \mathbb{A}) = \sum_{j=1}^{D} \mathsf{P}(\boldsymbol{X} \in \mathbb{A}_k \mid X_k > u_k)\, \mathsf{P}(X_k > u_k), \qquad u_k = \inf\{x_k : \boldsymbol{x} \in \mathbb{A}_k\}.$$

Assuming that the limit distribution holds exactly above $X_k > u_k$, the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are estimated under the working assumption that

$$\boldsymbol{X}_{-k} \mid X_k = x \sim \boldsymbol{\alpha} x + x^{\boldsymbol{\beta}} \boldsymbol{Z}, \qquad \boldsymbol{Z} \sim \mathsf{No}_{D-1}\{\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)\},$$

with nuisance parameters $\boldsymbol{\mu} \in \mathbb{R}^{D-1}$ and $\boldsymbol{\sigma}^2 \in \mathbb{R}_+^{D-1}$. The corresponding pseudo-likelihood is given in Example 2.35. Vectors of residuals can be obtained by using the estimated parameters $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, setting

$$\widehat{z}_i = \{(\boldsymbol{x}_{i,-j} - \widehat{\boldsymbol{\alpha}} x_{i,j}) / x_{i,j}^{\widehat{\boldsymbol{\beta}}}\}_{i=1}^{n_{u_j}} \tag{2.27}$$

for the n_{u_j} exceedances $\{\boldsymbol{X} : x_{i,j} > u_j, i = 1, \ldots, n\}$. The empirical distribution of the residuals $\boldsymbol{Z}$, or a smoothed version thereof, is used for extrapolation.

Suppose one is interested in estimating the probability of falling inside a risk region $\mathcal{A} \subset \sqcup_{k \in K} \mathscr{A}_k$ for $K \subset \{1, \ldots, D\}$. Using the model, extremes are simulated through Algorithm 2.1 by first drawing new values of $x_j \sim \mathsf{Exp}(1) + u_j$ and then by adding residuals drawn uniformly from the empirical distribution of $\boldsymbol{Z}$. The probability of falling in $\mathbb{A}_i$ is obtained empirically as the fraction of simulated points falling in risk region. New observations can also be obtained through Algorithm 2.1 by using the probability integral transform and mapping simulated samples back to the data scale.

The inferential approach for the conditional extremes is semiparametric and so is subject to the curse of dimensionality: the empirical distribution $\widehat{z}_{|j}$ may approximate the true residual poorly in high dimensions. Another major criticism is the lack of self-consistency: the conditional distributions for $X_1 \mid X_2 > u$ and $X_2 \mid X_1 > u$ typically yield incompatible models for $\mathsf{P}(X_1, X_2 \mid X_1 > u, X_2 > u)$. Liu and Tawn (2014) propose some additional constraints on the parameters to somewhat salvage the issue, which can also be solved by assuming exchangeability. Alternatives include averaging over simulation or defining disjoint regions, keeping points for $X_1 > u$ only if they fall in the region $\{X_1 > X_2 \mid X_1 > u\}$; (Lugrin, 2018, § 6.2) consider the latter approach to only consider points once in the likelihood, resulting in a proper likelihood model which discontinuous on the boundary $X_1 = X_2$.

2.4.1 Conditional spatial extremes

The Heffernan–Tawn conditional extremes model can be extended to the spatial setting (Tawn et al., 2018). Consider a (intrinsically) stationary spatial process X, a conditioning location $\boldsymbol{s}_0 \in \mathcal{S}$ and D additional sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D \in \mathcal{S}$ at which measurements are available. We consider scaling functions $a(x; \boldsymbol{h}) : \mathbb{R} \to (0, \infty)$ and $b(x, \boldsymbol{h}) : \mathbb{R} \to \mathbb{R}$ such that $b(x, \boldsymbol{0}) = x$, both of which are dependent on the value of $X(\boldsymbol{s}_0) = x$ and the distance between sites, $\boldsymbol{h}_j = \boldsymbol{s}_j - \boldsymbol{s}_0$. The spatial conditional extremes model assumes that, for any $x > 0$ and $\boldsymbol{z} \in \mathbb{R}^D$ (Wadsworth and Tawn, 2018),

$$\frac{\mathsf{P}\left(X(\boldsymbol{s}_0) > x + u, \dfrac{X(\boldsymbol{s}_j) - b\{X(\boldsymbol{s}_0), \boldsymbol{h}_j\}}{a\{X(\boldsymbol{s}_0), \boldsymbol{h}_j\}} \leq z_j, j = 1, \ldots, D\right)}{\mathsf{P}\{X(\boldsymbol{s}_0) > u\}} \xrightarrow{\mathsf{w}} \mathsf{P}\{Z(\boldsymbol{s} \leq \boldsymbol{z})\} \exp(-x), \quad u \to \infty.$$

The residual process Z must satisfy $Z(\boldsymbol{s}_0) = 0$ almost surely for identifiability. If the dependence functions a and b depend on $\boldsymbol{h}$ only through its norm, the joint distribution of pairs is exchangeable, and this resolves the problem of self-consistency of conditional distributions. For asymptotically dependent processes, such as $X(\boldsymbol{s}_0)$-Pareto processes, the scaling functions satisfy $b(x, \boldsymbol{h}) = x$ and $a(x, \boldsymbol{h}) = 1$ for all $\boldsymbol{h}$ and it makes sense to consider $b(x, \boldsymbol{h}) \approx x$ for $\boldsymbol{h} \approx \boldsymbol{0}$. Wadsworth and Tawn (2018) propose parametric models for a and b: lag-asymptotic dependence is obtained by taking $b(x, \boldsymbol{h}) = x$ if $\|\boldsymbol{h}\| \leq d_{\max}$ and $b(x, h)$ decaying monotonically beyond $d_{\max}$. For example, if $d_{\max} = 0$, one could consider as scale function $a = x^{\beta(\boldsymbol{h})}$ for $\beta(\boldsymbol{h}) \in [0, 1)$ with $\beta(\boldsymbol{0}) = 0$. The residual vector $\boldsymbol{Z}$ from the multivariate model of Heffernan and Tawn (2004) is replaced by an (intrinsically) stationary spatial process Z conditioned on $Z(\boldsymbol{s}_0) = 0$ almost surely, the most natural candidate being a Gaussian process with variogram $\gamma(\boldsymbol{h})$. We refer to § 3 of Wadsworth and Tawn (2018) for more details on specification of the normalizing functions, the residual process and the marginal transformation.

Estimation of the probability of lying in risk regions beyond the range of the data proceeds as before through simulation; the second step in Algorithm 2.1 is replaced by sampling independently from $X(\boldsymbol{s}_0) \sim u + \mathsf{Exp}(1)$ from the $\mathsf{Z} \mid Z(\boldsymbol{s}_0) = 0$. The new observations on the standardized scale are obtained as $X(\boldsymbol{s}_j) = \widehat{b}[X(\boldsymbol{s}_0), \boldsymbol{s}_j - \boldsymbol{s}_0] + \widehat{a}[X(\boldsymbol{s}_0), \boldsymbol{s}_j - \boldsymbol{s}_0] Z(\boldsymbol{s}_j)$ for $j = 1, \ldots, D$,

where the parameters of the scaling functions a and b are obtained through composite likelihood (Example 2.31). If Z belongs to a location-scale family, the last two steps correspond to forward sampling in a random location-scale model. Algorithms for simulation are described in Section 2.5.6.

For generating extremal scenarios, the standardized simulated observations must be mapped to the data scale; preliminary marginal transformation to standardized margins is standard in the copula literature, but spatial models are needed to impute the value of the process at unobserved sites. One benefit of the conditional spatial extremes approach is the ability to model negative dependence.

There is much flexibility in the potential specification of the components of the model, and the dependence model becomes fully parametric relative to the multivariate methodology outlined in Section 2.4 that uses a pseudo-likelihood and only uses the empirical distribution of scale residuals $\boldsymbol{Z}$. The spatial model may be expressed as a mixture, which makes it attractive for further development. Because extreme events are counted multiple times if they correspond to exceedances of a threshold in more than one variable, approximate Bayesian inference using composite likelihood would need to be adjusted to account for this multiplicity (Ribatet et al., 2012).

2.5 Simulation

Conditional and unconditional simulation of max-stable processes, conditional extremes and generalized ℓ-Pareto processes play a key role in inference, since they can be used to create a catalogue of extreme events at a set of sites of interest. Goodness-of-fit diagnostics can also be designed based on the ability to correctly replicate events at holdout sites.

2.5.1 Extremal functions, sub-extremal functions and hitting scenario

The de Haan representation provides a constructive definition of max-stable models and will serve as a basis for simulations. The functions appearing in the representation $Z^*(\boldsymbol{s}) = \max_{i\in\mathbb{N}^+} \zeta_i W(\boldsymbol{s})$ can be viewed as points of a marked Poisson point process $\Phi = \{\varphi_i\}_{i\in\mathbb{N}^+}$ whose elements are $\varphi_i \equiv \zeta_i \mathrm{W}(\mathcal{S})$. The points of the Poisson process of intensity $\zeta^{-2}\mathrm{d}\zeta$ can be sampled in decreasing order and are independent of the marks $\mathrm{W}(\mathcal{S})$.

Suppose we observe or wish to sample the max-stable process on a compact subset $D \subset \mathcal{S}$ at D sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$. Assuming regularity, the maximum at each location, $\max_{i\in\mathbb{N}^+} \varphi_i(\boldsymbol{s}_j)$, will almost surely be realized by a unique function. We call the collection of such functions extremal functions and denote them by $\varphi^+_{\boldsymbol{s}_j}$. We say $\varphi^+ \in \Phi$ is extremal at $\boldsymbol{s}_0$ if and only if, for all $\varphi \in \Phi \setminus \{\varphi\}$, $\varphi^+(\boldsymbol{s}_0) > \varphi(\boldsymbol{s}_0)$. One realization of the point process may contribute to the max-stable process at a single location or at multiple locations. The partition $\boldsymbol{\Pi}$, called the hitting scenario of the set $\{1, \ldots, D\}$ in Wang and Stoev (2011), encodes the information

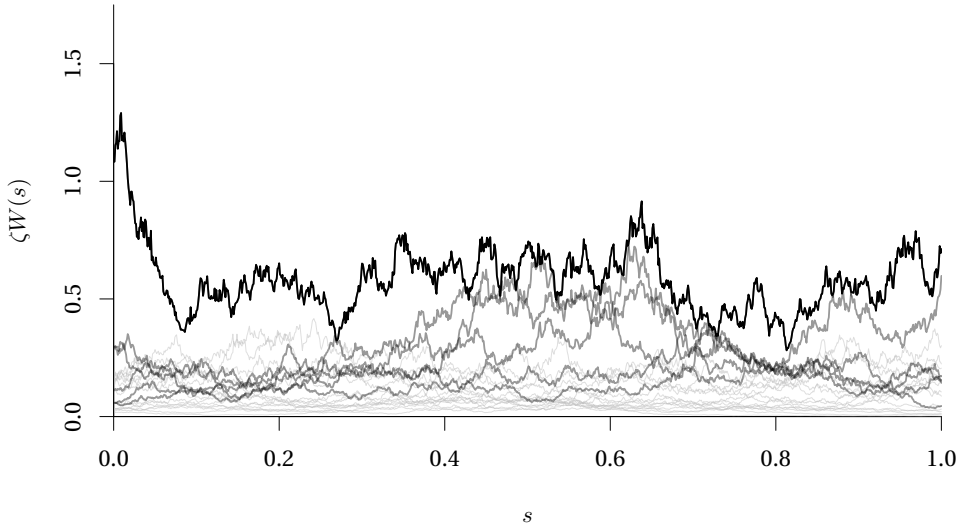

Figure 2.1 – *Sub-extremal functions (grey), extremal functions (dark grey) and Brown–Resnick max-stable process (black) with power variogram* $\gamma(h) = \sigma^2 \|h\|_2^\alpha$ *on* $[0,1]$*, simulated using Algorithm 2.3.*

about which extremal function contributes at which sites, allowing us to define a collection of unique extremal functions $\varphi^+_{\varpi_j}$ with cardinality $|\boldsymbol{\Pi}| = k \leq D$. We shall also need to consider the collection of sub-extremal functions which do not contribute to the max-stable process at location $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$. Figure 2.1 shows a simulation from a Brown–Resnick max-stable process (black), along with the extremal and sub-extremal functions appearing in its construction. To recapitulate, we have:

- the hitting scenario, that is, the random partition $\boldsymbol{\Pi} = (\varpi_1, \ldots, \varpi_k)$. This is such that the indices of any two distinct locations $\boldsymbol{s}_j, \boldsymbol{s}_k$ belong to the same component ϖ_0 if and only if $\varphi^+_{\boldsymbol{s}_j} = \varphi^+_{\boldsymbol{s}_k}$, so $\boldsymbol{\Pi}$ accounts for the possible repeated contributions of an extremal function to the max-stable process;

- the extremal point process $\Phi^+ = \{\varphi^+_{\varpi_j}\}_{j=1}^k$, the collection of functions that contribute to the extrema at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$;

- the sub-extremal point process $\Phi^- = \Phi \setminus \Phi^+$, whose points consist of random functions falling below the realized value of the extremal point process at sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$, meaning that $\varphi^-(\boldsymbol{s}_j) < Z^*(\boldsymbol{s}_j)$ for $j = 1, \ldots, D$.

Algorithm 2.2 Exact simulation for uniformly bounded processes (Schlather, 2002)

Require: distribution of the spectral functions W, $F_{\boldsymbol{W}}$, bound C such that $W(\boldsymbol{s}) \leq C$ for any $\boldsymbol{s} \in \mathcal{S}$.

1: set $\boldsymbol{Z} = \mathbf{0}_D$
2: simulate $\zeta^{-1} \sim \mathsf{Exp}(1)$;
3: **while** $\min \boldsymbol{Z} \leq C\zeta$ **do**
4: simulate $\boldsymbol{W} \sim F_{\boldsymbol{W}}$ and $E \sim \mathsf{Exp}(1)$;
5: set $\boldsymbol{Z} \leftarrow \max\{\boldsymbol{Z}, \zeta \boldsymbol{W}\}$;
6: set $\zeta^{-1} \leftarrow \zeta^{-1} + E$;
7: **return** $\boldsymbol{Z}$.

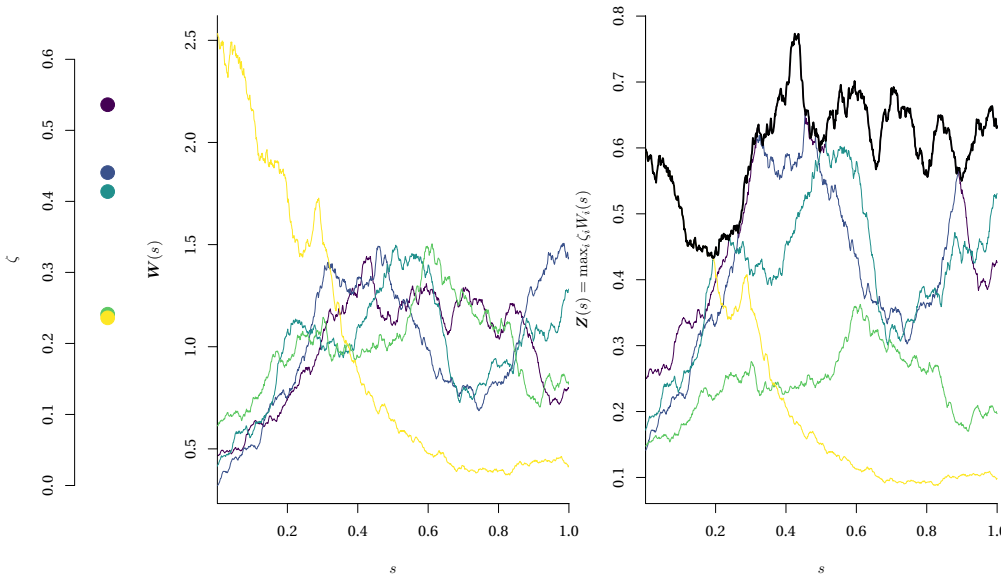

Figure 2.2 – *Construction of a max-stable process: ordered realizations $\zeta_1 > \zeta_2 > \cdots$ from a Poisson process of intensity $\zeta^{-2}\mathrm{d}\zeta$ (left), spectral functions $W_i(s)$ (middle) and the resulting product $Z(s) = \max_{i=1}^{\infty} \zeta W_i(s)$ with the max-stable process bordered in black (right). Only extremal functions, i.e., those contributing to the max-stable process, are shown.*

2.5.2 Unconditional simulations of max-stable processes

Unconditional simulation of max-stable processes appears daunting because the de Haan spectral representation presupposes taking pointwise maximum over an infinite collection of variables. An approximate simulation, based on truncating the spectral representation, would consist of sampling a large but finite number of replicates, arranging the points ζ in decreasing order. However, this approach leads to approximate samples if the extremal functions are unbounded. When $\mathsf{W}(\mathcal{S})$ is uniformly bounded over $\mathcal{S}$, meaning $W(\boldsymbol{s}) \leq C$ for any $\boldsymbol{s} \in \mathcal{S}$ and $C \in \mathbb{R}_+$ or bounded at $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$, it is possible to simulate exactly using Algorithm 2.2.

Provided that one can simulate from the angular distribution over $\rho_{1,1}(\cdot)$, one can use Algorithm 2.2 with $\boldsymbol{W} = D\boldsymbol{S}, \boldsymbol{S} \sim \rho_1(\cdot)$, as $\boldsymbol{W}$ is bounded by $C = D$ by definition. The algorithm is

further illustrated in Figure 2.2. Many multivariate spectral distributions on the l_1 sphere admit a representation as transformed Dirichlet mixtures (Boldi, 2009; Belzile and Nešlehová, 2017), so in principle sampling from these is straightforward. If not, an alternative stochastic representation of the shape of the spectral functions may lead to the desired boundedness property, as the following example shows.

Example 2.19 (Extremal Student)
Thibaud and Opitz (2015) use Algorithm 2.2 to sample from the extremal Student max-stable process at D sites, relying on the alternative spectral representation

$$\boldsymbol{Z} = \max_{i \in \mathbb{N}^+} \zeta_k \frac{2\pi^{1/2}\Gamma\{(\nu+D)/2\}}{\Gamma\{(\nu+1)/2\}\Gamma(D/2)} (\mathbf{A}\boldsymbol{U}_i)_+^{\nu},$$

where $\mathbf{A}$ is the Cholesky root of the correlation matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{U} \sim \mathsf{U}\big(\{\boldsymbol{x} \in \mathbb{R}_+^D : \|\boldsymbol{x}\|_2 = 1\}\big)$ is a uniform vector on the l_2 sphere, meaning that $\|(\mathbf{A}\boldsymbol{U}_i)_+^{\nu}\|_\infty \leq 1$ for any $i \in \mathbb{N}^+$. The bounding constant C increases with both the tail index ν and the dimension D, so this scheme will work well in moderate dimensions with more heavy-tailed data. This also provides some insight as to why it is so hard to sample from the Brown–Resnick process, as it is the limiting model when $\nu \to \infty$.

□

Rather than generating a very large number of spectral functions, it is better to focus solely on generating the extremal functions. If the distribution of the latter is known, recent algorithms developed by Dombry et al. (2016) allow one to perform exact simulations.

Proposition 2.32 (Distribution of extremal functions at $\boldsymbol{s}_0$)
The random variables $Z^*(\boldsymbol{s}_0)$ and $\varphi^+_{\boldsymbol{s}_0}/Z^*(\boldsymbol{s}_0)$ are independent and the distribution function of $\varphi^+_{\boldsymbol{s}_0}/Z^*(\boldsymbol{s}_0)$ is (Dombry and Eyi-Minko, 2013, Proposition 4.2)

$$\mathsf{P}\left(\varphi^+_{\boldsymbol{s}_0}/Z^*(\boldsymbol{s}_0) \in B\right) = \int_{\mathscr{F}_0^+} \mathbf{1}_{\{\mathsf{W}/W(\boldsymbol{s}_0)\in B\}} W(\boldsymbol{s}_0)\nu(\mathrm{d}\mathsf{W}), \quad B \in \mathbb{B}(\mathscr{F}_0^+),$$

where ν denotes the measure of the spectral functions W.

Proof
Recall that $\Phi = \{\varphi\}$ is a marked point process whose elements are random functions $\varphi_i = \zeta_i \mathsf{W}_i \in \mathscr{F}_0^+$ and whose marks are W_i, that $\mathsf{Z}^* = \max_{i\in\mathbb{N}^+}\varphi_i$ has limit measure Λ, and that $Z^*(\boldsymbol{s}_0)$ has a unit Fréchet distribution. The extremal function $\varphi^+_{\boldsymbol{s}_0}$ is equal to Z^* at $\boldsymbol{s}_0$ and no other extremal (or sub-extremal) function can be larger at $\boldsymbol{s}_0$. Thus, we can write for any $x \in \mathbb{R}_+$ and any Borel set B

$$\begin{aligned}
\mathsf{P}\left(Z^*(\boldsymbol{s}_0) \leq x, \varphi^+_{\boldsymbol{s}_0}/Z^*(\boldsymbol{s}_0) \in B\right) &= \mathsf{E}\left(\sum_{\varphi\in\Phi} \mathbf{1}_{\{\varphi^+_{\boldsymbol{s}_0}/\varphi(\boldsymbol{s}_0)\in B\}} \mathbf{1}_{\{\varphi^+(\boldsymbol{s}_0)\leq x\}} \mathbf{1}_{\{\varphi^+_{\boldsymbol{s}_0}(\boldsymbol{s}_0)>\varphi(\boldsymbol{s}_0)\ \forall\ \varphi\in\Phi\setminus\{\varphi^+_{\boldsymbol{s}_0}\}\}}\right) \\
&= \int_{\mathscr{F}_0^+} \mathsf{E}\left(\mathbf{1}_{\{\varphi(\boldsymbol{s}_0)\leq x, Z^*(\boldsymbol{s}_0)\geq\varphi(\boldsymbol{s}_0), \varphi/\varphi(\boldsymbol{s}_0)\in B\}}\right) \Lambda(\mathrm{d}\varphi) \\
&= \int_{\mathscr{F}_0^+} e^{-1/\varphi(\boldsymbol{s}_0)} \mathbf{1}_{\{\varphi/\varphi(\boldsymbol{s}_0)\in B, \varphi(\boldsymbol{s}_0)\leq x\}} \Lambda(\mathrm{d}\varphi)
\end{aligned}$$

$$= \mathsf{E}\left(\int_0^\infty e^{-1/\{\zeta W(\boldsymbol{s}_0)\}} \mathbf{1}\{\zeta W(\boldsymbol{s}_0) \le x\} \mathbf{1}\{\mathsf{W}/W(\boldsymbol{s}_0) \in B\} \zeta^{-2} \mathrm{d}\zeta\right)$$
$$= e^{-1/x} \mathsf{E}\left(W(\boldsymbol{s}_0) \mathbf{1}\{\mathsf{W}/W(\boldsymbol{s}_0) \in B\}\right).$$

The second line follows from the Slivnyak–Mecke formula (2.2), followed by a change of variable to pseudo-polar coordinates.

■

For any $\boldsymbol{s}_0 \in \mathcal{S}$, the distribution of the scaled extremal function, $\varphi^+_{\boldsymbol{s}_0}/Z^*(\boldsymbol{s}_0)$ is denoted by $P_{\boldsymbol{s}_0}$ and is supported on the set $\{\varphi \in \mathcal{F} : \varphi(\boldsymbol{s}_0) = 1\}$.

Consider the sequence of point processes $\Phi^+_k = \{\varphi^+_{\boldsymbol{s}_j}\}_{j=1}^k$ and the associated sub-extremal functions $\Phi^- = \Phi \setminus \Phi^+_k$. The sets of extremal functions are nested, since $\Phi^+_k \subset \Phi^+_{k+1}$ by definition. Since an extremal function can contribute to many locations, the point process Φ^+_{k+1} is either equal to Φ^+_k or else includes a new function. The latter then provides the maximum at $\boldsymbol{s}_{k+1}$ and must fall below the other extremal functions at all sampled locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$. The new function φ^+_{k+1} satisfies $\varphi^+_{k+1}(\boldsymbol{s}_j) < \max_{l=1}^k \varphi^+_l(\boldsymbol{s}_j)$ for $j = 1, \ldots, k$ and $\varphi^+_{k+1}(\boldsymbol{s}_{k+1}) > \max_{l=1}^k \varphi^+_l(\boldsymbol{s}_{k+1})$. Let

$$\widetilde{\Phi}_{k+1} = \Phi^-_k \cap \{\varphi \in \mathcal{F}^+_0 : \varphi(\boldsymbol{s}_{k+1}) > \max \Phi^+_k(\boldsymbol{s}_{k+1})\}.$$

The event $\widetilde{\Phi}_{k+1} = \emptyset$ indicates that no sub-extremal function in Φ^-_k exceeds $\max \Phi^+_k$ at $\boldsymbol{s}_{k+1}$. Conditional on Φ^+_k, $\widetilde{\Phi}_{k+1}$ is a Poisson point process on $\mathcal{F}^+_0$ with intensity

$$\prod_{j=1}^k \mathbf{1}\{\varphi(\boldsymbol{s}_j) \le \max_{\phi \in \Phi_k} \phi(\boldsymbol{s}_j)\} \mathbf{1}\{\varphi(\boldsymbol{s}_{k+1}) > \max_{\phi \in \Phi_k} \phi(\boldsymbol{s}_{k+1})\} \Lambda(\mathrm{d}\varphi),$$

and (Dombry et al., 2016, Theorem 2)

$$\varphi^+_{\boldsymbol{s}_{k+1}} \overset{\mathrm{d}}{=} \mathbf{1}\{\widetilde{\Phi}_{k+1} \neq \emptyset\} \operatorname*{arg\,max}_{\varphi \in \widetilde{\Phi}_{k+1}} \varphi(\boldsymbol{s}_{k+1}) + \mathbf{1}\{\widetilde{\Phi}_{k+1} = \emptyset\} \operatorname*{arg\,max}_{\varphi \in \Phi^+_k} \varphi(\boldsymbol{s}_{k+1}).$$

One can sequentially sample the extremal functions at site k from $P_{\boldsymbol{s}_{k+1}}$, discarding any realization that exceeds Φ^+_k at $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$; Algorithm 2.3 summarizes the procedure.

Dombry et al. (2016), who suggested Algorithm 2.3, also derived the distribution of the extremal function for many commonly used models presented in Sections 2.2.2 and 2.2.3.

Example 2.20 (Brown–Resnick process)

For Brown–Resnick processes, the distribution $P_{\boldsymbol{s}_0}$ is equal in distribution to the log-Gaussian process

$$Y(\boldsymbol{s}) = \exp\left\{X(\boldsymbol{s}) - X(\boldsymbol{s}_0) - \gamma(\boldsymbol{s} - \boldsymbol{s}_0)\right\}, \qquad \boldsymbol{s} \in \mathcal{S},$$

where $X(\boldsymbol{s})$ is a Gaussian process with variogram $\gamma(\cdot)$ and stationary increments.

□

Remark 2.6

The alternative expression for the intensity function of the point process in Example 2.6

Algorithm 2.3 Exact simulation at D sites based on extremal functions (Dombry et al., 2016)

Require: Distribution of extremal functions $P_{\boldsymbol{s}_1}, \ldots, P_{\boldsymbol{s}_D}$.
simulate $\zeta^{-1} \sim \mathsf{Exp}(1)$ and $Y \sim P_{\boldsymbol{s}_1}$;
set $\boldsymbol{Z} = \zeta \boldsymbol{Y}$;
for $j = 2, \ldots, D$ **do**
    simulate $\zeta^{-1} \sim \mathsf{Exp}(1)$;
    while $Z_j < \zeta$ **do**
        simulate $\boldsymbol{Y}_j \sim P_{\boldsymbol{s}_j}$;
        if $\max_{i=1}^{j-1} \zeta Y_i < Z_i$ **then**
            set $\boldsymbol{Z} \leftarrow \max\{\boldsymbol{Z}, \zeta \boldsymbol{Y}\}$;
        simulate $E \sim \mathsf{Exp}(1)$;
        $\zeta^{-1} \leftarrow \zeta^{-1} + E$;
return $\boldsymbol{Z}$.

Algorithm 2.4 Sampling from the angular density on the $\|\cdot\|_1$ sphere

Require: Distribution of extremal functions $P_{\boldsymbol{s}_1}, \ldots, P_{\boldsymbol{s}_D}$.
simulate j_0 uniformly from $\{1, \ldots, D\}$;
simulate $\boldsymbol{Y} \sim P_{j_0}$;
return $\boldsymbol{S} \leftarrow \boldsymbol{Y} / \|\boldsymbol{Y}\|_1$.

follows from the equivalent spectral representation for the Brown–Resnick max-stable process, $\mathsf{Z}^* = \max_{i=1}^{\infty} \zeta_i \mathsf{Y}_i$, where ζ_i is a point process with intensity $\zeta^{-2} \mathrm{d}\zeta$ and $\mathsf{Y}_i \sim P_{\boldsymbol{s}_0}$, so $Y(\boldsymbol{s}_0) = 1$ almost surely (Dombry et al., 2016, Proposition 2 and Remark 1).

Example 2.21 (Extremal Student process)
Consider an extremal Student process with spectral representation (2.16) and correlation function $\rho(\cdot)$. The distribution $P_{\boldsymbol{s}_0}$ coincides with that of T_+^{ν}, with T a Student process with $\nu + 1$ degrees of freedom and location function $\mu(\boldsymbol{s}) = \rho(\boldsymbol{s}_0, \boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{S}$ and scale function $\widetilde{\rho}(\boldsymbol{s}_j, \boldsymbol{s}_k) = \{\rho(\boldsymbol{s}_j, \boldsymbol{s}_k) - \rho(\boldsymbol{s}_0, \boldsymbol{s}_j)\rho(\boldsymbol{s}_0, \boldsymbol{s}_k)\}/(\nu + 1)$, for $\boldsymbol{s}_j, \boldsymbol{s}_k \in \mathcal{S}$.

□

Dombry et al. (2016) note that if one is able to sample from the distribution of the scaled extremal functions it is easy to simulate from the corresponding spectral density based on the mixture representation. If the latter is unknown, but the distribution of the extremal function is known, then Dombry et al. (2016) suggest using Algorithm 2.4. The converse is also true: given a sample from the spectral density, one can divide the sample by the value of the j_0 component so that the latter is almost surely 1, to simulate from $P_{\boldsymbol{s}_0}$. Algorithm 2.3 requires fewer simulations on average than using a combination of Algorithms 2.2 and 2.4; simulations in Dombry et al. (2016) seem to indicate that the number of simulations for Algorithm 2.3 also has lower variance.

Recent work focuses on simulation algorithms whose cost does not grow linearly, see, e.g., Liu et al. (2016) and Oesting et al. (2018), but they appear tailored to specific models and require implementation of bespoke code. Refined methods exist that use the lack of uniqueness of the

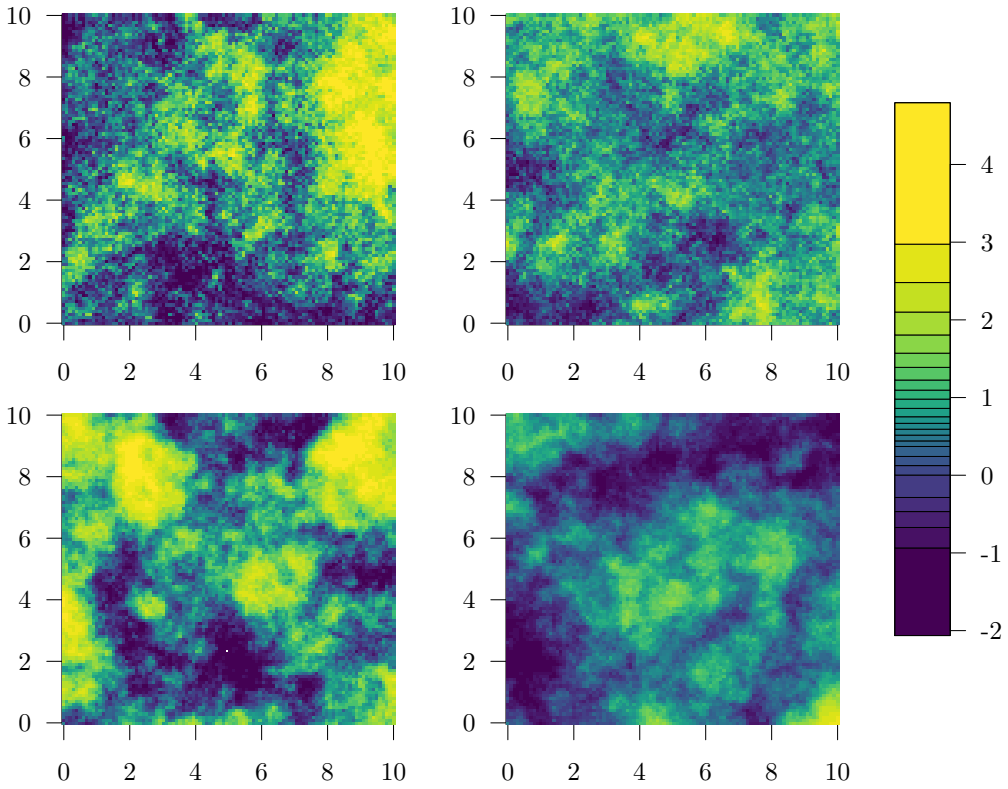

Figure 2.3 – *Exact simulations from Brown–Resnick max-stable processes with power semi-variogram* $\gamma(h) = (\|\boldsymbol{h}\|/\lambda)^{\alpha}$ *on* $[0,10]^2$ *with parameters* $\alpha = 0.7, \lambda = 1$ *(top left),* $\alpha = 0.7, \lambda = 3$ *(top right),* $\alpha = 1.2, \lambda = 1$ *(bottom left) and* $\alpha = 1.2, \lambda = 3$ *(bottom right). The data were simulated using the* R *package SpatialExtremes and transformed to Gumbel margins.*

processes $W(\mathcal{S})$, giving rise to the same max-stable process family by choosing candidate families suitably. The optimal choice of spectral functions (in the sense of minimizing the number of simulated candidate functions) is the so-called normalized spectral representation (Oesting, Schlather and Zhou, 2018). However, it is not known how to simulate from the normalized spectral representation; likewise the bounding constant appearing in these construction is often unknown and thus must be estimated.

For spectral functions based on Gaussian processes, the computational bottleneck is typically due to inversion of the covariance matrix or the Cholesky decomposition, which require $O(D^3)$ flops; see Appendix C.2 for a literature review. More efficient algorithms, notably the circulant embedding method, can be used for regular grids (Dietrich and Newsam, 1993; Wood and Chan, 1994). Figures 2.3 and 2.4 show exact simulations for the Brown–Resnick and extremal Student process on a 100×100 grid of points equally spaced on $[0,10]^2$.

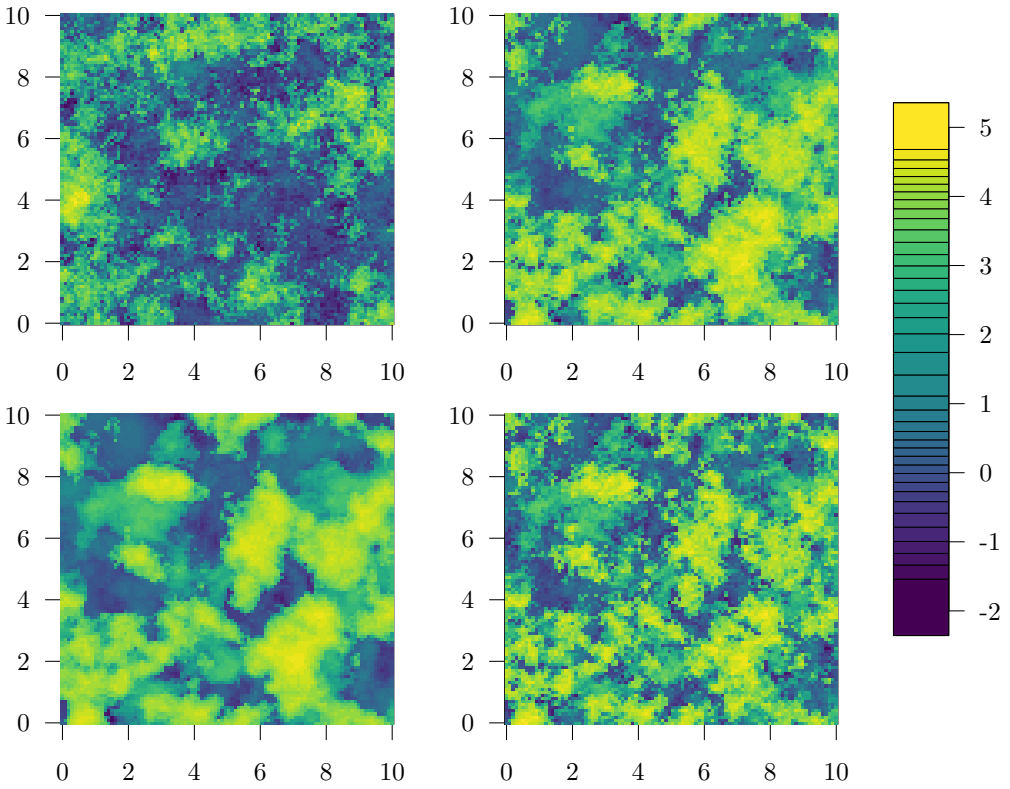

Figure 2.4 – *Exact simulations from extremal Student max-stable processes with* $\nu = 1$ *(Schlather's model) with correlation function* $\rho(h) = \exp\{-(\|\boldsymbol{h}\|/\lambda)^{\alpha}\}$ *on* $[0,10]^2$ *with parameters* $\alpha = 0.75, \lambda = 1$ *(top left),* $\alpha = 1, \lambda = 1$ *(top right),* $\alpha = 1.5, \lambda = 1$ *(bottom left) and* $\alpha = 1, \lambda = 0.5$ *(bottom right). The data were simulated using the* R *package RandomFields and transformed to Gumbel margins.*

2.5.3 Conditional simulation of max-stable processes

Supposing we observe a realization from a simple max-stable random field $Z(\mathcal{S})$ at $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$ and wish to simulate possible trajectories of $Z(\mathcal{S})$ conditional on the observed values $Z(\boldsymbol{s}_1) = z_1, \ldots, Z(\boldsymbol{s}_k) = z_k$, denoted hereafter $\boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}$. Provided that Λ is regular, one can derive the conditional distribution of $Z(\mathcal{S}) \mid \boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}$ using the notions of extremal and sub-extremal functions discussed on page 110 (Dombry and Eyi-Minko, 2013).

Since the event that a given realization of a max-stable process equals specified values at a fixed collection of sites is a Λ-null set (cf. Oesting et al., 2015), we treat conditional simulation in the multivariate framework, working with a predefined finite set of locations $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D\}$ and conditioning on the k sites in $\mathcal{S}$, without loss of generality the first ones. Algorithm 2.5 can be used to draw samples from the max-stable vector $\boldsymbol{Z}_{(k+1):D}$ conditional on the values $\boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}$.

Algorithm 2.5 Conditional simulations from max-stable processes (Dombry and Eyi-Minko, 2013)

Require: intensity measure Λ, conditioning set $\boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}$.
1: draw a hitting scenario $\boldsymbol{\Pi}$ conditional on $\boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}$ from $\mathsf{P}(\boldsymbol{\Pi} \in \cdot \mid \boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k})$;
2: sample the extremal functions Φ^+ conditional on the observed values of the process $\boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}$ and the partition $\boldsymbol{\Pi}$ from $\mathsf{P}\left(\varphi_j^+ \in \cdot \mid \boldsymbol{Z}_{1:k} = \boldsymbol{z}_{1:k}, \boldsymbol{\Pi} = \boldsymbol{\varpi}\right)$;
3: sample sub-extremal functions from a Poisson point process with intensity measure $\Lambda(\mathrm{d}\phi)\prod_{j=1}^k \mathbf{1}\{\phi(\boldsymbol{s}_j) < z_j\}$;
4: set $Z_i = \max_{j=1}^k \varphi_j^+(\boldsymbol{s}_i) \max_{\varphi\in\Phi^-} \varphi(\boldsymbol{s}_i)$ for $i = k+1, \ldots, D$;
5: **return** $\boldsymbol{Z}_{(k+1):D}$.

Conditional distributions and decomposition of max-stable processes

Our exposition follows Dombry et al. (2013) and Dombry et al. (2015). We encode the hitting scenario $\boldsymbol{\Pi}$ through a vector of sets: for each partition, ϖ_j denotes the jth component of $\boldsymbol{\varpi}$, the set of indices of the locations that appear in the jth hitting scenario component $\varpi_j \coloneqq \{i \in \{1,\ldots,k\} : \varphi^+_{\boldsymbol{s}_i} = \varphi^+_{\varpi_j}\}$. We write $\boldsymbol{z}_{\varpi_j^c}$ for the sub-vector whose components are in the complement of ϖ_j, $\{1,\ldots,k\} \setminus \varpi_j$. We use the notation $<_{\boldsymbol{s}}$ to indicate that the inequality holds pointwise for any $\boldsymbol{s} \in \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ and arguments of the intensity function $\lambda(\cdot)$ are understood to be ordered according to the conditioning set. The hitting scenario $\boldsymbol{\Pi} \mid \boldsymbol{Z}_{1:k} = \boldsymbol{z}$ has a discrete distribution on the set $\mathscr{P}_D$ of all non-empty partitions of $\{1,\ldots,k\}$, whose cardinality is the kth Bell number.

The point processes of extremal and sub-extremal functions conditional on $\boldsymbol{Z}_{1:k} = \boldsymbol{z}$,

$$\Phi^- = \{\varphi \in \Phi : \varphi(\boldsymbol{s}_i) < z_i, i = 1,\ldots,k\}, \quad \Phi^+ = \bigcup_{i=1}^k \{\varphi \in \Phi : \varphi(\boldsymbol{s}_i) = z_i\},$$

are independent (Dombry et al., 2013). Since $\Phi^- \perp\!\!\!\perp \{\boldsymbol{\Pi}, \Phi^+\} \mid \boldsymbol{Z}_{1:k} = \boldsymbol{z}$, we can simulate sub-extremal functions from a Poisson point process with intensity measure $\Lambda(\mathrm{d}\phi)\prod_{j=1}^k \mathbf{1}\{\phi(\boldsymbol{s}_j) < z_j\}$.

The max-stable vector $\boldsymbol{Z}_{1:k}$ has distribution

$$\begin{aligned}
\mathsf{P}(\boldsymbol{Z}_{1:k} \in \mathrm{d}\boldsymbol{z}) &= \sum_{\boldsymbol{\varpi}\in\mathscr{P}_D} \mathsf{P}(\boldsymbol{\Pi} = \boldsymbol{\varpi}, \boldsymbol{Z}_{1:k} \in \mathrm{d}\boldsymbol{z}) \\
&= \sum_{\boldsymbol{\varpi}\in\mathscr{P}_D} \mathsf{E}\left(\prod_{j=1}^{|\boldsymbol{\varpi}|} \mathbf{1}\{\varphi^+_{\varpi_j} \in \mathrm{d}\boldsymbol{z}_{\varpi_j}\}\mathbf{1}\{\varphi^+_{\boldsymbol{s}_k} < \boldsymbol{z}_{\varpi_j}, k \in \varpi_j^c\}\mathbf{1}\{\max\Phi^- <_{\boldsymbol{s}} \boldsymbol{z}\}\right) \\
&= \exp\{-V(\boldsymbol{z})\} \sum_{\boldsymbol{\varpi}\in\mathscr{P}_D} \left\{\prod_{j=1}^{|\boldsymbol{\varpi}|} \int \mathbf{1}\{\boldsymbol{u}_j < \boldsymbol{z}_{\varpi_j^c}\}\lambda_{\varpi_j,\varpi_j^c}(\boldsymbol{z}_{\varpi_j}, \boldsymbol{u}_j)\mathrm{d}\boldsymbol{u}_j\right\}\mathrm{d}\boldsymbol{z}.
\end{aligned}$$

This relation arises from consideration of all possible partitions and the density of extremal functions. The passage to the last line follows from the Slivnyak–Mecke formula, the independence of subextremal functions conditional on the values of $\boldsymbol{Z}_{1:k}$ and the assumption of regularity. Each term in the product corresponds to a different extremal function, which must

attain value z_j if it is in the partition ϖ_j at location $\boldsymbol{s}_j$ or else fall below the other extremal functions.

One can find the probability mass function of the hitting scenario conditional on the max-stable process using the law of total probability (Dombry et al., 2013),

$$\mathsf{P}\left(\boldsymbol{\Pi}=\boldsymbol{\omega} \mid \boldsymbol{Z}_{1:k}=\boldsymbol{z}\right)=\frac{\prod_{j=1}^{|\boldsymbol{\omega}|}\int \mathbf{1}\{\boldsymbol{u}_j<\boldsymbol{z}_{\varpi_j^{\mathsf{c}}}\}\lambda_{\varpi_j,\varpi_j^{\mathsf{c}}}(\boldsymbol{z}_{\varpi_j},\boldsymbol{u_j})\mathrm{d}\boldsymbol{u}_j}{\sum_{\boldsymbol{\omega}'\in\mathscr{P}_D}\prod_{j=1}^{|\boldsymbol{\omega}'|}\int \mathbf{1}\{\boldsymbol{u}_j<\boldsymbol{z}_{\varpi_j'^{\mathsf{c}}}\}\lambda_{\varpi_j',\varpi_j'^{\mathsf{c}}}(\boldsymbol{z}_{\varpi_j'},\boldsymbol{u_j})\mathrm{d}\boldsymbol{u}_j},$$

and under the assumption of regularity of Λ, we can express the intensity function in terms of the conditional intensity function $\lambda_{\varpi_j,\varpi_j^{\mathsf{c}}}(\boldsymbol{z}_{\varpi_j},\boldsymbol{u_j})=\lambda_{\varpi_j}(\boldsymbol{z}_{\varpi_j})\lambda_{\varpi_j^{\mathsf{c}}\mid\varpi_j}(\boldsymbol{u_j}\mid\boldsymbol{z}_{\varpi_j})$, yielding

$$\mathsf{P}\left(\boldsymbol{\Pi}=\boldsymbol{\omega} \mid \boldsymbol{Z}_{1:k}=\boldsymbol{z}\right)\overset{\boldsymbol{\omega}}{\propto}\prod_{j=1}^{|\boldsymbol{\omega}|}w_{\varpi_j}=\prod_{j=1}^{|\boldsymbol{\omega}|}\lambda_{\varpi_j}(\boldsymbol{z}_{\varpi_j})\int \mathbf{1}\{\boldsymbol{u}_j<\boldsymbol{z}_{\varpi_j^{\mathsf{c}}}\}\lambda_{\varpi_j^{\mathsf{c}}\mid\varpi_j}(\boldsymbol{u_j}\mid\boldsymbol{z}_{\varpi_j})\mathrm{d}\boldsymbol{u}_j,$$

where $\overset{\boldsymbol{\omega}}{\propto}$ indicates proportionality with respect to $\boldsymbol{\omega}$, i.e., the conditional probability mass function is given up to normalizing constants. Closed-form expressions for the conditional intensity of Brown–Resnick and extremal Student processes were given in Example 2.5 and Example 2.7. The extremal functions are conditionally independent given $\boldsymbol{\omega}=\boldsymbol{\omega}$ and (Dombry et al., 2015, Theorem 1.2.5)

$$\mathsf{P}\left(\varphi_j^{+}(\boldsymbol{s})\in\mathrm{d}\varphi \mid \boldsymbol{Z}_{1:k}=\boldsymbol{z},\boldsymbol{\Pi}=\boldsymbol{\omega}\right)=\Lambda\{\mathrm{d}\varphi\mid\varphi(\boldsymbol{s}_i)=z_i,i\in\varpi_j,\varphi(\boldsymbol{s}_l)<z_l,l\in\varpi_j^{\mathsf{c}}\}.$$

The conditional distribution of $\varphi_j^{+}(\boldsymbol{s})$ is thus obtained from the joint vector, conditioning on the fact that all other extremal functions must lie below $\boldsymbol{z}_{\varpi_j}$ at sites $\{\boldsymbol{s}_k,k\in\varpi_j\}$. It follows that the conditional random process $\mathsf{Z}(\mathcal{S})\mid\boldsymbol{Z}_{1:k}$ is not max-stable (Dombry et al., 2013).

The cardinality of the state space containing the set of partitions $\mathscr{P}_k$ is the kth Bell number, B_k. The integration constant appearing in the conditional distribution of the hitting scenario is computationally demanding since it requires the calculation of a high-dimensional integral for each of these B_k combinations. Even for $k\approx 20$, it is impossible to store the output in memory. To circumvent this problem, Dombry et al. (2013) turned to Markov chain Monte Carlo for inference. A convenient choice to deal with the hitting scenario is to update the partition to which $\boldsymbol{s}_j$ belongs sequentially for $j=1,\ldots,k$ rather than perform block update sampling from $\mathscr{P}_k$. Dombry et al. (2013) suggest a random scan Gibbs sampler (Liu, Wong and Kong, 1995) conditioning on $\boldsymbol{Z}_{1:k}=\boldsymbol{z}$ and the hitting scenario at all locations but the jth site. At each iteration of the Gibbs sampler, a location j is selected uniformly at random from $\{1,\ldots,k\}$ and the jth location is reassigned to one of the $|\boldsymbol{\omega}|$ existing clusters (including $\boldsymbol{\omega}$) if the cluster containing site j consisted of the singleton j, or else $|\boldsymbol{\omega}|+1$ clusters, potentially defining a new cluster containing only location $\boldsymbol{s}_j$. A Metropolis–Hastings step is used to determine whether or not to accept the new move to a partition $\boldsymbol{\omega}'$ in which only component j is reallocated, with acceptance ratio $\min\{1,\prod_{i=1}^{|\boldsymbol{\omega}|}w_{\varpi_i}/\prod_{l=1}^{|\boldsymbol{\omega}'|}w_{\varpi_l'}\}$. All but four of the weights w_{ϖ_i} cancel. The cost of scanning through these moves is thus linear and for each proposed

Algorithm 2.6 Accept-reject algorithm

Require: proposal density $g(\boldsymbol{x})$ with $f \ll g$, $c \geq \sup_{\boldsymbol{x}\in\mathrm{supp}(F)} f(\boldsymbol{x})/g(\boldsymbol{x})$.
1: **repeat**
2: sample $\boldsymbol{X} \sim g$;
3: sample independently $U \sim \mathsf{U}(0,1)$;
4: **until** $cU < f(\boldsymbol{X})/g(\boldsymbol{X})$
5: **return** $\boldsymbol{X}$

move, the most costly part is the evaluation of the integrals giving the weights w_{ϖ_i}. The price to pay for going from block to sequential updates is potentially slower mixing of the Markov chains.

2.5.4 Unconditional simulation of ℓ-Pareto processes

In principle, simulation from ℓ-Pareto processes is simpler than simulation from the associated max-stable process, since individual events correspond to particular rescaled spectral functions. Specifically, realizations from ℓ-Pareto processes consist of points from $\Phi = \{\varphi_i\}$, where $\varphi_i = \zeta_i \mathsf{W}_i$, falling in the risk region $\ell(\varphi) > u$.

The easiest method for simulation is through rejection sampling: suppose that one is interested in sampling $\boldsymbol{X} \sim F$; if $\boldsymbol{X}$ is difficult to sample from, but has a density $f(\boldsymbol{x})$ that can be easily computed, an alternative is to sample proposals G assuming there exists a bound c such that $f(\boldsymbol{x}) \leq cg(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathrm{supp}(F)$; this is summarized in Algorithm 2.6 (cf. Devroye, 1986, p. 42). The function g must have heavier tails and larger peaks than f and the upper bound of the likelihood ratio, c, should be as small as possible to maximize the acceptance rate.

Consider two risk functionals, ℓ_1 and ℓ_2, and suppose we want to simulate from the ℓ_1-Pareto process, but know only how to sample from the ℓ_2-Pareto process. For ℓ-Pareto processes, we rely on the stochastic representation (2.25), $P\mathsf{S}/\ell(\mathsf{S})$ for P a Pareto variable and S a random function with angular distribution $\rho_{\ell,\mathrm{ang}}$. We can use the change of measure argument in Proposition 2.28 and simulate from the point process on a large domain that includes the region of interest, using rejection sampling to retain only points falling in the risk region (de Fondeville and Davison, 2018). With the densities $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$ equal to $\lambda(\boldsymbol{x})/\Lambda\{\boldsymbol{x} : \ell_1(\boldsymbol{x}) > u_1\}$ and $\lambda(\boldsymbol{x})/\Lambda\{\boldsymbol{x} : \ell_2(\boldsymbol{x}) > u_2\}$ respectively, the bounding constant c in Algorithm 2.6 is $\Lambda\{\boldsymbol{x} : \ell_2(\boldsymbol{x}) > u_2\}/\Lambda\{\boldsymbol{x} : \ell_1(\boldsymbol{x}) > u_1\}$; the likelihood ratio need not be evaluated, since any point sampled from $g(\boldsymbol{x})$ such that $\ell_1(\boldsymbol{x}) > u_1$ is automatically accepted.

Method 2.22 (Rejection sampling for ℓ-Pareto processes)
Suppose one can simulate components ω from an (unnormalized) spectral measure $\rho^*_{\ell_1,\mathrm{ang}}$ so as to generate from the point process above $\{\boldsymbol{x} : \ell_2(\boldsymbol{x}) > u_2\}$. Then, provided that $\ell_1(\boldsymbol{x}) > 0$ $\rho_{\ell_2,\mathrm{ang}}$-almost everywhere over a threshold u_1, we can generate points from $\rho^*_{\ell_2,\mathrm{ang}}$ and devise an accept-reject scheme based on Proposition 2.28; the acceptance rate is obtained from

Equation (2.24).

$$p(\omega) = \min\left[1, \frac{\Lambda(f \in \mathscr{F} \setminus \mathscr{C}_{\ell_2} : \ell_1(f) \geq u_1)}{\Lambda(f \in \mathscr{F} \setminus \mathscr{C}_{\ell_2} : \ell_2(f) \geq u_2)} \left\{\frac{\ell_2(\omega)}{\ell_1(\omega)}\right\}^{\alpha}\right],$$

Choosing u_1 as large as possible, $\sup_{u_1}\{\boldsymbol{x} : \ell_1(\boldsymbol{x}) > u_1\} \subseteq \{\boldsymbol{x} : \ell_2(\boldsymbol{x}) > u_2\}$, leads to higher efficiency.

For some risk functionals, exact simulation from the angular measure is possible. For example, Algorithm 2.4 shows how to simulate from the sphere associated with the l_1 norm $\|\cdot\|_1$ by drawing $\boldsymbol{S}$ from a balanced mixture of extremal functions $P_{\boldsymbol{s}_1}, \ldots, P_{\boldsymbol{s}_D}$. This means that we can simulate from the l_1 angular distribution and sample for functionals such as max, min or $l_2 = \|\cdot\|_2$, with respectively $u^*_{\max} = 1$, $u^*_{\min} = D$ and $u^*_{l_2} = 1$. The acceptance ratio decreases with D, but is generally large enough to provide fast simulations provided the spatial dependence is very strong, so that most points are far from the axes.

While simulating from the angular distribution associated to the l_1-norm is the most obvious candidate for an accept-reject sampling algorithm, other choices may be available (cf. Oesting et al., 2019). The next example illustrates accept-reject for elliptical extremes using samples from the angular components associated to the Malahanobis norm.

Example 2.23 (Multivariate generalized Pareto with elliptic distributions)
A zero mean elliptic vector admits the stochastic representation $\boldsymbol{X} = R\mathbf{A}\boldsymbol{U}$, where R is a radial variable, $\mathbf{A}$ is the Cholesky root of a correlation matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$ and $\boldsymbol{U} \sim \mathsf{U}\{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x}^\top\boldsymbol{x} = 1\}$ is a vector sampled uniformly on the ℓ_2 sphere. Maxima of multivariate elliptic vectors with a radial component R which is regularly varying with index ν are attracted to extremal Student max-stable processes (Example 2.4).

Thibaud and Opitz (2015) show that $\mathbf{A}\boldsymbol{U}$ corresponds to samples from the spectral measure associated to the Mahalanobis norm, ρ_Σ. Let $R \sim \mathsf{Par}(\nu)$. The conditional random vector $\boldsymbol{Y} = R\mathbf{A}\boldsymbol{U} \,\big|\, \max_{j=1}^D (R\mathbf{A}\boldsymbol{U})_j \geq 1$ follows a multivariate generalized Pareto with elliptic extremal distribution characterized by the tail index ν, scale $\boldsymbol{\Sigma}$ and threshold $\mathbf{1}_D$ and can be simulated using an accept-reject scheme. The angular component is then $(\mathbf{A}\boldsymbol{U})^\nu_+ / \max\{(\mathbf{A}\boldsymbol{U})^\nu_+\}$. The vector $\mathrm{sgn}(\boldsymbol{Y})|\boldsymbol{Y}|^\nu$ has tail index 1, since R^ν is unit Pareto if R is ν-Pareto.

□

Remark 2.7
The accept-reject approach of Opitz (2013b) and Thibaud and Opitz (2015) differs from the proposal of de Fondeville and Davison (2018) only in the choice of the norm for the angular measure. The acceptance rate is the ratio of the measure of the sets, which is $V(\boldsymbol{u})/D$ for the ρ_1 norm and $V(\boldsymbol{u})\pi^{-1/2}\Gamma(\nu/2+1/2)\Gamma(D/2)\{2\Gamma(D/2+\nu/2)\}^{-1}$ for the $\rho_{\boldsymbol{\Sigma}}$ norm. Simulation from ρ_1 in Example 2.23 is always at least twice as efficient as sampling from the Mahalanobis angular distribution when $\nu \geq 2$. When $\nu = 2$, the relative acceptance ratio is 2 regardless of D and the only cases for which using ρ_Σ is preferable is when $\nu < 2$ and D is large, and even in this case the ratio of acceptance rates is close to unity.

Algorithm 2.7 Composition sampling for standard Pareto processes based on $\ell = \max$

1: sample an index j in $\{1,\ldots,D\}$ with probability $\psi_i / V(\mathbf{1}_D)$, $i = 1,\ldots,D$
2: sample a realization w_j from the marginal distribution of W_j
3: conditional on this value, simulate from $\mathsf{P}\left(\boldsymbol{W}_{-j} | \boldsymbol{W}_{-j} \leq w_j\right)$
4: set $\boldsymbol{\omega}_{\max} \leftarrow \boldsymbol{w}/w_j$
5: simulate $R \sim \mathsf{Par}(1)$
6: **return** $Z \leftarrow R\boldsymbol{\omega}_{\max}$.

For most parametric models utilized in the literature, one knows how to simulate from the spectral distribution $P_{\boldsymbol{s}_0}$ of the rescaled extremal function $\phi^+(\boldsymbol{s}_0)/Z(\boldsymbol{s}_0)$, which corresponds to an exceedance at site $\boldsymbol{s}_0$ (Dombry et al., 2016). We propose an alternative simulation technique when the distribution $\boldsymbol{S}$ can be represented as a mixture of extremal functions, as in the case of Algorithm 2.4. Sampling from the mixture is done in two steps, by first drawing the index of the mixture component i from $\{1,\ldots,D\}$, each with associated weight ψ_j and secondly simulating $\boldsymbol{S} \sim P_{\boldsymbol{s}_i}$. The rescaled angular vector is simply $\boldsymbol{S}/\ell(\boldsymbol{S})$. Ho (2018) proposed such an approach for simulating from multivariate generalized Pareto Brown–Resnick processes, but the procedure is more general and can be applied if the risk functional can be decomposed via indicators and linear combinations of $X(\boldsymbol{s}_j)$ $(j = 1,\ldots,D)$; this is the case for weighted maxima, minima, averages and projections.

Proof

We give the proof of Algorithm 2.7 for the multivariate Pareto distribution, corresponding to the case $\ell(f) = \max_{i=1}^D f(\boldsymbol{s}_i)$, but the case $\ell(\mathsf{X}) = \min_{j=1}^D X(\boldsymbol{s}_j)$ is analogous.

Let Λ be a -1-homogeneous measure that factorises in pseudo-polar and angular components,

$$\Lambda(A) = \int_{\mathscr{F}} \int_0^\infty \mathbf{1}\{\zeta \max_{j=1}^D f(\boldsymbol{s}_j) > u\} \mathbf{1}\{f(\boldsymbol{s})/\|f(\boldsymbol{s})\|_{\mathrm{ang}} \in A\} \zeta^{-2} \mathrm{d}\zeta \nu(\mathrm{d}f).$$

If we restrict the point process to the set $\left\{f \in \mathscr{F} : \max_{j=1}^D \{f(\boldsymbol{s}_j)\} > u\right\}$, the normalized intensity defines a probability measure on the latter

$$\begin{aligned}
&\int_{\mathscr{F}} \int_0^\infty \mathbf{1}\{\zeta \max_{j=1}^D f(\boldsymbol{s}_j) > u\} \mathbf{1}\{f(\boldsymbol{s})/\|f(\boldsymbol{s})\|_{\mathrm{ang}} \in A\} \zeta^{-2} \mathrm{d}\zeta \nu(\mathrm{d}f) \\
&\quad = \frac{1}{u} \int_{\mathscr{F}} \max_{j=1}^D f(\boldsymbol{s}_j) \mathbf{1}\{f(\boldsymbol{s})/\|f(\boldsymbol{s})\|_{\mathrm{ang}} \in A\} \nu(\mathrm{d}f) \\
&\quad = \frac{1}{u} \sum_{j=1}^D \int_{\mathscr{F}} f(\boldsymbol{s}_j) \mathbf{1}\{f(\boldsymbol{s}_j) > f(\boldsymbol{s}_i), i = 1,\ldots,D, i \neq j\} \mathbf{1}\{f(\boldsymbol{s})/\|f(\boldsymbol{s})\|_{\mathrm{ang}} \in A\} \nu(\mathrm{d}f) \\
&\quad = \frac{1}{u} \sum_{j=1}^D \int_{\mathscr{F}} \mathbf{1}\{f(\boldsymbol{s}_i) < 1, i = 1,\ldots,D, i \neq j\} P_{\boldsymbol{s}_j}(\mathrm{d}f),
\end{aligned}$$

since by definition rescaled extremal functions satisfy

$$\int_{\mathscr{F}} f(\boldsymbol{s}_j) \mathbf{1}\{f/\|f\|_{\mathrm{ang}} \in A\} \nu(\mathrm{d}f) = \int_{\mathscr{F}} \mathbf{1}\{f/\|f\|_{\mathrm{ang}} \in A\} P_{\boldsymbol{s}_j}(\mathrm{d}f)$$

and $P_{\boldsymbol{s}_j}$ is supported on the set $\{f \in \mathscr{F} : f(\boldsymbol{s}_j) = 1\}$. The normalization constant that makes Λ a probability measure is $V(u\mathbf{1}_D)$, so we must also normalize the remaining integral term to obtain a conditional probability distribution. The normalizing constant is

$$\psi_j := \int_{\mathscr{F}} \mathbf{1}\{f(\boldsymbol{s}_i) < 1, i = 1, \ldots, D, i \neq j\} P_{\boldsymbol{s}_j}(\mathrm{d}f).$$

These are precisely the terms appearing in the calculation of the exponent measure in eq. (2.13). In particular, $\sum_{j=1}^{D} \psi_j / V(u\mathbf{1}_D) = 1$ so the contribution of ψ_j to the likelihood cancels.

■

Algorithm 2.7 requires sampling truncated components $\boldsymbol{W}$; this can be achieved for Brown–Resnick and extremal Student processes using Algorithm 3.2, which is covered in the next chapter.

The computational bottleneck of Algorithm 2.7 is the calculation of the exponent measure, which typically involves high-dimensional integrals that need to be evaluated numerically. However, and contrary to the accept-reject schemes described before, only the set-up cost is high. For risk regions defined by $\ell(\mathsf{X}) = \min_{j=1}^{D} X(\boldsymbol{s}_j)$, the composition sampling simulation algorithm is particularly advantageous. If the mixing weights cannot be derived analytically, an alternative is to resort to Monte Carlo methods by simulating repeatedly from the scaled extremal functions and calculating the empirical average for the indicator set.

We provide two examples of computations for the weights of the mixture.

Example 2.24 (Mixture weights for bivariate Brown–Resnick Pareto distribution)
For simplicity, we consider a bivariate Brown–Resnick model; the calculations readily extend to higher dimensions. The derivations are obtained along the lines of Huser and Davison (2013).

By construction, the Brown–Resnick process is of the form $W(\boldsymbol{s}_i) = \exp\{X(\boldsymbol{s}_i) - \gamma(\boldsymbol{s}_i)\}$ for $i = 1, 2$ with $\{X(\boldsymbol{s})\}_{\boldsymbol{s} \in \mathcal{S}}$ intrinsically stationary Gaussian processes with semi-variogram γ with $X(\boldsymbol{o}) = 0$. We adopt the shorthand notation $X_i = X(\boldsymbol{s}_i)$, etc.

Let $\boldsymbol{X} \sim \mathsf{No}_2(\mathbf{0}_2, \boldsymbol{\Sigma})$ with $\Sigma_{ii} = 2\gamma_i$ $(i = 1, 2)$ and $\Sigma_{1,2} = \gamma_1 + \gamma_2 - \gamma_{1,2}$; we have $W_1 > W_2$ if and only if $X_1 - \gamma_1 + \gamma_2 > X_2$ and we can consider the change of variable $x_2^* = x_2 - x_1 + \gamma_1 - \gamma_2$. Using the law of iterated expectation and variance, the marginal distribution of X_2^* is Gaussian with expectation

$$\mathsf{E}\left(X_2^*\right) = \mathsf{E}_{X_1}\left(\mathsf{E}_{X_2|X_1}\left(X_2 - X_1 + \gamma_1 - \gamma_2\right)\right) = \mathsf{E}_{X_1}\left(\frac{\gamma_1 + \gamma_2 - \gamma_{1,2}}{2\gamma_1} X_1 - X_1\right) + \gamma_1 - \gamma_2 = \gamma_1 - \gamma_2,$$

and variance

$$\begin{aligned}\mathsf{Var}\left(X_2^*\right) &= \mathsf{Var}_{X_1}\left(\mathsf{E}_{X_2|X_1}\left(X_2 - X_1 + \gamma_1 - \gamma_2\right)\right) + \mathsf{E}_{X_1}\left(\mathsf{Var}_{X_2|X_1}\left(X_2 - X_1 + \gamma_1 - \gamma_2\right)\right) \\ &= \frac{(-\gamma_1 + \gamma_2 - \gamma_{1,2})^2}{(2\gamma_1)^2} \mathsf{Var}_{X_1}(X_1) + 2\gamma_2 - \frac{(\gamma_1 + \gamma_2 - \gamma_{1,2})^2}{2\gamma_1} = 2\gamma_{1,2},\end{aligned}$$

and thus

$$\mathsf{P}\left(W_1 > W_2\right) = \mathsf{P}\left(X_2^* < 0\right) = \Phi\left\{-\frac{\gamma_1 - \gamma_2}{(2\gamma_{1,2})^{1/2}}\right\}.$$

If $\gamma_1 = \gamma_2$, each component is maximal with probability 1/2.

□

Example 2.25 (Scaled extremal Dirichlet model)
Consider the setup of Example 2.15; for the $\max_{i=1}^D$ risk functional, the simplest way to obtain the weights $\boldsymbol{w}$ is to use Monte Carlo methods, sampling $\boldsymbol{D} \sim \mathsf{Dir}(\boldsymbol{\alpha})$ and returning the index of the component-maximum for $D_j^\kappa / c(\alpha_j, \kappa)$, i.e., $\mathbf{1}\{j \in \{1,\ldots,D\} : D_j^\kappa / c(\alpha_j,\kappa) = \max_{i=1}^D D_i^\kappa / c(\alpha_i,\kappa)\}$. The mixing weights can be obtained cheaply at any desired level of accuracy. Since the extremal profiles are independent, it suffices to simulate marginal components from the scaled Gamma distribution in the second step and to use an accept-reject algorithm to truncate the components below unity.

□

Method 2.26 (Markov chain Monte Carlo methods for Pareto process simulation)
Dombry and Ribatet (2015) propose a Markov chain Monte Carlo procedure for sampling from a ℓ-Pareto process, by simulating from the spectral functions $W(\boldsymbol{x})$ and running a Metropolis–Hastings algorithm. Sample $\boldsymbol{W}_1$ and set $\boldsymbol{Y}_1 = \boldsymbol{W}_1$ and accept new proposals with probability $\min(\{\ell(\boldsymbol{W}_n)/\ell(\boldsymbol{Y}_{n-1})\}^\nu, 1)$. The sequence of D-vectors for the random fields $\{\boldsymbol{Y}_i / \|\boldsymbol{Y}_i\|_{\text{ang}}\}_{i \in \mathbb{N}^+}$ is a reversible Markov chain whose stationary distribution is $\rho_{\ell,\text{ang}}$. This scheme is also proposed in Rootzén et al. (2018) for the max risk functional and by Oesting et al. (2019).

2.5.5 Unconditional simulation of generalized ℓ-Pareto processes

We can use representation (2.26) to derive an accept-reject algorithm (Algorithm 2.8) and obtain unconditional simulations from a given parametric family of generalized ℓ-Pareto processes with potentially different shape parameters $\xi \in \mathbb{R}^D$, provided the distribution of the extremal function, $P_{\boldsymbol{s}_0}$, is known. To this effect, we need u' such that

$$\left\{\boldsymbol{y} \in \mathcal{F}_0 : \ell\left(\boldsymbol{\tau}\frac{\boldsymbol{y}^{\boldsymbol{\xi}} - \mathbf{1}_D}{\boldsymbol{\xi}} + \boldsymbol{\eta}\right) > u\right\} \subset \left\{\boldsymbol{y} \in \mathcal{F}_0 : \|\boldsymbol{y}\|_1 > u'\right\}. \tag{2.28}$$

For example, if ℓ is $\max_{j=1}^D$, $u' = \min_{j=1}^D \{1 + \xi_j(u - \eta_j)/\tau_j\}^{1/\xi_j}$, whereas we can take $u' = \sum_{j=1}^D \{1 + \xi_j(u-\eta_j)/\tau_j\}^{1/\xi_j}$ if $\ell(\boldsymbol{x}) = \min_{j=1}^D x_j$. If we condition on exceedances at a single site $\boldsymbol{s}_0$, we can directly sample $\boldsymbol{W}$ from $P_{\boldsymbol{s}_0}$ and a Pareto variable above $u' = \{1 + \xi_j(u-\eta_j)/\tau_j\}^{1/\xi_j}$, in which case the acceptance rate is unity.

Example 2.27 (Bounds for generalized ℓ-Pareto processes with sum risk functional)
The choice of threshold for eq. (2.28) when $\ell(\boldsymbol{X}) = \sum_{j=1}^D X_j$ is not straightforward and the geometry depends on $\boldsymbol{\xi}$, which we assume may be unequal. The trivial bound is $u' = D$, which

is an option if for example $u - \sum_{j=1}^{D} \eta_j < 0$. If $\boldsymbol{\xi} > \mathbf{0}_D$, then

$$\left\{ \boldsymbol{y} \geq \mathbf{1}_D : \sum_{j=1}^{D} \frac{\tau_j}{\|\boldsymbol{\tau}\|_1} \frac{y_j^{\xi_j} - 1}{\xi_j} > \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right\} \subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \max_{j=1}^{D} \frac{y_j^{\xi_j} - 1}{\xi_j} > \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right\}$$
$$\subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \max_{j=1}^{D} y_j^{\xi_j} > 1 + \min_{j=1}^{D} \xi_j \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right\},$$

so we can take

$$u' = \left(1 + \min_{j=1}^{D} \xi_j \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right)^{\min_{j=1}^{D} \xi_j^{-1}}.$$

Similar calculations for $\boldsymbol{\xi} < \mathbf{0}_D$ give

$$\left\{ \boldsymbol{y} \geq \mathbf{1}_D : \sum_{j=1}^{D} \frac{\tau_j}{\|\boldsymbol{\tau}\|_1} \frac{y_j^{\xi_j} - 1}{\xi_j} > \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right\}$$
$$\subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : 1 - \min_{j=1}^{D} y_j^{-|\xi_j|} > \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right\}$$
$$\subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \max_{j=1}^{D} y_j^{-|\xi_j|} < 1 - \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right\}$$
$$\subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \min_{j=1}^{D} y_j > \left(1 - \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right)_+^{\min_{j=1}^{D} |\xi^{-j}|} \right\},$$

yielding the bound

$$u' = D \left(1 - \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right)_+^{\min_{j=1}^{D} |\xi_j^{-1}|}.$$

If $\boldsymbol{\xi} = \mathbf{0}_D$, we can use the relationship between geometric and arithmetic means to get

$$\left\{ \boldsymbol{y} \geq \mathbf{1}_D : \sum_{j=1}^{D} \tau_j \log(y_j) + \eta_j > u \right\} = \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \prod_{j=1}^{D} y_j^{\tau_j / \|\boldsymbol{\tau}\|_1} > \exp\left(\frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right) \right\}$$
$$\subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \left(\prod_{j=1}^{D} y_j \right)^{1/D} > \exp\left(\frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right)^{1/D} \right\}$$
$$\subset \left\{ \boldsymbol{y} \geq \mathbf{1}_D : \sum_{j=1}^{D} y_j > D \exp\left(\frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1} \right)^{1/D} \right\},$$

but the bound is not tight. If $\boldsymbol{\xi}$ contains both positive and negative components, we can adapt

Algorithm 2.8 Accept-reject algorithm for generalized ℓ-Pareto processes (de Fondeville, 2018, p. 104)

Require: thresholds u, u' satisfying eq. (2.28).
1: initialize $\boldsymbol{Y} \leftarrow \mathbf{0}_D$
2: **while** $\ell(\boldsymbol{Y}) < u'$ **do**
3: simulate $R/u \sim \mathsf{Par}(1)$
4: simulate $\boldsymbol{W}$ from the $\|\cdot\|_1$ spectral density (Algorithm 2.4)
5: $\boldsymbol{Y} \leftarrow \boldsymbol{\tau}\{(R\boldsymbol{W})^{\boldsymbol{\xi}} - \mathbf{1}_D\}/\boldsymbol{\xi} + \boldsymbol{\eta}$
6: **return** $\boldsymbol{Y}$

the first bound:

$$\begin{aligned}
&\left\{\boldsymbol{y} \geq \mathbf{1}_D : \sum_{j=1}^{D} \tau_j \log(y_j) + \eta_j > u\right\} \\
&\quad \subset \left\{\boldsymbol{y} \geq \mathbf{1}_D : \max\left\{2 - \max_{j:\xi_j<0} y_j^{\xi_j}, \max_{k:\xi_k>0} y_j^{\xi_k}\right\} > 1 + \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1}\right\} \\
&\quad \subset \left\{\boldsymbol{y} \geq \mathbf{1}_D : \max_{j=1}^{D} y_j^{|\xi_j|} > 1 + \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1}\right\},
\end{aligned}$$

given that $2 \leq y^{-|\xi|} + y^{|\xi|}$ for $y \geq 1$ and the function is increasing in y; as a result

$$u' = \left(1 + \min_{j=1}^{D} |\xi_j| \frac{u - \sum_{j=1}^{D} \eta_j}{\|\boldsymbol{\tau}\|_1}\right)^{\min_{j=1}^{D} |\xi_j|^{-1}}.$$

□

Conditional simulation of ℓ-Pareto processes is unexplored, but is intimately related to the conditional intensity of the normalized point process and depends on the parametric model employed for the extremal profile $\mathsf{W}(\mathcal{S})$. For Brown–Resnick and extremal Student processes, the conditional distributions are log-Gaussian and Student and we can leverage properties of conditional elliptical distributions (cf. Proposition C.2 and Proposition C.3) for simulating from the latter. As for unconditional simulations, simulated points are kept only if they fall in the risk region defined by ℓ.

2.5.6 Unconditional simulation from the conditional extremes model

The distribution corresponding to the conditional model for $\mathsf{X} \mid X(\boldsymbol{s}_0) > u_0$ is denoted $Q_0(u_0)$. Should one be interested in simulating extreme events corresponding to extremes at a given site, i.e., $\{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}} \mid \max_{j=1}^{D} X(\boldsymbol{s}_j)/u_j > 1$, it is possible to use the mixture representation. Keef et al. (2013) propose to use empirical proportions to estimate the probability of a component being largest, but an undesirable feature of this approach that no extreme event is ever simulated at a site if no exceedance was observed there. We propose a novel alternative based

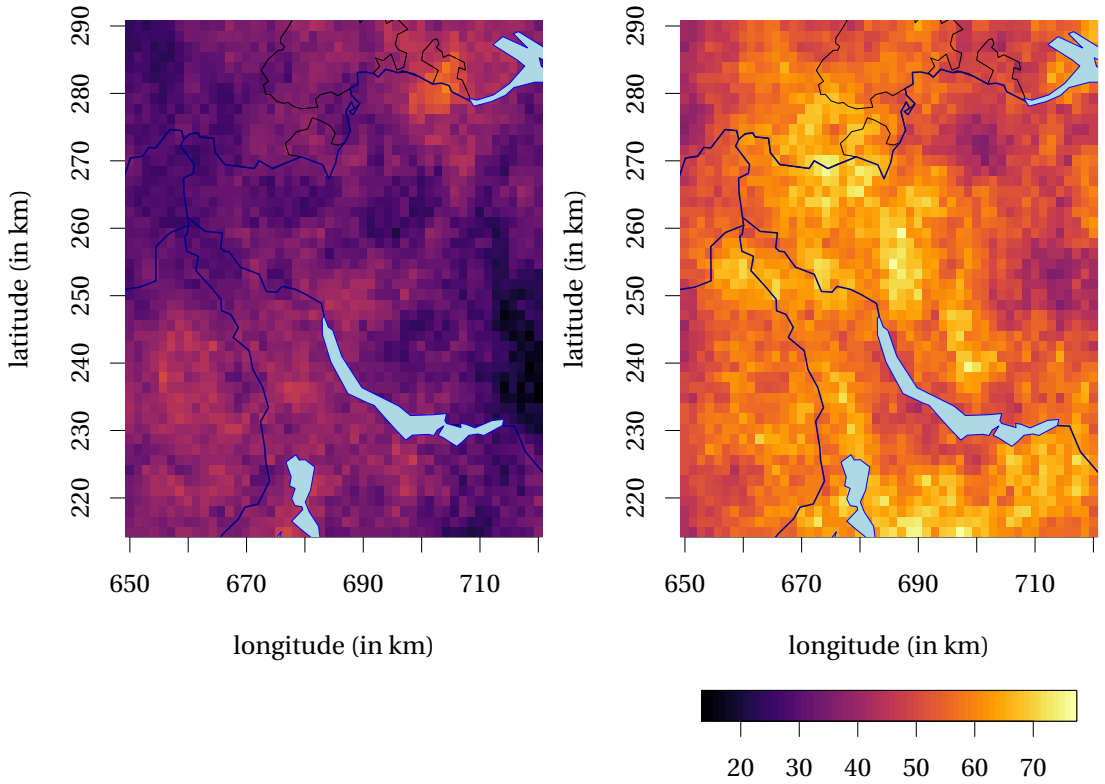

Figure 2.5 – *Unconditional simulation from a generalized ℓ-Pareto Brown–Resnick process with power semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/4)^{0.7}$ and location and scale parameters $\boldsymbol{\eta}, \boldsymbol{\tau}$ satisfying $\mathsf{E}(\boldsymbol{\eta}) = 26, \mathsf{E}(\boldsymbol{\tau}) = 9$ and $\xi = 0.1$ over canton Zürich. The risk functional are $\ell(\boldsymbol{Z}) = \max_{j=1}^{D} Z(\boldsymbol{s}_j)$ (left) and $\ell(\boldsymbol{Z}) = Z(\boldsymbol{s}_0)$ (right), both with a functional threshold of $u = 50$ units.*

on a decomposition slightly different than that proposed in Wadsworth and Tawn (2018), who suggest importance sampling.

We allow for potentially different thresholds u_j, since events of a given intensity on the data scale may be mapped to different quantiles on the standardized scale. We can decompose the event

$$\begin{aligned}
\pi_k &= \mathsf{P}\left(\frac{X(\boldsymbol{s}_k)}{u_k} = \max_{j=1}^{D}\left\{\frac{X(\boldsymbol{s}_j)}{u_j}\right\} \,\middle|\, \max_{j=1}^{D}\left\{\frac{X(\boldsymbol{s}_j)}{u_j}\right\} > 1\right) \\
&= \sum_{i=1}^{D} \mathsf{P}\left(\frac{X(\boldsymbol{s}_k)}{u_k} = \max_{j=1}^{D}\left\{\frac{X(\boldsymbol{s}_j)}{u_j}\right\}, \max_{j=1}^{D}\left\{\frac{X(\boldsymbol{s}_j)}{u_j}\right\} > 1 \,\middle|\, \frac{X(\boldsymbol{s}_i)}{u_i} > 1\right) \frac{\mathsf{P}\left(X(\boldsymbol{s}_i) > u_i\right)}{\mathsf{P}\left(\max_{j=1}^{D}\left\{X(\boldsymbol{s}_j)/u_j\right\} > 1\right)} \\
&\propto \sum_{i=1}^{D} \mathsf{P}\left(\frac{X(\boldsymbol{s}_k)}{u_k} = \max_{j=1}^{D}\left\{\frac{X(\boldsymbol{s}_j)}{u_j}\right\} \,\middle|\, \frac{X(\boldsymbol{s}_i)}{u_i} > 1\right) \mathsf{P}\left(X(\boldsymbol{s}_i) > u_i\right)
\end{aligned}$$

using the law of total probability; the term $\mathsf{P}\{\max_{j=1}^{D} X(\boldsymbol{s}_j)/u_j > 1\}$ needs not be evaluated as it is common to every π_k. We can estimate $\mathsf{P}\left(X(\boldsymbol{s}_i) > u_i\right)$ through the marginal survival function. The conditional probability $\mathsf{P}\left(X(\boldsymbol{s}_k)/u_k = \max_{j=1}^{D} X(\boldsymbol{s}_j)/u_j\right)$ can be estimated by

Algorithm 2.9 Unconditional simulation from the conditional extremes model given $\max_{j=1}^{D} X(\boldsymbol{s}_j)/u_j > 1$

Require: $\widehat{\boldsymbol{\theta}}$, thresholds u_j, marginal distribution functions $\widetilde{F}_j$ for $j = 1, \ldots, D$.

function CONDITIONAL SIMULATION($Q_k(u_j)$)

sample $X(\boldsymbol{s}_k) \sim \mathsf{Exp}(1) + u_k$;

sample $\{X(\boldsymbol{s}_j)\}_{j=1,\ldots,D,j\neq k} \mid X(\boldsymbol{s}_j) \leq X(\boldsymbol{s}_k)$, where

$$X(\boldsymbol{s}_j) = \widehat{b}[X(\boldsymbol{s}_k), \boldsymbol{s}_j - \boldsymbol{s}_k] + \widehat{a}[X(\boldsymbol{s}_k), \boldsymbol{s}_j - \boldsymbol{s}_k]Z(\boldsymbol{s}_j);$$

initialize $\boldsymbol{\pi} \leftarrow \mathbf{0}_D$;

for $j = 1, \ldots, D$ **do**

sample B replications $\boldsymbol{X}_b \sim Q_j(u_j)$ at sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$;

$$\pi_k \leftarrow \pi_k + \{1 - \widetilde{F}_j(u_j)\}B^{-1}\sum_{b=1}^{B} \mathbf{1}\{X_b(\boldsymbol{s}_k)/u_k = \max_{j=1}^{D} X_b(\boldsymbol{s}_j)/u_j\};$$

set $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}/\|\boldsymbol{\pi}\|_1$;

sample $k \sim \mathsf{Mult}(\boldsymbol{\pi})$;

return $\{\boldsymbol{X}(\boldsymbol{s}_j)\}_{j=1}^{D} \sim Q_k(u_j)$:

simulating unconditionally from the model for $X(\boldsymbol{s}_k) > u_k$ a large number of realizations and calculating the proportion of samples in which $X(\boldsymbol{s}_k)$ is largest; this is formalized in Algorithm 2.9.

Note that, if Z is a Gaussian process, we can easily simulate from $\boldsymbol{X}_{-k} \mid \boldsymbol{X}_{-k} \leq X_k, X_k > u_j$: suppose $\{Z(\boldsymbol{s}_j)\}_{j=1}^{D} \equiv \boldsymbol{Z} \sim \mathsf{No}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ unconditionally. We can easily obtain the distribution of $\boldsymbol{Z}_{-k} \mid Z_k = 0 \sim \mathsf{No}_{D-1}(\boldsymbol{\mu}_{|k}, \boldsymbol{\Sigma}_{|k})$, say, through Proposition C.2. The Gaussian distribution is a location-scale family, hence for $\boldsymbol{a}_{|k}$ and $\boldsymbol{b}_{|k}$ vectors corresponding to $a(X_k, \boldsymbol{s}_j - \boldsymbol{s}_k)$ for $j = 1, \ldots, D$, $j \neq k$, etc., $\boldsymbol{X}_{-k} = \boldsymbol{b}_{|k} + \boldsymbol{a}_{|k}\boldsymbol{Z}_{-k}$ is also multivariate Gaussian and we can perform conditional simulations subject to $\boldsymbol{X}_{-k} \mid \boldsymbol{X}_{-k} \leq X_k$ using an efficient accept-reject algorithm based on the minimax exponential tilting algorithm of Botev (2017).

An algorithm to perform conditional simulations is described in Wadsworth and Tawn (2018), § 5.3, assuming a common marginal model, the parameters of which are incorporated in the composite likelihood through the probability integral transform and estimated together with $\boldsymbol{\theta}$.

2.6 Likelihoods and estimating functions

The different extreme value paradigms are most easily reconciled using the Poisson point process formulation outlined in Section 2.5.1. We follow Huser et al. (2016) and restrict attention to the D-variate case, even if observations are understood to arise from spatial processes.

2.6.1 Max-stable process likelihood

We derive first the likelihood of a max-stable random vector. The exponent measure can be expressed as

$$V(\boldsymbol{z}) = \Lambda([\mathbf{0}_D, \boldsymbol{z}]^{\mathsf{c}}) = \int_0^{\infty} \mathsf{P}\left(\zeta \boldsymbol{W} \not\leq \boldsymbol{z}\right) \zeta^{-2} \mathrm{d}\zeta, \qquad \boldsymbol{z} \in \mathbb{R}_D^+ \setminus \{\mathbf{0}_D\}.$$

This probability can be re-expressed using the inclusion-exclusion formula (cf. Wadsworth and Tawn, 2014). Assuming Λ is regular and making the change of variable $r = \zeta^{-1}$, we obtain by taking partial derivatives on both sides the formula

$$-V_{1:D}(\boldsymbol{z}) = \lambda(\boldsymbol{z}) = \int_0^{\infty} r^D f_{\boldsymbol{W}}(r\boldsymbol{z}) \mathrm{d}r, \qquad \boldsymbol{z} \in \mathbb{R}_D^+ \setminus \{\mathbf{0}_D\}.$$

The likelihood for a simple max-stable process $\boldsymbol{Z}$ can be obtained by differentiating the distribution function $\exp\{-V(\boldsymbol{z})\}$ with respect to each $z_1, \ldots, z_D$, yielding

$$\mathscr{L}(\boldsymbol{z}) = -\exp\{-V(\boldsymbol{z})\} \sum_{\varpi \in \mathscr{P}_D} \prod_{j=1}^{|\varpi|} V_{I_j}(\boldsymbol{z}),$$

where $\mathscr{P}_D$ is the power set of all non-empty subsets of $\{1, \ldots, D\}$ and I_j is the set of indices in the jth set of the partition ϖ. The number of terms in the likelihood is the Dth Bell number, $B_D = \sum_{k=1}^{D} S(D,k)$ where $S(D,k)$, the Stirling numbers of the second kind, give the numbers of partitions of D elements into k parts. In dimension $D = 10$, there are $B_{10} = 115975$ distinct likelihood contributions and the calculations are prohibitive.

The likelihood of the max-stable likelihood is complicated because it incorporates the sum of all possible hitting scenarios. Knowledge of the hitting scenario simplifies considerably the likelihood, since then

$$\mathsf{P}\left(\boldsymbol{\Pi} = \boldsymbol{\varpi}, \boldsymbol{Z} \in \mathrm{d}\boldsymbol{z}\right) = \exp\{-V(\boldsymbol{z})\} \left\{ \prod_{j=1}^{|\boldsymbol{\varpi}|} \int \mathbf{1}\{\boldsymbol{u}_j < \boldsymbol{z}_{\varpi_j^{\mathsf{c}}}\} \lambda(\boldsymbol{z}_{\varpi_j}, \boldsymbol{u}_{\boldsymbol{j}}) \mathrm{d}\boldsymbol{u}_j \right\} \mathrm{d}\boldsymbol{z}$$

and one avoids the costly marginalization. This idea was proposed in Stephenson and Tawn (2005), who suggest incorporating the information about the partition if occurrence times are recorded; the joint distribution of the empirical partition Π_n and the component-wise maxima $\boldsymbol{M}_n$ as $n \to \infty$ is

$$\mathscr{L}(\boldsymbol{z}) = -\exp\{-V(\boldsymbol{z})\} \prod_{j=1}^{|\varpi_n|} V_{I_j}(\boldsymbol{z}).$$

With $\lambda(\boldsymbol{z}) = -\partial^D V(\boldsymbol{z}) / \prod_{j=1}^{D} \partial z_j$, we can see that, for regular models, the partial derivatives of

the exponent measure satisfy (Huser et al., 2019)

$$-\frac{\partial^{|\boldsymbol{\omega}|}V(\boldsymbol{z})}{\prod_{l\in\omega_j}\partial z_l} = \int \mathbf{1}\{\boldsymbol{u}_j < \boldsymbol{z}_{\omega_j^c}\}\lambda(\boldsymbol{z}_{\omega_j}, \boldsymbol{u_j})\mathrm{d}\boldsymbol{u}_j;$$

we thus obtain the expression in Stephenson and Tawn (2005).

In practice, we replace the limiting partition with the empirical one. The likelihood is biased unless $n \gg D$ since the empirical partition also needs to converge to the limiting hitting scenario; for weakly dependent processes, use of the observed partition may induce bias. Wadsworth (2015) shows that this is due to the lack of convergence of the empirical partition $\boldsymbol{\Pi}_n$ to the asymptotic partition $\boldsymbol{\Pi}$ and points out that the bias is stronger if D is large relative to n and if dependence is weak, suggesting an alternative bias-corrected likelihood that includes a second order bias correction. Thibaud and Opitz (2015) propose to impute the partition using a Gibbs sampler, while Huser et al. (2019) use a stochastic expectation-maximisation algorithm; the E-step for the missing partition uses a Monte-Carlo estimator, where approximate draws are obtained from the Gibbs sampler of Dombry et al. (2013).

2.6.2 Likelihood for ℓ-Pareto processes and generalizations

In order to be able to obtain the likelihood of the ℓ-Pareto process in a multivariate context, we need the condition $\ell(\boldsymbol{X}) > u$ to be determined by the finite D-dimensional distributions, excluding *de facto* risk functionals such as $\ell(X) = \sup_{\boldsymbol{s}\in\mathcal{S}} X(\boldsymbol{s})$. Threshold exceedances are defined on the risk region $\mathscr{A}_{\ell,u} = \{\boldsymbol{y} \in \mathbb{R}_+^D \setminus \{\mathbf{0}_D\} : \ell(\boldsymbol{y}) > u\}$ and the empirical point process of ℓ-exceedances converges in distribution to a Poisson process with intensity function $\lambda(\boldsymbol{x})$ on $\mathscr{A}_{\ell,u}$. Likelihoods for threshold exceedances combine the intensity function of the point process, $\lambda(\boldsymbol{x})$, a normalizing constant giving the measure of the risk region $\mathscr{A}_{\ell,\boldsymbol{u}}$ and the Jacobian for the marginal transformation of the observations to the unit Pareto scale. One can use the likelihood of the limiting Poisson process model for exceedances in a region $\mathscr{A}_{\boldsymbol{u}}$, $\exp\{-\Lambda(\mathscr{A}_{\boldsymbol{u}})\}\prod_{j=1}^{N_u}\lambda(x_i)$. If we consider the conditional density of the ℓ-exceedance instead, the Poisson contribution vanishes. One can include a likelihood contribution for the (random) number extreme observations N_u out of n exceeding u independent of the conditional distribution if we observe the number of observations falling outside the risk region, either Poisson (Wadsworth and Tawn, 2014) or binomial (Thibaud and Opitz, 2015). The Fisher information is then a direct sum of the contribution for N_u and that of the exceedances conditional on N_u (Huser et al., 2016); see Example 2.29 for an illustration.

Proposition 2.33 (Likelihood of the generalized ℓ-Pareto process)
Assuming data are observed at a finite collection of sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$, the likelihood of the generalized ℓ-Pareto process for one exceedance $\{\boldsymbol{X}_i : \ell(\boldsymbol{X}_i) \geq u\}$ is

$$f_u^{\ell}(\boldsymbol{x}_i) = \frac{\lambda\left\{\left(1+\xi\frac{\boldsymbol{x}_i-\boldsymbol{\eta}}{\boldsymbol{\tau}}\right)^{1/\xi}\right\}}{\Lambda\left\{\boldsymbol{Z}\in\mathbb{R}_+^D : \ell\{\boldsymbol{\tau}(\boldsymbol{Z}^{\xi}-1)/\xi+\boldsymbol{\eta}\}\geq u\right\}}\prod_{j=1}^{D}\tau^{-1}(\boldsymbol{s}_i)\left\{1+\xi\frac{x_{ij}-\eta(\boldsymbol{s}_i)}{\tau(\boldsymbol{s}_i)}\right\}^{1/\xi-1}.$$

The likelihood corresponds to the intensity renormalized by the measure of the risk region plus a Jacobian term.

Remark 2.8 (Identifiability)
The ℓ-Pareto process likelihood is identifiable only up to a scale multiple. If we define the scale $\boldsymbol{\varsigma} = \kappa\boldsymbol{\tau}$ for $\kappa > 0$, then

$$\left(1+\xi\frac{\boldsymbol{x}-\boldsymbol{\eta}}{\boldsymbol{\varsigma}}\right)^{1/\xi} = \kappa^{-1/\xi}\left(1+\xi\frac{\boldsymbol{x}-\boldsymbol{\phi}}{\boldsymbol{\tau}}\right)^{1/\xi},$$

where $\boldsymbol{\phi} = \boldsymbol{\eta} + \boldsymbol{\varsigma}(1/\kappa - 1)/\xi$. Since $\Lambda(\cdot)$ is a -1-homogeneous measure,

$$\kappa^{1/\xi}\Lambda\left\{\mathsf{Z}\in\mathscr{F}_0 : \ell\{\tau(\mathsf{Z}^\xi-1)/\xi+\phi\}\geq u\right\} = \Lambda\left\{\mathsf{Z}\in\mathscr{F}_0 : \ell\{\varsigma(\mathsf{Z}^\xi-1)/\xi+\eta\}\geq u\right\}.$$

if ℓ is a positive homogeneous functional. Since we assume the existence of derivatives of orders $1 < k \leq D$, $\lambda_{1:k}$ is $-(1+k)$-homogeneous and the constant $\kappa^{1/\xi}$ cancels with the contribution from the Jacobian. Fixing one of the marginal parameters (whether location or scale) restores identifiability.

Suppose $\boldsymbol{\tau} = \mathbf{1}_D$ and $\boldsymbol{\eta} = \mathbf{0}_D$: the multivariate generalized Pareto distribution (Rootzén and Tajvidi, 2006) is an important special case of the ℓ-Pareto process with $\ell(\boldsymbol{X}) = \max_{j=1}^D X_i$.

Proposition 2.34 (Likelihood of the multivariate generalized Pareto)
Let $q(\boldsymbol{y}) = \left(1+\boldsymbol{\xi}\boldsymbol{y}/\boldsymbol{\sigma}\right)_+^{1/\boldsymbol{\xi}}$ for N_u exceedances. The distribution function of a multivariate generalized Pareto distribution for a given exponent measure $V(\cdot)$, threshold $\boldsymbol{u}$ and marginal parameters $\boldsymbol{\xi}\in\mathbb{R}^D, \boldsymbol{\sigma}\in\mathbb{R}_+^D$, is

$$F(\boldsymbol{y}) = \frac{V[\min\{q(\boldsymbol{y}), q(u\mathbf{1}_D)\}] - V\{q(\boldsymbol{y})\}}{V\{q(u\mathbf{1}_D)\}}, \qquad \boldsymbol{y}\not\leq\boldsymbol{u},$$

and the density is

$$f(\boldsymbol{y}) = \frac{-V_{1:D}\{q(\boldsymbol{y})\}}{V\{q(\boldsymbol{u})\}}\prod_{j=1}^{N_u}\sigma_j^{-1}\left(1+\xi\frac{y_i}{\sigma_j}\right)_+^{1/\xi-1} \qquad \boldsymbol{y}\not\leq\boldsymbol{u}.$$

This is the conditional density given that observations $\boldsymbol{y}$ lie in the set $[\mathbf{0},\boldsymbol{u}]^{\mathsf{c}}$. The normalization constant in the denominator is therefore the measure of the set $[\mathbf{0},\boldsymbol{u}]^{\mathsf{c}}$.

Most of the results presented so far follow from Definition 2.15, whereby the limit measure is -1-homogeneous. This implies that the exceedances are tail equivalent to a unit Pareto random variable and the scale varies with the threshold, but is linked to the unknown scaling function a_n. We have also presented a generalization, whereby normalized variables have limit measure Λ as given Remark 2.4. The model for $\boldsymbol{W}$ is thus derived on the standardized scale and the intensity is simply given by the composition $\Lambda\circ T$.

We may wish to derive the max-stable attractor of a particular parametric model. For example, with a scale mixture model of the form $R\boldsymbol{X}$, where $R\in\mathsf{RV}(\xi^{-1})$ and $\xi > 0$, the limiting

distribution follows from an application of Breiman's lemma (Ho, 2018). Rootzén et al. (2018) use the same argument as in Remark 2.4 for the D-variate case, but model $\boldsymbol{V} \stackrel{\mathrm{d}}{=} \boldsymbol{\sigma}/\boldsymbol{\xi}\boldsymbol{W}^{\boldsymbol{\xi}}$ for $\boldsymbol{\xi} > \mathbf{0}_D$, leading to the so-called R representation of the multivariate generalized Pareto vector, $\boldsymbol{V}\zeta^{-\boldsymbol{\xi}} - \widetilde{\boldsymbol{\sigma}}/\boldsymbol{\xi}$.

An appealing choice of risk functional for the ℓ-Pareto process is the l_1 norm, since for $\mathscr{A}_{\Sigma,\boldsymbol{u}} = \{\boldsymbol{y} \in \mathbb{R}_+^D \setminus \{\mathbf{0}_D\} : \|\boldsymbol{y}/\boldsymbol{u}\|_1 > 1\}$, $\Lambda(\mathscr{A}_{\Sigma,u}) = \|\mathbf{1}_D/\boldsymbol{u}\|_1$, which is model independent and simplifies to D/u when $\boldsymbol{u} = u\mathbf{1}_D$. The name 'spectral likelihood' is sometimes attached to the associated likelihood (cf. Engelke et al., 2015). Unfortunately, no such results exist for the generalized ℓ-Pareto processes when $\xi \neq 1$.

Extremal Student processes have mass on the boundary of $\mathscr{A}_{\Sigma,\boldsymbol{u}}$ and transformation of the observations to the unit Pareto scale ensures that all of them exceed $\mathbf{1}_D$, which is inconsistent with the fact that some components have been truncated to zero. Thibaud and Opitz (2015) propose the use of a censored likelihood to reduce the bias inherent to this marginal transformation.

2.6.3 Censored likelihoods

In practice, and since the limiting model may be a poor approximation if not all components are large, censoring is advised. Censored likelihoods are obtained by integrating the intensity function with respect to the components which are censored. We provide examples for specific families for which explicit expressions exist. For intrinsically stationary fields, one can condition on the set of observations at any site for which the process is observed and is uncensored.

Example 2.28 (Censored likelihood for Brown–Resnick model based on projections (1)) Asadi et al. (2015) give the density of the censored Brown–Resnick Pareto process associated to the projection $P_{\boldsymbol{s}_0}$, satisfying $Y(\boldsymbol{s}_0) > u_0$, assuming the data are unit Pareto distributed. Suppose that $X_i > u_i$ for $i \in \mathscr{A} \subseteq \{1,\ldots,D\} \setminus \emptyset$ and $X_i < u_i$ for $i \in \mathscr{A}^{\mathsf{c}}$. Assume without loss of generality that the variables $\boldsymbol{X}$ are reordered so that the $k = |\mathscr{A}|$ exceedances are labelled $1{:}k$ and that we condition on X_1. Let $\phi_l(\cdot;\boldsymbol{\Sigma})$ and $\Phi_l(\cdot;\boldsymbol{\Sigma})$ denote the density and distribution functions of a Gaussian variable with mean $\mathbf{0}_l$ and covariance matrix $\boldsymbol{\Sigma}$, with the convention that $\phi_0 = 1, \Phi_0 = 1$. The censored likelihood is then

$$\mathscr{L}_{\mathscr{A}}(\boldsymbol{x}) = x_1^{-1}\left(\prod_{i=1}^{k} x_i^{-1}\right)\phi_{k-1}(\widetilde{\boldsymbol{x}}_{1:k-1};\mathbf{0}_{k-1}\boldsymbol{\Sigma}_{1:k-1,1:k-1})\Phi_{D-k}\left(\mathbf{0}_{k-1};\boldsymbol{\mu}_{\mathscr{A}}\boldsymbol{Q}^{-1}_{k:D-1,k:D-1}\right),$$

where $\widetilde{x}_i = \log(x_{i+1}/x_1) + 2\lambda^2_{i+1,1}$ $(i = 1,\ldots,D-1)$ and $\widetilde{u}_i = \log(u_{i+1}/x_1) + 2\lambda^2_{i+1,1}$. The conditional mean and variance are

$$\begin{aligned}\boldsymbol{\mu}_{\mathscr{A}} &= \widetilde{\boldsymbol{u}}_{k:D-1} - \boldsymbol{Q}^{-1}_{k:D-1,k:D-1}\boldsymbol{Q}_{k:D-1,1:k-1}\widetilde{\boldsymbol{x}}_{1:k-1}\\ &= \widetilde{\boldsymbol{u}}_{k:D-1} + \boldsymbol{\Sigma}_{k:D-1,1:k-1}\boldsymbol{\Sigma}^{-1}_{1:k-1,1:k-1}\widetilde{\boldsymbol{x}}_{1:k-1}\end{aligned}$$

and the covariance matrix is

$$\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1} = \left\{2\left(\lambda_{i,1}^2 + \lambda_{j,1}^2 - \lambda_{i,j}^2\right)\right\}_{i,j=2}^{D}.$$

The expression for the conditional mean is incorrectly reported in eq. 29 of Asadi et al. (2015). □

Example 2.29 (Censored likelihood of Brown–Resnick model (2))
An alternative formulation based on the Poisson process was proposed by Wadsworth and Tawn (2014), who use $\ell = \max_{j=1}^{D}$. Wadsworth and Tawn (2014) base their inference on the covariance matrix of the (non)-stationary random field characterized by the covariance function C as in where $\Sigma_{i,j} = C(\boldsymbol{s}_i, \boldsymbol{s}_j)$; recall that the semivariogram and the covariance function are linked via the relation $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = C(\boldsymbol{s}_i, \boldsymbol{s}_i) + C(\boldsymbol{s}_j, \boldsymbol{s}_j) - 2C(\boldsymbol{s}_i, \boldsymbol{s}_j)$ under weak stationarity. If in addition the process is stationary, $C(\boldsymbol{s}, \boldsymbol{s}) = \sigma$ for any $\boldsymbol{s} \in \mathcal{S}$ and $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sigma^2\{1 - \rho(\boldsymbol{h})\}$ for a correlation function ρ.

Let $\boldsymbol{\Sigma}$ be the $D \times D$ covariance matrix and $\boldsymbol{\sigma} = \mathrm{diag}(\boldsymbol{\Sigma})$ the vector of site-wise variance. The likelihood contribution for a partly censored observation is

$$\begin{aligned}
-V_{1:k}(\boldsymbol{x}_{1:k}, \boldsymbol{u}_{(k+1):D}) = {} & \frac{\Phi_{D-k}\left(\log(\boldsymbol{u}_{(k+1):D}) - \boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k\right)\left|\boldsymbol{\Sigma}_{1:k}^{-1}\right|^{1/2}}{(2\pi)^{(k-1)/2}(\mathbf{1}_k^\top \boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k)^{1/2}\prod_{i=1}^{k} x_i} \\
& \times \exp\left\{-\frac{1}{2}\left(\frac{1}{4}\boldsymbol{\sigma}_k^\top \mathbf{A}_k \boldsymbol{\sigma}_k + \frac{\boldsymbol{\sigma}_k^\top \boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k - 1}{\mathbf{1}_k^\top \boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k}\right)\right\} \\
& \times \exp\left[-\frac{1}{2}\left\{\log(\boldsymbol{x}_{1:k})^\top \mathbf{A}_k \log(\boldsymbol{x}_{1:k}) + \log(\boldsymbol{x}_{1:k})^\top\left(\frac{2\boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k}{\mathbf{1}_k^\top \boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k} + \mathbf{A}_k\boldsymbol{\sigma}_k\right)\right\}\right],
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_k &= \boldsymbol{\Sigma}_{1:k}^{-1} - \frac{\boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k\mathbf{1}_k^\top\boldsymbol{\Sigma}_{1:k}^{-1}}{\mathbf{1}_k^\top\boldsymbol{\Sigma}_{1:k}^{-1}\mathbf{1}_k}, \qquad \boldsymbol{\Gamma}_k^{-1} = \mathbf{K}_{01}^\top\boldsymbol{\Sigma}^{-1}\mathbf{K}_{01}, \\
\boldsymbol{\mu}_k &= -\boldsymbol{\Gamma}\left(\mathbf{K}_{01}^\top\mathbf{A}_D\mathbf{K}_{10}\log(\boldsymbol{x}_{1:k}) + \mathbf{K}_{01}^\top\frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}\right).
\end{aligned}$$

The intensity is $-V_{I_i}(\boldsymbol{x}_i^{\mathrm{c}})/V(\boldsymbol{u})$ for threshold exceedances with partly censored observations and the likelihood of the Poisson process is (Wadsworth and Tawn, 2014)

$$\mathscr{L}_{\mathrm{WT}} \propto \exp\{-nV(\boldsymbol{u})\}\prod_{i=1}^{N_u} -V_{I_i}(\boldsymbol{x}_i^{\mathrm{cens},*})\prod_{j=1}^{D}\mathbf{1}\{x_{i,j} > u_j\}\left|J_t^{\mathrm{p}}(x_{i,j})\right|,$$

where $x_{i,j}^{\mathrm{cens},*} = \max\{x_{i,j}^*, u_j\}$ or

$$x_{i,j}^{\mathrm{cens},*} = \begin{cases} x_{i,j}^*, & x_{i,j}^* > u_j^* \\ u_j^*, & x_{i,j}^* \leq u_j^* \end{cases} \quad (j = 1, \ldots, D).$$

Since the Poisson component for N_u exceedances is $\exp\{-V(\boldsymbol{u})\}V(\boldsymbol{u})^{N_u}/N_u!$, the term $V(\boldsymbol{u})^{N_u}$

cancels from the density. The remaining contribution is from observations in $(\mathbf{0}, \boldsymbol{u}]$, which gives a factor of $\exp\{-(n-N_u)V(\boldsymbol{u})\}$.

Asadi et al. (2015) use the censored likelihood of Thibaud and Opitz (2015) with a binomial contribution for the number of exceedances N_u (provided $\boldsymbol{u}$ is large enough that $V(\boldsymbol{u}) < 1$):

$$\mathscr{L}_{\mathrm{AE}} \propto \{1-V(\boldsymbol{u})\}^{n-N_u} \prod_{i=1}^{N_u} -V_{I_i}\left(\boldsymbol{x}_i^{\mathrm{cens},*}\right) \prod_{j=1}^{D} \mathbf{1}\{x_{i,j}>u_j\} \left|J_t^{\mathrm{p}}(x_{i,j})\right|.$$

By the Poisson approximation to the binomial distribution, the two contributions are nearly equivalent.

□

2.6.4 Estimating functions

In practice, it is customary to use a two-stage approach: fit marginal parameters first using the independence likelihood and transform the observations to a standardized scale, then fit the dependence model. This simplifies the calculation by breaking the optimization into two problems, each with fewer parameters. Practitioners often use the independence likelihood as a working model to fit the marginals of multiple sites simultaneously, imposing smoothness restrictions through priors on the shape or fixing it to be constant over the domain. In applications the observations are not standardized and will need to be transformed to a common scale, for example unit Fréchet or Pareto, in which case observations are replaced by standardized one. This technique is also popular in the copula literature because margins can be estimated nonparametrically using the empirical distribution function $\widetilde{F}(x) = \mathrm{rank}(x)/(n+1)$ for a sample of size n. An alternative in larger samples is to use a semi-parametric estimate, fitting a generalized Pareto $\mathsf{GP}(\tau_j, \xi_j)$ to marginal exceedances above u_j and using the estimated distribution (Coles and Tawn, 1994, § 2.1)

$$\check{F}_j(x) = \begin{cases} \widetilde{F}_j(x), & x \leq u_j, \\ 1-\{1-\widetilde{F}_j(u_j)\}\left\{1+\widehat{\xi}\left(\frac{x-u_j}{\widehat{\tau}_j}\right)\right\}_+^{-1/\widehat{\xi}_j}, & x > u_j. \end{cases} \tag{2.29}$$

Two-stage estimation leads to incorrect uncertainty assessment and erroneous statistical inference unless uncertainty is properly accounted for.

If we assume that the threshold exceedance $X_j - u_j$ follows approximately a generalized Pareto distribution with parameters η_j, ξ_j and $\lambda_j = \mathsf{P}\left(X_j > u_j\right)$, then the Jacobian of the transformation from generalized Pareto to unit Fréchet (t^{f}) or unit Pareto (t^{p}) variates is

$$\mathrm{J}_j^{\mathrm{f}}(x_j) = \frac{\lambda v_j^{1+\xi_j}}{\eta_j(1-\lambda_j v_j)\log^2(1-\lambda_j v_j)}, \qquad \mathrm{J}_j^{\mathrm{p}}(x_j) = \frac{v_j^{\xi_j-1}}{\eta_j \lambda_j}, \qquad j = 1, \ldots, D,$$

with $v_j = \{1+\xi(x_j-u_j)/\eta_j\}^{-1/\xi_j}$.

Composite likelihoods

Composite likelihoods enjoy many properties of likelihoods, though not asymptotic optimality. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample an unknown multivariate distribution with continuous univariate margins that is in the maximum domain of attraction of a multivariate extreme-value distribution H.

A popular choice for modelling multivariate block maxima or threshold exceedances is the pairwise composite log-likelihood l_C (Padoan et al., 2010). The multivariate tail model for threshold exceedances can be fitted using the censored likelihood $\mathscr{L}(\boldsymbol{X};\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathscr{L}_i(\boldsymbol{X}_i;\boldsymbol{\theta})$, where

$$\mathscr{L}_i(\boldsymbol{X}_i;\boldsymbol{u},\boldsymbol{\theta}) = \left.\frac{\partial^{m_i} \exp\{-V(\boldsymbol{y})\}}{\partial y_{j_1} \cdots \partial y_{j_{m_i}}}\right|_{\boldsymbol{y}=t^{\mathrm{f}}\{\max(\boldsymbol{X}_i,\boldsymbol{u})\}} \prod_{k=1}^{m_i} \mathrm{J}^{\mathrm{f}}_{j_k}(X_{ij_k}), \quad i = 1, \ldots, n. \tag{2.30}$$

In this expression, the indices $j_1, \ldots, j_{m_i}$ are those of the components of $\boldsymbol{X}_i$ exceeding the thresholds $\boldsymbol{u}$ and for $\boldsymbol{x} \geq \boldsymbol{u}$, $t^{\mathrm{f}}(\boldsymbol{x}) = (t^{\mathrm{f}}_1(x_1), \ldots, t^{\mathrm{f}}_D(x_D))$, where $t^{\mathrm{f}}_j(x_j) = -1/\log\{\widetilde{F}_j(x_j;\theta_j)\}$ and

$$\mathrm{J}^{\mathrm{f}}_j(x_j) = \frac{\nu_j}{\sigma_j}\left(1 + \xi_j \frac{(x_j - u_j)}{\sigma_j}\right)^{-\xi_j - 1} \frac{1}{[\log\{\widetilde{F}_j(x_j;\sigma_j,\xi_j)\}]^2 \widetilde{F}_j(x_j;\sigma_j,\xi_j)}, \qquad j = 1, \ldots, D. \tag{2.31}$$

When D is large, one can also maximize the likelihood in Smith et al. (1997) that uses the tail approximation $\bar{F}(\boldsymbol{x}) \approx 1 - V(\boldsymbol{x})$. In either case, $V(\boldsymbol{x})$ and the higher-order partial derivatives of $V(\boldsymbol{x})$ need to be computed. Ledford and Tawn (1996) use the bivariate max-stable likelihood for threshold exceedances,

$$\ell_C(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{D-1} \sum_{k=j+1}^{D} \Big(\log\big[g\{t_j(x_{ij}), t_k(x_{ik}); \boldsymbol{\theta}, t_j(u_j), t_k(u_k)\}\big] + \mathbf{1}\{x_{ij} > u_j\} \log\{J_j(x_{ij})\} + \mathbf{1}\{x_{ik} > u_k\} \log\{J_k(x_{ik})\}\Big),$$

where, with $V_j = \partial V(y_j, y_k)/\partial y_j$, etc. and t_j and J_j as in eq. (2.31), we have, for all $j = 1, \ldots, D$,

$$g(y_j, y_k; \boldsymbol{\theta}, u_j, u_k) = \begin{cases} \exp\{-V(u_j, u_k)\}, & y_j \leq u_j, y_k \leq u_k, \\ -V_j(y_j, u_k) \exp\{-V(y_j, u_k)\}, & y_j > u_j, y_k \leq u_k, \\ -V_k(u_j, y_k) \exp\{-V(u_j, y_k)\}, & y_j \leq u_j, y_k > u_k, \\ \{V_j(y_j, y_k) V_k(y_j, y_k) - V_{jk}(y_j, y_k)\} \exp\{-V(y_j, y_k)\}, & y_j > u_j, y_k > u_k. \end{cases}$$

Ledford and Tawn (1996) motivate the choice of copula by the lower bias of the estimator in the case of near-independent variables, but the model cannot readily be extended to higher-order composite likelihood without a combinatorial explosion of the number of terms.

Uncertainty assessment is performed in the same way as for general estimating equations (White, 1982). Let $h(\boldsymbol{\theta})$ denote an unbiased estimating function and define the variability

matrix **J**, the sensitivity matrix **H** and the Godambe information matrix **G**, as

$$\mathbf{J} = \mathsf{E}\left(\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\top}\right), \qquad \mathbf{H} = -\mathsf{E}\left(\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\top}}\right), \qquad \mathbf{G} = \mathbf{H}\mathbf{J}^{-1}\mathbf{H}. \tag{2.32}$$

The maximum composite likelihood estimator is strongly consistent and asymptotically normal, centered at the true parameter $\boldsymbol{\theta}$ with covariance matrix $\mathbf{G}^{-1}$. The composite likelihood information criterion must be adjusted to account for model specification and is the Takeuchi information criterion, defined as

$$\mathsf{CLIC} = -2\ell_c(\widehat{\boldsymbol{\theta}}) + 2\operatorname{tr}(\mathbf{J}\mathbf{H}^{-1}).$$

Davison and Gholamrezaee (2011) propose scaling the latter, yielding $\mathsf{CLIC}^* = c\mathsf{CLIC}$, to make it comparable to the value of the Akaike information criterion for independent observations. The constant c will depend on the dimension of the likelihood; it is $(n-1)^{-1}$ for pairwise likelihood. Further adjustments are needed in case of missing values.

Hypothesis testing proceeds in the usual way. Suppose we are interested in testing for a particular value of a parameter of interest $\boldsymbol{\psi}$. Specifically, partitioning $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ into a q-dimensional parameter of interest $\boldsymbol{\psi}$ and a $(p-q)$-dimensional nuisance parameter $\boldsymbol{\lambda}$, along with the corresponding partitions **H**, **J** and **G**. Let $\widehat{\boldsymbol{\theta}}_C = (\widehat{\boldsymbol{\psi}}_C, \widehat{\boldsymbol{\lambda}}_C)$ denote the maximum composite likelihood parameter estimates and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}_0} = (\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}_0})$ the restricted parameter estimates under the hypothesis that the restricted model is adequate. The asymptotic distribution of the composite likelihood ratio test statistic $2\{\ell_C(\widehat{\boldsymbol{\theta}_C}) - \ell_C(\widehat{\boldsymbol{\theta}_0})\}$ is that of $c_1 Z_1 + \cdots + c_q Z_q$, where $Z_1, \ldots, Z_q$ are independent χ^2_1 variables and c_i are the eigenvalues of the $q \times q$ matrix $(\mathbf{H}_{\boldsymbol{\psi}\boldsymbol{\psi}} - \mathbf{H}_{\boldsymbol{\psi}\boldsymbol{\lambda}}\mathbf{H}^{-1}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}\mathbf{H}_{\boldsymbol{\lambda}\boldsymbol{\psi}})\mathbf{G}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}$ (Kent, 1982). Alternatively, one could also use a curvature adjustment (Ribatet et al., 2012, § 2.2) and modify the statistic so that its asymptotic distribution is χ^2. Specifically, the composite log-likelihood is evaluated at $\widehat{\boldsymbol{\theta}}_c + \mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}_0})\mathbf{M}_{\mathrm{G}}(\widehat{\boldsymbol{\theta}_0})(\widehat{\boldsymbol{\theta}_0} - \widehat{\boldsymbol{\theta}})$, where $\mathbf{M}(\boldsymbol{\theta})^{\top}\mathbf{M}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{M}_{\mathrm{G}}(\boldsymbol{\theta})^{\top}\mathbf{M}_{\mathrm{G}}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})$. The matrices $\mathbf{M}_{\mathrm{G}}$ and **M** can be retrieved using singular value decomposition.

The sensitivity matrix **H** is obtained from the Hessian matrix at the maximum composite likelihood estimate. The variability matrix **J** can be estimated directly, cf. Padoan et al. (2010, p. 266), but often it is preferable to estimate $\mathbf{G}^{-1}$ by the empirical covariance of B nonparametric bootstrap replicates and compute **J** through eq. (2.32). Varin (2008) and Varin et al. (2011) survey composite likelihood inference.

Example 2.30 (Composite likelihood for river flow on the Isar)
The following appeared in Belzile and Nešlehová (2017). We illustrate the use of censored likelihood estimation for peaks-over-threshold by fitting the scaled extremal Dirichlet model on a trivariate sample of daily river flows of the river Isar in southern Germany; this dataset is a subset of that analyzed in Asadi et al. (2015). We selected data measured at Lenggries (upstream), Pappinger Au (in the middle) and Munich (downstream). To ensure stationarity of the series and as the most extreme events occur during the summer, we restricted our attention to June, July and August. Since the sites measure the flow of the same river, dependence at

	σ_1	σ_2	σ_3	ξ_1	ξ_2	ξ_3
Scaled Dirichlet	$123_{\,7}$	$84_{\,5}$	$68_{\,4}$	$0.05_{\,0.04}$	$-0.03_{\,0.04}$	$0.02_{\,0.04}$
Neg. logistic	$117_{\,7}$	$86_{\,5}$	$70_{\,4}$	$0.08_{\,0.04}$	$-0.05_{\,0.04}$	$0_{\,0.04}$
Logistic	$117_{\,7}$	$87_{\,5}$	$70_{\,4}$	$0.08_{\,0.04}$	$-0.05_{\,0.04}$	$0_{\,0.04}$
Coles–Tawn	$114_{\,7}$	$84_{\,5}$	$68_{\,4}$	$0.12_{\,0.04}$	$-0.02_{\,0.04}$	$0.04_{\,0.04}$
Marginal	$129_{\,14}$	$95_{\,11}$	$76_{\,9}$	$-0.01_{\,0.08}$	$-0.15_{\,0.08}$	$-0.08_{\,0.08}$

Table 2.1 – *Generalized Pareto parameter estimates and standard errors (subscripts) for Example 2.30 using pairwise composite likelihood estimation for five models (scaled Dirichlet, logistic, negative logistic, Coles–Tawn and independence likelihood).*

	α_1	α_2	α_3	κ
Scaled Dirichlet	$0.8_{\,0.3}$	$1.6_{\,0.8}$	$2_{\,1.2}$	$-0.3_{\,0.1}$
Neg. logistic	1	1	1	$0.36_{\,0.02}$
Logistic	1	1	1	$-0.28_{\,0.01}$
Coles–Tawn	$3.3_{\,0.5}$	$10_{\,3}$	$13_{\,4}$	1

Table 2.2 – *Dependence parameters estimates and standard errors (subscripts) for Example 2.30 using pairwise composite likelihood estimation for four models (scaled Dirichlet, logistic, negative logistic, Coles–Tawn).*

extreme levels is likely to be present. Directionality of the river may lead to asymmetry in the asymptotic dependence structure, suggesting that the scaled extremal Dirichlet model may be well-suited for these data. Furthermore, as the Coles–Tawn, logistic and negative logistic models are nested within this family, their adequacy can be assessed through likelihood ratio tests. We used the changepoint score test of Northrop and Coleman (2014) and the simultaneous parameter stability plot of Wadsworth (2016) (not shown here) to select the thresholds $\boldsymbol{u} = (u_1, u_2, u_3)$, set at the 92% marginal percentile of each series. River flows exhibit temporal dependence and clusters of extreme values; we estimated the extremal index using the weighted least square and maximum likelihood estimator of Süveges (2007), suggesting values between 0.5 and 0.3. As such, we used the runs estimator of O'Brien (1987) with $r = 3$; since the series are multivariate, we keep only the cluster maxima for running windows of overlapping days. Set $\boldsymbol{\theta} = (\boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \kappa)$, where $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}$ are the marginal parameters of the generalized Pareto distribution and κ and $\boldsymbol{\alpha}$ are the parameters of the scaled extremal Dirichlet model. To estimate $\boldsymbol{\theta}$, we employed the pairwise composite log-likelihood l_C in Section 2.6.4 and fitted the scaled extremal Dirichlet model (cf. Example 2.17) as well as the logistic, negative logistic and Coles–Tawn extremal Dirichlet models to cluster maxima. The estimates of the marginal generalized Pareto parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}$ are given in Table 2.1. As the estimates were obtained by maximizing ℓ_C, their values depend on the fitted model; site-wise quantile-quantile plots indicate good marginal fit of the scaled extremal Dirichlet model (Belzile and Nešlehová, 2017).

The estimates of the dependence parameters $\boldsymbol{\alpha}$ and κ are given in Table 2.2. Whether the

asymmetry is significant can be assessed through composite likelihood ratio tests. To this end, consider a partition of $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ into a q dimensional parameter of interest $\boldsymbol{\psi}$ and a $3d+1-q$ dimensional nuisance parameter $\boldsymbol{\lambda}$, and the corresponding partitions of the sensibility, variability and Godambe information matrices $\mathbf{H}$, $\mathbf{J}$ and $\mathbf{G}$. Since the Coles–Tawn extremal Dirichlet, negative logistic and logistic models are nested within the scaled extremal Dirichlet family, we test for a restriction to these simpler models: the respective approximate P-values of 0.003, 0.74 and 0.78 suggest that while the Coles–Tawn extremal Dirichlet model is clearly not suitable, there is insufficient evidence to discard the logistic and negative logistic models. The effects of possible model misspecification are visible for the Coles–Tawn extremal Dirichlet model, as the values of α_1, α_2 and α_3 reported in the last line of Table 2.2 are very large and the shape parameter estimates in Table 2.1 are significantly larger than those obtained for the independence likelihood.

□

Example 2.31 (Composite likelihood for conditional spatial extremes)
Parameter estimates can be obtained through the composite likelihood for standardized data $\mathbf{X}$ observed at sites $\boldsymbol{s}_j$, $j = 1, \ldots, D$ with $f_{\boldsymbol{Z}}$ the density function of Z; denoting the parameters of the normalizing functions and the residual process by $\boldsymbol{\theta}$, we have

$$\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{j=1}^{D} \prod_{i=1}^{n} f_Z \left(\left\{ \frac{\boldsymbol{x}_{i,-j} - b(x_{i,j}, \boldsymbol{s}_k - \boldsymbol{s}_j)}{a(x_{i,j}, \boldsymbol{s}_k - \boldsymbol{s}_j)} \right\}_{k \neq j} \right) \prod_{\substack{j=1 \\ j \neq k}}^{D} a(x_{i,j}, \boldsymbol{s}_k - \boldsymbol{s}_j)^{-1} \mathbf{1}_{\{x_{i,j} > u_i\}}.$$

An observation vector may correspond to an exceedance at multiple sites concurrently, so would contribute multiple times to the likelihood.

□

Gradient score

Other estimating equations could be used to circumvent the calculation of $V(\boldsymbol{x})$ and its partial derivatives. An interesting alternative in the framework of proper scoring rules is the gradient score of Hyvärinen (Hyvärinen, 2005), adapted for peaks-over-threshold inference by de Fondeville and Davison (2018). The gradient score is

$$\delta_w(\boldsymbol{x}) = \sum_{i=1}^{D} \left(2 w_i(\boldsymbol{x}) \frac{\partial w_i(\boldsymbol{x})}{\partial x_i} \frac{\partial \log h(\boldsymbol{x})}{\partial x_i} + w_i^2(\boldsymbol{x}) \left[\frac{\partial^2 \log h(\boldsymbol{x})}{\partial x_i^2} + \frac{1}{2} \left\{ \frac{\partial \log h(\boldsymbol{x})}{\partial x_i} \right\}^2 \right] \right),$$

for a differentiable weighting function $w(\boldsymbol{x})$, unit Pareto observations $\boldsymbol{x}$ and density $h(\boldsymbol{x})$. The parameter estimates are obtained as the solution to $\arg\max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \delta_w(\boldsymbol{x}_i) \mathbf{1}_{\{\ell(\boldsymbol{x}_i/\boldsymbol{u}) > 1\}}$, where $\boldsymbol{\theta}$ is the vector of parameters of the model and ℓ is a differentiable risk functional, usually the ℓ_p norm for some $p \in \mathbb{N}^+$. Although the gradient score is not asymptotically the most efficient estimating equation, this makes little practical difference in high-dimensional settings in which the number of parameters is small relative to the number of sites. The weighting function $w(\boldsymbol{x})$ can be chosen to reproduce approximate censoring (de Fondeville and Davison, 2018).

Example 2.32 (Gradient score for the Brown–Resnick process)
Assuming unit Pareto margins and using the intensity function of the Poisson point process in Wadsworth and Tawn (2014), de Fondeville and Davison (2018) derived the gradient score

$$\nabla_x \log\lambda(\boldsymbol{x}) = -\mathbf{A}\log(\boldsymbol{x}) \odot \boldsymbol{x}^{-1} - \boldsymbol{x}^{-1} \odot \left(\frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}{\mathbf{1}_D^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}_D} + \mathbf{1}_D + \frac{\mathbf{A}^{-1}\boldsymbol{\sigma}}{2} \right), \tag{2.33}$$

where $\mathbf{A} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{1}_D\mathbf{1}_D{}^\top\boldsymbol{\Sigma}^{-1}/\mathbf{1}_D{}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D$ and $\boldsymbol{\sigma} = \operatorname{diag}(\boldsymbol{\Sigma})$ as in Example 2.29 and where $\odot$ denotes the Hadamard product (entrywise product) between two matrices. The diagonal terms of the Hessian matrix $\mathbf{H}$ with entries $\mathrm{H}_{ii} = \partial^2 \log\lambda(\boldsymbol{x})/\partial x_i^2$, are

$$\begin{aligned}
\operatorname{diag}(\mathbf{H}) &= -\operatorname{diag}(\mathbf{A})\left[\left\{1-\log(\boldsymbol{x})\right\}\boldsymbol{x}^{-2}\right] + \left[\{\mathbf{A}-\operatorname{diag}(\mathbf{A})\}\log(\boldsymbol{x})\right] \odot \boldsymbol{x}^{-2} \\
&\quad + \left(\frac{\boldsymbol{x}^{-2}}{2}\right) \odot \left(\frac{2\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D} + 2\mathbf{1}_D + \mathbf{A}^{-1}\boldsymbol{\sigma} \right) \\
&= -\operatorname{diag}(\mathbf{A}) \odot \boldsymbol{x}^{-2} + \left\{\mathbf{A}\log(\boldsymbol{x})\right\} \odot \boldsymbol{x}^{-2} + \left(\frac{\boldsymbol{x}^{-2}}{2}\right) \odot \left(\frac{2\boldsymbol{\Sigma}^{-1}\mathbf{1}_D}{\mathbf{1}_D^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}_D} + 2\mathbf{1}_D + \mathbf{A}^{-1}\boldsymbol{\sigma} \right).
\end{aligned}$$

□

Example 2.33 (Gradient score for the extremal Student process)
With the notation of Example 2.7 and assuming again that the marginal parameters are known, we suppose without loss of generality that the components of the vector of observations $\boldsymbol{x}$ are ordered from largest to smallest and consider first the case where some components of $\boldsymbol{x}$ are truncated at zero, so that $\boldsymbol{x}_{1:k} > \mathbf{0}_k, \boldsymbol{x}_{(k+1):D} = \mathbf{0}_{D-k}\}$. Recall the intensity of the D-dimensional process,

$$\log\lambda^+(\boldsymbol{x}) \overset{\boldsymbol{x}}{\propto} -\left(\frac{\nu+D}{2}\right)\log\left\{\boldsymbol{x}_{1:k}^{1/\nu\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_{1:k}^{1/\nu}\right\} + \frac{1-\nu}{\nu}\log(\boldsymbol{x}_{1:k})^\top\mathbf{1}_D + \log\left\{\mathsf{St}_{D+\nu}(\boldsymbol{x}_{(k+1):D}^{1/\nu};\boldsymbol{\eta}_k,\boldsymbol{\Omega}_k)\right\}.$$

The last term is the distribution function of a $(D-k)$-dimensional Studen variable whose location vector $\boldsymbol{\eta}_k$ and scale matrix $\boldsymbol{\Omega}_k$ are both functions of $\boldsymbol{x}_{1:k}$. For the $(D-k)$-dimensional censored component, the score is easily seen to be zero by the chain rule. However, the contribution of the $(D-k)$ dimensional distribution function term is non-zero for the components $\boldsymbol{x}_{1:k}$ and must be evaluated numerically by Monte Carlo methods. If $\boldsymbol{x} \in \mathbb{R}_+^D$ has only strictly positive components, the score is

$$\nabla_{\boldsymbol{x}}\log\lambda(\boldsymbol{x}) = -\frac{(\nu+D)\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^{1/\nu}\odot\boldsymbol{x}^{1/\nu-1}}{\nu a(\nu,\boldsymbol{x})} + \frac{1-\nu}{\nu}\boldsymbol{x}^{-1},$$

where $a(\nu,\boldsymbol{x}) = \boldsymbol{x}^{1/\nu\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^{1/\nu}$. The diagonal elements of the Hessian are

$$\begin{aligned}
\operatorname{diag}(\mathbf{H}_{1:k}) = -\frac{\nu+D}{\nu}\Bigg[&\frac{\left(\frac{2}{\nu}-1\right)\operatorname{diag}(\boldsymbol{\Sigma}^{-1})\odot\boldsymbol{x}^{2/\nu-2} + \left(\frac{1}{\nu}-1\right)\left\{\boldsymbol{\Sigma}^{-1}-\operatorname{diag}(\boldsymbol{\Sigma}^{-1})\right\}\boldsymbol{x}^{1/\nu}\odot\boldsymbol{x}^{1/\nu-2}}{a(\nu,\boldsymbol{x})} \Bigg] \\
&+ \frac{2(\nu+D)}{\nu^2 a(\nu,\boldsymbol{x})^2}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^{1/\nu}\odot\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^{1/\nu}\odot\boldsymbol{x}^{2/\nu-2} - \frac{1-\nu}{\nu}\boldsymbol{x}^{-2}.
\end{aligned}$$

□

Example 2.34 (Gradient score of scaled extremal Dirichlet family)
Consider the intensity function of the scaled extremal Dirichlet max-stable vector given in Example 2.16 with unit Fréchet observations. Straightforward calculations give the gradient and the diagonal terms of the Hessian for the gradient score.

$$\frac{\partial \log\{\lambda(\boldsymbol{x})\}}{\partial x_i} = -\frac{(\bar{\alpha}+\kappa)c(\alpha_i,\kappa)^{1/\kappa}x_i^{1/\kappa-1}}{\kappa\sum_{j=1}^{D}\{c(\alpha_j,\kappa)x_j\}^{1/\kappa}} + \left(\frac{\alpha_i}{\kappa}-1\right)\frac{1}{x_i},$$
$$\frac{\partial^2 \log\{\lambda(\boldsymbol{x})\}}{\partial x_i^2} = \frac{(\bar{\alpha}+\kappa)c(\alpha_i,\kappa)^{1/\kappa}x_i^{1/\kappa-1}}{\kappa\sum_{j=1}^{D}\{c(\alpha_j,\kappa)x_j\}^{1/\kappa}}\left[\frac{(\kappa-1)}{\kappa x_i} + \frac{c(\alpha_i,\kappa)^{1/\kappa}x_i^{1/\kappa-1}}{\kappa\sum_{j=1}^{D}\{c(\alpha_j,\kappa)x_j\}^{1/\kappa}}\right] - \frac{\alpha_i-\kappa}{\kappa x_i^2},$$

where $c(\alpha,\kappa)=\Gamma(\alpha+\kappa)/\Gamma(\alpha)$. These expressions can be inserted into eq. (2.33) with a suitable weighting function.

□

Pseudo-likelihood

In the Heffernan and Tawn conditional extremes model, the model parameters are estimated under the working assumption that the residual vector follows a $\mathsf{No}_D(\boldsymbol{\mu}, \mathrm{diag}\{\boldsymbol{\sigma}^2\})$ distribution.

Example 2.35 (Pseudo-likelihood for the conditional extremes model)
Consider the Heffernan–Tawn conditional extremes model presented in Section 2.4. The pseudo maximum likelihood estimator can be obtained from the empirical data through constrained optimization using the estimating function

$$\mathscr{L}(\boldsymbol{\alpha}_{|k},\boldsymbol{\beta}_{|k},\boldsymbol{\mu}_{|k},\boldsymbol{\sigma}_{|k};\boldsymbol{x}^{\mathrm{g}}) = \prod_{i=1}^{n}\prod_{\substack{j=1\\ j\neq k}}^{D}\phi\left(\frac{x_{i,j}^{\mathrm{g}} - a_{j|k}(x_{i,k}^{\mathrm{g}}) - b_{j|k}(x_{i,k}^{\mathrm{g}})\mu_{j|k}}{b_{j|k}(x_{i,k}^{\mathrm{g}})\sigma_{j|k}}\right)^{\mathbf{1}\{x_{i,k}^{\mathrm{g}}>u_k\}}.$$

□

2.6.5 Nonparametric estimation of the angular measure

In low dimensions, empirical estimation of the angular measure given in Equation (2.22) is an interesting avenue for inference. Consider a collection of independent and identically distributed vectors $\boldsymbol{X}$ with continuous marginal distributions that are in the max-domain of attraction of a max-stable distribution. One can transform the observations $\boldsymbol{X}_i$ to the unit Pareto scale $\boldsymbol{X}^{\mathrm{p}}$ using the probability integral transform, whereby $u_{ij}=\{1-\mathrm{rank}(x_{ij})/(n+1)\}^{-1}$, or else using the semi-parametric estimator of eq. (2.29). Once a vector $\boldsymbol{X}_i^{\mathrm{p}}$ has been obtained, we can form pseudo-angles by taking $r_i=\sum_{j=1}^{D}x_{ij}^{\mathrm{p}}$ and $w_{ij}=x_{ij}^{\mathrm{p}}/r_i$ for $i=1,\ldots,D-1$.

The angular distribution $\rho_1(\mathrm{d}\boldsymbol{w})$ given in eq. (2.11) asymptotically satisfies the moment constraint $\int_{\mathbb{S}_1}\omega_j\rho_1(\mathrm{d}\boldsymbol{\omega})=1/D$ for $j=1,\ldots,D$; this constraint can be imposed through empirical likelihood (Einmahl and Segers, 2009). We propose a novel extension that allows for estimation

of the angular measure (2.22) of ℓ-Pareto processes.

Let ℓ be a homogeneous risk functional; the collection of angles $\{\boldsymbol{w}_i\}_{i=1}^N$ for which $\ell(\boldsymbol{X}^{\mathrm{p}}) > u$ should be approximately distributed according to the angular distribution $\rho_{\ell,\mathrm{ang}}$ of eq. (2.22); we can use the change of measure formula (2.24) with $\alpha = 1$ to derive weights $\{a_i\}_{i=1}^N$ with $a_i = \ell(\boldsymbol{w}_i)$ to adjust for the fact that inference is performed using observations that approximately follow ρ_1.

The corresponding maximum empirical likelihood estimator of the spectral measure $\rho_{\ell,1}(\mathrm{d}\boldsymbol{w})$ solves

$$\max_{\boldsymbol{p}\in[0,1]^k} \prod_{j=1}^n a_j p_j \left(\sum_{i=1}^n a_i p_i\right)^{-1}, \qquad \sum_{i=1}^N p_i = 1, \quad \sum_{i=1}^N p_i \boldsymbol{w}_i = \boldsymbol{D}^{-1}.$$

The empirical probabilities $\widehat{p}_j$ associated to the unique observations $\boldsymbol{w}_j \in \mathbb{S}_1$ can be obtained through the Lagrange multiplier method for an arbitrary known weighting function a_i,

$$\mathcal{L} = \sum_{j=1}^n \log(a_j p_j) - n\log\left(\sum_{j=1}^n a_j p_j\right) + n\gamma\left(\sum_{j=1}^n p_j - 1\right) + n\boldsymbol{\lambda}^\top\left(\sum_{i=1}^n p_j \boldsymbol{w}_j - \boldsymbol{D}^{-1}\right),$$

where $\gamma, \boldsymbol{\lambda}$ are Lagrange multiplier coefficients. The solution of the Lagrangian yields $\gamma = -\boldsymbol{\lambda}^\top \boldsymbol{D}^{-1}$ and

$$p_k = \frac{s}{na_k} \frac{1}{1 + \boldsymbol{\kappa}^\top(\boldsymbol{w}_k - \boldsymbol{D}^{-1})/a_k},$$

where $s = \sum_{i=1}^n a_i p_i$ and $\boldsymbol{\kappa} = -\boldsymbol{\lambda} s$ is a dummy variable. If we let $\boldsymbol{b}_k = (\boldsymbol{w}_k - \boldsymbol{D}^{-1})/a_k$, the constraints are such that

$$1 = \sum_{i=1}^n p_i = \frac{s}{n}\sum_{i=1}^n \frac{1}{a_i}\frac{1}{1+\boldsymbol{\kappa}^\top\boldsymbol{b}_i}, \qquad \boldsymbol{D}^{-1} = \sum_{i=1}^n p_i \boldsymbol{w}_i = \frac{s}{n}\sum_{i=1}^n \frac{\boldsymbol{w}_k}{a_i}\frac{1}{1+\boldsymbol{\kappa}^\top\boldsymbol{b}_i},$$

from which we deduce that

$$s = \frac{n}{\sum_{i=1}^n a_i^{-1}(1+\boldsymbol{\kappa}^\top\boldsymbol{b}_i)^{-1}}.$$

The vector $\boldsymbol{\kappa}$ solves the optimization problem

$$\sum_{i=1}^n \frac{\boldsymbol{b}_i}{1+\boldsymbol{\kappa}^\top\boldsymbol{b}_i} = 0,$$

which is the gradient with respect to $\boldsymbol{\kappa}$ of

$$f(\boldsymbol{\kappa}) = \sum_{i=1}^n \log\left(1+\boldsymbol{\kappa}^\top\boldsymbol{b}_i\right). \tag{2.34}$$

It thus remains to find the minimum of eq. (2.34). As $-f$ is convex over the set $\{\boldsymbol{\kappa} \in \mathbb{R}^D : 1 +$

$\boldsymbol{\kappa}\boldsymbol{b}_i > 0, i = 1,\ldots,n\}$, the optimum solution will be found via iterated least-squares, replacing log by a self-concordant quartic approximation that will guarantee convergence to the global minimum provided that $\boldsymbol{D}^{-1}$ lies in the convex hull of $\{\boldsymbol{w}_i\}$ (Owen, 2013).

de Carvalho et al. (2013) proposes to use Euclidean likelihood to calculate the optimal weights subject to a mean constraint. They use the objective function $-\sum_{i=1}^N (Np_i - 1)^2/2$ subject to the constraints $\sum_{i=1}^N p_i = 1$ and $\sum_{i=1}^N p_i \boldsymbol{w}_i = D^{-1}\mathbf{1}_D$, the main benefit of which compared to empirical likelihood is the existence of a closed-form solution that can be obtained by least squares. A drawback is that the weights p_i could be negative, even if this is rare in practice.

Consider a set of unique observations $\boldsymbol{w}_j \in \mathbb{S}_1$, each weighted by a known function $a(\boldsymbol{w}_j)$; the objective function of the Euclidean likelihood is $\prod_{j=1}^n a_j p_j / \left(\sum_{i=1}^n a_i p_i\right)$ with constraints $p_j \geq 0, \sum_{j=1}^n p_j = 1$ and $\sum_{j=1}^n p_j \boldsymbol{w}_j = \boldsymbol{D}^{-1}$. The Lagrange function is

$$\mathcal{L} = \sum_{j=1}^n \log(a_j p_j) - n\log\left(\sum_{j=1}^n a_j p_j\right) + n\gamma\left(\sum_{j=1}^n p_j - 1\right) + n\boldsymbol{\lambda}^\top\left(\sum_{i=1}^n p_j \boldsymbol{w}_j - \boldsymbol{D}^{-1}\right).$$

Define $q_i = a_i p_i, s = \sum_{i=1}^n q_i$ and $r = \sum_{i=1}^n q_i^2$; differentiating the Lagrangean with respect to p_j and setting the result to zero gives deduce $\gamma = \boldsymbol{\lambda}^\top \boldsymbol{d}^{-1}$ and substituting this value yields

$$\sum_{i=1}^n (nq_i - s)(\mathrm{I}_{i=j}s - q_i) = nq_j s - nr = -\frac{s^3}{a_j n}\boldsymbol{\lambda}^\top(\boldsymbol{w}_j - \boldsymbol{d}^{-1})$$

and so

$$q_j = \frac{r}{s} - \frac{s^2}{a_j n^2}\boldsymbol{\lambda}^\top(\boldsymbol{w}_j - \boldsymbol{d}^{-1}),$$

where s is the solution of the cubic equation

$$s^3 \sum_{i=1}^n \frac{\boldsymbol{\lambda}^\top(\boldsymbol{w}_i - \boldsymbol{d}^{-1})}{a_i n^2} + s^2 - rn = 0.$$

The constraints for the mean and the sum of the weights yield

$$\sum_{j=1}^n \frac{r}{sa_j}(\boldsymbol{w}_j - \boldsymbol{d}^{-1}) - \sum_{j=1}^n \frac{s^2}{n^2 a_j^2}(\boldsymbol{w}_j - \boldsymbol{d}^{-1})\boldsymbol{\lambda}^\top(\boldsymbol{w}_j - \boldsymbol{d}^{-1}) = \mathbf{0}.$$

These derivations show that the solution of the Euclidean likelihood for weighted observations cannot be found by ordinary least squares.

A nonparametric estimator of the spectral measure (2.22) is $\widehat{H}(w) = \sum_{i=1}^N \widehat{p}_i \mathrm{I}\{w_i \leq w\}$, where $\widehat{p}_i$ solves either the empirical or Euclidean likelihood problems with a mean constraint (de Carvalho et al., 2013; Einmahl and Segers, 2009). Since the resulting spectral distribution is discrete (which is problematic in simulations), de Carvalho et al. (2013) suggest fitting a Dirichlet kernel to observations, with parameters $\nu\boldsymbol{w}_i (i = 1,\ldots,D)$ subject to the constraint $\|\nu\boldsymbol{w}_i\|_1 = 1$. The

"bandwidth" tuning parameter ν is chosen via cross-validation as the solution of (personal communication with Michał Warchoł)

$$\arg\max_{\nu\in\mathbb{R}} \sum_{i=1}^{n} \log\left\{ \sum_{\substack{k=1\\k\neq i}}^{n} p_{k,-i} f(\boldsymbol{w}_i; \nu \boldsymbol{w}_k) \right\},$$

where f is the density of the Dirichlet distribution, $p_{k,-i}$ is the Euclidean weight obtained from estimating the Euclidean likelihood problem without observation i, so $\sum_{k=1,k\neq i}^{n} p_{k,-i} = 1$. A drawback of this approach is that there is no closed-form expression for the leave-one-out procedure and one needs to perform n distinct optimizations.

2.7 Dependence measures and diagnostics

We focus here on some (typically bivariate) summaries of dependence that are used as goodness-of-fit diagnostics. Since the focus of this work is on parametric modelling of spatial extremes, we will be interested in the ability of the fitted model to reproduce the dependence structure and empirical estimates of the dependence measures can be used as diagnostics of model adequacy and goodness-of-fit.

2.7.1 Coefficients of tail dependence

To classify processes according to their tail behaviour, we consider the following notion.

Definition 2.35 (Tail correlation coefficient and asymptotic independence)
For a D-dimensional absolutely continuous random vector $\boldsymbol{X}$ with distribution function F and marginal distribution functions F_i $(i = 1, \ldots, D)$, the tail correlation coefficient χ is

$$\chi \equiv \lim_{v\to 1} \chi(v) = \lim_{v\to 1} \frac{\mathsf{P}\{F_1^{-1}(X_1) > v, \ldots, F_D^{-1}(X_D) > v\}}{1-v}.$$

The tail correlation coefficient χ lies in $[0,1]$; the upper bound is achieved by the comonotonic copula $C(\boldsymbol{v}) = \min(\boldsymbol{v})$, whereas the lower bound is reached in the bivariate case by countermonotonic vectors, whose copula is $C(\boldsymbol{v}) = \max\{0, u_1 + u_2 - 1\}$. In the case of independence, $\chi(v) = (1-v)^{D-1}$. If $\chi > 0$, the survival copula decays to zero at the same rate as the marginal survivor function and we say that $\boldsymbol{X}$ is asymptotically dependent and, if $\chi = 0$, asymptotically independent.

We consider empirical estimators of the tail correlation coefficient introduced in Definition 2.35. Let C denote the copula of $\boldsymbol{X}$, $C(\boldsymbol{v}) = F\{F_1^{-1}(v_1), \ldots, F_D^{-1}(v_D)\}$; the bivariate coefficient of upper tail dependence is (Coles et al., 1999)

$$\chi = \lim_{v\to 1} \mathsf{P}\{F_1(X_1) > v \mid F_2(X_2) > v\} = \lim_{v\to 1} \frac{\mathsf{P}\{F_1(X_1) > v \mid F_2(X_2) > v\}}{\mathsf{P}\{F_2(X_2) > v\}}$$

$$
\begin{aligned}
&= \lim_{v \to 1} \frac{\mathsf{P}[\min_i\{F_i(X_i) > v\}]}{1-v} \\
&= 2 - \lim_{v \to 1} \frac{1 - C(v,v)}{1-v}.
\end{aligned} \tag{2.35}
$$

The bivariate empirical estimator is usually defined as $2 - \log\{C(v\mathbf{1}_2\}/\log(v)$ by making the approximation $1-v \sim -\log(v), 1 - C(v,v) \sim -\log\{C(v,v)\}$ for $v \ll 1$. The resulting estimand is constant for bivariate extreme value distributions.

We can estimate the D-variate coefficient of upper tail dependence through Equation (2.35). Suppose that an n-sample from $\boldsymbol{X}$ is available, denoted by the $n \times D$ matrix $\mathbf{X}$ with (i,j)th entry $x_{i,j}$. An estimator of $\chi(v)$ is obtained by replacing the probability of exceedances of the structure variable $p_v = \mathsf{P}[\min_{j=1}^D \{F_j(X_j) > v\}]$ by its empirical counterpart

$$
\widehat{p}_v = n^{-1} \sum_{i=1}^n \mathbf{1}\{\min_{j=1}^D \{\mathrm{rank}(x_{ij})/(n+1)\} > v\},
$$

where $\mathrm{rank}(x_{ij}) = \sum_{k=1}^n \mathbf{1}\{x_{ij} \geq x_{kj}\}$. Assuming absolute continuity, $\mathbf{1}\{\min_i\{F_i(X_i)\}>v\} \sim \mathsf{Bin}(n, p_v)$ and an estimator of the variance of $\widehat{\chi}(v) = \widehat{p}_v/(1-v)$ can be obtained by the delta-method as $n^{-1}\widehat{p}_v(1-\widehat{p}_v)/(1-v)^2$. Alternatively, any empirical estimator of the survival copula, such as the beta copula (Segers et al., 2017),

$$
\widehat{C}^{\beta}(\boldsymbol{v}) = n^{-1} \sum_{i=1}^n \prod_{j=1}^D \sum_{s=\mathrm{rank}(x_{ij})}^n \binom{n}{s} v_j^s (1-v_j)^{n-s}, \tag{2.36}
$$

could be employed.

If we transform the marginal distributions to unit Pareto, with $X_i^{\mathrm{p}} = \{1 - F_i(X_i)\}^{-1}$ and $x = (1-v)^{-1}$, then $\mathsf{P}\big(\min_{j=1}^D X_j^{\mathrm{p}} > x\big) \sim \chi/x$. While $\chi = 0$ for any asymptotically independent process, such processes may exhibit different penultimate tail behaviour; Gaussian random vectors with correlation matrix different from $\mathbf{1}_D\mathbf{1}_D^\top$ are examples of asymptotically independent processes.

If $\chi = 0$, the tail correlation coefficient says nothing about the rate of decay of dependence. To better characterize the joint rate of decay of the survivor function, write

$$
\mathsf{P}\left(\min_{j=1}^D X_j^{\mathrm{p}} > x\right) = L(x)x^{-1/\eta}, \qquad \eta \in (0,1],
$$

with $L(x) \in \mathsf{RV}_0$ a slowly varying function, meaning $\lim_{x\to\infty} L(tx)/L(x) = 1$ for any $t > 0$ (Ledford and Tawn, 1996, § 5). When $\chi > 0$, $L(x) \to \chi$ as $x \to \infty$ and $\eta = 1$. The coefficient of tail dependence, η, takes values in $(0, D^{-1})$ if the variables are negatively associated, $\eta = D^{-1}$ for independent variables and $\eta \in (D^{-1}, 1]$ if the variables exhibit positive association. In the multivariate setting, the coefficients η_C for subsets $C \subset \{1,\ldots,D\}$ satisfy $\eta_C \leq \eta$.

With unit Pareto margins, the structural variable $T^{\mathrm{p}} = \min_{j=1}^D X_j^{\mathrm{p}}$ is such that, for large u

(Ledford and Tawn, 1996, eq. 5.6),

$$\mathsf{P}\left(T^{\mathrm{p}} > u + t \mid T^{\mathrm{p}} > u\right) \sim \frac{L(u+t)}{L(u)}(1 + t/u)^{-1/\eta},$$

so we can estimate η by fitting a generalized Pareto distribution with shape η and scale ηu to exceedances of T^{p} above u. If we transform the data to exponential margins instead and define $T^{\mathrm{e}} = \min_{j=1}^{D} X_j^{\mathrm{e}}$, then

$$\mathsf{P}\left(T^{\mathrm{e}} > u + t \mid T^{\mathrm{e}} > u\right) \approx \exp(-t/\eta), \qquad u \to \infty,$$

and the maximum likelihood estimator of the scale parameter η coincides with the Hill estimator,

$$\widehat{\eta}^{\mathrm{e}} = n_u^{-1} \sum_{i=1}^{n} \left(\min_{j=1}^{D} x_{ij}^{\mathrm{e}} - u\right), \qquad n_u = \#\left\{i : \min_{j=1}^{D} x_{ij}^{\mathrm{e}} > u\right\}.$$

It is enlightening to consider marginal transformation of the data, as illustrated in Figure 2.6: with unit Pareto margins, extrapolation is done on rays from the origin and the probability of lying in the set rA (grey hashed) can be estimated based on the relation $\mathsf{P}\left(\boldsymbol{Y}^{\mathrm{p}} \in rA\right) \sim r^{-1/\eta}\mathsf{P}\left(\boldsymbol{Y}^{\mathrm{p}} \in A\right)$ for an extreme set A. The right plot shows the same data transformed to exponential margins, where now $\mathsf{P}\left(\boldsymbol{Y}^{\mathrm{e}} \in r + A\right) \sim \exp(-r/\eta)\mathsf{P}\left(\boldsymbol{Y}^{\mathrm{e}} \in A\right)$ and extrapolation is done by translating the set A along the 45° degree line. In both cases, the probability of the less extreme set A can be estimated empirically.

Alternative estimators of the tail dependence coefficient η reviewed in § 9.5.2 of Beirlant et al. (2004) may work better when convergence of the ratio $L(u+t)/L(u) \to 1$ is slow.

One drawback of the Ledford and Tawn approach is that both components must decay at the same rate. Wadsworth and Tawn (2013) propose to look at different extrapolation paths by replacing the multivariate regular variation by a collection of univariate regular variation assumptions, assuming that for $\boldsymbol{\beta} \in \mathbb{R}_+^D \setminus \{\mathbf{0}_D\}$,

$$\mathsf{P}\left(Y^{\mathrm{e}} > \boldsymbol{\beta}\log(t)\right) = t^{-\kappa(\boldsymbol{\beta})}L(t;\boldsymbol{\beta}), \qquad t \to \infty,$$

where $L(t;\boldsymbol{\beta})$ is a slowly varying function and κ is a non-decreasing homogeneous function of order 1, termed the angular dependence function, that depends on $\boldsymbol{\beta}$ through the angle $\boldsymbol{\omega} = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_1 \in \mathbb{S}_1$. Extrapolation can be performed as before upon rays at angles $\boldsymbol{\omega}$. Under positive quadrant dependence, κ is convex and pointwise estimates can be obtained by ignoring the slowly-varying function and estimating the tail index, with $\boldsymbol{Y}^{\mathrm{e}}/\boldsymbol{\beta}$ as input data. Further properties of the angular dependence function, given in Wadsworth and Tawn (2013), can be exploited to obtain more efficient estimates across a range of angles but the slowly-varying function changes with $\boldsymbol{\omega}$ and this is likely to impact estimators of κ.

Another widely used tail dependence coefficient is $\bar{\chi} = 2\eta - 1$, which gives $\bar{\chi} \in (-1, 1]$ (Coles

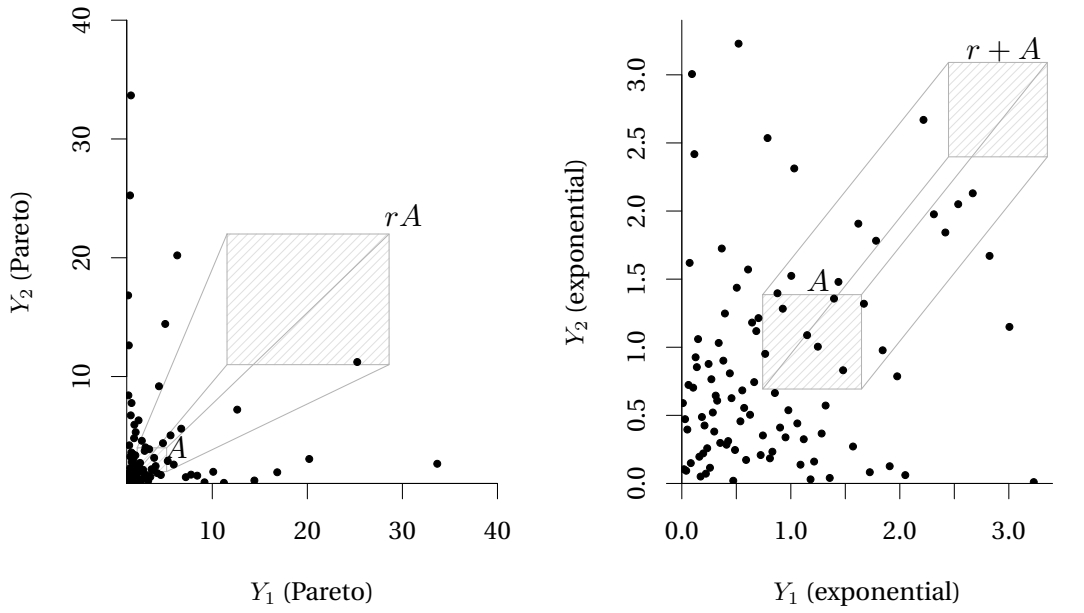

Figure 2.6 – *Empirical extrapolation of the tail probability based on hidden regular variation in Pareto margins (left) and exponential margins (right).*

et al., 1999). An estimator of the latter is $\bar{\chi}(v) = 2\log(1-v)/\log\{\bar{C}(v\mathbf{1}_D)\} - 1$ for $v \in (0,1)$ and $\bar{C}$ an estimator of the survival copula. For the empirical estimator, we obtain approximate pointwise standard errors through the delta-method.

Example 2.36 (Detection of asymptotic dependence and extrapolation)
Figure 2.7 shows empirical estimates of η and χ for daily summer rainfall measurements at four weather stations in canton Zürich (Zürich Fluntern, Zürich Kloten, Dietikon and Effretikon) for the period 1962–2012. The pairwise distance between stations is less than 25 km. Rainfall extremes tend to be due to mixture of events (large scale depression systems), but are often localized as the intensity increases. The empirical coefficient of tail dependence $\widehat{\eta}$ is stable below 0.8. This is contrast with estimates obtained through Hill or maximum likelihood estimates of η (not shown), which are unstable and tend to one. The estimate of χ decreases towards zero, incompatible with asymptotic dependence and typical in applications.

□

2.7.2 Extremogram

Davis and Mikosch (2009) discuss an analog of the correlogram designed for extremes, the so-called extremogram, based on the tail empirical process. For sets $A, B \subset \mathbb{R}^m$ bounded away from the origin,

$$\varrho_{AB}(h) = \lim_{x\to\infty} \mathsf{P}\left(Z(\boldsymbol{h}) \in xB \mid Z(\boldsymbol{o}) \in xA\right), \quad h \in \mathbb{R}^m.$$

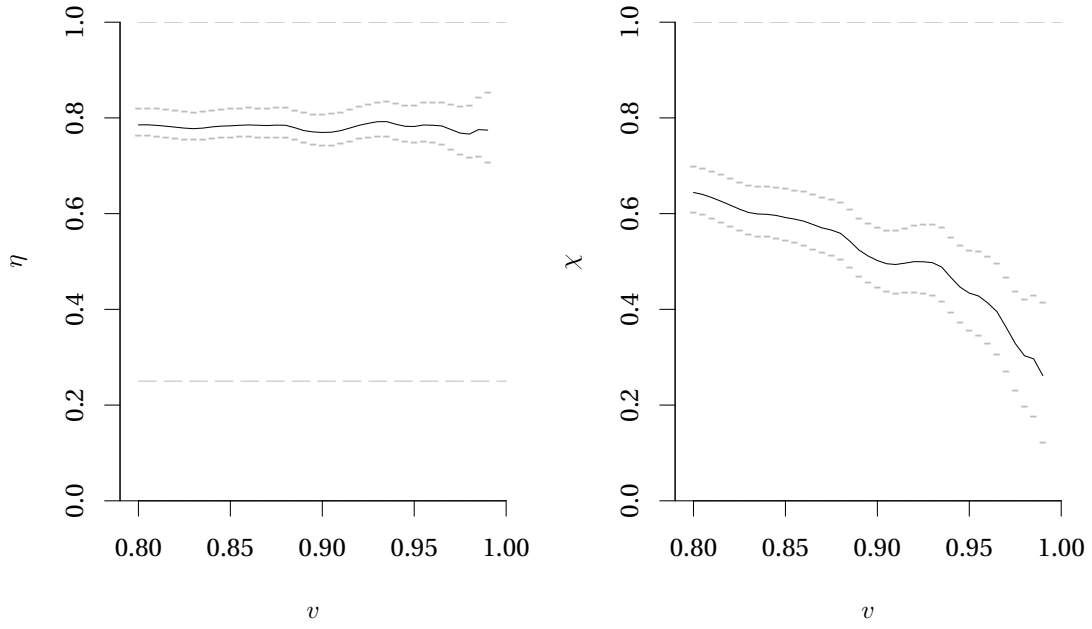

Figure 2.7 – *Empirical tail dependence (left) and tail correlation coefficients estimated using the Beta copula* (2.36)*, with pointwise asymptotic 95% confidence intervals. The dashed line at 0.25 in the left panel indicates the theoretical η for independence.*

In the unidimensional setting with $m = 1$ and with sets of the form $A = B = (z, \infty)$, the extremogram reduces to the pairwise χ coefficient of Ledford and Tawn (1996). Max-stable processes have positive pairwise dependence, i.e., $\chi(\boldsymbol{h}) \geq 0$. The function $\chi(\boldsymbol{h})$ for sites $\boldsymbol{s}_j$, $\boldsymbol{s}_k$ at lag distance $\boldsymbol{h}$ and $\mathsf{Z}^*(\mathcal{S})$ a simple max-stable process is

$$\chi(\boldsymbol{h}) = \lim_{z \to \infty} \mathsf{P}\left(Z^*(\boldsymbol{s}_j) > z \mid Z^*(\boldsymbol{s}_k) > z\right).$$

Proposition 2.36 (Extremal coefficient for ℓ-Pareto processes)
de Fondeville and Davison (2018) proposed a variant of χ that targets risk exceedances of the form

$$\varrho(v; \boldsymbol{s}_j, \boldsymbol{s}_k, \ell) = \mathsf{P}\left(Y^*(\boldsymbol{s}_j) > v \mid Y^*(\boldsymbol{s}_k) > v, \ell\{\mathsf{Z}^*(\mathcal{S})\} > u\right).$$

Whenever $\mathscr{A}_u := \{\boldsymbol{Y}^* : \ell(\boldsymbol{Y}) > u\} \subseteq \{\boldsymbol{Y}^* : Y^*(\boldsymbol{s}_j) > v, Y^*(\boldsymbol{s}_k) > v\}$, then $\varrho(v; \boldsymbol{s}_j, \boldsymbol{s}_k, \ell) = 2 - \theta(\boldsymbol{s}_j, \boldsymbol{s}_k)$.

Proof
Consider a pairwise vector $\boldsymbol{Y}^*$ with unit Pareto margins defined over the risk region $\mathscr{A}_u$, since $\mathsf{P}\left(Y^*(\boldsymbol{s}_k) > v, \ell\{\mathsf{Y}^*(\mathcal{S})\} > u\right) = 1/\{v\Lambda(\mathscr{A}_u)\}$. We have

$$\varrho(v; \ell) = \frac{\mathsf{P}\left(Y^*(\boldsymbol{s}_j) > v, Y^*(\boldsymbol{s}_k) > v, \ell\{\mathsf{Y}^*(\mathcal{S})\} > u\right)}{\mathsf{P}\left(Y^*(\boldsymbol{s}_k) > v, \ell\{\mathsf{Y}^*(\mathcal{S})\} > u\right)}$$

$$
\begin{aligned}
&= v\Lambda(\mathscr{A}_u)\left\{\mathsf{P}\left(Y^*(\boldsymbol{s}_j)>v\right)+\mathsf{P}\left(Y^*(\boldsymbol{s}_k)>v\right)-\mathsf{P}\left(Y^*(\boldsymbol{s}_j)\leq v, Y^*(\boldsymbol{s}_k)\leq v\right)\right\}\\
&= v\Lambda(\mathscr{A}_u)\left[2\{v\Lambda(\mathscr{A}_u)\}^{-1}-\frac{V_{jk}(v\mathbf{1}_2)}{\Lambda(\mathscr{A}_u)}\right],
\end{aligned}
$$

using the -1-homogeneity of V, which gives $\Lambda\{\boldsymbol{z}:\boldsymbol{z}\in[\mathbf{0},v\mathbf{1}_2)^{\mathsf{c}}\}=V_{jk}(v\mathbf{1}_2)$.

■

An empirical estimator is obtained by thresholding the data at a large quantile, for example $u_j=F_{\boldsymbol{s}_j}^{-1}(1-p)$, and calculating the fraction of points falling in the region defined by the marginal thresholds and the functional threshold u associated to the risk functional ℓ,

$$
\widehat{\varrho}(\boldsymbol{s}_j,\boldsymbol{s}_k,\ell)=\frac{\sum_{i=1}^n \mathbf{1}\{\ell(\boldsymbol{y}_i)>u\}\mathbf{1}\{y_{ij}>u_j\}\mathbf{1}\{y_{ik}>u_k\}}{\sum_{i=1}^n \mathbf{1}\{\ell(\boldsymbol{y}_i)>u\}\mathbf{1}\{y_{ij}>u_j\}};
$$

the ratio estimator $\widehat{\varrho}(\boldsymbol{s}_j,\boldsymbol{s}_k,\ell)$ provides empirical estimates of the probability of joint marginal and risk exceedances. Davis et al. (2012) advocate the use of a stationary block bootstrap procedure for uncertainty assessment of the extremogram.

2.7.3 Extremal coefficient

The extremal coefficient is the exponent measure evaluated at $\mathbf{1}_D$, $\theta_{1:D}:=V(\mathbf{1}_D)$. If $\boldsymbol{Z}^*$ is a D-dimensional simple max-stable vector, $\mathsf{P}(\boldsymbol{Z}^*<z\mathbf{1}_D)=\exp(-\theta_{1:D}/z)$ for any $z\in\mathbb{R}$, owing to homogeneity of V and $\chi_{ij}=2-\theta_{ij}$. The extremal coefficient was introduced by Buishand (1984) in the multivariate setting and in Smith (1990),Coles (1993) and Schlather and Tawn (2003) for max-stable processes. If C is the extreme value copula of $\boldsymbol{Z}^*$, then $\boldsymbol{\theta}=\log\{C(u\mathbf{1}_D)\}/\log(u)$ for any $u\in(0,1)$ and we can use this as a diagnostic of max-stability, replacing C by an empirical estimator of the copula and checking whether or not the estimator is constant over $u\in(0,1)$.

The range of admissible values for the extremal coefficient at D sites is $[1,D]$, with 1 corresponding to comonotonicity and D to independence; this led Schlather and Tawn (2003) to say that the extremal coefficient measures the "effective number of independent variables". For a subset $\mathrm{C}\subset\{1,\ldots,D\}$, the extremal coefficient is

$$
\theta_{\mathrm{C}}=\int_{\mathbb{S}_1}\max_{i\in\mathrm{C}} w_i\rho_1(\mathrm{d}\boldsymbol{w}).
$$

The bivariate extremal function $\theta(\boldsymbol{s}_1,\boldsymbol{s}_2)=\theta(\boldsymbol{h})$ for $\boldsymbol{h}=\boldsymbol{s}_1-\boldsymbol{s}_2$ is the most used summary of spatial extremal dependence. The dependence between sites can usually be expected to weaken as the distance between them increases. If Z^* is a stationary max-stable process, then $2-\theta(\boldsymbol{h})$ is positive definite, is nondifferentiable at the origin (unless $\theta(\boldsymbol{h})$ is constant and equal to 1), has at most one jump (at the origin) in dimension $D\geq 2$ and is continuous elsewhere (Schlather and Tawn 2003, Theorem 3). For isotropic processes, $\theta(\boldsymbol{h})$ reduces to a univariate function $\theta(\|\boldsymbol{h}\|):\mathbb{R}\to[1,2]$. The positive-definitiveness of $2-\theta(\boldsymbol{h})$ allows one to use it as

a covariance function. In the spatial setting, further conditions are derived by Cooley et al. (2006) using properties of the variogram, since $2-\theta(\boldsymbol{h})$ must be a positive definite function. The extremal coefficient function $\theta(\boldsymbol{h})$ satisfies the constraints

1. $\theta(\boldsymbol{h}_1+\boldsymbol{h}_2) \leq \theta(\boldsymbol{h}_1)\theta(\boldsymbol{h}_2)$;
2. $\theta(\boldsymbol{h}_1+\boldsymbol{h}_2)^\tau \leq \theta(\boldsymbol{h}_1)^\tau + \theta(\boldsymbol{h}_2)^\tau - 1$, $0 \leq \tau \leq 1$;
3. $\theta(\boldsymbol{h}_1+\boldsymbol{h}_2)^\tau \geq \theta(\boldsymbol{h}_1)^\tau + \theta(\boldsymbol{h}_2)^\tau - 1$, $\tau \leq 0$.

Example 2.37 (Extremal coefficient of the Brown–Resnick process)
The pairwise extremal coefficient $\theta(\boldsymbol{h})$ is $2\Phi(\{\gamma(\boldsymbol{h})\}^{1/2})$ for a semivariogram function $\gamma(\cdot)$ (Engelke et al., 2015). For the Smith storm process, the semivariogram is replaced by the Malahanobis distance and $\gamma(\boldsymbol{h}) = 2\Phi(\{\boldsymbol{h}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{h}/2\}^{1/2})$.

□

Example 2.38 (Extremal coefficient of the extremal Student process)
The pairwise extremal coefficient of the extremal Student process with ν degrees of freedom and correlation function $\rho(\cdot)$ is (Ribatet, 2013, p. 130)

$$\theta(\boldsymbol{h}) = 2\mathsf{St}\left(\left[\frac{(\nu+1)\{1-\rho(\boldsymbol{h})\}}{1+\rho(\boldsymbol{h})}\right]^{1/2}; \nu+1\right) \leq 2\mathsf{St}\left\{(\nu+1)^{1/2}; \nu+1\right\}, \tag{2.37}$$

where $\mathsf{St}(\cdot;\nu)$ denotes the Student distribution function with ν degrees of freedom.

□

Proposition 2.37 (Schlather's estimator of the extremal coefficient)
Suppose the data matrix consists of D-variate random vectors which are assumed to be distributed according to a unit Fréchet distribution. For any collection $\mathrm{C} \subset \{1,\ldots,D\}$, the variable $\max_{j\in\mathrm{C}} Z_j$ is in the max domain of attraction of a Fréchet variable with survival function $H(z) = 1-\exp(-\theta_\mathrm{C}/z)$; assuming this holds exactly for large observations above a threshold u, Schlather and Tawn (2003) propose a censored likelihood estimator. The contribution to the likelihood for observations falling below u is $G(u)$, while that of the n_u exceedances z_i, $i=1,\ldots,n_u$ is $g(z_i)$. The Schlather and Tawn likelihood is

$$\ell_\mathrm{C}(\theta_\mathrm{C};\boldsymbol{z}) = \mathrm{card}\left\{i : \max_{j\in\mathrm{C}}\{\widetilde{z}_{i,j}\} > u\right\}\log(\theta_\mathrm{C}) - \theta_\mathrm{C}\sum_{i=1}^{n}\left(\max\left\{u, \max_{j\in\mathrm{C}}\widetilde{z}_{i,j}\right\}\right)^{-1},$$

where $\widetilde{z}_{i,j} = z_{i,j}n^{-1}\sum_{i=1}^n 1/z_{i,j}$ are the Fréchet observations rescaled by the component-wise harmonic mean. The so-called Schlather estimator corresponds to the maximum likelihood estimator $\widehat{\theta}_\mathrm{C}$. Thibaud and Opitz (2015) propose an alternative estimator of $\theta(h)$ for ℓ-Pareto processes with max-risk functional in § 2 of the Supplementary Material.

The extremal coefficients satisfy certain ordering properties: let $\mathscr{P}_D$ be the collection of all non-empty subsets of the indices $\{1,\ldots,D\}$ and let C_D be any subset of $\{1,\ldots,D\}$. For any

$1 < m \leq D$, the extremal coefficient satisfies the inequalities

$$\max\left\{\sum_{\substack{\mathrm{C}\in\mathscr{P}_m\setminus\{\mathrm{C}_m\}\\ \mathrm{c}\subseteq\mathrm{C}}}(-1)^{|\mathrm{C}\setminus\mathrm{c}|}\theta_{\mathrm{C}}\right\} \leq \theta_{\mathrm{C}_m} \leq \min\left\{\sum_{\substack{\mathrm{C}\in\mathscr{P}_m\setminus\{\mathrm{C}_m\}\\ \mathrm{c}\subseteq\mathrm{C}}}(-1)^{|\mathrm{C}\setminus\mathrm{c}|+1}\theta_{\mathrm{C}}\right\}.$$

These bounds are sharp. Ensuring self-consistency of the estimator translates into a linear program which is costly to solve except in low dimensions. A simpler constraint applies for bivariate diagnostics. The pairwise extremal coefficients $\{\theta_{jk}\}$, $j,k \in \{1,\ldots,D\}$ are self-consistent whenever the matrix with entries $1-2(\theta_{jk}-1)^2$ is positive definite.

If observations arise from a stationary isotropic max-stable random field, the Schlather estimator can be plotted against distance to form the analog of the sample variogram cloud. A function $\widehat{\theta}(h)$ is obtained by taking

$$\widehat{\theta}(h) = 1 \vee \frac{\sum_{(j,k)\in B_h}(\mathrm{card}\{T_j\cap T_k\})^{1/2}\widehat{\theta}_{jk}}{\sum_{(j,k)\in B_h}(\mathrm{card}\{T_j\cap T_k\})^{1/2}} \wedge 2,$$

where T_j is the index set of data for the jth site, $\widehat{\theta}_{jk}$ is estimated using complete pairs and B_h is a bin of width 2ϵ, defined as

$$B_h := \{(j,k) : \|\boldsymbol{s}_j - \boldsymbol{s}_k\| \in [h-\epsilon, h+\epsilon], \mathrm{card}\{T_j\cap T_k\} > 1\}.$$

This smoothing procedure for the binned observations allows one to deal with missing values and handle the unequal variances due to the different numbers of pairs per bin. It also reduces the level of noise in the cloud of points.

Proposition 2.38 (Smith's estimator of the extremal coefficient)
A simpler estimator of the extremal coefficient is obtained by noting that, if $\boldsymbol{Z}^*$ has unit Fréchet margins, the extremal coefficient $\theta(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_D)$ is the scale parameter of $1/\max\{\max_{j=1}^D Z^*(\boldsymbol{s}_j)\} \sim \mathsf{Exp}(\theta(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_D))$. An estimator of the pairwise extremal coefficient is therefore the reciprocal arithmetic mean of the observations (Smith, 1990; Coles et al., 1999; Erhardt and Smith, 2012),

$$\widehat{\theta}_{\mathrm{C}} = \frac{n}{\sum_{i=1}^n \min_{j\in\mathrm{C}} z_i^{*-1}(\boldsymbol{s}_j)},$$

which is equivalent to the Schlather estimator without reweighting and a zero threshold.

Proposition 2.39 (F-madogram estimator of the extremal coefficient)
Let F_i be the marginal distribution at site $\boldsymbol{s}_i$ and define the F-madogram as

$$\nu_F(\boldsymbol{s}_j,\boldsymbol{s}_k) = \frac{1}{2}\mathsf{E}\left[\left|F_j\{Z(\boldsymbol{s}_j)\} - F_k\{Z(\boldsymbol{s}_k)\}\right|\right].$$

Assuming that the data arise from an isotropic stationary max-stable process, one can write

ν_F as a univariate function in $h = \|\boldsymbol{s}_j - \boldsymbol{s}_k\|$, and simple calculations yield

$$\theta(\boldsymbol{h}) = \frac{1+2\nu_F(h)}{1-2\nu_F(h)}. \tag{2.38}$$

A natural estimator of ν_F for n complete pairs of observations is the rank-based statistic

$$\widehat{\nu}_F(\boldsymbol{s}_j, \boldsymbol{s}_k) = \frac{1}{2n^2}\sum_{i=1}^{n}\left|\text{rank}(x_{ij}) - \text{rank}(x_{ik})\right|,$$

where $\text{rank}(x_{ij} = \sum_{l=1}^{n}\mathbf{1}\{x_{ij}\geq x_{lj}\}$. The estimator $\widehat{\nu}_F(\boldsymbol{s}_j, \boldsymbol{s}_k) \in [1,3]$ by construction, pairwise extremal coefficients larger than 2 indicating negative dependence between pairs. Alternatively, one could use the empirical distributions $\widetilde{F}_j$, $\widetilde{F}_k$ estimated using all observations, set

$$\widetilde{\nu}_F(\boldsymbol{s}_j, \boldsymbol{s}_k) = \frac{1}{2n}\sum_{i=1}^{n}\left|\widetilde{F}_j(x_{ij}) - \widetilde{F}_k(x_{ik})\right|,$$

and estimate $\widetilde{\theta}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ based on eq. (2.38), truncating the pairwise estimates to fall in the interval $[1,3]$. Due to its simplicity, the F-madogram based extremal coefficient estimator is a widely used diagnostic that has superseded the competing estimators.

Example 2.39 (Swiss daily rainfall)
We compute the extremal coefficient for a rainfall dataset from MeteoSwiss which contains daily cumulated rainfall records from 1863 until 2016 at 163 stations over Switzerland. Estimates for the three different methods, based on yearly maxima, are displayed in Figure 2.8. The Schlather–Tawn estimator is highly sensitive to the choice of threshold, with higher values leading to stronger dependence estimates because of the censoring scheme. All estimates broadly indicate a decrease of dependence with distance, but it is hard to conclude whether events become asymptotically independent as h increases, due to the small size of the domain.

□

Proposition 2.40 (λ-madogram estimator)
An extension proposed in Naveau et al. (2009) gives different weight to the pairs. The λ-madogram is defined for $\lambda \in (0,1)$ as

$$\nu_F(h,\lambda) = \frac{1}{2}\mathsf{E}\left(\left|F_j^{\lambda}(Z(\boldsymbol{s}_j)) - F_k^{1-\lambda}(Z(\boldsymbol{s}_k))\right|\right).$$

If the data are in the max-domain of attraction of a simple multivariate extreme value distribution, then

$$\nu_F(h,\lambda) = \frac{V_h(\lambda,1-\lambda)}{1+V_h(\lambda,1-\lambda)} - \frac{3}{2(1+\lambda)(2-\lambda)}.$$

Naveau et al. (2009) propose the empirical estimator

$$\widehat{\nu}_F(\boldsymbol{s}_j, \boldsymbol{s}_k, \lambda) = \frac{1}{2n}\sum_{i=1}^{n}\left|\widetilde{F}_j^{\lambda}(x_{ij}) - \widetilde{F}_k^{1-\lambda}(x_{ik})\right|,$$

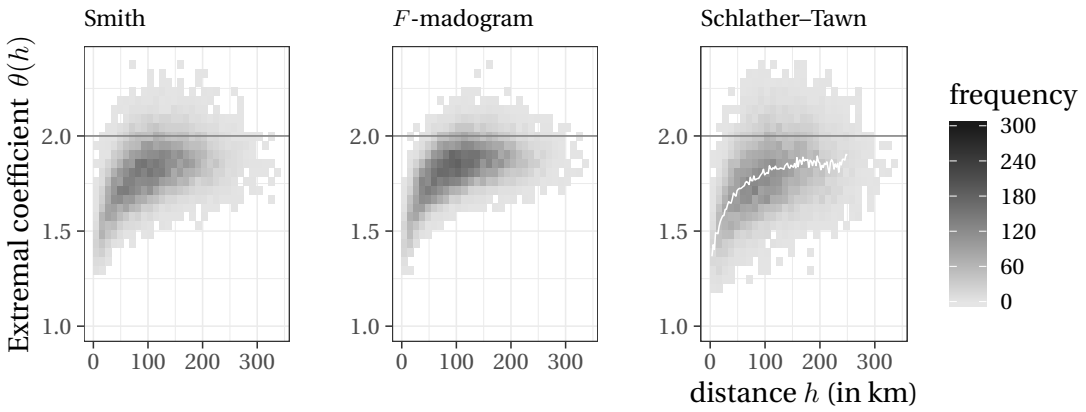

Figure 2.8 – *Empirical extremal coefficient cloud for the Swiss rainfall data, calculated using Smith's estimator (left), the F-madogram (middle) and Schlather's estimator thresholded at the 50% percentile displayed alongside the binned estimate average curve in white (right).*

along with a boundary-corrected estimator. The λ-madogram provides a non-parametric estimator of the exponent measure V and allows one to consider directions other than $v\mathbf{1}_D$, similar in spirit to Wadsworth and Tawn (2013).

2.7.4 Extremal concurrence probability

Concurrence of extreme events arises naturally in the spectral construction of max-stable processes for which all the information is observed. Dombry et al. (2018) introduced the extremal concurrence probability $p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_D)$ as the limiting probability of the sample concurrent probability at fixed locations $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_D \in \mathcal{S}$ as $n\to\infty$. Extremal concurrence corresponds to the hitting scenario $\boldsymbol{\Pi} = \{\varpi\}$, where $\varpi = \{1,\ldots,D\}$; this means that all the values of the max-stable process stem from a single extremal function. Recall that the marked point process has elements $\varphi_i = \zeta_i \boldsymbol{W}_i$ and the max-stable process $\boldsymbol{Z}^* = \max_{i\geq 1}\varphi_i$; we thus have

$$p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_D) = \mathsf{P}\left(\exists i \in \mathbb{N}^+ : \varphi(i)(\boldsymbol{s}_j) = Z^*(\boldsymbol{s}_j), j = 1,\ldots,D\right).$$

The pairwise extremal concurrence probability shares strong connections with the extremal coefficient and many properties of $\theta(\cdot)$ exist in a similar form for $p(\cdot)$, see the Corollary in Dombry et al. (2018). We can obtain a diagnostic plot akin to the variogram cloud by mapping the pairwise extremal concurrence probability against the pairwise distance $h = \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$ for $p(h) \in [0,1]$; the bounds correspond respectively to independence and comonotonicity. The extremal concurrence probability can be used to estimate the spatial dependence relative to a site of interest, or else integrated over a domain to estimate the area of a concurrence cell.

Remarkably, the bivariate extremal concurrence probability $p(h)$ equals Kendall's τ for max-stable vectors and an estimate of the latter can be obtained in $\mathrm{O}\{n\log(n)\}$ operations (Chris-

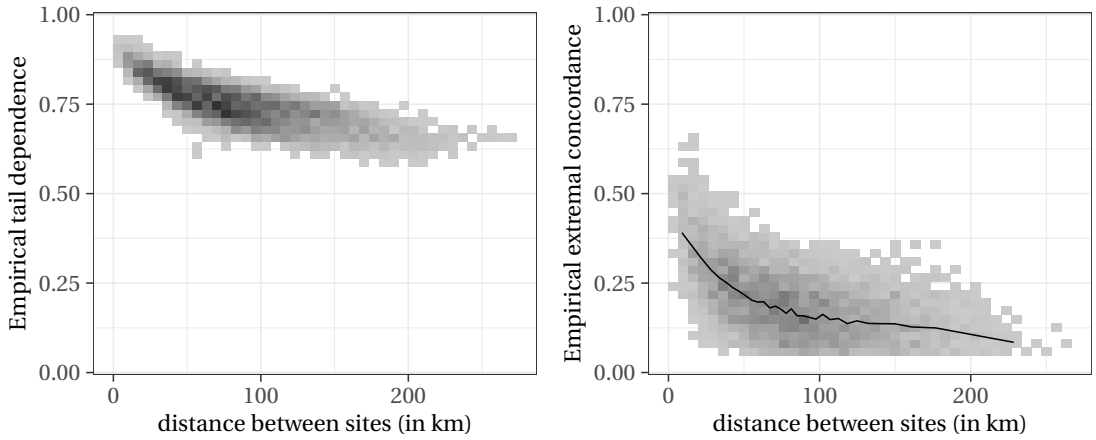

Figure 2.9 – *Empirical tail dependence coefficient cloud, $\eta(h)$ for the Aar basin rainfall data, calculated using the empirical estimator for η based on observations for the months of May to September (left). Empirical concurrence probability $p(h)$ estimated using Kendall's τ based on yearly maxima for the same series, with superimposed line giving average for binned estimates based on 30 bins (right).*

tensen, 2005). Unfortunately, analytical expressions for $p(h)$ are difficult to obtain for most processes of interest and one may need to resort to Monte Carlo methods.

Example 2.40 (Extremal concurrence probability of Brown–Resnick process)
Generally, we can relate the exponent measure to the extremal concurrence probability via

$$p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k) = \mathsf{E}_{\boldsymbol{Y}^*}\left(\{V(\boldsymbol{Y}^*)\}^{-1}\right),$$

where $\boldsymbol{Y}*$ is an independent vector identically distributed as $\boldsymbol{Y}$.

The pairwise extremal concurrence probability function is (Dombry et al., 2018)

$$p(\boldsymbol{o}, \boldsymbol{h}) = \mathsf{E}\left(\left(\Phi(Z) + \exp\left[\gamma(\boldsymbol{h}) - \{2\gamma(\boldsymbol{h})\}^{1/2} Z\right] \Phi\left[\{2\gamma(\boldsymbol{h})\}^{1/2} - Z\right]\right)^{-1}\right),$$

where $Z \sim \mathsf{No}(0,1)$ with associated distribution function Φ. One can approximate this by Monte Carlo integration, possibly using antithetic variables to reduce the variability of the estimate.

□

Example 2.41 (Extremal dependence for rainfall)
We consider spatial estimates of the tail dependence coefficient $\eta(h)$ and extremal concordance probability $p(h)$ for 105 stations located in the Aar watershed, reported in Figure 2.9. The measurements are daily cumulated rainfall over the period May 1930–October 2014. As the distance between sites increases, there is a clear decrease in dependence and the estimated concordance probability drops sharply. Both measures suggest asymptotic independence at large lags.

□

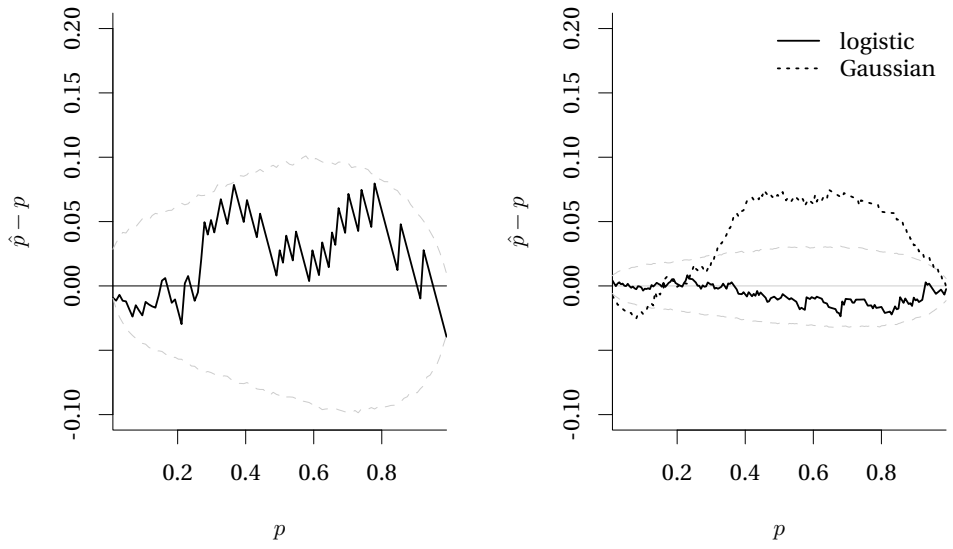

Figure 2.10 – *Tukey's probability-probability plot for the test of max-stability with nonparametric bootstrap 99% confidence region for the null hypothesis. The plot show the discrepancy between estimated and theoretical probability $\widehat{p} - p$ as a function of the percentile p. Statistic for $D = 10$ sites of the Swiss rainfall data based on 103 yearly maxima (left) and test for simulated 10-dimensional Gaussian data with covariance $\mathbf{I}_n + \mathbf{1}_D\mathbf{1}_D^\top/2$ and logistic max-stable model with $\alpha = 0.5$ (right).*

2.8 Model assessment and diagnostics

Most model diagnostics rely on projections of the data onto univariate summaries using functionals. For example, Kiriliouk et al. (2019) suggest the use of a quantile-quantile plot for the multivariate generalized Pareto distribution based on aggregated data; the pseudo-polar decomposition gives a radial variable that is generalized Pareto distributed.

Proposition 2.41 (Projection of multivariate generalized Pareto)
If $\boldsymbol{X} \sim \mathsf{MGP}(\boldsymbol{\sigma}, \xi\mathbf{1}_D, \Lambda)$, then for $\boldsymbol{a} \in [\mathbf{0}_D, \infty_D)$, the conditional distribution $\boldsymbol{aX} \mid \boldsymbol{aX} > 0 \sim \mathsf{GP}(\boldsymbol{a}^\top\boldsymbol{\sigma}, \xi)$ (Rootzén et al., 2018, prop. 9.1).

Example 2.42 (Test of max-stability)
If $\boldsymbol{Z}$ is a max-stable process with standard Gumbel margins, with distribution function $G(\boldsymbol{z}) = \exp[-V\{\exp(\boldsymbol{z})\}]$, then for any subset of size $J \subseteq \{1, \ldots, D\}$, the variable $Z_J = \max_{j\in J} Z(\boldsymbol{s}_j)$ has distribution function $H_J(z) = \exp\{-\exp(z - \mu_J)\}$, where $\mu_J = \log\{V_J(\mathbf{1})_{|J|}\}$ and $0 \leq \mu_J \leq \log(|J|)$. This follows from the homogeneity of the exponent measure. Gabda et al. (2012) propose a probability-probability plot based on fitting μ_J through maximum likelihood with the parameter constraints for each set of $|J|$ stations. The data are obtained by pooling replications of the max-stable field and selecting all subsets of size $|J|$ if $\binom{D}{|J|}$ is small, or else a limited number of stations among those that display the higher dependence so as to maximize the power of the test (independence being a special case of max-stability). Uncertainty quantification is performed using a nonparametric bootstrap, as shown in Figure 2.10.

□

The two main methods for comparing models from different parametric families are information criteria based and score based; we review the latter.

2.8.1 Proper scoring rules

Scoring rules (Gneiting and Raftery, 2007) are functionals that assign a penalty based on the predictive performance of a forecast distribution $G \in \mathscr{F}$ for an observation $\boldsymbol{y} \sim F$, viz. $S(G, y) : \mathscr{F} \times \text{supp}(F) \mapsto \mathbb{R} \cup \{\infty\}$.

A scoring rule S is proper if, for $Y \sim F$ if $\mathsf{E}_F\{S(F, Y)\} \leq \mathsf{E}_F\{S(G, Y)\}$ for all $F, G \in \mathscr{F}$; the score is strictly proper if equality holds only if $F = G$; in particular, this implies faithfulness, meaning the forecaster couldn't improve his score by deviating from the data generating mechanism should he know the latter. By convention, smaller values of the scoring rule indicate better predictive distributions (the latter is said to be negatively oriented); the numerical value is only useful for comparisons between models.

Popular choices of proper scoring rule for continuous random variables are the log score, $S(G, y) = -\log\{g(y)\}$ where $g = \partial G(x)/\partial x$ is the density associated to G, and the continuous ranked probability score (CRPS),

$$S(G, y) = \int_{\mathbb{R}} [G(z) - \mathbf{1}\{y \geq z\}]^2 \, \mathrm{d}z.$$

For categorical outcomes with K categories, alternatives are the Brier score $\sum_{k=1}^{K} [G(z_k) - \mathbf{1}\{y = z_k\}]^2$ or the quantile score $2[\mathbf{1}\{y < \widehat{q}(\tau)\} - \tau]\{\widehat{q}(\tau) - \tau\}$ where $\widehat{q}(\tau)$ is the estimate of the τth quantile. The CRPS is the integral of the Brier scores for binary indicators of exceedance. It may seem natural to consider weighting scoring rules to focus on particular regions of interest of the forecast distribution, but this can result in improper scoring rules: for example, if $w(z) = \mathbf{1}\{z > r\}$, the score is not proper unless the predictive density is strictly positive (Lerch et al., 2017). Threshold-weighted versions exist, such as

$$S_{\text{twCRPS}}(G, y) = -\int_{\mathbb{R}} w(z) \log |G(z) - \mathbf{1}\{y > z\}| \mathrm{d}z.$$

Scores are normally evaluated on hold-out data, and expectations are replaced by averages. If the data generating mechanism is known, the ideal forecast can be computed but otherwise one resorts to climatological forecasts for benchmarking, where the latter are typically based on empirical distributions of the quantity of interest (Gneiting et al., 2007).

While calculation of scores is performed using point estimates in frequentist inference, the output of Bayesian inference procedures is usually draws from the posterior, from which posterior predictive samples can be obtained and there are alternative ways to compute scores. One could estimate the density of the posterior predictive draws and smooth these values, or

instead use the mixture-of-parameters distribution,

$$F_0(x \mid \boldsymbol{y}) = \int_\Theta F_p(x \mid \boldsymbol{\theta}) \mathrm{d}p(\boldsymbol{\theta} \mid \boldsymbol{y}),$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the posterior distribution and F_p the conditional predictive distribution. Krüger et al. (2017) recommend using the mixture-of-parameters distribution.

2.9 Summary

This chapter reviews functional extremes, with a focus on spatial parametric extreme value models. Under the assumption of regular variation, we obtain weak convergence of pointwise maxima or exceedances of some risk functional to a non-degenerate distribution on regions bounded away from cones. The point process representation unifies the different models under a common umbrella, connecting max-stable processes, conditional extremes and generalized ℓ-Pareto processes and extends the univariate convergence results presented in Chapter 1. The spectral representation of max-stable processes, given by the de Haan representation, allows us to derive the intensity function associated to Λ under the assumption of absolute continuity. The derivation of extremal functions proves a fruitful avenue for simulation algorithms and for likelihood based inference. We have also reviewed some commonly used diagnostic tools.

The main new contribution is the mev package (Belzile et al., 2019), in which most of the features listed in this chapter are implemented (Appendix D). Other original material includes derivations for the scaled extremal Dirichlet model (Examples 2.15 and 2.16), a comparison of the accept-reject methods for extremal Student (Remark 2.7), the remark about composition sampling (Algorithm 2.7) and the use of minimax exponential tilting algorithm (Section 3.2.2) for elliptical models and the use of Monte Carlo algorithms to compute the weights, the bounds for accept-reject algorithm for the generalized sum-Pareto risk functionals of Example 2.27, the alternative unconditional simulation algorithm for the spatial conditional extremes model in Algorithm 2.9, derivation of the gradient score in Examples 2.33 and 2.34 and the weighted nonparametric estimators of the angular measure for min and max risk functionals using empirical and Euclidean likelihood in Section 2.6.5.

Many of the ideas in this chapter have been incorporated in the mev package, including algorithms for simulating ℓ-Pareto processes, in conjunction with simulation algorithms for conditional Gaussian and Student vectors presented in Chapter 3. Semi-parametric modelling of low-dimensional extremes based on the pseudo-polar decomposition for arbitrary regions, using empirical likelihood methods is also implemented.

The assumption of regular variation is key for inference since it is the basis for extrapolation, but cannot easily be verified in practice. The conditional extremes model of Heffernan–Tawn is based on weaker assumptions than the ℓ-Pareto processes, but the lack of self-consistency has prevented its wider use. The spatial extension is very interesting, and the new formulation

based on mixtures could provide a theoretically-justified but tractable model. Because it is based on weaker assumptions and can model negative dependence, the need for inferential censoring is less prominent. Extensions of the methodology seem a fruitful avenue for future research.

The field of spatial extremes has been undergoing rapid development in recent years, yet much remains to be done. Max-stable processes, which were among the first models to be considered, are of limited practical interest because the events they describe are not often the ones of interest; counterexamples include invasive species outbreaks (Thibaud et al., 2016). Threshold models are gaining traction in the literature because they can describe single events, but inference remains challenging because observations are often far from being extreme in every component. Few parametric models exist for extremes; we have focused on two families that are derived from elliptical model, namely the Brown–Resnick and extremal Student models. Extensions that account for skewness exist (Beranger et al., 2017), but parameter estimation for such models is challenging. The main computational bottleneck in all cases is due to censoring, and workarounds are needed should one wish to truly tackle high-dimensional datasets like those encountered in other areas of spatial statistics. Pairwise composite likelihoods are among the most widely used tools for inference, because of the high computational cost associated with censored likelihood estimation. Development of alternative methods, such as the gradient score, require preliminary marginal transformation and further work is needed to obtain proper uncertainty quantification, often through a bootstrap. This problem is not unique to gradient scoring, as most estimation methods rely on multi-stage approaches. In the next chapter, we attempt to quantify this uncertainty from a likelihood-based perspective, looking at both censored and uncensored likelihood estimation for max-Pareto processes within Bayesian hierarchical models.

3 Bayesian hierarchical modelling of generalized ℓ-Pareto processes

Multidimensional inference for exceedances has only recently become prominent in the literature (de Fondeville and Davison, 2018; Rootzén et al., 2018) and few papers concern applications. Kiriliouk et al. (2019) fit a multivariate generalized Pareto model with structured components to model rainfall accumulation to study landslides, whereas de Fondeville (2018) use a two-step approach to fit a spatio-temporal model to storms over Europe. The class of generalized ℓ-Pareto processes comprises the only non-degenerate limits for functional threshold exceedances under the assumption of generalized regular variation. Their form is simpler than that of the corresponding max-stable processes, but their use is less widespread, notably due to the lack of closed-form expressions for the marginal distributions, which precludes semi-parametric inference based on copulas. The marginal distribution of a generalized ℓ-Pareto process can be expressed in terms of the measure of risk sets, which must be evaluated numerically using the limiting intensity measure of the point process and generalized ℓ-Pareto processes are threshold stable.

Due to the lack of ordering of extremes in dimension $D > 2$ (Barnett, 1976), there is no unique definition of what constitutes an extreme. We define exceedances as observations that are large in the sense that a scalar functional exceeds a high threshold. This means that some of its components may be small (or even zero in the case of rainfall), leading to a discrepancy between the asymptotic model for threshold exceedances and the observed realizations. A common way to resolve this is to marginally censor any observation falling below a threshold, as the limiting generalized Pareto process distribution is a poor approximation to these points. The drawback of this approach is that the likelihood contribution of the censored components often involves calculation of high-dimensional integrals of elliptical distributions. We review the most commonly used Monte Carlo method for estimating these integrals, namely separation of variables, and compare it with an alternative based on minimax exponential tilting. The latter is more efficient for rare event estimation, but requires solving a linear program; simulations indicate that for one-tailed regions, minimax exponential tilting is not competitive. The algorithm is however useful for conditional simulations, which arise in data augmentation schemes.

For threshold exceedances, the main computational bottleneck is due to censoring. Since

most spatial models are based on elliptical distributions, difficulties in calculating high-dimensional elliptical distribution functions hinder large-scale applications and Bayesian models. Moreover, the spatial extent of some natural phenomena, such as rainfall episodes, can become very localized as the intensity of the event increases; use of ℓ-Pareto processes has been criticized because the models are threshold-stable and cannot accommodate this behaviour (Tawn et al., 2018), even if they can capture independence at large distances. Proposals based on scale mixture models have been used to accommodate both regimes (Yuen and Guttorp, 2014; Huser et al., 2017; Huser and Wadsworth, 2019); the tail behaviour of such models can be derived using Breiman's lemma. There are drawbacks to using scale mixture models: the dependence structure must be the same everywhere, inference for marginal parameters is decoupled from the dependence structure (using copulas or estimating marginal parameters in a preliminary step), and use of scale mixtures requires evaluation of high-dimensional integrals to marginalize over the random scale in the presence of censoring, which limits their applicability. In the Bayesian framework, the scale parameter can be imputed with the censored observations, but the percentage of censoring in extreme value problems is higher than in classical settings, since most observations have components that fall below their marginal thresholds. Imputation of censored observations, conditional on current parameter values, is possible in elliptical models but leads to slow mixing for Markov chain Monte Carlo methods because it affects the curvature of the likelihood.

With a generalized ℓ-Pareto process, different scale parameters can lead to very few exceedances site-wise, so having a marginal threshold equal to the functional threshold leads to few, or even no, marginal threshold exceedances. Since there are potentially $2D+1$ marginal parameters to estimate assuming a common shape parameter, designing efficient proposals for the location and scale parameters is tricky in the absence of information about the gradient of the log-likelihood with respect to the parameters. Another difficulty is the support constraints, which are problematic if $\xi < 0$. Recall the stochastic representation of the generalized ℓ-Pareto process given in Equation (2.26),

$$\mathsf{Z} = \boldsymbol{\tau}\frac{\mathsf{X}^{\xi}-1}{\xi} + \boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is a vector of location parameters, $\boldsymbol{\tau}$ is a vector of scale parameters and ξ is the shape parameter. If we consider the shifted process $\mathsf{Z}-\boldsymbol{u}$ and censor observations falling below zero, the marginal scale parameters now correspond to $\boldsymbol{\sigma} = \boldsymbol{\tau} + \xi(\boldsymbol{u}-\boldsymbol{\eta})$ and coincide with those of the generalized Pareto distributions for marginal exceedances above $\boldsymbol{u}$.

The generalized ℓ-Pareto process arises from exceedances of $\{\boldsymbol{X} - \ell(\boldsymbol{b}_n)\}/\ell(\boldsymbol{a}_n)$, so each site has different marginal location and scale parameters asymptotically. For generalized max-Pareto processes with a common shape parameter, an homogeneous risk functional such as $\ell(\mathsf{X}) = \max_{j=1}^{D} X(\boldsymbol{s}_j)$ leads to modelling the process $\mathsf{X} - \mathsf{b}_n$. The parameter b_n is a threshold function, and we then set the functional threshold to zero and $\boldsymbol{\eta} = \mathbf{0}_D$; this avoids situations in which most exceedances occur at a single site, because the location parameter is much larger.

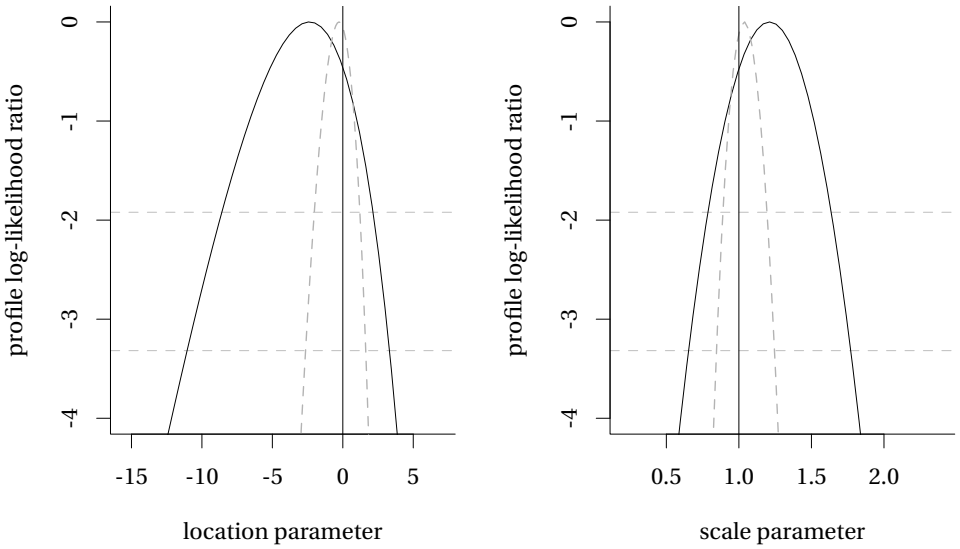

Figure 3.1 – *Profile log-likelihood ratio for the location (left) and scale (right) parameters for a sample of size $n = 200$ from a $D = 20$ dimensional logistic max-Pareto process, with $\boldsymbol{\eta} = \mathbf{0}_D, \boldsymbol{\tau} = \mathbf{1}_D$, $\xi = 0.1\mathbf{1}_D$, $\alpha = 0.7$. Full likelihood (dashed grey line) and censored likelihood (solid), inferentially censoring observations falling below their marginal 20 percentile (black).*

The location parameter (effectively a marginal threshold) is fixed at a pointwise quantile of the process, much like in the univariate setting. This has two consequences: its uncertainty is ignored and interpolation will be necessary to extrapolate at other sites where the quantile level is not observed. While this may seem unsatisfactory, it also remove the identifiability constraint discussed in Remark 2.8, simplifies the support constraints and reduces the number of parameters by at least $D-1$. Moreover, it is unclear whether the location parameter could be estimated solely based on the censored observations: Figure 3.1 shows the profile likelihoods for η and τ for a sample of size $n = 200$ in dimension $D = 20$ with a common location and scale parameter for exceedances over $u = 40$ for a logistic max-Pareto process model. The profile likelihood ratio for the location parameter shows that the range of potential values is quite broad even when observations are censored marginally at their 20th percentile; in contrast, the full likelihood for the uncensored data is much more peaked.

We propose a Bayesian hierarchical model formulation to model threshold exceedances based on generalized ℓ-Pareto processes. Two-stage estimation for the parameters of the latter is customary, yet ignores the marginal uncertainty. The study illustrates challenges related to inferential censoring and demonstrates that, while data augmentation schemes could be envisioned for commonly used parametric models, they are unlikely to succeed given the high proportion of non-extreme points. Pseudo-marginal Metropolis–Hastings algorithms are used to speed up computations and make inference feasible. The method is illustrated on rainfall extremes in Switzerland: our data application shows evidence of model misspecification, which appears when performing joint estimation of marginal and dependence parameters,

despite flexible models being employed for the latter. We compare the added uncertainty associated with marginal parameters with results from two-stage estimation and report results from simulation studies with a logistic model.

Our aim is threefold. First, we compare the efficiency of different methods for estimating the likelihood. Second, we look at the loss of information resulting from censoring and the relative efficiency of the censored and full likelihoods for multivariate generalized Pareto distributions. Third, we construct a joint model using the pseudo-marginal method to estimate all parameters simultaneously and compare this approach with the two-step method used by Thibaud (2014).

3.1 Basics of Markov chain Monte Carlo methods

We begin with a review of Markov chain Monte Carlo methods. Most of the notions presented are standard and can be found in Gelman et al. (2013), Robert and Casella (2005) and Brooks et al. (2011). In Bayesian statistics, the target of inference is the posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}, \tag{3.1}$$

where $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood function and $p(\boldsymbol{\theta})$ the prior; we denote the unnormalized posterior by $\pi(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. The normalizing constant of the posterior distribution, the denominator on the right-hand side of eq. (3.1), is typically intractable, and numerical schemes that circumvent its calculation are needed. Markov chain Monte Carlo (MCMC) methods proceed by simulating states from a Markov chain whose stationary distribution is the posterior distribution of interest. New values for the parameters of the Markov chain, termed proposals, are generated according to a transition kernel; the decision rule for accepting a proposal does not depend on the (unknown) normalizing constant, making these schemes attractive. The samples from the posterior, which are autocorrelated, can be used to compute expectations of functionals of interest.

The simplest approach to simulating Markov chain proposals is the Metropolis–Hastings algorithm (cf. Robert and Casella, 2005, Chapter 7). Starting from a current value of the Markov chain $\boldsymbol{\theta}^{(b)}$, a proposal $\boldsymbol{\theta}^*$ is drawn from a transition kernel $q\big(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b)}\big)$ chosen to ensure that the Markov chain is irreducible, aperiodic and reversible. A proposal is accepted with probability $\min(1, \alpha)$, where

$$\alpha = \frac{\pi\big(\boldsymbol{\theta}^*\big)}{\pi\big(\boldsymbol{\theta}^{(b)}\big)} \frac{q\big(\boldsymbol{\theta}^{(b)}; \boldsymbol{\theta}^*\big)}{q\big(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b)}\big)},$$

so proposed states that lead to higher posterior densities are systematically accepted. If the proposal is rejected, we set $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^{(b)}$ and otherwise $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^*$. The most common choices of kernel are symmetric kernels, such as $q\big(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b)}\big) \sim \mathsf{No}(\boldsymbol{\theta}^{(b)}, \boldsymbol{\Sigma})$, in which case the kernel density ratio cancels from the acceptance ratio α and the algorithm is termed a Metropolis

algorithm. An alternative is to use state-dependent proposals: if we sample the proposal from a kernel centred at $\boldsymbol{\theta}^{(b)}$, the algorithm accomplishes a random walk that allows the chain to explore the state space. One could also consider a proposal based on a discretization of the Langevin diffusion equation: the Metropolis adjusted Langevin algorithm (MALA) kernel uses information about the gradient of the unnormalized log-posterior $\log\{\pi(\boldsymbol{\theta})\}$, using (Brooks et al., 2011, p. 99)

$$q\left(\boldsymbol{\theta}^*;\boldsymbol{\theta}^{(b)}\right) \sim \mathsf{No}\left(\boldsymbol{\theta}^{(b)} + 0.5\boldsymbol{\Sigma}\nabla_{\boldsymbol{\theta}}\log\{\pi(\boldsymbol{\theta})\}, \boldsymbol{\Sigma}\right). \tag{3.2}$$

A drawback of this kernel is that the curvature of the proposal does not adapt to the state and that it requires analytical knowledge of the gradient. The choice of the covariance matrix (often the identity or an empirical estimate of the covariance matrix obtained from the history of the chain) is open.

Suppose we can split the state vector $\boldsymbol{\theta}$ into components θ_i and $\boldsymbol{\theta}_{-i}$, and furthermore assume that one can simulate from the conditional distribution $\pi(\theta_i \mid \boldsymbol{\theta}_{-i})$. The Gibbs sampler takes the transition kernel $q(\theta_i^*;\theta_i^{(b)}) = \pi(\theta_i^* \mid \boldsymbol{\theta}_{-i})$, in which case the Metropolis–Hastings acceptance ratio at step b is

$$\alpha = \frac{\pi\left(\boldsymbol{\theta}_{-i}^{(b)}\right)\pi\left(\theta_i^* \mid \boldsymbol{\theta}_{-i}^{(b)}\right)}{\pi(\boldsymbol{\theta}_{-i}^{(b)})\pi(\theta_i^{(b)} \mid \boldsymbol{\theta}_{-i}^{(b)})} \frac{\pi(\theta_i^{(b)} \mid \boldsymbol{\theta}_{-i}^{(b)})}{\pi(\theta_i^* \mid \boldsymbol{\theta}_{-i}^{(b)})} = 1$$

and every proposal is accepted.

Gibbs sampling requires the user to be able to simulate from each conditional distribution $\pi(\theta_i \mid \boldsymbol{\theta}_{-i})$. Markov chain Monte Carlo algorithms in which components are updated one at a time or in block using Metropolis–Hastings and Gibbs sampling are termed Metropolis-within-Gibbs; compared to simultaneous updates, they often yield Markov chains with better mixing properties and higher acceptance rates, even if the correlation between posterior draws increases and exploration of the state space is slow.

The choice of the proposal covariance matrix $\boldsymbol{\Sigma}$ for the transition kernel of a D-dimensional parameter vector is crucial. The adaptive random walk Metropolis–Hastings algorithm uses as transition kernel $\mathsf{No}(\mathbf{0}_D, \boldsymbol{\Sigma})$, where the proposal is $\boldsymbol{\theta}_p \sim \mathsf{No}(\boldsymbol{\theta}_c, k\boldsymbol{\Sigma} + \varepsilon\mathbf{I})$ at each step; in the long run, adaptation decreases with iterations and, provided that the chain converges to its stationary distribution, the empirical covariance $\boldsymbol{\Sigma} = \mathsf{Cov}\left(\{\boldsymbol{\theta}^{(b)}\}_{b=1}^{B}\right)$ with ridge penalty $\varepsilon\mathbf{I}$ added is a sensible choice (Haario et al., 2001). The proposal covariance matrix $\boldsymbol{\Sigma}$ is typically multiplied by the scaling factor $k = 2.38^2/D$, which is optimal for certain classes of Gaussian models. One problem with using the empirical covariance in high dimensions is that adaptation is slow, so mixing is often poor because the chain must make small steps. The burn-in phase consists of an initial stretch of the iterations that is not used for inference, but rather to allow convergence to the stationary distribution.

If parameters are constrained, any proposal outside the admissible region would be rejected

Algorithm 3.1 Metropolis–Hastings algorithm

Initialize $\boldsymbol{\theta}^{(0)}$
for $b = 1, \ldots, B$ **do**
  simulate $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b-1)})$
  compute $\alpha = \pi(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^{(b-1)}; \boldsymbol{\theta}^*) \{\pi(\boldsymbol{\theta}^{(b-1)}) q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b-1)})\}^{-1}$
  sample $U \sim \mathsf{U}(0,1)$
  if $\log(U) < \log(\alpha)$ **then**
    set $\boldsymbol{\theta}^{(b)} \leftarrow \boldsymbol{\theta}^*$
  else
    set $\boldsymbol{\theta}^{(b)} \leftarrow \boldsymbol{\theta}^{(b-1)}$
return $\{\boldsymbol{\theta}^{(b)}\}_{b=1}^{B}$

by the sampler. This is however inefficient; one could instead sample the parameters $\boldsymbol{\theta}^{(b)}$ on an unconstrained space and include a Jacobian for the transformation or use truncated proposals for marginal constraints.

Rue and Held (2005) suggest expanding the (conditional) log-posterior $\ell(\theta)$ for single parameters in a second-order Taylor series around θ_0 and taking the transition kernel for the proposal in a Metropolis–Hastings algorithm to be

$$q(\theta^*; \theta_0) \sim \mathsf{No}\left(\theta_0 - \omega \frac{\ell'(\theta_0)}{\ell''(\theta_0)}, -\frac{1}{\ell''(\theta_0)}\right), \qquad \omega \leq 1, \tag{3.3}$$

provided that the Hessian $\ell''(\theta)$ is negative definite at θ_0. The mean of the transition kernel of the proposal θ^* corresponds to a Newton step, so the mean should converge to the mode of the posterior if the objective function is roughly quadratic. The Metropolis ratio includes the transition kernel $q(\theta^*; \theta_0) \sim \mathsf{No}\big(\theta_0 - \omega\ell'(\theta_0)/\ell''(\theta_0), -1/\ell''(\theta_0)\big)$, so it also requires a Laplace approximation of the (unnormalized) conditional log-posterior density function at the proposal, viz. $q(\theta_0 \mid \theta^*) \sim \mathsf{No}\big(\theta^* - \omega\ell'(\theta^*)/\ell''(\theta^*), -1/\ell''(\theta^*)\big)$.

3.1.1 Data augmentation and pseudo-marginal methods

In many problems, the likelihood $p(\boldsymbol{y}; \boldsymbol{\theta})$ is intractable or costly to evaluate and auxiliary variables are introduced to simplify calculations, as in the expectation-maximization algorithm. The Bayesian analog is data augmentation (cf. Tanner and Wong, 1987), which we present succinctly: let $\boldsymbol{\theta} \in \Theta$ be a vector of parameters and consider auxiliary variables $\boldsymbol{u} \in \mathbb{R}^k$ such that $\int_{\mathbb{R}^k} p(\boldsymbol{u}, \boldsymbol{\theta}; \boldsymbol{y}) \mathrm{d}\boldsymbol{u} = p(\boldsymbol{\theta}; \boldsymbol{y})$, i.e., the marginal distribution is that of interest, but evaluation of $p(\boldsymbol{u}, \boldsymbol{\theta}; \boldsymbol{y})$ is cheaper. The data augmentation algorithm consists in running a Markov chain on the augmented state space $(\Theta, \mathbb{R}^k)$, simulating in turn from the conditionals $p(\boldsymbol{u}; \boldsymbol{\theta}, \boldsymbol{y})$ and $p(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y})$ with new variables chosen to simplify the likelihood. If simulation from the conditionals is straightforward, we can use data augmentation to speed up calculations or improve mixing. For more details, see Chapter 10 of Brooks et al. (2011).

Extreme value models for extremes are doubly intractable: the normalizing constant for the posterior is unknown, but the likelihood derived using the point process often involves the exponent measure $V(\boldsymbol{u})$ or the measure of the risk region $\Lambda(\mathscr{A}_u)$ which must be estimated using Monte Carlo.

Andrieu and Roberts (2009) show that, when only an unbiased estimator of the likelihood is available, a Markov chain Monte Carlo algorithm can nevertheless be constructed that targets the posterior of interest; this approach is termed pseudo-marginal and can be viewed as a data augmentation scheme: let $\boldsymbol{\theta}$ denote the parameter vector, let $p(\boldsymbol{u})$ denote the marginal distribution of the auxiliary variables and let $\widehat{p}(\boldsymbol{\theta} \mid \boldsymbol{u}; \boldsymbol{y})$ stand for the (Monte Carlo) likelihood estimator. The acceptance probability in the Metropolis–Hastings algorithm with transition kernel $q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(b-1)})p(\boldsymbol{u})$ from $\boldsymbol{\theta}^{(b-1)}$ is $\min\{1, \alpha\}$, where α is the posterior ratio

$$\begin{aligned}\alpha &= \frac{\widehat{p}(\boldsymbol{\theta}^* \mid \boldsymbol{u}^*; \boldsymbol{y})p(\boldsymbol{u}^*)p(\boldsymbol{\theta}^*)}{\widehat{p}(\boldsymbol{\theta}^{(b-1)} \mid \boldsymbol{u}^{(b-1);\boldsymbol{y}})p(\boldsymbol{\theta}^{(b-1)})p(\boldsymbol{u}^{(b-1)})} \frac{q(\boldsymbol{\theta}^{(b-1)}; \boldsymbol{\theta}^*)p(\boldsymbol{u}^{(b-1)})}{q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b-1)})p(\boldsymbol{u}^*)} \\ &= \frac{\widehat{p}(\boldsymbol{\theta}^* \mid \boldsymbol{u}^*; \boldsymbol{y})p(\boldsymbol{\theta}^*)}{\widehat{p}(\boldsymbol{\theta}^{(b-1)} \mid \boldsymbol{u}^{(b-1);\boldsymbol{y}})p(\boldsymbol{\theta}^{(b-1)})} \frac{q(\boldsymbol{\theta}^{(b-1)}; \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^{(b-1)})};\end{aligned}$$

if the proposal is accepted, the new value for the chain is $\{\boldsymbol{\theta}^*, \widehat{p}(\boldsymbol{\theta}^* \mid \boldsymbol{u}^*; \boldsymbol{y})\}$, so the value of the estimated likelihood function is stored from one iteration to the next. If $\int_{\boldsymbol{u}} p(\boldsymbol{\theta}^* \mid \boldsymbol{u}; \boldsymbol{y})p(\boldsymbol{u})\mathrm{d}\boldsymbol{u} = p(\boldsymbol{\theta}^*; \boldsymbol{y})$, i.e., the marginal distribution after integrating out the auxiliary variables is the target of interest, the pseudo-marginal Markov chain Monte Carlo will produce draws from the correct stationary distribution asymptotically as $b \to \infty$. This happens if the estimator of the likelihood is unbiased.

The Monte Carlo sample size, or the size of the auxiliary variable, determines the precision of the estimator $\widehat{p}$. If the latter is too large, the mixing of the Markov chain will be poorer, but the alternative is costly updates. Doucet et al. (2015) investigate the inefficiency of the scheme under additive Gaussian noise and suggest taking a standard deviation of about 1.68, but their derivations require stringent theoretical assumptions.

3.1.2 Hamiltonian Monte Carlo

In high dimensions, random walk Metropolis–Hastings is ineffective because the number of potential directions for the proposals to explore grows exponentially in the dimension, but the posterior mass is typically concentrated on a small region. The gradient of the posterior is not parametrization-invariant and thus, to efficiently explore the state-space, sampling algorithms must account for the geometry of the posterior $\pi(\boldsymbol{\theta})$.

Hamiltonian Monte Carlo, introduced in Duane et al. (1987), is an auxiliary variable method: a so-called momentum particle $\boldsymbol{u} \in \mathbb{R}^p$ is introduced and a Markov chain is run for $(\boldsymbol{\theta}, \boldsymbol{u})$ with joint density $p(\boldsymbol{\theta})p(\boldsymbol{u} \mid \boldsymbol{\theta}) \propto \exp\{-H(\boldsymbol{\theta}, \boldsymbol{u})\}$, where $H(\cdot, \cdot)$ is termed the Hamiltonian function. The deterministic evolution of the parameters over time is described by the partial derivatives

of H,

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \frac{\partial H}{\partial u_i}, \quad \frac{\mathrm{d}u_i}{\mathrm{d}t} = -\frac{\partial H}{\partial \theta_i}, \qquad i = 1, \ldots, p. \tag{3.4}$$

The Hamiltonian equations (3.4) leaves the Hamiltonian invariant (Neal, 2011, § 5.2.2.2), so only level sets $\{(\boldsymbol{\theta}, \boldsymbol{u}) : H(\boldsymbol{\theta}, \boldsymbol{u}) = H(\boldsymbol{\theta}^*, \boldsymbol{u}^*)\}$ are explored. Perturbations of the momentum at random times are introduced to resolve this.

The conditional density $p(\boldsymbol{u} \mid \boldsymbol{\theta})$ is arbitrary, but is usually taken to be Gaussian with covariance $\mathbf{M}$, often diagonal. There is no closed-form solution for the Hamiltonian equations, so numerical integration is used. Euler's method is inadequate because the trajectories it generates diverge from the exact solution, since error accumulates as the integration period increases (Neal, 2011, § 5.2.3.1). An efficient alternative in the context of Hamiltonian Monte Carlo is the Verlet integrator, or leapfrog method. With $p(\boldsymbol{u} \mid \boldsymbol{\theta}) \sim \mathsf{No}(\mathbf{0}, \mathbf{M})$, the Hamiltonian equations are $\mathrm{d}\boldsymbol{\theta}/\mathrm{d}t = \mathbf{M}^{-1}\boldsymbol{u}$ and $\mathrm{d}\boldsymbol{u}/\mathrm{d}t = \nabla p(\boldsymbol{\theta})$ and the leapfrog steps are

$$\begin{aligned} \boldsymbol{u}^{(t+\varepsilon/2)} &= \boldsymbol{u}^{(t)} + \varepsilon \nabla p(\boldsymbol{\theta}^{(t)})/2, \\ \boldsymbol{\theta}^{(t+\varepsilon/2)} &= \boldsymbol{\theta}^{(t)} + \varepsilon \mathbf{M}^{-1} \boldsymbol{u}^{(t+\varepsilon/2)}, \\ \boldsymbol{u}^{(t+\varepsilon)} &= \boldsymbol{u}^{(t+\varepsilon/2)} + \varepsilon \nabla p(\boldsymbol{\theta}^{(t+\varepsilon)})/2. \end{aligned} \tag{3.5}$$

The leapfrog method does L consecutive updates of size ε, starting and finishing with a half-update of the momentum using the new values simulated along the trajectory to compute the Hamiltonian equations and the first two steps of eq. (3.5) yield a Langevin move (3.2). Leapfrog steps are usually stable, but may also diverge in regions of high curvature; since they diverge quickly towards infinity, this behaviour can be easily diagnosed and can be used for troubleshooting. The final value $(\boldsymbol{\theta}^{(t+L\varepsilon)}, \boldsymbol{u}^{(t+L\varepsilon)})$ is accepted using a Metropolis–Hastings step that corrects for the approximation error. The divergence of the vector field defined by the Hamiltonian equations (3.4) is zero, so the Hamiltonian equations (3.4) are volume-preserving. This implies that the determinant of the Jacobian of the transformation is unity, so we do not need to account for changes in the Metropolis–Hastings step. Other key properties are symplecticness and reversibility of the chain, both of which are useful for deriving the validity of the algorithm. Other than the matrix $\mathbf{M}$, the parameters L and ε must be tuned to ensure good properties. The NUTS algorithm (Homan and Gelman, 2014) implemented in Carpenter et al. (2017) removes the need for manual tuning of these parameters. Neal (2011) and Betancourt (2017) provide comprehensive explanations of Hamiltonian Monte Carlo.

3.1.3 Monitoring convergence and measures of efficiency

An important (and difficult) question for practical Bayesian inference is to assess convergence of a Markov chain Monte Carlo algorithm. Despite theoretical guarantees of convergence to the stationary distribution, samplers may be slow, may fail to explore the space or else be too inefficient for the output to be usable. If we assume that the Markov chain has reached

stationarity, then we would prefer samplers with good mixing properties: too large a proposal variance will lead to sticky Markov chains and the sampler will have a low acceptance rate because of the inefficient proposals, whereas too small a variance leads to slow exploration of the space; this is a Goldilocks principle, where we want the variance to be just right, not too small nor too large. Performance indicators for the Markov chain can be monitored graphically, for example through trace plots and running mean plots.

Measures of efficiency of a sampler should quantify information available in the posterior draws and how well the sampler explores the target distribution. Under stationarity, the draws $\{\boldsymbol{\theta}^{(b)}\}$ are autocorrelated samples from π: based on a central limit theorem for Markov chains, an estimator of the variance of the average of some functional $g(\boldsymbol{\theta})$ is

$$\mathsf{Var}\left(B^{-1}\sum_{b=1}^{B} g(\boldsymbol{\theta}^{(b)})\right) \approx B^{-1}\mathsf{Var}_{\pi}\left(g(\boldsymbol{\theta})\right)\tau_g,$$

where τ_g, the integrated autocorrelation time, is

$$\tau_g = 1 + 2\sum_{i=1}^{\infty} \mathsf{Cor}\{g(\boldsymbol{\theta}^0), g(\boldsymbol{\theta}^{(i)})\}.$$

The default estimation method for τ_g is to fit an autoregressive model using the Yule–Walker equations to sample autocorrelation, with the order of the AR process selected via AIC. A correlogram can serve as a graphical diagnostic: geometric decay indicates that samples are approximately independent when the autocorrelation reaches zero.

A quantitative measure of the quality of the sample is the effective sample size, $\mathsf{ESS} = B/\widehat{\tau}_g$, where B is the number of MCMC runs and $\widehat{\tau}$ is the estimated integrated autocorrelation time (Rosenthal, 2011, p. 95). A better way to compare methods is computational efficiency, which is the effective sample size divided by the total computing time needed to obtain B samples.

The speed at which the Markov chain explores the state space gives another indication of quality, with faster mixing chains being preferred. The jumping distance is $(B-1)^{-1}\sum_{b=1}^{B-1}\|\boldsymbol{\theta}^{(b+1)} - \boldsymbol{\theta}^{(b)}\|^2$, where $\boldsymbol{\theta}^{(b)}$ is the value of the chain at iteration b. If the moves are not accepted, the values of the chain are repeated and, likewise, small moves lead to small jumps.

Another way to assess convergence of a MCMC sample is to run parallel chains from different starting values, much like for optimization routines where we want to ensure convergence to a global mode in cases with multimodal objective functions. While this approach is wasteful (in the sense that each chain must be run for sufficiently long to allow it to reach stationarity), it is nevertheless helpful for diagnostics of (lack of) convergence, as evidenced by different empirical summaries. The potential scale reduction factor of Gelman and Rubin (1992), denoted $\widehat{R}$, compares the variability within and between chains: if M parallel chains are run for B iterations and we denote by $\theta^{(b,m)}$ the bth iteration of the mth chain (cf. Gelman et al.,

2013, p.284), then

$$\bar{\theta}_m = B^{-1}\sum_{b=1}^{B}\theta^{(b,m)}, \quad \bar{\theta} = M^{-1}\sum_{m=1}^{M}\bar{\theta}_m,$$
$$S_{\mathrm{w}}^2 = M^{-1}\sum_{m=1}^{M}(B-1)^{-1}\sum_{b=1}^{B}(\theta^{(b,m)}-\bar{\theta}_m)^2, \qquad S_{\mathrm{b}}^2 = B(M-1)^{-1}\sum_{m=1}^{M}(\bar{\theta}_m-\bar{\theta})^2,$$

where $\bar{\theta}_m$ is the mean in chain m, $\bar{\theta}$ is the global mean, S_{w}^2 is the within-chain variance and and S_{b}^2 is the between-chains variance. A total variance σ^2 can be estimated consistently as $B\to\infty$ using $\widehat{\sigma}^2 = \{(B-1)S_{\mathrm{w}}^2 + S_{\mathrm{b}}^2\}/B$ and the potential scale reduction factor is $\widehat{R} = (\widehat{\sigma}^2/S_{\mathrm{w}}^2)^{1/2}$.

If the chains have all reached stationarity, the ratio of the estimated variance to the within-variance should be 1. The scale reduction factor requires finite variance, and fails to capture scenarios where the means differ while the variance is equal; Vehtari et al. (2019) suggest a more robust version using rankits in place of observations. A comparative review of diagnostics of convergence for Markov chain Monte Carlo algorithms can be found in Cowles and Carlin (1996).

3.1.4 Adaptive Markov chain Monte Carlo

The efficiency of the Metropolis–Hastings algorithm, or a variant thereof, depends largely on the choice of the covariance matrix of the proposal $\boldsymbol{\Sigma}$. This can be specified based on preliminary optimization as the negative inverse Hessian of posterior distribution, but optimization may not be straightforward in high-dimensional hierarchical models.

An alternative is to update the $\boldsymbol{\Sigma}$ on the fly during the burn-in period, using the history of the chain to select a covariance matrix that leads to a good acceptance rate and has lower asymptotic variance. Intuitively, if the chains have converged to their stationary distribution, then Monte Carlo estimators of the moments can be used to tune $\boldsymbol{\Sigma}$. Special care must be taken to ensure that the adapted chain converges to the posterior distribution, but also to avoid being stuck in local modes or regions of the state space with low posterior mass. Under the hypothesis of diminishing adaptation and containment (cf. Rosenthal, 2011, p.104), ergodicity is guaranteed. Diminishing adaptation can be obtained by any scheme in which adaptation reduces (for example if one uses the empirical covariance of the chain as proposal), or if we stop adapting. Comprehensive reviews can be found in Andrieu and Thoms (2008) and in Rosenthal (2011).

3.2 Numerical evaluation of elliptical distribution functions

Censored likelihoods for the extremal Student and the Brown–Resnick processes include the conditional intensities presented in Examples 2.6 and 2.7, which both contain high-

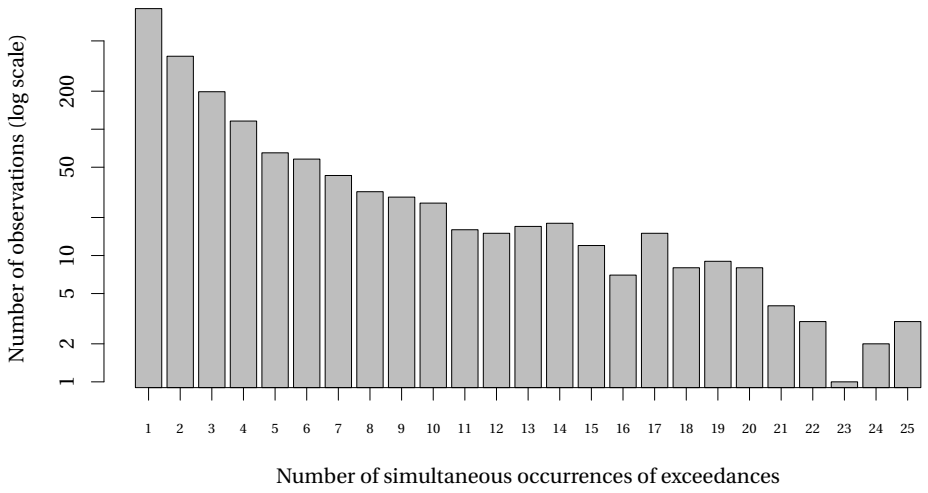

Figure 3.2 – *Number of non-censored components per observation for a random subsample of size 2000 of the Swiss rainfall data at 25 sites.*

dimensional distribution functions of Gaussian and Student-t vectors. The censoring pattern changes for each observation and the main computational bottleneck for estimation of the likelihood is the numerical evaluation of these distribution functions, hence the need to obtain efficient Monte Carlo estimators. For example, on the Zürich rainfall dataset with 42 sites and keeping 1142 exceedances, the average time for one evaluation of the censored likelihood is 53.5 seconds, of which 2.5 seconds is for the exponent measure. In high dimensions, this problem is exacerbated because more components are censored: Figure 3.2 shows the number of exceedances for 2000 observations at a subset of $D = 25$ sites from the Swiss rainfall data described in Section 3.6. The number of exceedances decreases roughly exponentially in the dimension, so most observations have only a few non-censored components.

Although the Monte Carlo estimator for $V(\boldsymbol{u})$ is unbiased, $V(\boldsymbol{u})^{-1}$ and $\exp\{-V(\boldsymbol{u})\}$ are biased and so is the likelihood estimator (de Fondeville and Davison, 2018). In such cases, the pseudo-marginal algorithm targets a biased version of the posterior distribution. Since in practice the bias is small and the precision can be increased, we disregard this. Thibaud et al. (2016) report numerical evidence indicating that the effect of the bias is negligible.

3.2.1 Separation of variables

The Gaussian distribution function must be evaluated numerically. Efficient algorithms based on Monte-Carlo methods are available for this purpose; we consider below only the case where the covariance matrix $\boldsymbol{\Sigma}$ is of full rank. Consider the integral

$$\Phi_D(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \int \mathbf{I}_{\boldsymbol{a} \leq \boldsymbol{x} \leq \boldsymbol{b}} \phi_D(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{x}.$$

We may assume without loss of generality that the distribution is centred, since $\Phi_D(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \Phi_D(\boldsymbol{a}-\boldsymbol{\mu}, \boldsymbol{b}-\boldsymbol{\mu}; \mathbf{0}_D, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky decomposition of the covariance matrix, with $\mathbf{L}$ lower triangular. Making the change of variables $\boldsymbol{x} \mapsto \mathbf{L}\boldsymbol{y}$, so that $\boldsymbol{Y} \sim \mathsf{No}_D(\mathbf{0}_D, \mathbf{I}_D)$, reduces the problem to calculation of $\mathsf{P}(\boldsymbol{a} \leq \mathbf{L}\boldsymbol{Y} \leq \boldsymbol{b})$. The triangular system of equations leads to a sequential decomposition of the region of integration,

$$\begin{aligned}
\widetilde{a}_i &:= L_{11}^{-1} a_1 \leq y_1 \leq L_{11}^{-1} b_1 =: \widetilde{b}_i, \\
\widetilde{a}_j &:= \frac{a_j - \sum_{i=1}^{j-1} L_{ji} y_i}{L_{jj}} \leq y_j \leq \frac{b_j - \sum_{i=1}^{j-1} L_{ji} y_i}{L_{jj}} =: \widetilde{b}_j, \qquad j = 2, \ldots, D.
\end{aligned}$$

If instead we start with a precision matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ and let $\mathbf{T}\mathbf{T}^\top = \mathbf{Q}$, with $\mathbf{T}$ lower triangular, we can represent the joint distribution of $\boldsymbol{X}$ using an inhomogeneous autoregressive process (Rue and Held, 2005, Theorem 2.7),

$$X_j \mid \boldsymbol{X}_{j+1:D} = \boldsymbol{x}_{j+1:D} \sim \mathsf{No}\left(-T_{jj}^{-1} \sum_{i=j+1}^{D} T_{ij} x_j, T_{jj}^{-2}\right).$$

Let $\eta_j = T_{jj}^{-1} \sum_{i=j+1}^{D} T_{ij} x_j / T_{ii}$; the Cholesky decomposition of the precision matrix leads to the backward triangular system of equations

$$T_{jj} a_j + \sum_{i=j+1}^{D} T_{ij}(x_i - \eta_i)/T_{ii} \leq x_j \leq T_{jj} b_j + \sum_{i=j+1}^{D} T_{ij}(x_i - \eta_i)/T_{ii}, \quad j = D-1, \ldots, 1. \tag{3.6}$$

and $T_{DD} a_D \leq x_D \leq b_D T_{DD}$, where $x_i \sim \mathsf{TNo}(a_i, b_i; -\eta_i/T_{ii}, T_{ii}^{-2})$, a truncated Gaussian variable, can be simulated by generating first $z_i \sim \mathsf{TNo}(T_{ii} a_i + \eta_i, T_{ii} b_i + \eta_i; 0, 1)$ then setting $x_i = (z_i - \eta_i)/T_{ii}$. If furthermore the process is Markovian, some of the entries T_{ij} will be zero and dedicated routines for sparse matrices can be used to accelerate the calculations.

We can represent a Student t variable as a Gaussian scale mixture, $\boldsymbol{Z} \overset{\mathrm{d}}{=} \nu^{1/2} \boldsymbol{X}/R \sim \mathsf{St}(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\nu > 0$ is the degrees of freedom parameter and $R^2 \sim \chi^2_\nu$ is an independent chi-square variable. This yields a decomposition equivalent to eq. (3.6), conditional on $R = r$, obtained by replacing a_i by $a_i r \nu^{-1/2}$ and b_i by $b_i r \nu^{-1/2}$.

Writing

$$\Phi_D(\boldsymbol{a}, \boldsymbol{b}, \mathbf{0}_{\boldsymbol{D}}, \boldsymbol{\Sigma}) = \int_0^\infty \int_{\widetilde{a}_i}^{\widetilde{b}_i} \cdots \int_{\widetilde{a}_D}^{\widetilde{b}_D} \phi_D(\mathbf{0}, \mathbf{I}_D) \mathrm{d}\boldsymbol{y} f_R(r) \mathrm{d}r$$

suggests that a sensible Monte Carlo estimator can be obtained by sequential importance sampling upon making a change of variable $y_i = \Phi^{-1}(u_i)$ (Genz and Bretz, 2009). Consider the importance sampling density

$$g_{\mathrm{sov}}(r, \boldsymbol{y}) = f_R(r) f(y_1 \mid r) \prod_{j=1}^{D} f(y_j \mid r, y_1, \ldots, y_{j-1}) = f_R(r) \prod_{j=1}^{D} \frac{\phi(y_j) \mathbf{1}\{\widetilde{a}_j \leq y_j \leq \widetilde{b}_j\}}{\Phi(\widetilde{b}_j) - \Phi(\widetilde{a}_j)}.$$

Algorithm 3.2 Accept-reject algorithm for truncated Student vectors (Botev and L'Écuyer, 2017, Alg. 2)

Require: Cholesky lower root $\boldsymbol{L}$, bounds $\boldsymbol{a}$, $\boldsymbol{b}$, $r^*, \boldsymbol{y}^*; \widehat{\boldsymbol{\mu}}, \widehat{\eta}$

repeat
  sample $R \sim \mathsf{TNo}(\widehat{\eta}, 1; 0, \infty)$;
  for $j = 1, \ldots, D$ **do**
    sample $X_i \sim \mathsf{TNo}(\widehat{\mu}_i, 1; \widetilde{a}_i, \widetilde{b}_i)$;
  sample independently $U \sim \mathsf{U}(0,1)$;
until $U < \exp\{\psi(R, \boldsymbol{X}; \widehat{\eta}, \widehat{\boldsymbol{\mu}}) - \psi(r^*, \boldsymbol{y}^*; \widehat{\boldsymbol{\mu}}, \widehat{\eta})\}$
return $\sqrt{\nu}\boldsymbol{X}/R$.

The Geweke–Hajivassiliou–Keane estimator of the integral, commonly referred to as the separation of variables (SOV) estimator, is $T^{-1}\sum_{t=1}^{T}\prod_{j=1}^{D}\{\Phi(\widetilde{b}_j^{(t)}) - \Phi(\widetilde{a}_j^{(t)})\}$. The separation of variables algorithm is described on p. 50 of Genz and Bretz (2009) using a randomized quasi-Monte Carlo procedure with points simulated from randomized and periodized Kronecker sequence of the form $\{K_N = i\mathbf{v} \mod 1, i = 1, \ldots, N\}$ with $\mathbf{v}$ taken to be the Richtmyer point set $v_i = p_i^{1/2}$, where p_i is the ith prime number.

3.2.2 Minimax exponential tilting

Botev (2017) and Botev and L'Écuyer (2017) propose an exponential tilted version of the sequential importance sampling density g_{sov},

$$g_{\text{met}}(r, \boldsymbol{y}; \eta, \boldsymbol{\mu}) = \frac{\phi(r-\eta)}{\Phi(\eta)} \prod_{j=1}^{D} \frac{\phi(y_j - \mu_j)\mathbf{1}\{\widetilde{a}_j \le y_j \le \widetilde{b}_j\}}{\Phi(\widetilde{b}_j - \mu_j) - \Phi(\widetilde{a}_j - \mu_j)},$$

i.e., conditional on the previous one, each variable is truncated Gaussian with mean μ_j and the radial parameter $R \sim \mathsf{TNo}(\eta, 1; 0, \infty)$, see Algorithm 3.2. The tilting parameters $(\boldsymbol{\mu}, \eta)$ are chosen to minimize the worst-case scenario, solving the saddlepoint program

$$\inf_{\eta, \boldsymbol{\mu}} \sup_{(r,\boldsymbol{y}): r\boldsymbol{a} \le \mathbf{L}\boldsymbol{y} \le r\boldsymbol{b}} \psi(r, \boldsymbol{y}; \eta, \boldsymbol{\mu}), \tag{3.7}$$

where $\psi(r, \boldsymbol{y}; \eta, \boldsymbol{\mu}) = \log\{f_R(r)\phi_D(\boldsymbol{y})/g_{\text{met}}(r, \boldsymbol{y}; \eta, \boldsymbol{\mu})\}$ is the log likelihood ratio of the truncated Gaussian density relative to the minimax exponential tilting proposal g_{met}. With this choice of importance sampling density, the optimal pair $\eta, \boldsymbol{\mu}$ can be found via a convex optimization program with $2D-1$ parameters. For Gaussian samples, $R = 1$ almost surely and a similar, albeit simplified, version of the likelihood ratio and of the convex program can be run (Botev, 2017, § 3.1 and Algorithm 2); Algorithm 3.2 is modified accordingly by removing R, ν and η.

The value of ψ evaluated at the solution of eq. (3.7), denoted $c = \psi(r^*, \boldsymbol{y}^*; \widehat{\boldsymbol{\mu}}, \widehat{\eta})$, provides a upper bound of the likelihood ratio that can be used to design a very efficient accept-reject algorithm in dimension $D \le 100$, based on sequential sampling from g_{met}.

For the Gaussian case, the minimax exponential tilting (MET) estimator has good theoretical guarantees for its variance in the rare-event scenario (Theorem 1 Botev, 2017). For the Student distribution, minimax exponential tilting is superior to separation of variables when the correlation is negative, but its benefit is less clear when the dependence is strong.

3.2.3 Variable reordering

Several heuristic strategies can be used to improve the precision or reduce the computing time of the integrals. Gibson et al. (1994) suggest a greedy reordering algorithm that aims to minimize the variance of the integral, i.e.,

$$\int_{\mathbf{0}_D}^{\mathbf{1}_D} \prod_{j=1}^{D} \left[\Phi\left\{\widetilde{b}_j(\boldsymbol{u}_{1:(j-1)}\right\} - \Phi\left\{\widetilde{a}_j(\boldsymbol{u}_{1:(j-1)})\right\}\right]^2 \mathrm{d}\boldsymbol{u} - \Phi_D(\boldsymbol{a}, \boldsymbol{b}, \mathbf{0}_{\boldsymbol{D}}, \boldsymbol{\Sigma})^2.$$

Variables are reordered so that the innermost variable (whose expectation is computed conditionally on all the other variables) is largest; in particular, any index for which $a_i = -\infty$, $b_i = \infty$ should be moved outermost and marginalized over. The algorithm builds the permutation vector $\boldsymbol{\pi}$ and the lower triangular Cholesky root $\mathbf{L}$ sequentially. There are two options for the objective function: the first is to minimize the average expectation in the innermost loop, as advocated by Gibson et al. (1994). The second is to minimize the average variance of the truncated Gaussian in the innermost loop, as advocated in Genz and Bretz (2009). These strategies are not available for precision matrices unless we invert them. If the precision is sparse, it is computationally advantageous to reorder the entries to induce sparsity in the Cholesky factor $\mathbf{T}$. For example, Bolin and Lindgren (2015) use the CAMD approximate minimum degree ordering of Amestoy et al. (1996); the reordering gives a two or three-fold increase in the speed of the calculations when the dimension of the precision matrix $D \approx 2500$. The variable reordering presented in Algorithm 3.3, in contrast, aims at minimizing variance for a fixed number of simulations.

An estimator of the variance of the (quasi) Monte-Carlo estimators can be obtained by running the procedure M times. If $I_{N,i}$ denotes the estimate from batch i obtained from N, then $I_M = M^{-1}\sum_{i=1}^{M} I_{N,i}$ with squared standard error estimator

$$\widehat{\sigma}_M^2 = \frac{1}{M(M-1)} \sum_{i=1}^{M} (I_{N,i} - I_M)^2.$$

Genz and Bretz (2009) recommend multiplying $\widehat{\sigma}_M$ by $\Phi^{-1}(0.99)$ to obtain a robust estimator. There is no theoretical justification of this ad-hoc inflation and simulation studies have shown that estimates of $\widehat{\sigma}_M^2$ can be very small even when the true error is large (Botev, 2017). The SOV estimator performs adequately for strongly correlated variables.

Example 3.1 (Comparison of algorithms for multivariate Gaussian distribution function) We consider the accuracy, bias, variable and the execution time of four numerical estimators of the multivariate Gaussian distribution function that are available in R . The first is from

Algorithm 3.3 Variable reordering (Genz and Bretz, 2009, § 4.1.3)

Require: lower bound $\boldsymbol{a}$, upper bound $\boldsymbol{b}$, covariance matrix $\boldsymbol{\Sigma}$.

1: **function** EXPECTATION OF $\mathsf{TNo}(0,1;a,b)$

$$\mathsf{E}(a,b) := \frac{\phi(a)-\phi(b)}{\Phi(b)-\Phi(a)}$$

2: **function** VARIANCE OF $\mathsf{TNo}(0,1;a,b)$

$$\mathsf{V}(a,b) := 1 + \frac{a\phi(a)-b\phi(b)}{\Phi(b)-\Phi(a)} - \left(\frac{\phi(a)-\phi(b)}{\Phi(b)-\Phi(a)}\right)^2$$

3: $\boldsymbol{a} \leftarrow \boldsymbol{a}\,\mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}$, $\boldsymbol{b} \leftarrow \boldsymbol{b}\,\mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}$
4: $\boldsymbol{\pi} \leftarrow 1:D$
5: $m = \min_{i\in 1:D} \mathsf{V}(a_i,b_i)$ or $m = \min_{i\in 1:D}\{\Phi(b_i)-\Phi(a_i)\}$
6: $\mathrm{swap}(\pi_m,\pi_1)$
7: $\mathbf{L}_{1,\cdot} \leftarrow \boldsymbol{\Sigma}_{\cdot,\pi_1}\Sigma_{\pi_1\pi_1}^{-1/2}$
8: $\mathrm{swap}(\mathbf{L}_{1,\cdot},\mathbf{L}_{\pi_1,\cdot})$
9: $\mathrm{L}_{1,1} = \Sigma_{\pi_1,\pi_1}^{1/2}$
10: **for** $j \in 2,\ldots,D$ **do**
11: $\quad \mu_{j-1} \leftarrow \mathsf{E}(\widetilde{a}_{\pi_{j-1}},\widetilde{b}_{\pi_{j-1}})$
12: $\quad$ **for** $k \in j+1,\ldots,D$ **do**
13: $\quad\quad u_k \leftarrow (\Sigma_{\pi_k,\pi_k} - \mathbf{L}_{k,1:j}\mathbf{L}_{k,1:j}^\top)^{-1}(b_{\pi_k} - \mathbf{L}_{k,1:k}^\top\boldsymbol{\mu}_{1:j})$
14: $\quad\quad l_k \leftarrow (\Sigma_{\pi_k,\pi_k} - \mathbf{L}_{k,1:j}\mathbf{L}_{k,1:j}^\top)^{-1}(a_{\pi_k} - \mathbf{L}_{k,1:k}^\top\boldsymbol{\mu}_{1:j})$
15: $\quad m = \min_{k\in(j+1):D} \mathsf{V}(a_k,b_k)$ or $m = \min_{k\in(j+1):D}\{\Phi(b_k)-\Phi(a_k)\}$
16: $\quad \mathrm{L}_{j,j} = \Sigma_{\pi_m,\pi_m} - \mathbf{L}_{j,1:(j-1)}\mathbf{L}_{j,1:(j-1)}^\top$
17: $\quad$ **for** $k \in j+1,\ldots,D$ **do**
18: $\quad\quad \mathrm{L}_{k,j} = \mathrm{L}_{j,j}^{-1}\left(\Sigma_{\pi_k,\pi_m} - \mathbf{L}_{j,1:(j-1)}\mathbf{L}_{\pi_k,1:(j-1)}^\top\right)$

the mvtnorm package and relies on Fortran routines and quasi Monte Carlo integration with a randomized Korobov lattice rule $\boldsymbol{v}(h) = \{h^i \mod p\}_{i=0}^{k-1}$, for a prime p, given value $1 \leq h \leq \lfloor p/2 \rfloor$ and k an integer that depends on the prime p (Genz and Bretz, 2009, p. 47) with scrambling achieved through a baker transform and antithetic variables. The function is hardcoded to use 25000 variables with an absolute tolerance of 10^{-3}. The second, coded in C++ and provided in the mvPot package, also uses the separation of variables estimator with quasi Monte Carlo integration, but the lattice rule is different (Nuyens and Cools, 2006). The last two procedures, from the TruncatedNormal package, are implemented in pure R and are based on the minimax exponential tilting algorithms of Botev and L'Ecuyer (2015), using quasi Monte Carlo (with a digital net using Sobol sequences) or Monte Carlo.

We consider a Gaussian process with power variogram $\gamma(\boldsymbol{h}) = (\|h\|_2/2)^{3/2}$ with $D = 64$ sites defined on a regular grid, $\{1,\ldots,8\}^2$. We condition on a site, say the kth component, to get a mean vector $\boldsymbol{\gamma}(\boldsymbol{s}_i - \boldsymbol{s}_k)$ and a covariance matrix $\boldsymbol{\Sigma}^{(k)}$ with entries $\Sigma_{ij}^{(k)} = \gamma_{i,k} + \gamma_{j,k} - \gamma_{i,j}$ for $i,j \in \{1,\ldots,D\} \setminus \{k\}$. We assess the performance of minimax exponential tilting versus that of separa-

D	10	15	20	30	50	100	200	400
mvtnorm	0.98	1.36	1.22	1.16	1.32	1.48	1.58	1.64
MET (QMC)	2.40	2.68	2.54	2.72	3.08	3.44	3.78	4.20
MET (MC)	1.82	1.92	1.84	1.98	2.36	3.08	3.66	4.40

Table 3.1 – *Execution time for an evaluation of the Gaussian distribution function Φ_D relative to the execution time of mvPot routine for varying D.*

tion of variables by computing the 63-dimensional integral $\Phi_{D-1}\{\mathbf{0}_{D-1}, \boldsymbol{\Sigma}^{(k)}; -\infty_{D-1}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\}$. We consider numbers of replicates from 500 to 5000 in increments of 500, rounded down to the nearest prime number, because the method in mvPot requires prime numbers; the point estimate is the average of 12 replicates, on which the standard error of the mean is based. We consider the relative estimator error associated to the 63-dimensional integration as a function of the number of Monte Carlo samples used to approximate the integral.

The relative error for each procedure is reported in Figure 3.3; the entries labelled MET refer to the functions in the TruncatedNormal package. In this low-dimensional context, the quasi Monte Carlo methods have a relative error of roughly a third of the Monte Carlo minimax exponential tilting estimator. The estimated standard error for the latter is extremely small, even when the estimated probabilities have a higher relative error, which decays at rate $\mathrm{O}(n^{-1/2})$. The theoretical upper bound for the error of the sample randomized quasi Monte Carlo mean estimator is $\mathrm{O}\{\log(n)^D/n\}$, but in practice the average appears to be much lower and quasi Monte Carlo estimation is competitive for moderate dimensions.

We also report the computing time relative to that of the function in mvPot in Table 3.1. The minimax exponential tilting is slower by a factor of about three relative to the mvtnorm implementation because it is programmed in R and involves solving a preliminary convex optimization problem (eq. (3.7)).

We consider next the estimation of $\mathsf{P}(\boldsymbol{X} \leq \mathbf{0}_D)$ for $\boldsymbol{X} \sim \mathsf{No}_D(\mathbf{0}_D, 0.5\,\mathrm{diag}(\mathbf{1}_D) + 0.5\mathbf{1}_D\mathbf{1}_D^\top)$; this has value $1/(D+1)$. The minimax exponential tilting estimator offers theoretical guarantees on the upper bound of the error, however, and is more accurate in high dimensions, as it is designed to deal with rare event estimation. The relative bias of the various estimators, displayed in Figure 3.4, shows that the separation of variables relative bias increases with the dimension.

□

The integrals we solve when computing the exponent measure of ℓ-Pareto processes are right-truncated and separation of variables remains competitive for the dimensions under consideration, i.e., $D \approx 20$. Moreover, the current implementation of minimax exponential tilting is not competitive, as it is coded on R code, but could be as fast if coded in C++.

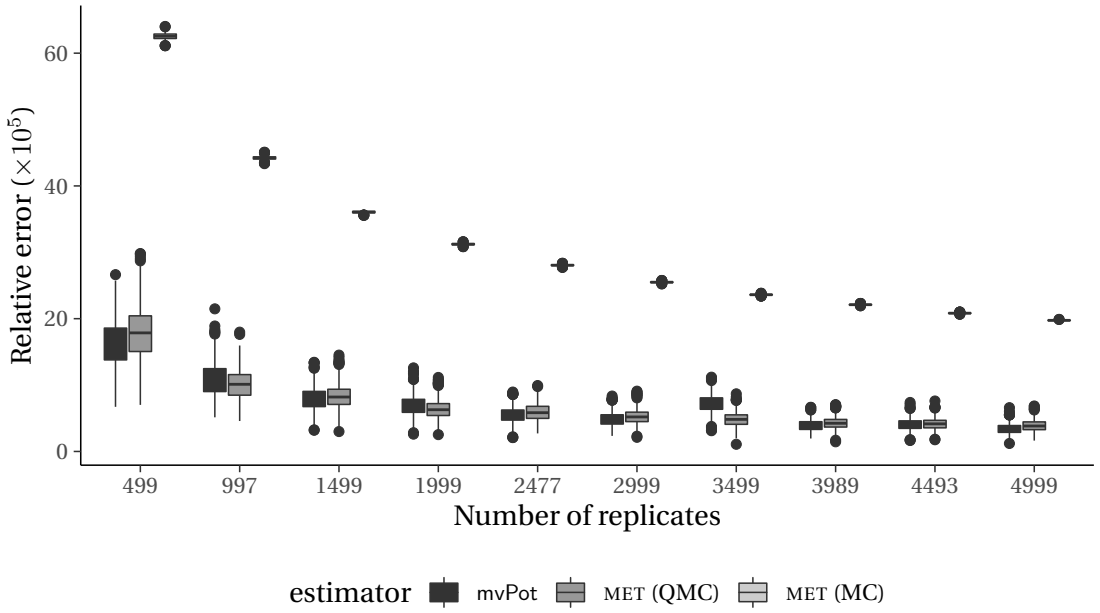

Figure 3.3 – *Relative error of (quasi) Monte Carlo estimators as a function of the number of replicates. Separation of variables estimator from the mvPot package (left) and minimax exponential tilting quasi Monte-Carlo (middle) and Monte Carlo (right) estimators from the TruncatedNormal package, based on 1000 replications. The relative error of the Monte Carlo minimax exponential tilting estimator is more than three times that of the two quasi Monte Carlo estimators and its uncertainty decreases at rate $n^{-1/2}$.*

3.3 Penultimate models for extremes based on scale mixtures

Pareto processes are not the only models that have been considered in recent years for modelling threshold exceedances: Gaussian scale mixtures have also gained traction. Elliptical random fields admit the stochastic representation $R\mathsf{W}$, where $R \sim F_R$ is a positive radial variable and W, independent of R, is a Gaussian process with zero mean. If the tail index of R is $\xi \leq 0$, then the limiting dependence model for threshold exceedances will be independence, and otherwise $R\mathsf{W}$ will be asymptotically dependent by Breiman's lemma (Lemma 2.17). Motivated by the desire to obtain models that can accommodate both asymptotic dependence and independence, Huser et al. (2017) considered Gaussian scale mixtures. Earlier attempts include Yuen and Guttorp (2014), who considered Gaussian scale mixture model with a generalized Pareto radius and Opitz (2016), who considered Laplace random fields to model asymptotically independence. Choosing a radial variable different from Pareto may allow more flexibility, but in the end models are either asymptotically dependent or asymptotically independent for a given F_R.

Consider $t = 1, \ldots, T$ independent temporal replications of a spatial process at D sites $\{\boldsymbol{s}_j\}_{j=1}^{D}$ at which at least one component exceeds a predetermined threshold. We label the observations on the original scale y_{tj}, for $j = 1, \ldots, D$. To each site, we associate a marginal threshold u_j

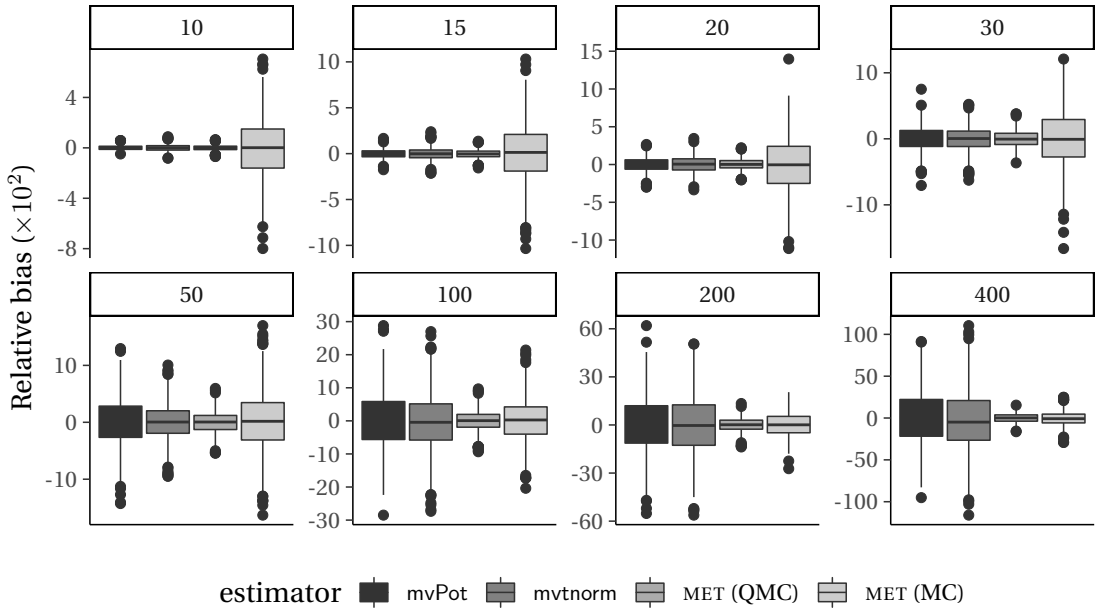

Figure 3.4 – *Box-and-whiskers plots of the relative bias of (quasi) Monte Carlo estimators as a function of the dimension of the integral, D, for the integration problem* $\mathsf{P}(\boldsymbol{X} \leq \mathbf{0}_D)$ *with* $\boldsymbol{X} \sim \mathsf{No}_D(\mathbf{0}_D, 0.5\,\mathrm{diag}(\mathbf{1}_D) + 0.5\mathbf{1}_D\mathbf{1}_D^\top)$*, based on 1000 replications. The MET estimators refer to minimax exponential tilting implemented in the TruncatedNormal package.*

above which observations are extreme and let the marginal probability of exceedance of this threshold be ζ_{u_j} Let $\mathbb{A}_t = \{j \in 1,\ldots,D : y_{tj} > u_j\}$ and $\mathbb{U}_t = \{j \in 1,\ldots,D : y_{tj} \leq u_j\}$ denote the set of observations above and under the threshold, respectively. The $N-T$ observation vectors $\boldsymbol{y} \leq \boldsymbol{u}$ are fully censored.

The drawback of using scale mixture models is the same as for Pareto processes if we use censored likelihood estimation. The density of the D-variate Gaussian scale mixture is is easily derived (cf. Fang et al., 1990, § 2.2.3): for a Gaussian random vector $\boldsymbol{W}_t$ with covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{Q}$, the likelihood contribution given that variables in $\mathbb{A}_t$ are observed is

$$\begin{aligned}&\frac{\partial^{|\mathbb{A}_t|}}{\partial \boldsymbol{w}_{\mathbb{A}_t}} \int_0^\infty \Phi_D(\boldsymbol{w}/r, \boldsymbol{Q}^{-1}) f_R(r)\mathrm{d}r \\ &\quad = \int_0^\infty \Phi_{|\mathbb{U}_t|}(\boldsymbol{w}_{\mathbb{U}_t}/r + \boldsymbol{Q}_{\mathbb{U}_t}^{-1}\boldsymbol{Q}_{\mathbb{U}_t,\mathbb{A}_t}\boldsymbol{w}_{\mathbb{A}_t}/r, \boldsymbol{Q}_{\mathbb{U}_t}^{-1}) \phi_{|\mathbb{A}_t|}(\boldsymbol{w}_{\mathbb{A}_t}/r, \boldsymbol{\Sigma}_{\mathbb{A}_t}) r^{-|\mathbb{A}_t|} f_R(r)\mathrm{d}r.\end{aligned}$$

The integrand contains a $\mathbb{U}_t$-dimensional distribution function that has to be calculated for each t and each observation may have a different censoring pattern, so the marginal precision matrix may need to be derived for each of the T observations. The conditional distribution $\boldsymbol{W}_{\mathbb{U}_t} \mid \boldsymbol{W}_{\mathbb{A}_t}$ is best expressed using the precision matrix, while the marginal $\boldsymbol{W}_{\mathbb{A}_t}$ has precision given by the Schur complement of $\boldsymbol{Q}_{\mathbb{U}_t}$, namely $\boldsymbol{\Sigma}_{\mathbb{U}_t}^{-1} = \boldsymbol{Q}_{\mathbb{A}_t} - \boldsymbol{Q}_{\mathbb{A}_t,\mathbb{U}_t}\boldsymbol{Q}_{\mathbb{U}_t}^{-1}\boldsymbol{Q}_{\mathbb{U}_t,\mathbb{A}_t}$. It is best to

resort to Markov chain Monte Carlo methods for inference using a data augmentation scheme.

To avoid having to test for asymptotic independence on the boundary of the parameter space, Huser and Wadsworth (2019) propose using a model arising as the power transform of a scale mixture of Gaussian of the form $R^{\delta}\mathsf{W}^{1-\delta}$, where $R \sim \mathsf{GP}(1,\xi)$ and W is a log-Gaussian process. Since $R^{\delta}W^{1-\delta}$ has the same copula as $\delta\log(R)+(1-\delta)\log(W)$, Huser and Wadsworth adopt a semi-parametric approach, using the empirical distribution for the margins and estimating the dependence parameters using the mixture representation; $\delta = 1/2$ represents the boundary between asymptotic dependence and independence and so a test for asymptotic independence is easy to derive. The censored likelihood estimation of the dependence parameter for the Gaussian model is as expensive to implement as alternatives.

Gaussian scale mixture models are expensive to fit and their dependence structure is complicated by the introduction of the scale. Since the latent radial variable can be viewed as auxiliary, it is easily handled in a Bayesian model. This does not resolve the issue of censoring. Based on the Huser et al. (2017) model, Shaby proposed a partial augmentation scheme incorporating a nugget term. The benefit of his construction is that it incorporates a marginal transformation. Let D denote the number of spatial locations and T the number of time points. Using the peaks-over-threshold method, we model $Y^*_{j,t} \sim \mathsf{GP}(\sigma_j,\xi_j)$ and transform these observations to be marginally distributed according to the Gaussian scale mixture, whose marginal distribution and density functions are respectively

$$F_{\mathsf{SM}}(y) = \int_0^{\infty} \Phi(y/r) f_R(r)\mathrm{d}r, \qquad f_{\mathsf{SM}}(y) = \int_0^{\infty} \phi(y/r) r^{-1} f_R(r)\mathrm{d}r.$$

The observed data on the transformed scale is $Y_{j,t} = \max\{Z_{j,t}, v^*_j\}$, where $v^*_j = F^{-1}_{\mathsf{SM}}(1-\zeta_j)$, $Z_{j,t} = F^{-1}_{\mathsf{SM}}\{F_{\mathsf{GP}_j}(Y^*_{j,t})\}$ and $\zeta_j = \mathsf{P}(Y^*_{j,t} > u_j)$ is the marginal site-wise probability of exceeding the threshold. We denote censoring by $\mathbb{C}_{j,t} = \mathbf{1}\{Y_{j,t}<u_j\}$.

The process $\boldsymbol{Z}$ is modelled as a latent Gaussian process with an additional nugget component. The hierarchical spatial linear model is

$$\begin{aligned} &Z_{j,t} = \beta_0 + \mu_{j,t} + \varepsilon_{j,t}, \quad (j=1,\ldots,D), && \boldsymbol{\mu}_t \sim \mathsf{No}_D(\mathbf{0}_D, \sigma^2_{t,\mu}\mathbf{R}_{\kappa,\phi}),\\ &\varepsilon_{j,t} \overset{\text{iid}}{\sim} \mathsf{No}(0,\sigma^2_{\varepsilon}), && \sigma^2_{\varepsilon} \sim \mathsf{IG}(a_{\varepsilon}, b_{\varepsilon}), \qquad \sigma_{t,\mu} \sim F_R(\boldsymbol{\theta}_R), \end{aligned}$$

where $\mathbf{R}$ is a correlation matrix associated to the random effect and $\kappa,\phi,\boldsymbol{\theta}_R$ are assigned prior distributions. By choosing a distribution F_R for the random scale of the Gaussian process, one can bridge asymptotic independence and asymptotic dependence.

A data augmentation scheme for $(Z_{j,t}, \mathbb{C}_{j,t}=1)$ allows one to fully exploit the structure of the Gaussian scale mixture model without having to compute a high-dimensional Gaussian integral. Indeed, $Z_{j,t}$ is conditionally a scalar truncated Gaussian variable, which is easily

sampled. The conditional posterior for the censored observations is

$$\begin{aligned} Z_{j,t} \mid \mu_{j,t}, \beta_0, \mathbb{C}_{j,t} = 1 &\sim \mathsf{TNo}(\beta_0 + \mu_{j,t}, \sigma_\varepsilon^2; \nu_j), \\ \sigma_\varepsilon^2 \mid \boldsymbol{Z}, \boldsymbol{\mu}, \beta_0 &\sim \mathsf{IG}(DT/2 + a_\varepsilon, \|\mathrm{vec}(\boldsymbol{Z}) - \mathrm{vec}(\boldsymbol{\mu}) - \beta_0 \mathbf{1}_{DT}\|^2/2 + b_\varepsilon), \\ \boldsymbol{\mu}_{,t} \mid \boldsymbol{Z}_{,t}, \beta_0, \phi, \kappa, \sigma_\varepsilon^2, \sigma_{\mu_t}^2 &\sim \mathsf{No}_D\left(\boldsymbol{\Xi}^{-1}\sigma_\varepsilon^{-2}(Z_{,t} - \beta_0 \mathbf{1}_D), \boldsymbol{\Xi}^{-1}\right), \end{aligned}$$

where $\boldsymbol{\Xi}^{-1} := \sigma_{\mu_t}^{-2}\mathbf{R}_{\kappa,\phi}^{-1} + \sigma_\varepsilon^{-2}\mathbf{I}_D$. Gibbs sampling is thus straightforward for the variables $Z_{j,t}, \sigma_\varepsilon^2$ and $\boldsymbol{\mu}_{,t}$. The update for $\boldsymbol{\mu}$ is simply a conjugate update for the Gaussian linear model. Without loss of generality, suppose that the k_t censored observations are stored as the first elements of the vector $\boldsymbol{\mu}_t$ and denote the k_t subvector by $\boldsymbol{\mu}_{ta}$. Then, conditional on the uncensored observations, we get

$$\boldsymbol{\mu}_{ta} \sim \mathsf{No}_{k_t}\left(\boldsymbol{\Xi}^{-1}\sigma_\varepsilon^{-2}(\boldsymbol{y}_a - \boldsymbol{\alpha}_a), \boldsymbol{\Xi}^{-1}\right),$$

where $\boldsymbol{\Xi} = \boldsymbol{\Phi}_a^\top \boldsymbol{\Phi}_a + \mathbf{D}_a^{-1}$ and $\alpha_{ti} = \log(R_t) - \gamma(\boldsymbol{s}_i)/2 - \sigma_\varepsilon^2/2$. Moving the mean components to the process level $\boldsymbol{\mu}_t$ would instead give

$$\boldsymbol{\mu}_{ta} \sim \mathsf{No}_{k_t}\left(\boldsymbol{\Xi}^{-1}\boldsymbol{\Phi}_a^\top\left\{\boldsymbol{\Phi}_a(\boldsymbol{y}_a - \boldsymbol{\alpha}_a) - \boldsymbol{\Phi}_b(\boldsymbol{\mu}_b - \boldsymbol{y}_b + \boldsymbol{\alpha}_b)\right\}, \boldsymbol{\Xi}^{-1}\right).$$

For the uncensored variables, $\log(Z_{ti}) = \log(R_t) + \mu_{ti}$, so the latter are determined once R_t is fixed. This means that the update for $\boldsymbol{\mu}_t$ should be done conditional on those values. For the censored variables, we have

$$Y_{ti} \mid \cdot \sim \mathsf{No}(\log(R_t) + \mu_{ti} - \gamma(\boldsymbol{s}_i)/2 - \sigma_\varepsilon^2/2, \sigma_\varepsilon^2), \qquad \boldsymbol{\mu}_t \sim \mathsf{No}_D(\mathbf{0}_D, \mathrm{C}_{\boldsymbol{\theta}}).$$

Shaby's construction, which includes both an intercept and a nugget, has two awkward features: each temporal replication has a random scale $\sigma_t^2 \sim F_R$, and the distribution of $Z_{j,t}$, the censored variable, also depends on the parameters of the F_R model. Consider the transformation $g_j(x) = F_{\mathsf{GP}_j}^{-1} \circ F_{\mathsf{SM}}(x)$ and assume that $Z_{j,t}$ is fully observed. The density of $\boldsymbol{Y}_{,t}^* = g(\boldsymbol{Z}_{,t}) = (g_1(Z_{1,t}), \ldots, g_d(Z_{D,t}))$ for $\boldsymbol{Y}_{,t}^* > \boldsymbol{u}$ is obtained after a change of variable as

$$f_{\boldsymbol{Y}_{,t}^*}(\boldsymbol{y}_{,t}^*) = f_{\boldsymbol{Z}_{,t}}\{g^{-1}(\boldsymbol{y}_{,t}^*)\} \prod_{j=1}^{D} \left| \frac{\partial g^{-1}(y_{j,t}^*)}{\partial y_{j,t}^*} \right|.$$

Since the $Z_{j,t}$ are conditionally independent given $\boldsymbol{\mu}, \sigma_\varepsilon^2, \beta_0$, the joint density is simply the product over all temporal and spatial replicates with $f_{\boldsymbol{Z}_{,t}} = \prod_{j=1}^{D} \sigma_\varepsilon \phi\{(z_{j,t} - \beta_0 - \mu_{j,t})/\sigma_\varepsilon\}$. The likelihood contribution is therefore

$$\ell_{j,t}(y_{j,t}^*) = \begin{cases} \Phi\left\{\dfrac{F_{\mathsf{SM}}^{-1}(\zeta_j) - \beta_0 - \mu_{j,t}}{\sigma_\varepsilon}\right\}, & \mathbb{C}_{j,t} = 1, \\ \dfrac{1}{\sigma_\varepsilon}\phi\left\{\dfrac{g^{-1}(y_{j,t}^*) - \beta_0 - \mu_{j,t}}{\sigma_\varepsilon}\right\} \dfrac{f_{\mathsf{GP}_j}(y_{j,t}^*)}{f_{\mathsf{SM}}(g^{-1}(y_{j,t}^*))}, & \mathbb{C}_{j,t} = 0. \end{cases}$$

Updating the parameters of the random scale F_R changes the observations $\boldsymbol{Y}$ and the thresholds on the transformed scale, which becomes $v_j = F_{\mathsf{SM}}^{-1}(u_j)$. It is possible that imputed observations $(Z_{s,t}, \mathbb{C}_{s,t} = 1)$ fall above the threshold after updating the parameters of the F_R model. This mismatch is due to the incompatibility between the marginal of $\boldsymbol{Z}$,which follows a scale mixture model, and the conditional distribution of $\boldsymbol{Z}$ given the random effect $\boldsymbol{\mu}$, which is Gaussian. Rather than performing censoring, Shaby's construction jitters censored variables, truncating them to ensure they fall below the threshold. The nugget thus successfully approximates censoring provided that its variance σ_ε^2 is large, and this choice also improves the mixing of the Markov chain. One unsatisfying feature of this hierarchical construction is that the observations above the threshold are measured with large error with the nugget, so the resulting field is rough. We do not pursue this approach further.

3.4 Bayesian linear model

In the rest of this thesis, we focus exclusively on generalized max-Pareto processes. Since data are collected only at a finite number of sites, spatial regression models with random effects can serve to interpolate estimates of the parameters of the generalized Pareto distribution of threshold exceedances. Such an approach was pioneered in the extreme value literature by Casson and Coles (1999), who used regression surfaces with the Poisson process likelihood. Cooley et al. (2007) combined generalized Pareto margins with Gaussian process priors for the scale, shape and exceedance rate, using distance between sites in a climate space, based on mean precipitation and elevation. These covariates must be available at every location for which predictions are requested.

Latent variable models (cf. Davison et al., 2012, § 4) allow one to borrow strength from neighbouring sites, but cannot be utilized directly for simulation of extreme events. Indeed, conditional on the marginal parameter vectors, all pointwise realizations are independent of those at the other sites, resulting in discontinuous spatial realizations. Latent variable models can be used to obtain return levels and such models can be fitted in high dimensions at little cost: for example, Geirsson et al. (2015) use a Gaussian Markov random field approximation for the covariance of the random effect for maxima, and Jalbert et al. (2017) fit a second order improper Gaussian Markov random field with scaling to account for temporal nonstationarity. The regression can be more or less complex, with covariates derived from climatic models; as an example, Dyrrdal et al. (2014) employ Bayesian model averaging with a conditional Bayes factor for model selection to determine optimal regression models, whereas other have used information criteria.

It is possible to include dependence structures: Sang and Gelfand (2010) employed a Gaussian copula to introduce dependence between generalized extreme value distributions and Fuentes et al. (2013) use a truncated Dirichlet process mixture of Gaussian copulas. Shaby and Reich (2012) and Reich and Shaby (2012) propose the use of the hierarchical kernel extreme value process for inference based on block maxima, but estimation is complicated due to the

dual role of the nugget parameter (Sebille, 2016); the stochastic construction is laid out in Stephenson (2009) and in Section 2.2.1. Use of max-stable processes has been hampered by the computational costs associated with evaluation of the likelihood: Ribatet et al. (2012) suggest an adjustment for composite likelihoods to approximately restore the asymptotic χ^2 distribution for the likelihood ratio, whereas Shaby (2014) propose an open-faced sandwich adjustment that adjusts samples *post hoc*. Such methods have rarely been used (but see Sharkey and Winter (2019) for an application to areal data with an improper conditional autoregressive model).

We model the marginal parameters $\{\log(\boldsymbol{\sigma}), \boldsymbol{\xi}\}$ using Gaussian processes, assuming that scale and shape parameters are independent a priori. Covariates are often incorporated in a generalized spatial linear model to model potential nonstationarity and pool information across space, but the specification of the response surface must be the same for both scale and shape to ensure that threshold stability is preserved (Eastoe and Tawn, 2009, § 2.2). If the shape ξ is constant (or involves spatial covariates that are constant in time), then the scale $\sigma_u(\boldsymbol{s})$ can only be a linear function of covariates, which is incompatible with a log-link transformation to ensure positivity; Gyarmati-Szabó et al. (2017) propose an alternative parametrization of the model as workaround. Let $\boldsymbol{\theta}$ denote either of the log-scale or the shape vectors and let $\mathbf{X}_\theta$ denote a $D \times p$ matrix of covariates and assume that inference is performed conditional on the correlation matrices $\mathbf{R}_\theta$ and $\mathbf{R}_{\boldsymbol{\beta}_\theta}$ and hyperparameters $a, b, \boldsymbol{\nu}_\theta$. The Bayesian linear model employs conjugate priors for the regression parameters, namely a Gaussian-inverse Gamma prior. Write the posterior as

$$p(\boldsymbol{\beta}_\theta, \tau^2 \mid \boldsymbol{\theta}) \propto p(\boldsymbol{\theta} \mid \boldsymbol{\beta}_\theta, \tau_\theta^2 \mathbf{R}_\theta) p(\boldsymbol{\beta}_\theta \mid \tau_\theta^2) p(\tau_\theta^2)$$

where

$$\boldsymbol{\theta} \sim \mathsf{No}_D(\boldsymbol{\beta}_\theta \mathbf{X}_\theta, \tau_\theta^2 \mathbf{R}_\theta), \qquad p(\boldsymbol{\beta}_\theta \mid \tau_\theta^2) \sim \mathsf{No}_p(\boldsymbol{\beta}_\theta; \boldsymbol{\nu}_\theta, \tau_\theta^2 \mathbf{R}_{\boldsymbol{\beta}_\theta}), \qquad p(\tau_\theta^2) \sim \mathsf{IG}(\tau_\theta^2; a_\theta, b_\theta). \quad (3.8)$$

The posterior is then proportional to

$$\begin{aligned} p(\boldsymbol{\beta}_\theta, \tau^2 \mid \boldsymbol{\theta}) \propto \tau_\theta^{-D} \exp\left\{ -\frac{1}{2\tau_\theta^2} (\boldsymbol{\theta} - \mathbf{X}_\theta \boldsymbol{\beta}_\theta)^\top \mathbf{R}_\theta^{-1} (\boldsymbol{\theta} - \mathbf{X}_\theta \boldsymbol{\beta}_\theta) \right\} \\ \times \frac{1}{2} \boldsymbol{\beta}_\theta (\tau_\theta^2)^{-a-1} \exp\left(-\frac{b}{\tau_\theta^2} \right) \tau_\theta^{-p} \exp\left\{ -(\boldsymbol{\beta}_\theta - \boldsymbol{\nu}_\theta)^\top \mathbf{R}_{\boldsymbol{\beta}_\theta}^{-1} (\boldsymbol{\beta}_\theta - \boldsymbol{\nu}_\theta) \right\}. \end{aligned}$$

By completing the square and grouping terms that depend on $\boldsymbol{\beta}_\theta$, we readily obtain the parameters of the Gaussian inverse gamma posterior. Simulation from this is straightforward using composition sampling: first sample τ_θ^2, then use this value to generate the new regression parameters $\boldsymbol{\beta}_\theta$, i.e.,

$$\tau_\theta^2 \mid \boldsymbol{\theta} \sim \mathsf{IG}\left(a_\theta + \frac{D+p}{2}, b + \frac{\boldsymbol{\theta}^\top \mathbf{R}_\theta^{-1} \boldsymbol{\theta} + \boldsymbol{\nu}_\theta^\top \mathbf{R}_{\boldsymbol{\beta}_\theta}^{-1} \boldsymbol{\nu}_\theta - \mathbf{m}^\top \mathbf{M} \mathbf{m}}{2} \right),$$

$$\boldsymbol{\beta}_\theta \mid \tau_\theta^2, \boldsymbol{\theta} \sim \mathsf{No}_p(\mathbf{Mm}, \tau_\theta^2 \mathbf{M})$$

with $\mathbf{M}^{-1} := \mathbf{X}_\theta^\top \mathbf{R}_\theta^{-1} \mathbf{X}_\theta + \mathbf{R}_{\boldsymbol{\beta}_\theta}^{-1}$ the precision matrix of $\boldsymbol{\beta}_\theta$ and $\mathbf{m} = \boldsymbol{\theta}^\top \mathbf{R}_\theta^{-1} \mathbf{X}_\theta + \mathbf{R}_{\boldsymbol{\beta}_\theta}^{-1} \boldsymbol{\nu}_\theta$.

The correlation matrix of the spatial random effects, $\mathbf{R}_\theta$, is usually taken to be exponential, i.e., $\mathsf{Cor}\{\boldsymbol{\theta}(s_i), \boldsymbol{\theta}(s_j)\} = \exp(-\|\boldsymbol{s}_i - \boldsymbol{s}_j\| / \varsigma)$, giving $\mathbf{R}_\theta = \exp(-\mathbf{D}/\varsigma)$ for $\mathbf{D}$ the matrix of pairwise distances between sites. The range hyperparameter ς is notably hard to estimate and the prior will tend to dominate the posterior of ς unless the number of sites is large.

Dyrrdal et al. (2014) fit a Bayesian hierarchical model with generalized extreme value margins and uses the transition kernel of eq. (3.3) in a Metropolis-within-Gibbs algorithm to update the parameters one by one. We adopt this idea for the range parameters of the exponential correlation $\mathbf{R}_\theta$, since gradients can be readily calculated analytically. With a $\Gamma(a_\varsigma, b_\varsigma)$ hyperprior for ς, the conditional log-posterior for ς is

$$f(\varsigma) := \log\{p(\varsigma \mid \boldsymbol{\beta}_\theta, \tau^2, \boldsymbol{\theta})\} \propto -\frac{\tau_\theta^{-2}}{2} (\boldsymbol{\theta} - \boldsymbol{\beta}_\theta)^\top \mathbf{R}_\theta^{-1} (\boldsymbol{\theta} - \boldsymbol{\beta}_\theta) - \frac{1}{2} \log|\mathbf{R}_\theta| + (a_\varsigma - 1)\log(\varsigma) - b_\varsigma \varsigma,$$

and the first two derivatives of the conditional posterior for ς are (Dyrrdal et al., 2014)

$$\begin{aligned} f'(\varsigma) &= \frac{\tau_\theta^{-2}}{2} (\boldsymbol{\theta} - \boldsymbol{\beta}_\theta)^\top \mathbf{R}_\theta^{-1} \dot{\mathbf{R}}_\theta \mathbf{R}_\theta^{-1} (\boldsymbol{\theta} - \boldsymbol{\beta}_\theta) - \frac{1}{2} \operatorname{tr}\left\{\mathbf{R}_\theta^{-1} \dot{\mathbf{R}}_\theta\right\} - b_\varsigma + (a_\varsigma - 1)\varsigma^{-1}, \\ f''(\varsigma) &= \frac{\tau_\theta^{-2}}{2} (\boldsymbol{\theta} - \boldsymbol{\beta}_\theta)^\top \mathbf{N}_\varsigma (\boldsymbol{\theta} - \boldsymbol{\beta}_\theta) - \frac{1}{2} \operatorname{tr}\left\{\mathbf{R}_\theta^{-1} \ddot{\mathbf{R}}_\theta - \left[\mathbf{R}_\theta^{-1} \dot{\mathbf{R}}_\theta^{-1}\right]^2\right\} - (a_\varsigma - 1)\varsigma^{-2}, \end{aligned}$$

where

$$\dot{\mathbf{R}}_\theta = \varsigma^{-2} \mathbf{D} \circ \mathbf{R}_\theta, \qquad \ddot{\mathbf{R}}_\theta = 2\varsigma^{-3} \mathbf{D} \circ \mathbf{R}_\theta + \varsigma^{-2} \mathbf{D} \circ \dot{\mathbf{R}}_\theta, \qquad \mathbf{N}_\varsigma = 2\mathbf{R}_\theta^{-1} \left[\dot{\mathbf{R}}_\theta \mathbf{R}_\theta^{-1}\right]^2 - \mathbf{R}_\theta^{-1} \ddot{\mathbf{R}}_\theta \mathbf{R}_\theta^{-1},$$

and $\circ$ denotes the Hadamard (i.e., element-wise) product. The Laplace approximation proposal has the drawback that the chains gets stuck if the proposal mean is negative; we thus sample from a truncated Gaussian to ensure that $\varsigma > 0$ and the step size of random walk is bounded; we also reduce $\omega < 1$ in the Newton step to improve mixing.

For a latent Gaussian model $\log(\boldsymbol{\sigma}) \sim \mathsf{No}_D(\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\sigma}}, \mathbf{Q}^{-1})$, the gradient and Hessian with respect to $\boldsymbol{\sigma}$ are

$$\begin{aligned} f'(\boldsymbol{\sigma}) &= -\boldsymbol{\sigma}^{-1} - \mathbf{Q}\{\log(\boldsymbol{\sigma}) - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\sigma}}\} \circ \boldsymbol{\sigma}^{-1}, \\ f''(\boldsymbol{\sigma}) &= \operatorname{diag}(\boldsymbol{\sigma}^{-2}) - \mathbf{Q} \circ \boldsymbol{\sigma}^{-1} \otimes \boldsymbol{\sigma}^{-1} + \operatorname{diag}\left[\mathbf{Q}\{\log(\boldsymbol{\sigma}) - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\sigma}}\} \circ \boldsymbol{\sigma}^{-2}\right]. \end{aligned}$$

3.5 Bayesian inference for generalized ℓ-Pareto processes

3.5.1 Data augmentation

The major difficulty for Bayesian inference for the ℓ-Pareto process is the lack of closed-form expressions for the risk region and the censored contribution, meaning that gradients are not available. This precludes the use of many more efficient algorithms for Markov chain Monte Carlo. An alternative is to use the point process representation to augment the likelihood with the inferentially censored components; the joint likelihood for the components above the threshold and the imputed values involves only the full intensity function.

Given current values of the state vector $\boldsymbol{\theta}$ and a vector of partly censored observations $(\boldsymbol{x}_{\mathbb{A}}, \mathbf{1}\{\boldsymbol{x}_{\mathbb{U}} \leq \boldsymbol{v}_{\mathbb{U}}\})$ and $\boldsymbol{\lambda}_u$ the percentage of observations falling above their site-wise threshold, we map observations to the standardized scale via the marginal transformation $\boldsymbol{y}_{\mathbb{A}} = (1 + \xi \boldsymbol{x}_{\mathbb{A}}/\boldsymbol{\sigma}_{\mathbb{A}})^{1/\xi}/\boldsymbol{\lambda}_u$, and similarly for the marginal censoring limits $\boldsymbol{v}^*_{\mathbb{U}} = (1 + \xi \boldsymbol{v}_{\mathbb{A}}/\boldsymbol{\sigma}_{\mathbb{A}})^{1/\xi}/\boldsymbol{\lambda}_u$.

Figure 3.5 illustrates the censoring of non-extreme observations and shows the risk region. Conditional simulation is mostly useful for censored observations: if the risk region is such that we can identify exceedances from a single observation (e.g., with functionals such as $\max_{j=1}^{D} X(\boldsymbol{s}_j)$ or $X(\boldsymbol{s}_0)$) that is uncensored, then additional observations can be simulated on the unit Pareto scale. These observations can be drawn from a truncated Gaussian distribution for the Brown–Resnick and from a truncated Student-t distribution for the extremal Student using the algorithm of Botev (2017) and Botev and L'Écuyer (2017) covered in Section 3.2.2; simulated samples below are then back-transformed to the data scale. censoring Conditional simulations would automatically fall in the risk region shown in grey in Figure 3.5, but below the marginal threshold, along the line corresponding to the non-censored component.

For other risk regions such as $\{\boldsymbol{X} \in \mathbb{R}^D : \sum_{j=1}^{D} X(\boldsymbol{s}_j) > u\}$, such sampling schemes must be embedded in an accept-reject algorithm to ensure that simulated points fall in the risk region after the marginal transformation; see Figure 3.5. For the sum risk functional, one would also need to ensure that at least one observation exceeds its marginal threshold, by taking, e.g., $u/D\mathbf{1}_D$ as the vector of marginal thresholds. Algorithm 3.4 can be used for data augmentation in the case of the Brown–Resnick process. For the extremal Student process, the augmented components could be assigned negative values, as is commonly done in Bayesian modelling of rainfall where zero components are modelled as truncated Gaussian on $[\mathbf{0}_D, \infty_D)$ and the augmented values are negative (cf. Sansó and Guenni, 1999).

While data augmentation is commonly used in spatial statistics (de Oliveira, 2005), the number of censored components is so large that the information from the imputed values dominates the likelihood and hinders the good mixing of standard Markov chain Monte Carlo algorithms. Efforts to leverage on this have been largely unsuccessful, notably because the doubly-intractable likelihood prevents use of methods such as Hamiltonian Monte Carlo that could lead to more efficient exploration of the space.

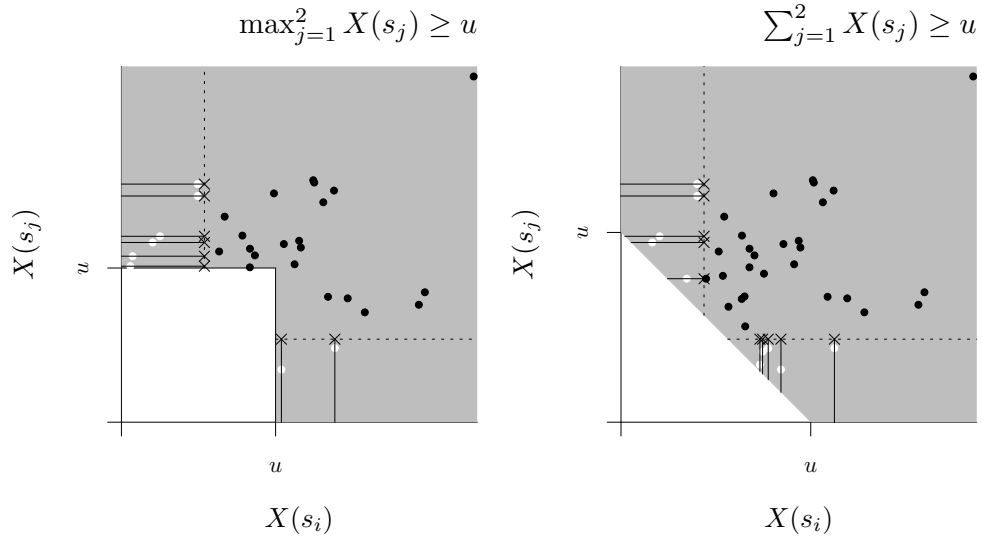

Figure 3.5 – *Illustration of inferential censoring for max (left) and sum (right) risk functionals in dimension $D = 2$. Points in white are censored, i.e., only the information that they fell below their marginal threshold (dashed line) is retained.*

Algorithm 3.4 Imputation of censored observations, Brown–Resnick generalized ℓ-Pareto process

Require: marginal thresholds $\boldsymbol{v} \leq u$, functional threshold u;
Require: marginal parameters $\boldsymbol{\lambda}_u, \boldsymbol{\eta}, \boldsymbol{\tau}, \xi$, semivariogram matrix $[\gamma_{ij}]_{i,j=1,\ldots,D}$;
Require: observation vector $\boldsymbol{y} \in \{\boldsymbol{y} \in \mathbb{R}^D : \ell(\boldsymbol{y}) > u, \boldsymbol{y} \not\geq \boldsymbol{v}, \boldsymbol{y} \not\leq \boldsymbol{v}\}$;
define $\mathbb{A} := \{j : y_j \geq v_j\}$, $\mathbb{U} := \{j : y_j < v_j\}$, $\mathbb{A}_1 = \min\{\mathbb{A}\}$
define $\mathbb{A}^- := \{j : (j+1) \in \mathbb{A}\}$, $\mathbb{U}^- := \{j : (j + \mathbf{1}\{j<\mathbb{A}_1\}) \in \mathbb{U}\}$
function MARGINAL TRANSFORMATION($\boldsymbol{x}$)
    return $H(\boldsymbol{x}) = [1 + \xi(\boldsymbol{x} - \boldsymbol{\eta})/\boldsymbol{\tau}]^{1/\xi}/\boldsymbol{\lambda}_u$
set $\boldsymbol{v}^* \leftarrow H(\boldsymbol{v})$, $\boldsymbol{u}^* \leftarrow H(u\mathbf{1}_D)$, $\boldsymbol{y}^* \leftarrow H(\boldsymbol{y})$
set $\boldsymbol{\Sigma}_{ij} \leftarrow \gamma_{i,\mathbb{U}_1} + \gamma_{j,\mathbb{U}_1} - \gamma_{i,j}$, $i, j \in \{1,\ldots,D\} \setminus \{\mathbb{U}_1\}$
set

$$\boldsymbol{\mu}_{\mathbb{U}} \leftarrow \log\left(\frac{\boldsymbol{v}_{\mathbb{U}}}{y^*_{\mathbb{A}_1}}\right) + \boldsymbol{\gamma}_{\mathbb{U},\mathbb{A}_1} - \boldsymbol{\Sigma}_{\mathbb{U}^-,\mathbb{A}_-}\boldsymbol{\Sigma}^{-1}_{\mathbb{A}^-,\mathbb{A}_-}\left\{\log\left(\frac{\boldsymbol{y}^*_{\mathbb{A}_{-1}}}{y^*_{\mathbb{A}_1}}\right) + \boldsymbol{\gamma}_{\mathbb{A}_{-1},\mathbb{A}_1}\right\}$$
$$\boldsymbol{\Sigma}_{\mathbb{U}} \leftarrow \boldsymbol{\Sigma}_{\mathbb{U}^-,\mathbb{U}^-} - \boldsymbol{\Sigma}_{\mathbb{U}^-,\mathbb{A}^-}\boldsymbol{\Sigma}^{-1}_{\mathbb{A}^-,\mathbb{A}^-}\boldsymbol{\Sigma}_{\mathbb{A}^-,\mathbb{U}^-}$$

repeat
    sample $\boldsymbol{z} \sim \mathsf{TNO}(\boldsymbol{\mu}_{\mathbb{U}}, \boldsymbol{\Sigma}_{\mathbb{U}}; -\infty_{|\mathbb{U}|}, \boldsymbol{v}_{\mathbb{U}})$
    set $\boldsymbol{y}^*_{\mathbb{U}} \leftarrow \exp\{\boldsymbol{z} + \log(\boldsymbol{y}^*_{\mathbb{A}_1})\}$
    set $\boldsymbol{y}_{\mathbb{U}} \leftarrow \boldsymbol{\tau}_{\mathbb{U}}(\boldsymbol{y}^{*\xi}_{\mathbb{U}} - 1)/\xi + \boldsymbol{\eta}_{\mathbb{U}}$
until $\ell(\boldsymbol{y}) > u$
return $\boldsymbol{y}$

3.5.2 Simulation study: logistic max-Pareto process

For most parametric families presented in Section 2.2.3, there are no closed-form expressions for the exponent measure $V(\boldsymbol{u})$ and the censored likelihood. One notable exception is the (simple) logistic model, for which $V_{1:k}$ and V are fully known. We use the model in simulations, partly because this allows for considerable speed-up, but also because optimization is easily performed in this four-parameter problem and this allows us to draw meaningful comparisons between the censored and full likelihoods. We fitted a Bayesian hierarchical model based on simulated data from a logistic generalized max-Pareto process with $\eta = 0$ and $\tau = 1$ and varying shape $\xi \in \{-0.15, 0, 0.15\}$ and dependence parameters $\alpha \in \{0.25, 0.5, 0.75\}$. We selected the 40th marginal percentile as threshold in order to assess the loss of information from censoring, since this corresponds roughly to the proportion of missing values when we look at the observations for which there is at least one exceedance. For identifiability, the first location parameter was set to zero. We ran two parallel chains using Hamiltonian Monte Carlo implemented in Stan (Carpenter et al., 2017), keeping 5000 iterations after burn-in of 750. Hamiltonian Monte Carlo is extremely efficient at exploring the state space. We selected independent vague priors for all parameters, with $p(\eta) \sim \mathsf{No}(0, 100)$, $p(\tau) \sim \mathsf{HNo}(0, 100)$, $p(\xi) \sim \mathsf{No}(0, 1)$ and $p(\alpha) \sim \mathsf{Be}(5/4, 5/4)$. For each dataset, we estimated the maximum a posteriori using constrained optimization using both censored and full likelihood in addition to the MCMC runs.

We consider varying sample sizes $n \in \{125, 250, 500\}$ and varying dimensions $D \in \{5, 10, 20\}$ and performed 1000 replications of the simulation. Simulations in which the algorithm did not converge (less than 0.05%) were discarded; we monitored the potential scale reduction factor $\widehat{R}$ and the effective sample size; the median of the latter is over 50% of the simulated values, and lowest effective sample size across all 64000 simulations is 1500 for a single parameter.

Because there are only four parameters, increasing the dimension usually has the same effect as increasing the sample size: both increase the precision of the estimators. For some combinations of the parameters, there is virtually no loss of information when using a censored likelihood relative to full likelihood for the location parameter.

Table 3.2 gives the relative root mean squared error for the median posterior estimates of the parameters. The stronger the dependence between sites (smaller values of α), the more precise the estimates of the location; this is likely due to the identifiability constraint. In contrast, the posterior median estimator of the scale and shape have lower variances when α is large and dependence weak. The relative efficiency of the censored likelihood for η decreases when ξ increases, which seems natural as the data have higher a variance and an infinite upper endpoint. Both estimators of the dependence parameter are nearly as efficient, except when $n = 500$ and $D = 5$, in which case the full likelihood is much sharper. In the case of weak dependence, a lot of information about the shape is lost and the efficiency is around one-third.

In most cases, censoring leads to unbiased but less efficient estimators, though with the notable exception that, with $\alpha = 0.25$ and $D = 5$, the censored likelihood shape estimator is

		par.	η			τ			ξ			α		
ξ_0	D	$n \mid \alpha_0$	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75
−0.15	5	125	90	90	93	41	86	82	49	62	47	58	95	85
		250	90	86	96	35	89	84	44	65	48	49	95	83
		500	88	88	95	27	85	85	34	59	46	38	95	83
	10	125	85	86	95	85	88	90	74	62	49	91	96	89
		250	87	88	94	91	91	92	80	62	51	94	96	90
		500	87	86	93	93	92	92	78	63	50	96	96	89
	20	125	84	87	92	97	94	89	79	64	55	100	98	89
		250	85	87	94	96	95	91	79	66	52	99	97	91
		500	83	86	93	97	94	92	80	67	52	99	98	90
0	5	125	86	81	81	39	83	82	56	49	31	58	95	84
		250	85	77	84	34	86	85	50	51	32	50	95	82
		500	84	80	82	26	83	86	42	47	29	38	95	82
	10	125	78	76	82	84	88	87	66	50	33	91	95	89
		250	80	79	79	90	90	87	69	48	32	94	95	90
		500	80	77	77	91	91	90	68	49	33	96	96	89
	20	125	77	77	76	96	94	83	65	49	34	100	97	89
		250	78	78	79	95	95	84	65	49	31	99	97	91
		500	76	77	77	96	95	86	66	49	32	99	97	90
0.15	5	125	81	69	42	37	81	80	62	44	31	59	94	83
		250	81	65	41	32	84	83	57	44	30	50	94	81
		500	79	68	41	24	81	83	52	43	27	39	94	82
	10	125	71	63	44	82	87	81	67	52	36	92	95	89
		250	74	66	40	88	90	79	67	49	33	94	95	90
		500	72	64	37	90	91	81	68	51	33	96	95	89
	20	125	70	64	39	95	94	76	71	60	38	99	97	89
		250	71	66	40	94	95	77	72	57	37	99	97	91
		500	69	64	37	95	95	77	70	56	39	99	97	90

Table 3.2 – *Relative root mean squared error of the censored likelihood versus the full likelihood (in %) for the logistic model based on* $B = 1000$ *replications.*

ξ_0	−0.15			0			0.15		
par. $\mid n$	125	250	500	125	250	500	125	250	500
τ	16	15	15	17	16	15	18	16	16
ξ	−4	−4	−4	−3	−3	−3	−2	−2	−2
α	−2	−2	−2	−2	−2	−2	−2	−2	−2

Table 3.3 – *Average bias (in %) of posterior median estimators based on* $B = 1000$ *replications using the censored likelihood for the logistic model with* $D = 5$ *and* $\alpha_0 = 0.25$.

negatively biased (Figure 3.6) and the estimator of the scale is positively biased, mimicking the classical tradeoff observed in Example 1.19 for the generalized Pareto distribution; see Table 3.3. This bias holds irrespective of the value of ξ_0 and persists even as n increases. Credible intervals have poor coverage properties because of the bias, as evidenced by the error rates in the first nine lines of Tables 3.5 to 3.7.

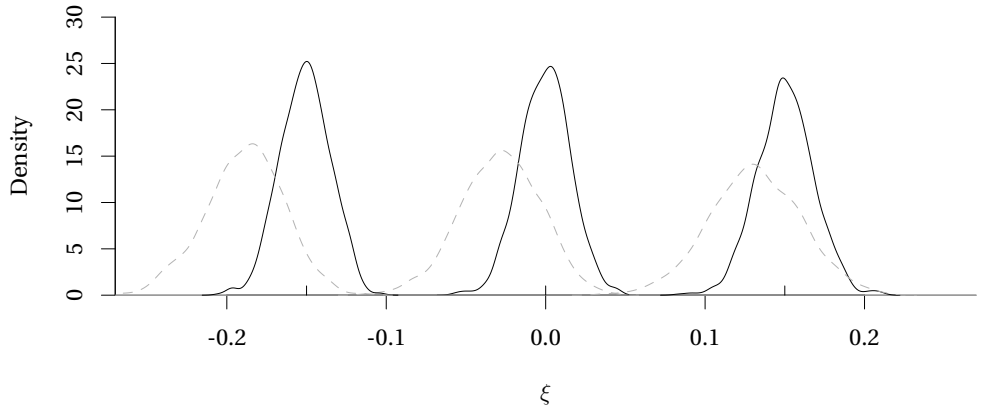

Figure 3.6 – *Density estimate of the posterior median based on $B = 1000$ replications for the shape parameter with $D = 5$ and $\alpha = 0.25$, with $\xi_0 = -0.15$ (left), $\xi = 0$ (middle) and $\xi = 0.15$ (right) based on the censored likelihood (grey dashed) and full likelihood (black full).*

Tables 3.4 to 3.7 report the one-sided nominal error rates for credible intervals. While the latter have no theoretical coverage guarantees, it is nevertheless reassuring that, with vague priors, the coverage is decent except in cases of strong bias discussed earlier.

Running a simulation study using the logistic model, with our choice of location and scale parameter, was convenient because it allowed us to get a large number of replicates and compare various scenarios. However, we set the marginal parameters to be equal, which is unrealistic for practical settings. Moreover, spatial extreme value models must be flexible enough to accommodate the spatial dependence, hence the need to consider models based on elliptical distributions.

α_0	D	ξ_0 | error rate	1	2.5	5	10	25	25	10	5	2.5	1
0.25	5	−0.15	0.6	1.5	4	8	22	24	9	5	1.5	0.8
		0	0	1	3	7	22	24	8	4	1.5	0.6
		0.15	0	1	2	6	22	23	8	4	1.5	0.4
	10	−0.15	1.2	3	5	9	24	24	9	5	2.5	1
		0	1.2	2.5	5	9	24	25	9	5	2.5	0.8
		0.15	1.4	2	4	9	23	25	9	5	3	0.8
	20	−0.15	2	3	6	11	23	26	10	5	2.5	1.4
		0	1.8	4	6	11	23	26	9	4	3	1.2
		0.15	1.6	3.5	6	11	24	26	9	5	2.5	1.4
0.50	5	−0.15	1.2	3	5	9	23	28	13	6	3	0.6
		0	1	3	4	9	22	28	12	6	3	1
		0.15	0.8	2.5	4	9	23	28	12	6	2.5	0.8
	10	−0.15	1.2	2.5	5	9	26	23	8	5	2	0.6
		0	0.6	2.5	5	10	26	23	9	4	2	0.6
		0.15	0.6	2	5	10	26	24	9	4	1.5	0.4
	20	−0.15	1	3	5	10	25	23	9	4	2	0.6
		0	0.6	2.5	5	9	25	24	9	4	2	0.4
		0.15	0.8	2.5	5	9	26	24	9	4	2	0.4
0.75	5	−0.15	0.6	2	4	10	25	26	11	5	2	1.6
		0	0.4	1.5	5	11	25	26	11	5	2.5	1.2
		0.15	0.4	2	5	11	25	26	11	5	3	1.2
	10	−0.15	1.2	3	6	11	26	24	9	4	1.5	0.6
		0	1.2	3	6	12	26	23	9	5	2	0.6
		0.15	1	3	6	11	26	23	8	4	1.5	0.4
	20	−0.15	0.8	3	5	9	26	25	9	4	2	1.2
		0	0.6	2	5	9	25	25	9	4	2	0.8
		0.15	1	2	4	8	24	24	10	5	2	1.2

Table 3.4 – *One-sided nominal error rate (in %) for lower and upper credible intervals for the location parameter η for the censored likelihood estimator with $n = 250$ based on $B = 1000$ replications.*

α_0	D	ξ_0 | error rate	1	2.5	5	10	25	25	10	5	2.5	1
0.25	5	−0.15	38.8	48.5	57	67	80	3	1	0	0	0
		0	39.6	49	58	68	81	3	1	0	0.5	0
		0.15	41	50.5	59	69	82	3	1	1	0	0
	10	−0.15	2.6	5	8	13	28	21	7	3	1	1
		0	3.2	5.5	8	13	29	21	7	4	1.5	0.8
		0.15	3.2	5.5	8	14	29	21	7	4	1	0.8
	20	−0.15	1	3	6	13	29	22	8	4	2	1
		0	1.2	3	7	13	30	21	8	4	2	1
		0.15	1.4	3	7	14	30	21	8	4	2	1
0.50	5	−0.15	1.4	3	6	13	29	20	6	2	1	0.2
		0	1.4	3	6	13	29	19	7	2	1	0
		0.15	1.4	3	7	13	30	19	7	2	1	0
	10	−0.15	1.2	2.5	6	12	28	19	8	3	1.5	0.4
		0	1.4	3	6	12	29	20	7	4	1.5	0.2
		0.15	1.2	3	6	13	30	20	7	3	1.5	0.4
	20	−0.15	2.2	3.5	7	14	32	19	7	4	1.5	0.6
		0	2.2	4	7	14	32	19	7	4	1.5	0.8
		0.15	2	4	7	13	31	19	7	4	1.5	0.8
0.75	5	−0.15	2	4	8	14	32	18	7	3	1	0.6
		0	2.2	5	8	14	33	18	7	3	1.5	0.4
		0.15	2.8	4	8	14	33	18	6	3	1.5	0.4
	10	−0.15	1.6	4	7	15	31	20	8	4	1.5	0.4
		0	1.6	5	8	15	30	22	8	4	2	0.4
		0.15	1.2	4	8	15	30	22	9	4	1.5	0.6
	20	−0.15	1.8	3.5	7	13	31	21	9	5	2	1
		0	1.2	3.5	6	13	30	22	8	4	1.5	0.8
		0.15	1.4	3.5	6	11	29	23	9	4	1.5	0.6

Table 3.5 – *One-sided nominal error rate (in %) for lower and upper credible intervals for the scale parameter τ for the censored likelihood estimator with $n = 250$ based on $B = 1000$ replications.*

α_0	D	ξ_0 | error rate	1	2.5	5	10	25	25	10	5	2.5	1
0.25	5	−0.15	0.2	0.5	1	2	5	69	48	36	25.5	17.2
		0	0	0.5	1	2	8	54	33	22	15	7
		0.15	0	1	2	4	14	42	22	13	6.5	2.6
	10	−0.15	0.4	1.5	4	8	21	28	12	7	3.5	1
		0	1	2	4	8	21	27	11	5	2	1
		0.15	1.4	2.5	4	8	23	25	10	4	2.5	0.8
	20	−0.15	0.4	1.5	3	7	21	30	13	6	3.5	1
		0	0.6	1.5	3	7	21	29	12	6	3	1.2
		0.15	0.8	1.5	3	8	21	27	11	6	3	1.2
0.50	5	−0.15	0.2	2	4	9	22	30	11	5	3	1.8
		0	0.8	2	4	10	24	28	11	5	2.5	1.4
		0.15	0.6	2	5	8	24	27	11	6	3	1.8
	10	−0.15	0.8	2	4	9	21	31	12	7	3.5	1.4
		0	0.8	2	4	9	21	30	12	7	3.5	1.2
		0.15	0.6	1.5	3	9	21	30	11	6	3.5	1
	20	−0.15	0.4	1.5	3	6	17	32	15	8	4.5	2.2
		0	0.4	1.5	4	7	20	32	13	7	4	1.8
		0.15	0.6	2	5	9	21	29	12	7	3.5	1.4
0.75	5	−0.15	1.4	2.5	4	8	22	26	11	6	2.5	0.6
		0	1.2	3	4	9	22	26	10	5	2.5	0.8
		0.15	0.6	2	5	9	24	25	9	4	2.5	1
	10	−0.15	0.6	2.5	4	8	18	32	13	7	3.5	1.4
		0	1	3	5	8	20	30	13	6	3	1.6
		0.15	1.6	3	5	9	22	29	12	6	3	1
	20	−0.15	0.8	2.5	4	8	20	31	15	8	4.5	2
		0	1.2	3	5	9	21	30	12	7	4	1.4
		0.15	1	3	5	10	23	28	12	6	3.5	1.2

Table 3.6 – *One-sided nominal error rate (in %) for lower and upper credible intervals for the shape parameter ξ for the censored likelihood estimator with $n = 250$ based on $B = 1000$ replications.*

α_0	D	ξ_0 | error rate	1	2.5	5	10	25	25	10	5	2.5	1
0.25	5	−0.15	0	0.5	1	2	6	72	56	45	37.5	27.2
		0	0	0.5	1	2	6	72	56	45	37.5	27
		0.15	0	0.5	1	2	6	72	55	44	37	26.8
	10	−0.15	1	2.5	5	9	23	27	12	7	4.5	2.6
		0	0.8	2.5	5	9	23	28	12	7	5	2.4
		0.15	0.8	2.5	5	10	23	28	12	7	4.5	2.6
	20	−0.15	1.4	2	5	9	22	27	12	6	2.5	1
		0	1.2	2.5	4	9	22	28	12	6	3	1.2
		0.15	1.2	2	4	9	22	28	12	6	3	1
0.50	5	−0.15	0.6	1	4	8	21	28	12	6	3.5	1.2
		0	0.6	1.5	4	8	20	28	12	6	3	1.4
		0.15	0.6	1.5	4	8	20	29	12	6	3	1.4
	10	−0.15	0.6	2	4	8	23	27	11	5	2.5	1.6
		0	0.6	2	4	8	23	27	11	5	2.5	1.6
		0.15	0.6	2	4	8	22	28	11	5	2.5	1.6
	20	−0.15	0.6	2	4	8	21	29	13	6	3.5	1.6
		0	0.8	2	4	8	21	30	13	6	3.5	2
		0.15	0.6	2	4	8	21	30	13	6	4	1.8
0.75	5	−0.15	0.4	1	3	9	23	30	13	6	3	1.2
		0	0.4	1	3	8	23	30	13	7	3	1.2
		0.15	0.4	1	3	8	22	31	14	7	3.5	1
	10	−0.15	0.6	2	4	8	21	31	13	6	3.5	1.4
		0	0.4	2	4	8	21	31	12	6	3.5	1.4
		0.15	0.6	2	4	8	21	31	12	7	3.5	1.4
	20	−0.15	0.6	1.5	4	9	21	30	11	6	3	1.6
		0	0.6	1.5	4	9	22	30	11	6	3	1.6
		0.15	0.6	1.5	4	9	21	30	11	6	3	1.4

Table 3.7 – *One-sided nominal error rate (in %) for lower and upper credible intervals for the dependence parameter α for the censored likelihood estimator with $n = 250$ based on $B = 1000$ replications.*

3.6 Simulation study: Bayesian hierarchical model

We conduct a simulation study to assess the performance of the Markov chain Monte Carlo algorithm that will be used in Section 3.7. We use data from a collection of MeteoSwiss rainfall time series to derive plausible parameter values and simulate realizations from a Brown–Resnick model with a power variogram $\gamma(\boldsymbol{h}) = (\|\mathbf{A}\boldsymbol{h}\|/\lambda)^{\alpha}$ and from an extremal Student model with a power exponential correlation function $\rho(\boldsymbol{h}) = \exp\{-\mathbf{A}\boldsymbol{h}\|/\lambda)^{\alpha}\}$, including geometric anisotropy in both. We obtained marginal parameters $(\boldsymbol{\sigma}, \xi)$ by fitting a generalized Pareto distribution to exceedances over the 90% of the site-wise rainfall series to all sites, with a common shape parameter or a site-specific shape. For the dependence structure, we used the extremogram (Section 2.7.2) to derive a scatter of pairwise estimates displayed in Figure 3.7. The model-based estimate (grey line) is obtained by minimizing the squared distance between the pairwise estimates $\widehat{\chi}(\boldsymbol{h})$ and the fitted curve: for example, for the Brown–Resnick process with power variogram γ, we found $\min_{\mathbf{A},\boldsymbol{\vartheta}} \|\widehat{\chi} - 2\Phi(\{\gamma(\mathbf{A}\boldsymbol{h};\boldsymbol{\vartheta})\})\|^2$ using constrained optimization. While we considered using the Schlather variogram model (B.1), the parameter estimates for the shapes (α, β) of the variogram model based on two-step modelling were nearly identical, i.e., $\widehat{\alpha} \approx \widehat{\beta}$, the extremogram curves were overlaid and preliminary runs from the Markov chain Monte Carlo algorithm yielded chains with a negative correlation of -0.98, indicating that one parameter was superfluous and the power variogram can adequately capture the dependence. There is strong indication of geometric anisotropy in the North-East direction; the parameter estimates of (2.3) are $a_2 = 2.15$ and $\varrho = 0.7$. The model-based extremograms for the Brown–Resnick and extremal Student models are nearly indistinguishable (Figure 3.7), except for short distances. While the extremal Student process is not mixing, meaning that $\lim_{h\to\infty} \theta(h) < 2$ (2.37) and does not lead to asymptotic independence even as the distance between site becomes infinite, it is difficult to detect asymptotic independence on a small region.

The simulation study considered four different models (Brown–Resnick versus extremal Student, common shape across the spatial domain or different shape parameters at each site) and, for each scenario, three datasets were generated. We selected 500 observations at 20 sites for inference because of the computational burden associated with the Markov chain Monte Carlo procedure. One easy way to assess whether the model is plausible is by verifying that summary statistics are well captured. Figure 3.8 shows the frequency at which a given site provides the maximum exceedance for the Swiss rainfall data at 25 sites, compared with the simulated data. The simulations show excellent agreement, with two values out of 25 falling outside the pointwise 95% confidence interval.

Description of the algorithm

We ran four parallel chains on each dataset, and estimated both Brown–Resnick and extremal Student models with common shape ξ in all cases, even if the underlying shape parameter is different for every site. To assess the effect of fixing the marginal parameters in a preliminary

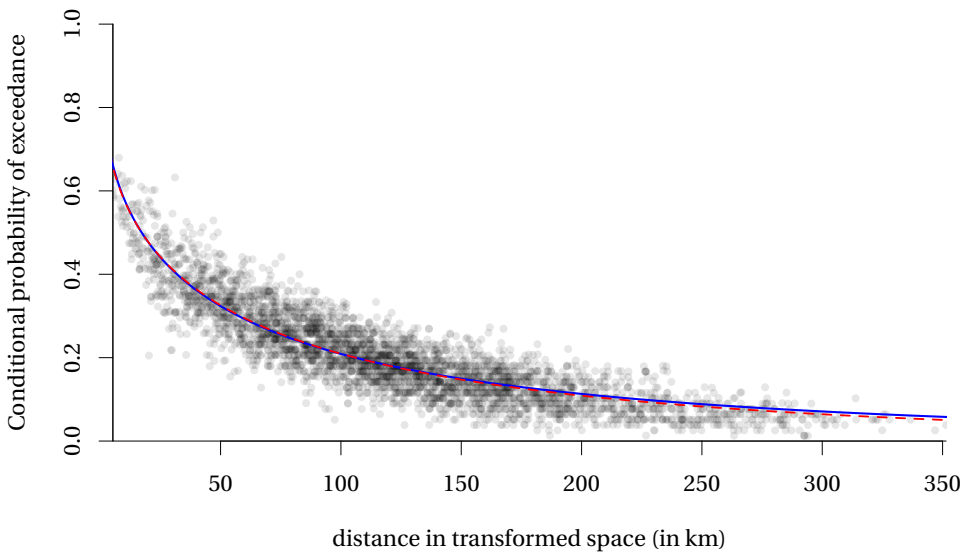

Figure 3.7 – *Extremogram for the Swiss rainfall data in the deformed space with the theoretical extremogram for the fitted Brown–Resnick process with power variogram (red dashed) and for the extremal Student process with power exponential correlation (blue full). The geometric anisotropy parameter estimates for the models are nearly identical and both curves are overlapping.*

step, as done by Thibaud (2014), we also ran latent Gaussian models to estimate scale and shape parameters, and then fixed these to the median of the posterior draws before running further MCMC steps for the dependence parameters alone. While plug-in estimators are often used in practice in frequentist estimation, they make little sense in the Bayesian framework: the marginal distribution of the ith component of a multivariate generalized Pareto model is typically not generalized Pareto distributed. Unlike in the data application, we did not include a spatial random effect and used a Bayesian linear model with independent components, meaning that the spatial dependence is captured through the mean only as the covariates include the spatial coordinates of the collection sites.

The extremal Student model with power exponential correlation function becomes a Brown–Resnick model with power variogram when the degrees of freedom $\nu \to \infty$ (Nikoloulopoulos et al., 2009). If we fit an extremal Student process to the Brown–Resnick model, the values of the marginal chain for the degrees of freedom are very large, sometimes larger than 10^5, and the chains mix poorly unless such values were observed during the adaptation phase during which the variance of the proposal is adapted tuned. The chains are sticky and we can observe very high correlation between the scale parameter and ν. Mixing would likely be improved by capping ν at 100, say, where extremal Student and Brown–Resnick models are indistinguishable. Because of the low sensitivity of the model, updates for ν are best performed on the log-scale and we adapt the algorithm for the data application to incorporate these features.

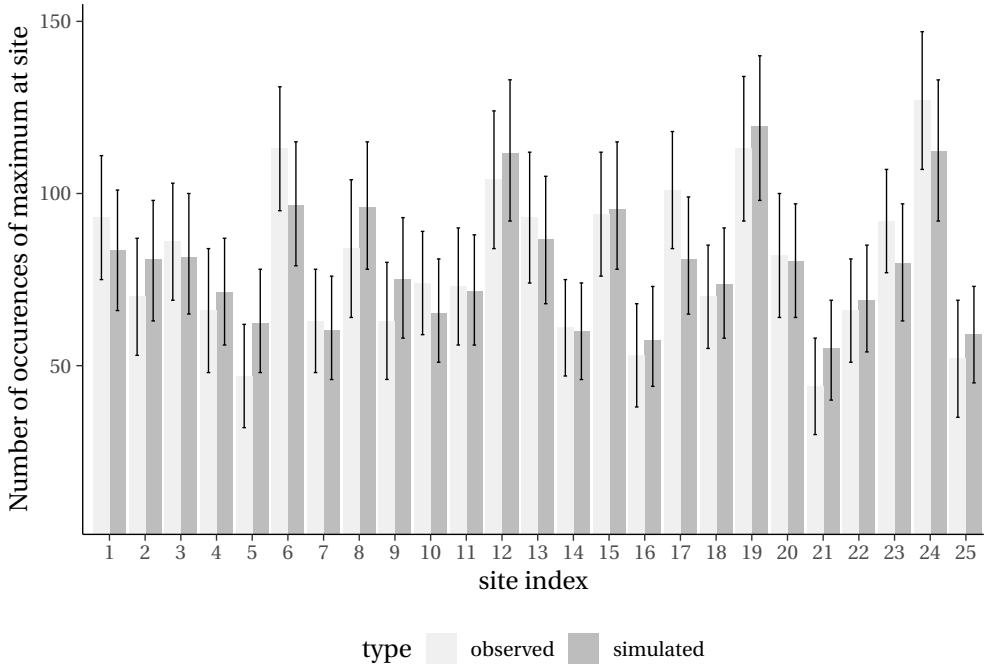

Figure 3.8 – *Bar plot of the frequency at which each of 25 sites has the largest exceedance for 2000 observations with 95% confidence intervals (vertical bars) for the Swiss rainfall data (pale grey) and simulated observations (dark grey) from a Brown–Resnick model generalized max-Pareto process with Schlather variogram. The marginal parameters used to simulate the data were estimated by fitting a generalized Pareto distribution at every site with a common shape parameter and the dependence parameters were obtained by using least squares and minimizing the distance between the empirical extremogram and the fitted curve.*

Estimation is challenging: as the model is doubly intractable, we can only calculate the gradient numerically through Monte Carlo, and this prevents use of more efficient algorithms such as Metropolis-adjusted Langevin or Hamiltonian Monte Carlo. We tried without success to implement the adaptive incremental mixture MCMC algorithm of Maire et al. (2019), which uses an independence sampler with a diffuse Gaussian proposal and adaptively adds Gaussian mixtures when observations sampled have both high posterior mass and low probability under the transition kernel. The proposal failed in the initialization phase, as the diffuse proposals completely missed the posterior mode and resulted in innefficient proposals; moreover, there is some evidence from the adaptive Metropolis-within-Gibbs output that the posterior is unimodal. Independence samplers have near zero acceptance rate in high dimensions and the mixture ignores the geometry. We also considered a Gaussian proposal for $\boldsymbol{\sigma}$ whose mean and variance are based on marginal generalized Pareto distribution, but the approach is less efficient than random walk Metropolis–Hastings updates because the proposed move may be systematically in the direction opposite to the posterior mode. Alternatives for pseudo-marginal methods include Lindsten and Doucet (2016), who transform auxiliary variables in order to derive a Hamiltonian Monte Carlo algorithm, but the approach is impractical because

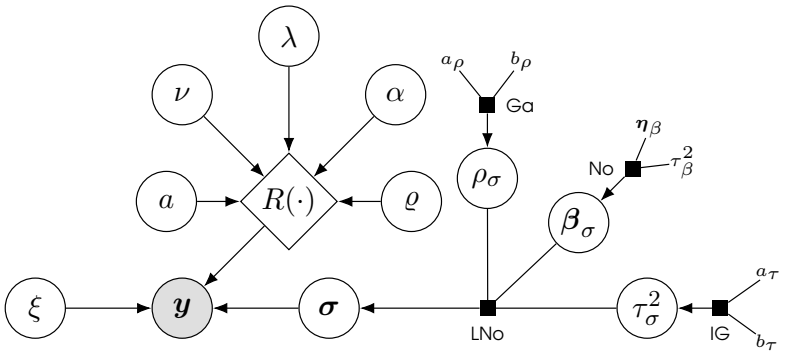

Figure 3.9 – *Directed acyclic graphical representation of the Bayesian model for the extremal Student multivariate generalized Pareto distribution with powered exponential correlation function $R(\cdot)$.*

of the quasi Monte Carlo scheme we use and is unlikely to work given the large number of augmented variables. We thus resorted to Metropolis-within-Gibbs: the marginal parameters $\boldsymbol{\sigma}$ were sampled in groups of 2 or 3 sites using a random scan (Liu et al., 1995) to reduce the number of likelihood evaluations. We used truncated Gaussian proposals for the scales and the shape parameters to ensure the proposals satisfied the boundary constraints of the generalized Pareto distribution marginally. The parameters of the Gaussian linear model for $\log(\boldsymbol{\sigma})$ were updated using Gibbs sampling and the sampler described in Section 3.4. A summary of the hierarchical model is given in Figure 3.9 and the prior specification is reported in Table 3.8; we used conjugate priors for the latent Gaussian model and vague priors for the scale and range parameters. The values of the shape parameter ξ reported in the literature are more often than not positive and small and Serinaldi and Kilsby (2014), using extensive records, report an average of around 0.1 for the shape of the generalized Pareto distribution fitted to more than 15 000 sites. Since the shape is hard to estimate, we use an informative prior for the latter and set $p(\alpha/2) \sim \mathsf{Be}(60,40)$ in the data application, reflecting previous studies (Davison et al., 2012; Thibaud and Opitz, 2015) and $p(\alpha) \sim \mathsf{No}(0,0.2)$ in the simulation study. The Beta prior is centered at 0.1 with approximate standard deviation of 0.05. The difference in log-likelihood between $\xi = 0.1$ and $\xi = 0.2$ is -10, whereas it increases to -30 if $\xi = 0.4$ relative to the mean of the prior. We standardized the covariates, namely longitude, latitude and log elevation, to have zero mean and unit standard deviation and likewise specified the prior for the reciprocal scale λ after standardizing by the average distance between sites. For the shape parameter of the variogram or power exponential correlation matrix, we used a Gaussian prior centered at zero, $\alpha/2 \sim \mathsf{Be}(4,4)$ to avoid fields that are too rough or too smooth, both of which are not observed in practice.

Adaptive methods are used to adjust the variance of the proposals within the Metropolis-within-Gibbs scheme: we tuned it to yield acceptance rates of about 0.44 for individual parameters and 0.234 for multivariate proposals, based on results for Gaussian distributions (Rosenthal, 2011). After exploring for 1000 iterations, we updated the covariance every 200 iterations by setting $\boldsymbol{\Sigma} \leftarrow \varepsilon\boldsymbol{\Sigma} + (1-\varepsilon)\{\mathsf{Cov}\left(\{\boldsymbol{X}^{(b)}\}_{b-1000}^{b}\right) + \varepsilon\mathbf{I}\}$ for $\varepsilon \in (0,1)$; the ridge $\varepsilon\mathbf{I}$ ensures

Brown–Resnick (1)	$\alpha/2 \sim \mathsf{Be}(2,2)$,	$\widetilde{d}/\lambda \sim \mathsf{HNo}(0,2^2)$,	$2-\beta \sim \mathsf{Ga}(2,2)$,
Brown–Resnick (2)	$\alpha/2 \sim \mathsf{Be}(2,2)$,	$\widetilde{d}/\lambda \sim \mathsf{HNo}(0,5^2)$,	
Extremal Student	$\alpha/2 \sim \mathsf{Be}(4,4)$,	$\widetilde{d}/\lambda \sim \mathsf{HNo}(0,5^2)$,	$\nu-1 \sim \mathsf{Ga}(3,3)$,
Shape and anisotropy	$\xi+1/2 \sim \mathsf{Be}(60,40)$,	$a-1 \sim \mathsf{HNo}(0,1.5^2)$,	$\varrho \sim \mathsf{U}(-\pi/2,\pi/2)$,
Latent Gaussian model	$\boldsymbol{\beta}_\sigma \sim \mathsf{No}_4(0,5\mathbf{I}_4)$,	$\tau_\sigma^2 \sim \mathsf{IG}(0.5,0.1)$,	$\rho_\sigma \sim \mathsf{Ga}(2,2/\widetilde{d})$.

Table 3.8 – *Prior specification for the Bayesian hierarchical model used in the data analysis, for the Brown–Resnick model with Schlather variogram (1) and with a power variogram (2) and the extremal Student model with a power exponential correlation function.* No*: Gaussian,* IG*: inverse gamma,* HNo*: half-Gaussian, i.e., truncated Gaussian on* $(0,\infty)$*,* HCa*: half-Cauchy, i.e., truncated Cauchy on* $(0,\infty)$*;* LNo*: log-Gaussian. The gamma distribution is parametrized in terms of shape and scale and we denote the median distance between sites by* $\widetilde{d}$*.*

that the covariance matrix does not collapse to $\mathbf{O}$. We experienced with alternatives proposed in Andrieu and Thoms (2008), but found our simple scheme to work better in practice while being relatively robust. The parameters of the variogram/correlation function, together with the degrees of freedom for the extremal Student process, were updated independently only for the first 3000 iterations and we adapted the marginal variances during that period. Some of the parameters are strongly correlated a posteriori, justifying the need for joint updates and we proposed from multivariate truncated Gaussian distribution. Adaptive proposals reduce manual tuning, but can lead to poor samplers, particularly if the chain has not reached stationarity during burn-in and this difficult to monitor.

Convergence assessment

We can assess the convergence of the algorithm by running standard tests and looking at trace plots. For all cases, including misspecified models in which the shape parameter is wrongly assumed to be constant in space and/or the dependence model is not that of the data generating mechanism, the model captures most of the dependence dynamics. This is unsurprising given the negligible differences between the fitted models reported in Figure 3.7. Except for the cases in which the extremal Student model was fitted to the Brown–Resnick and the estimated degrees of freedom are very large, the scale reduction factors were equal to 1 for every scenario. We also performed Geweke tests (Geweke, 1992), which failed for some parameters, but never for more than one of the four chains. Mixing is relatively good: Figure 3.10 shows trace plots for selected parameters from the posterior of an extremal Student model fitted to simulated data from the same model and all chains appear to have reached stationarity. The left panel of Figure 3.11 shows kernel density estimates for the four chains for dependence parameters, with true value marked by a black dot. All of the data generating parameters lie within the bulk of the posterior density estimates and the density curve overlap, indicating agreement between the posterior marginals estimated by each chain. The sample autocorrelation of the Markov chain (right of Figure 3.11) show geometric decay and suggest that samples at lag 100 are approximately independent. The parameters that are harder to sample are often highly correlated (Figure 3.12), but bivariate density plots suggest that the

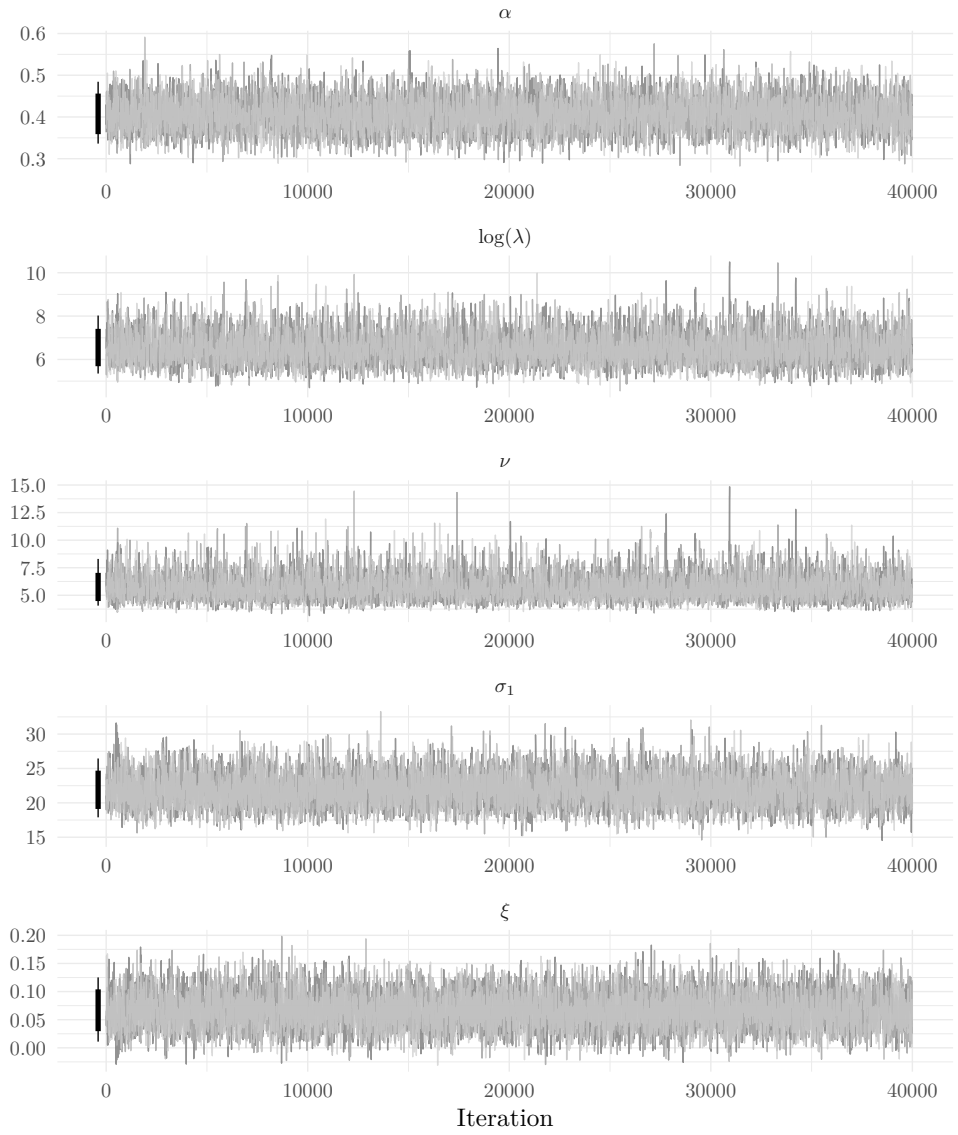

Figure 3.10 – *Traceplots of the posterior draws (post burn-in) from the extremal Student model based on four chains. From top to bottom: shape and log-scale for the powered exponential correlation function, degrees of freedom, scale parameter at the first site and common shape parameter. The bars on the left mark the 25% and 75% and 10% and 90% of the marginal posterior draws.*

joint posterior distribution is not degenerate.

The effective sample size varies a lot: the parameters of the linear model, updated using Gibbs, are nearly independent, but the effective sample size for the other parameters is between five and ten percent. Some of the dependence parameters of the variogram/correlation model, which are updated jointly, are more difficult to sample and the shape parameter α has effective sample size of 1.5% in the worst case scenario: with 160 000 draws, this is sufficient to estimate

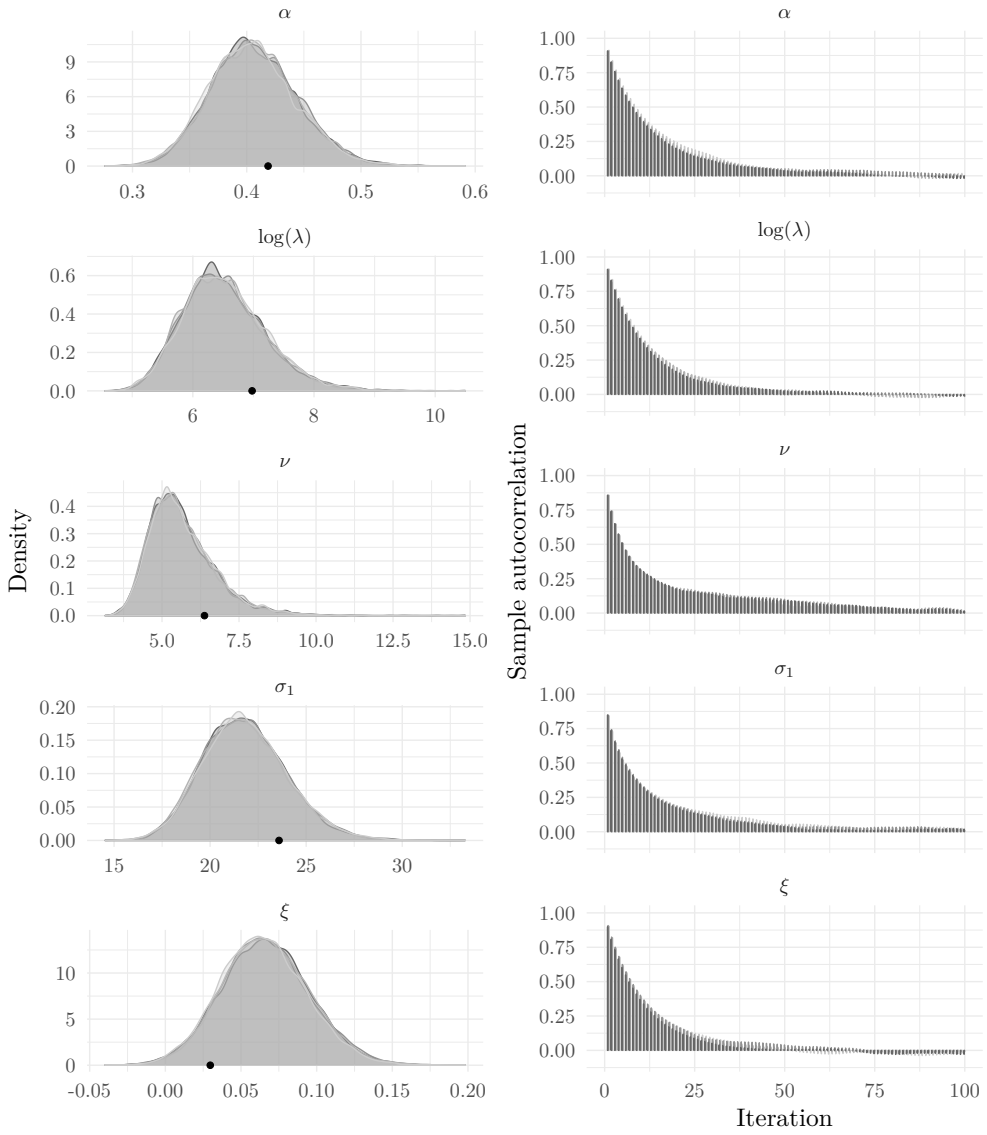

Figure 3.11 – *Density plots (left) and correlograms (right) for the posterior draws (post burn-in) from the extremal Student model based on four chains. From top to bottom: shape and log-scale for the powered exponential correlation function, degrees of freedom, scale parameter at the first site and common shape parameter. The true parameter are indicated by dots on the x-axis of the plots in the left panel.*

95% marginal credible intervals reliably.

Comparison between the parametric models based on the samples from the posterior is not straightforward. If the model is correctly specified, we can compare the parameter values used to generate the synthetic observations with the posterior quantiles. Tables 3.9 and 3.10 give the marginal posterior quantiles for runs from the correct model, which shows that both the two-stage approach and the full likelihood yield comparable output for the different

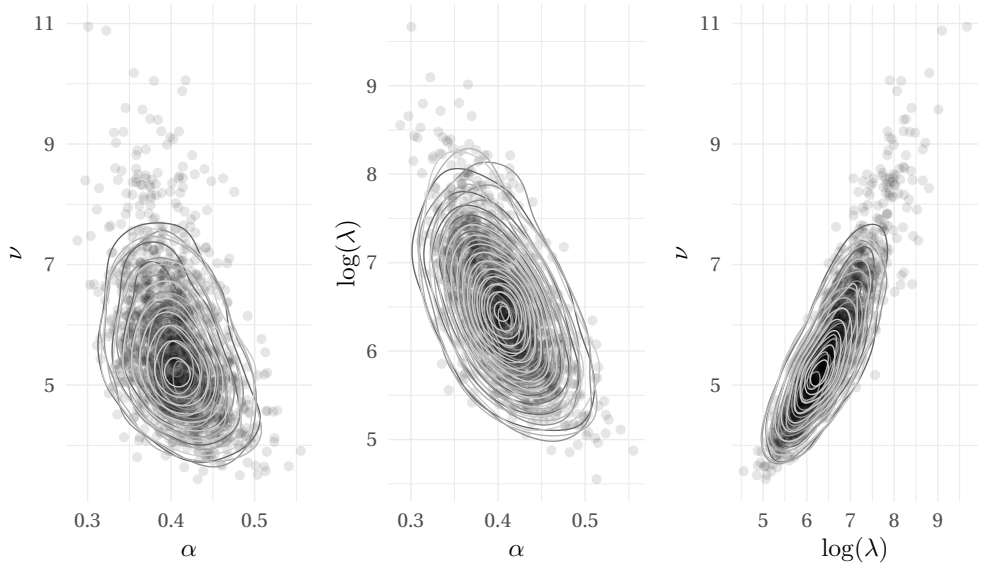

Figure 3.12 – *Random subsamples of the posterior draws (post burn-in) from the extremal Student model and bivariate kernel density (contour lines) for the four chains for the shape and log-scale for the powered exponential correlation function,* (α, λ) *and the degrees of freedom,* ν*.*

simulated datasets. The generalized Pareto likelihood should yield consistent estimators of the marginal parameters. Because there are more than 30 parameters in the models, we report the distribution of the marginal scale parameters $\boldsymbol{\sigma}$ graphically and focus on the more challenging dependence parameters and ξ. To avoid cluttering, we show density estimates of the posterior distribution of the scale parameters, standardized by dividing the draws by the value of $\boldsymbol{\sigma}$ used to generate the series. On the basis of this visual inspection, the distributions appear well calibrated for all methods, regardless of the dependence structure, but the estimated scale parameters for the extremal Student appear to be systematically lower than their counterparts from latent Gaussian models and the posterior distribution is more concentrated than for draws obtained by fitting a Brown–Resnick model. The draws from the latent Gaussian model are in line with the others; the distribution of $\boldsymbol{\sigma}$ is not very different whether we use a generalized Pareto distribution with common shape (dots) or the multivariate generalized Pareto model, suggesting that both carry the same amount of information about the marginal parameters.

Conditional predictions from the hierarchical model at holdout sites can be obtained by forward sampling: conditional on the current posterior draws from $\boldsymbol{\sigma}, \xi$, we transform the observation vector to the unit Pareto scale and simulate the value of the process at the holdout sites from a log-Gaussian process, the parameters being derived from the conditional intensity of the Brown–Resnick process. To back-transform observations to the observation scale, marginal parameters are sampled from the trend surface $\mathsf{No}(\mathbf{X}\boldsymbol{\beta}_\sigma, \tau_\sigma^2 R(\cdot, \rho_\sigma))$ conditional on the current values of $\boldsymbol{\sigma}$. For unconditional predictions, we replace the conditional simulation

		ξ		α		λ		a		ϱ	
rep.	% |	FULL	2S	FULL	2S	FULL	2S	FULL	2S	FULL	2S
1	2.5	0	−0.06	0.75	0.76	21	25.25	1.98	1.98	0.49	0.48
	10	0.02	−0.04	0.77	0.78	23	26.5	2.06	2.06	0.51	0.51
	50	0.04	0	0.81	0.81	26.75	28.5	2.25	2.23	0.56	0.56
	90	0.07	0.04	0.85	0.85	30.75	30.75	2.44	2.42	0.61	0.6
	97.5	0.09	0.06	0.87	0.87	33	32	2.55	2.53	0.63	0.63
2	2.5	−0.05	−0.06	0.67	0.68	17.5	21.25	2.17	2.12	0.53	0.53
	10	−0.04	−0.05	0.69	0.69	19.25	22.5	2.28	2.22	0.56	0.56
	50	−0.01	−0.01	0.72	0.73	23	24.75	2.48	2.42	0.6	0.6
	90	0.02	0.02	0.76	0.77	27	27	2.72	2.64	0.65	0.65
	97.5	0.03	0.04	0.78	0.79	29.5	28.25	2.85	2.77	0.67	0.67
3	2.5	0.02	−0.04	0.65	0.65	15.25	16.75	1.92	1.91	0.61	0.62
	10	0.03	−0.02	0.67	0.67	16.75	17.75	2.02	2	0.64	0.65
	50	0.06	0.01	0.7	0.7	20	19.75	2.21	2.19	0.7	0.71
	90	0.1	0.05	0.74	0.74	23.75	21.5	2.43	2.4	0.76	0.76
	97.5	0.12	0.07	0.76	0.76	26	22.75	2.56	2.52	0.79	0.8

Table 3.9 – *Posterior marginal quantiles using the full likelihood* (FULL) *and two-stage estimation* (2S) *across three replicate datasets of size* $n = 500$ *with* $D = 20$ *sites using a Brown–Resnick process with a correctly specified likelihood. The data generating process values are* $\xi = 0.03$, $\alpha = 0.73$, $\lambda = 22$, $a = 2.36$ *and* $\varrho = 0.61$.

		ξ		α		λ		ν		a		ϱ	
rep.	% |	FULL	2S	FULL	2S	FULL	2S	FULL	2S	FULL	2S	FULL	2S
1	2.5	−0.05	−0.06	0.35	0.35	0	0	6.5	6.3	1.74	1.71	0.51	0.51
	10	−0.03	−0.04	0.37	0.37	0	0	7.4	7.1	1.89	1.86	0.56	0.56
	50	0	0	0.41	0.41	50	50	10	9.6	2.23	2.19	0.67	0.67
	90	0.03	0.04	0.46	0.46	50	50	14.8	14.2	2.65	2.58	0.78	0.78
	97.5	0.05	0.06	0.48	0.48	100	150	18.7	18.3	2.91	2.81	0.84	0.84
2	2.5	−0.05	−0.02	0.39	0.38	0	0	4.8	4.9	1.55	1.52	0.41	0.41
	10	−0.03	0	0.41	0.4	0	0	5.3	5.4	1.67	1.64	0.47	0.48
	50	0.01	0.03	0.46	0.45	0	0	6.7	6.7	1.94	1.91	0.59	0.59
	90	0.05	0.07	0.51	0.5	50	50	8.9	8.9	2.27	2.22	0.7	0.72
	97.5	0.07	0.09	0.54	0.52	100	50	10.7	10.7	2.47	2.41	0.77	0.79
3	2.5	0.01	−0.02	0.34	0.33	0	0	4.1	4.7	1.99	1.91	0.48	0.47
	10	0.03	0	0.36	0.35	0	0	4.5	5.2	2.16	2.07	0.52	0.51
	50	0.07	0.03	0.4	0.39	0	0	5.5	6.5	2.55	2.43	0.61	0.61
	90	0.1	0.07	0.46	0.44	50	50	7	8.5	3.03	2.84	0.71	0.7
	97.5	0.13	0.1	0.48	0.46	50	50	8.3	10.2	3.32	3.09	0.76	0.76

Table 3.10 – *Posterior marginal quantiles using the full likelihood* (FULL) *and two-stage estimation* (2S) *across three replicate datasets of size* $n = 500$ *with* $D = 20$ *sites using an extremal Student process with a correctly specified likelihood. The data generating process values are* $\xi = 0.03$, $\alpha = 0.42$, $\lambda = 1073$, $\nu = 6.4$, $a = 2.36$ *and* $\varrho = 0.61$.

from the Brown–Resnick process by an unconditional simulation, which would allows us to compute the CRPS score, which we did not attempt due to lack of time.

The dependence parameters are meaningless on their own if the model is misspecified, so

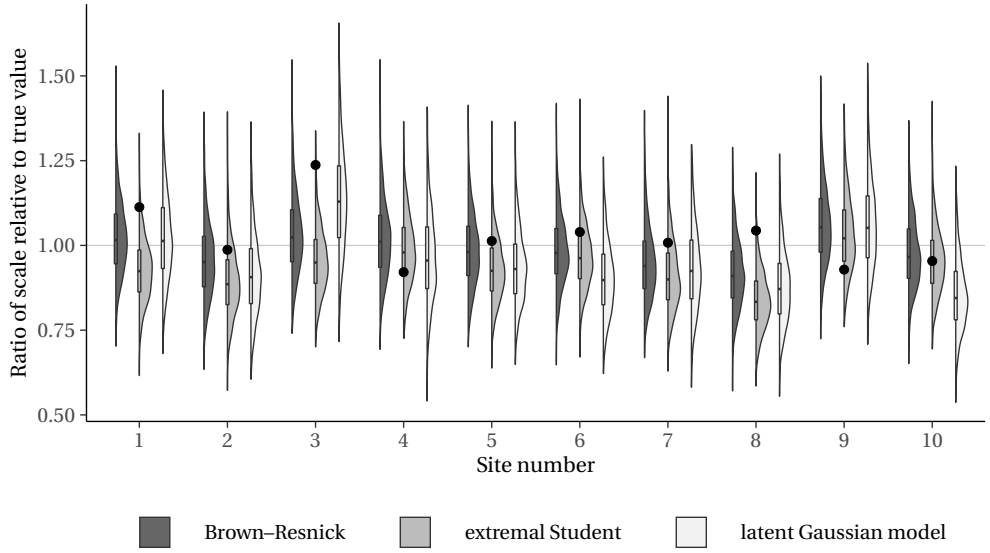

Figure 3.13 – *Kernel density estimates and boxplots for the posterior distribution of the marginal scale parameters at selected sites based on data simulated from an extremal Student process. From left to right: Brown–Resnick model (dark grey), extremal Student (grey) and latent Gaussian model (light grey). The dots on the left indicate the maximum likelihood estimates of the marginal shape obtained by fitting independent generalized Pareto distributions to each sites with a common shape parameter.*

we consider the spatial conditional probability of exceedance in Figure 3.14. There are few discernible differences: the credible bands for the extremogram curves obtained from the joint fit using the envelope method (Davison and Hinkley, 1997, pp. 153–154) are wider than the two-stage model ones. If the true model is extremal Student, but we fit a Brown–Resnick model, the joint credible interval does not capture the true curve.

The goal of the simulation study was to assess whether the Markov chain Monte Carlo used for the data applications in Section 3.7 converges to the stationary posterior distribution of interest. Since we know the data generating mechanism, it is possible to relate marginal parameter estimates of the posterior to the true values; we find that most parameters are adequately captured, as illustrated by Figure 3.13. The convergence diagnostics, such as Figure 3.11, indicate good mixing for the Markov chain sampler and all of the chains have converged to the same stationary distribution. The two-stage method and the full likelihood seem consistent and give similar results (Tables 3.9 and 3.10). The parameter estimates reported in the tables and graphically in Figure 3.13 show that, although the two-stage approach and the full likelihood do not have the same point estimates, the marginal posterior variance is approximately the same. This implies that, for a correctly specified model, little is gained from estimating margins and dependence structure simultaneously. The point estimates of the two-stage method need not give a local or global optimum, but the marginal adjustment as assessed by quantile-quantile plots based on the fit of a generalized Pareto distribution

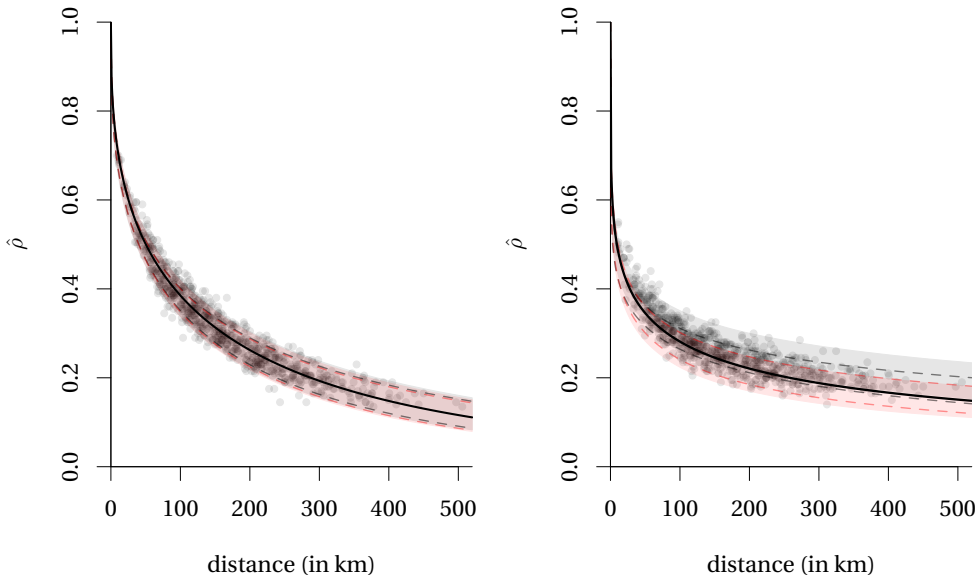

Figure 3.14 – *Extremogram of simulated data from a Brown–Resnick process (left) and from an extremal Student process (right), with 95% simultaneous credible intervals for the curve based on the fitted Brown–Resnick (grey shaded), extremal Student (red shaded). The credible intervals for the parameter estimated using the two-stage approach are given by dashed curves (same color code).*

to site-wise threshold exceedances is likely to be better. While analysis of the Markov chain Monte Carlo sampler on simulated data from the model yielded comparable output for the joint likelihood and the two-stage approaches (Tables 3.9 and 3.10), we have no guarantee of the optimality of the parameter estimates resulting from the two-stage procedure: joint modelling can lead to very different marginal estimates of $\boldsymbol{\sigma}, \xi$.

3.7 Data applications

3.7.1 Zürich rainfall

We fit a Bayesian hierarchical model to the Zürich rainfall data used in Davison et al. (2012) and in Thibaud and Opitz (2015). The latter consists of daily cumulated rainfall between June 1st and August 31st for the summer months recorded at 44 stations in central Switzerland between 1962 and 2012. The so-called Plateau is a small region and it may be difficult to capture weakening spatial dependence because the maximum distance between sites is small. The location of the measurement sites are indicated with crosses on Figure 3.15; we retain 20 sites for inference, leading to approximately 13 multivariate exceedances per year. Thibaud and Opitz (2015) used an extremal Student model, but fixed the margins to the maximum a posteriori from a latent Gaussian model using the Poisson likelihood for threshold exceedances given in Proposition 1.4 with Gaussian process priors on the parameters, as in eq. (3.8). Thibaud (2014) compared the results from running a latent Gaussian model with

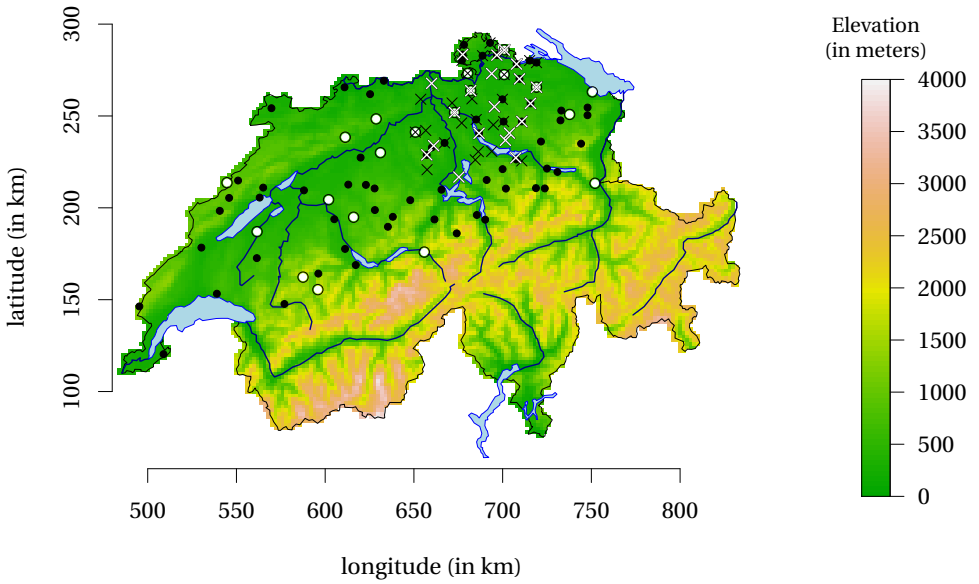

Figure 3.15 – *Location of the measurement stations of the Zürich rainfall dataset of Thibaud and Opitz (2015) (crosses) and of the Swiss rainfall dataset (circles), for which longer time series are available. Stations used to fit the model are colored in white and hold-out sites in black.*

	$\beta_{\sigma,0}$	$\beta_{\sigma,1}$	$\beta_{\sigma,2}$	$\beta_{\sigma,3}$	τ^2_σ	ρ_σ
5%	2.1	−0.07	−0.16	−0.06	0.01	18
median	2.3	0.02	−0.06	0	0.03	46
95%	2.5	0.12	0.04	0.05	0.05	95

Table 3.11 – *Posterior median with 90% credible interval for the regression parameters of the latent Gaussian model with a common shape parameter for the joint fit of the Brown–Resnick model with a power variogram to the Zürich dataset. The regression parameters $\boldsymbol{\beta}_\sigma$ correspond to the intercept, longitude, latitude and log altitude, standardized to have zero mean and unit variance.*

adjusted likelihood curvature following Ribatet et al. (2012); writing that the latter mix poorly, leads to point estimates for the latter centered at different values and with wider "credible intervals" for the adjusted likelihood.

The parameter estimates are given in Table 3.11 and Figure 3.16 shows an example of a simulated realization of $\sigma(\boldsymbol{s})$ over the domain based on the log-Gaussian hierarchical model (left panel) along with the median estimated surface from the latent Gaussian model. The log-Gaussian distribution is skewed to the right, which can lead to very large scale parameters. The default log link is not a panacea, but alternatives such as identity or reciprocal link functions require sampled components from truncated Gaussian components, which are more expensive to compute. If the parameters are far from zero and the variance is small, this may be an interesting alternative.

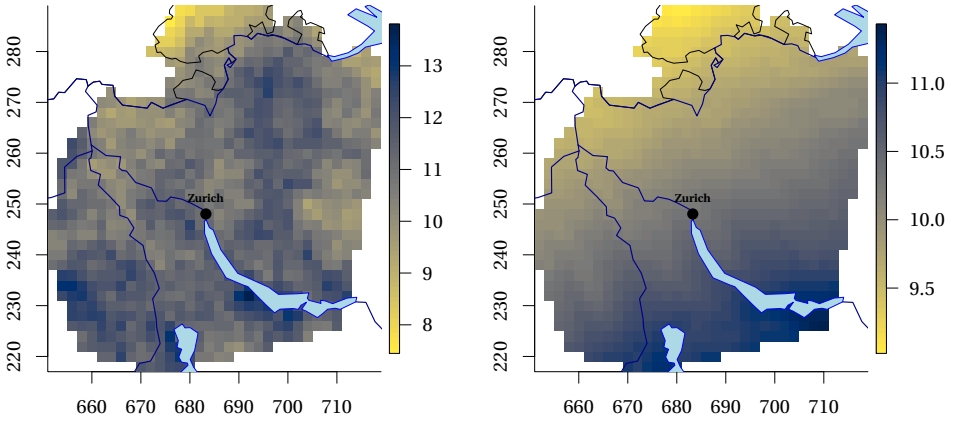

Figure 3.16 – *Simulation of the fitted log-Gaussian surface for the scale over the convex hull of the 20 sites used for inference (left) and pointwise estimate for the regression surface based on median posterior of* $\boldsymbol{\beta}_\sigma$.

We consider joint estimation based on a hierarchical model similar to the one described in Section 3.6 for 20 sites and consider marginal threshold exceedances above the 97.5% of the series, yielding a total of 588 observations for the Zürich data. To obtain sensible starting values, we fitted generalized Pareto distributions to marginal threshold exceedances with a common shape parameter and fitted a variogram model by minimizing the l_2 distance between the cloud of fitted tail correlation coefficient $\widehat{\chi}$ and the theoretical curve for the selected model. We ran 10 parallel chains for 55 000 iterations, discarding the first 10 000 for burn-in and stopping adaptation after 15000. Trace plots (not shown) are stable for all parameters and the marginal distribution of the parameters appear unimodal. The effective sample size for the scale parameters is on average 15000 for the Brown–Resnick models and about 5000 for the extremal Student model; those for the dependence parameters are reported in Tables 3.13 to 3.15. We used the separation of variables estimator for the Gaussian and Student distribution functions, averaging over 12 runs based on a sample of 307 observations, whereas we increase the precision for the Gaussian distribution functions appearing in the exponent measure by taking 499 observations to reduce the bias. The variance of the likelihood estimator is around 0.2 and a single evaluation of the log-likelihood at 20 sites takes about a second. While it may advisable to use parallel computing when the number of sites D is in the hundreds to split the evaluation of Gaussian distribution functions between cores, the communication costs make this option impractical with our cluster architecture.

To assess the sensitivity of the model to prior selection, we ran Markov chains with different priors on the Zürich data, doubling or halving the precision of the parameters. For the Bayesian linear model, we used the two-stage approach to reduce the computational burden. Changing the priors $a_\tau, b_\tau, a_\rho, b_\rho, \tau^2_\beta$ had no impact on the distribution of the hyperparameters $\boldsymbol{\beta}_\sigma, \rho_\sigma, \tau^2_\sigma$: the magnitude of the differences was about 0.01 for τ^2_σ and 1 for ρ_σ, whose posterior median is 144; all these changes are within Monte Carlo variability and hint to the fact that the likelihood

σ_1	ξ	α	λ^α	a	ϱ	ν
10.60	0.286	0.60	4.21	1.766	0.580	
10.62	0.286	0.61	4.24	1.764	0.581	
10.74	0.288	0.61	4.36	1.770	0.580	
10.57	0.287	0.60	4.18	1.764	0.582	
10.71	0.436	0.35	165.40	2.074	0.637	5.98
10.57	0.436	0.36	168.35	2.050	0.634	5.98
10.66	0.435	0.35	166.20	2.070	0.633	5.98
10.73	0.435	0.35	167.65	2.060	0.634	5.98
10.64	0.436	0.35	166.65	2.066	0.636	6.00

Table 3.12 – *Prior sensitivity analysis for the Zurich data: marginal posterior mean for the parameters of the Brown–Resnick model with power variogram (first four rows) and the extremal Student with power exponential correlation (last five rows). The priors relative to the model specification are* $\alpha/2 \sim \mathsf{Be}(1,1)$ *(first and fifth rows),* $\alpha/2 \sim \mathsf{Be}(4,4)$ *(second and sixth rows)* $\bar{d}/\lambda \sim \mathsf{HNo}(0, 2.5^2)$ *(third and seventh row),* $\bar{d}/\lambda \sim \mathsf{HNo}(0, 10^2)$ *(fourth and eigth rows),* $\nu - 1 \sim \mathsf{Ga}(3,3)$ *(ninth row).*

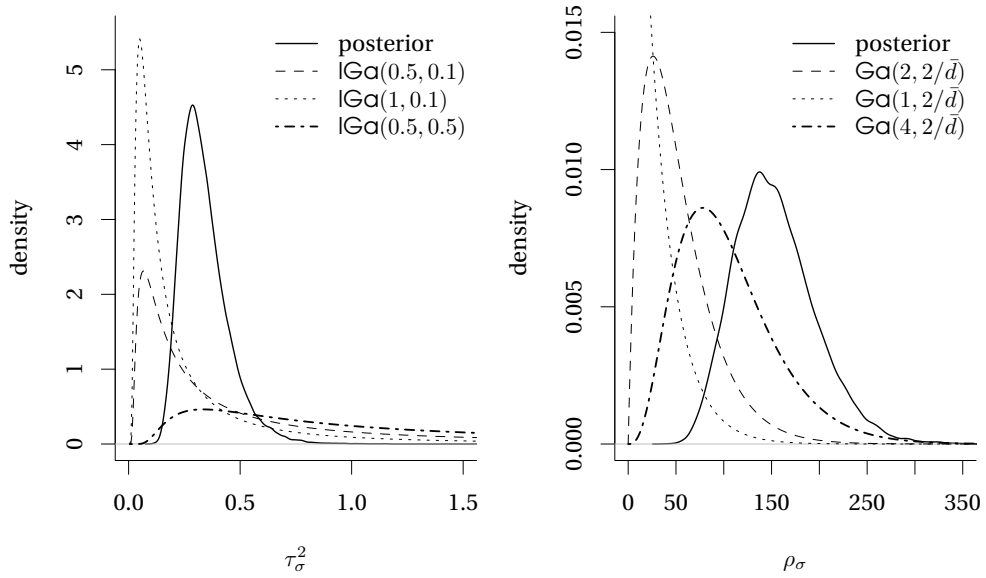

Figure 3.17 – *Posterior kernel density estimates with various priors density for the Zürich data with a latent Gaussian model for the variance (left) and scale parameters (right) of the exponential covariance model for the spatial random effect model; the prior used in the data analysis is the first specified in the legend (dashed).*

contribution dominates the posterior, with the prior having little to no impact except for ξ. Figure 3.17 shows that, despite very large changes in the priors, the posterior estimates are unaffected when the support of the prior and that of the posterior overlap. Table 3.12 shows the impact of changes in the prior specification for the dependence parameters of the Brown–Resnick and extremal Student models, relative to the specification outlined in Table 3.8 for the full model for selected updates. The only informative prior is that of the

method	$\mid \theta$	σ_1	ξ	α	λ	a	ϱ
full	5%	9.48	0.26	0.57	8.8	1.6	0.5
	median	10.57	0.29	0.6	10.7	1.76	0.58
	95%	11.82	0.32	0.64	13	1.94	0.66
	ESS(×0.01)	152	428	384	68	434	578
two-stage	5%	8.93	0.1	0.54	5.41	1.59	0.5
	median	9.97	0.13	0.57	6.08	1.75	0.58
	95%	11.12	0.17	0.61	6.79	1.93	0.66
	ESS(×0.01)	928	566	414	376	250	352

Table 3.13 – *Posterior median, 90% marginal credible intervals, and effective sample size (ESS) for the parameters of the Brown–Resnick model with a power variogram for the Zürich rainfall dataset.*

shape, whereby $p(\xi + 1/2) \sim \mathsf{Be}(60,40)$; since the sites share a common shape parameter, the truncation allows for some regularization, so as to avoid regions of the parameter space where the estimation of the log-likelihood becomes numerically unstable or the shape parameter values are implausible. The difference in posterior when the value of ξ increases is so large that the prior has a small impact. Despite the very large shape estimates, the Markov chain Monte Carlo sampler seem to have reached stationarity, even if mixing is poorer than for simulated data of Section 3.6. The latent Gaussian model however has lower autocorrelation and larger jump distance: we use a Mahalanobis distance, with a diagonal covariance matrix based on marginal variance of the posterior for the scale to make comparison meaningful. The jump distance is 2.22 for the 20 scale parameters for the latent Gaussian model with 22.5% of effective samples, compared to a jump distance of 1.06 and 2.2% of effective samples for the full Bayesian hierarchical model. Quantiles of the sample draws from the posterior are reported in Tables 3.13 to 3.15; note the very large difference with the estimates for the scale and the shape reported for the two-stage fit. For the Brown–Resnick model with the Schlather variogram, there is strong negative correlation between posterior draws for (α, β). A reparametrization of the model parameters would be advisable, but both parameters are constrained and the constraint linking (α, β) is nonlinear.

Marginal goodness-of-fit diagnostics for the training data and at holdout sites are given through marginal Bayesian quantile-quantile plots (Figure 3.18) for the 10 first sites, which show that the fitted model overpredicts the largest observations except sites 4 and 7. The high shape parameter leads to poor fit in the lower tail of the distribution of the extremes at the holdout sites. For the fourth holdout site shown in the third row of Figure 3.18, the extremes seem to originate from a mixture.

We also plot the fitted conditional probability of exceedance curve in Figure 3.19, including pairwise estimates $\widehat{\chi}$ for the 24 holdout sites. All models completely fail to capture the linear rate of decay of dependence with distance. The two-stage approach credible bands are too narrow and underestimate the dependence. Both models seem to adjust poorly to the data;

method	$\mid \theta$	σ_1	ξ	β	α	λ	a	ϱ
full	5%	9.66	0.26	0.57	0.14	9.43	1.62	0.51
	median	10.78	0.29	0.88	0.41	11.41	1.78	0.58
	95%	12.04	0.32	1.15	0.64	13.84	1.97	0.66
	ESS(×0.01)	144	388	182	172	56	406	560
two-stage	5%	8.93	0.1	0.46	0.13	5.64	1.62	0.5
	median	9.96	0.13	0.71	0.47	6.33	1.79	0.58
	95%	11.12	0.17	1.07	0.7	7.06	1.97	0.65
	ESS(×0.01)	896	560	98	92	128	210	344

Table 3.14 – *Posterior median, 90% marginal credible intervals, and effective sample size (ESS) for the parameters of the Brown–Resnick model with a Schlather variogram for the Zürich rainfall dataset.*

method	$\mid \theta$	σ_1	ξ	α	λ^α	a	ϱ	ν
full	5%	9.56	0.42	0.31	133.95	1.68	0.51	5.15
	median	10.67	0.44	0.35	166.16	2.04	0.63	5.96
	95%	11.92	0.45	0.4	206.81	2.51	0.76	6.97
	ESS(×0.01)	50	666	298	178	374	518	300
two-stage	5%	8.93	0.1	0.29	92.45	1.67	0.52	5.49
	median	9.96	0.13	0.33	110.18	2.04	0.65	6.14
	95%	11.12	0.17	0.37	131.36	2.52	0.78	6.9
	ESS(×0.01)	896	566	90	120	198	304	116

Table 3.15 – *Posterior median, 90% marginal credible intervals, and effective sample size (ESS) for the parameters of the extremal Student model with a power exponential correlation function for the Zürich rainfall dataset.*

this puzzling finding requires further investigation to better understand the causes of these discrepancies.

3.7.2 Swiss rainfall

We consider a complementary dataset with 86 time series from MeteoSwiss running between 1901 and 2016, keeping observations between May 1st and September 30th to ensure approximate stationarity, thresholding and censoring observations falling below the 98% and using the first 1000 observations threshold exceedances at 20 sites for inference to ensure reasonable computing time. All stations, shown with circles in Figure 3.15, are located north of the Alps.

The marginal posterior estimates for the shape parameter obtained from fitting the Brown–Resnick model to the Swiss rainfall data are much more reasonable, but the shape parameters are still high; marginal posterior estimates for the dependence parameters and for ξ are given in Tables 3.16 to 3.18. We only make the comparison with the two-stage approach for the Brown–

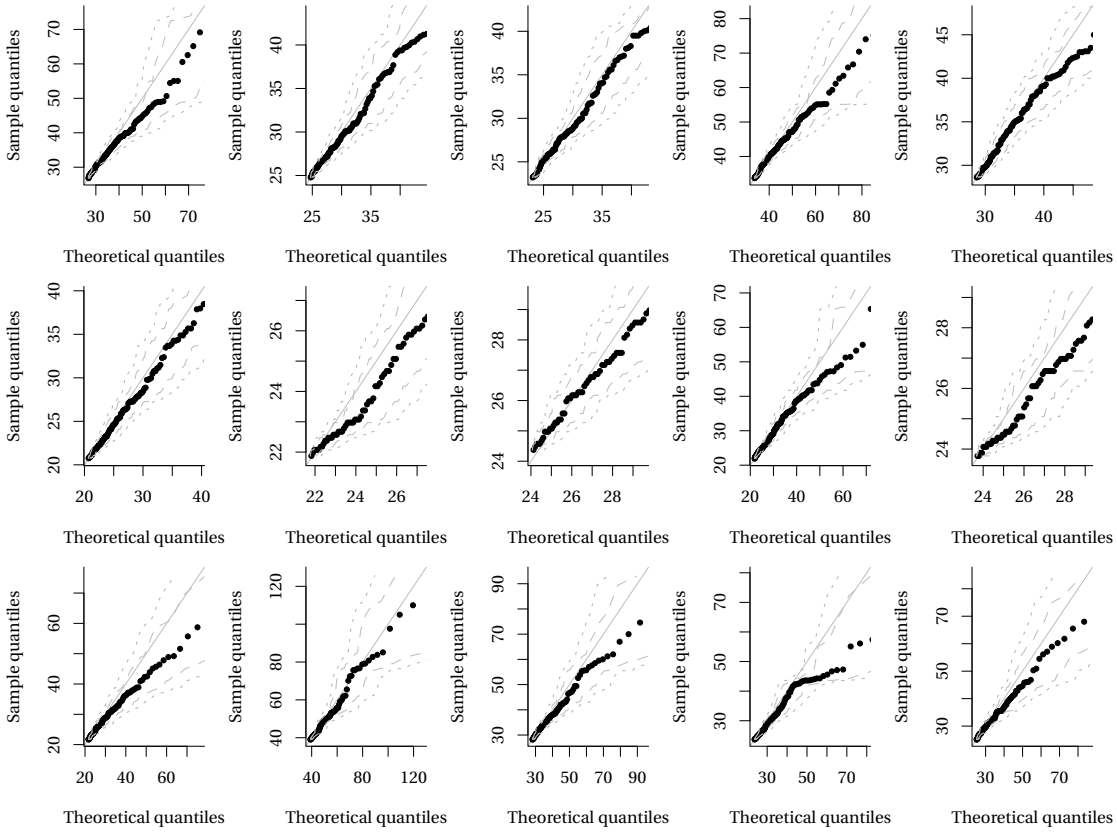

Figure 3.18 – *Quantile-quantile plots for marginal threshold exceedances above the 0.975%, based on the generalized Pareto distribution. The parameter values are posterior medians for* (σ_i, ξ) *for the first two rows (sites* $i = 1, \ldots, 10$*), which were used to estimate the model. The last row gives quantile-quantile plots for holdout data; the value of the scale parameter is obtained as the median from simulations from the latent Gaussian models, conditional on the posterior draws from the Bayesian linear model. The dashed (dotted) lines give approximate 95% pointwise (overall) credible intervals derived from the posterior predictive distribution.*

Resnick model with a power variogram. As for the Zürich data, we derive Bayesian analogs of quantile-quantile plots based on posterior draws for the 15 first sites, using all exceedances available at the station; it appears that the model nearly systematically overpredicts large events. The extremogram on Figure 3.22 shows that the dependence structure is well captured by a Brown–Resnick model and both variograms give indistinguishable results, capturing the decreasing dependence with distance, whereas the extremal Student underestimates the probability of joint occurrence of extremes. The estimated 95% credible interval for the curve from the two-stage fit is very narrow and lies below the cloud of empirical estimates of $\widehat{\chi}(h)$.

The left panel of Figure 3.20 shows the profile log-likelihood for the shape parameter ξ based on fitting independent generalized Pareto distribution to threshold exceedances at the 20 sites with a common shape parameter for the Swiss rainfall data and for simulated data; there is good agreement between the two methods. In contrast, the fit of the full Brown–Resnick

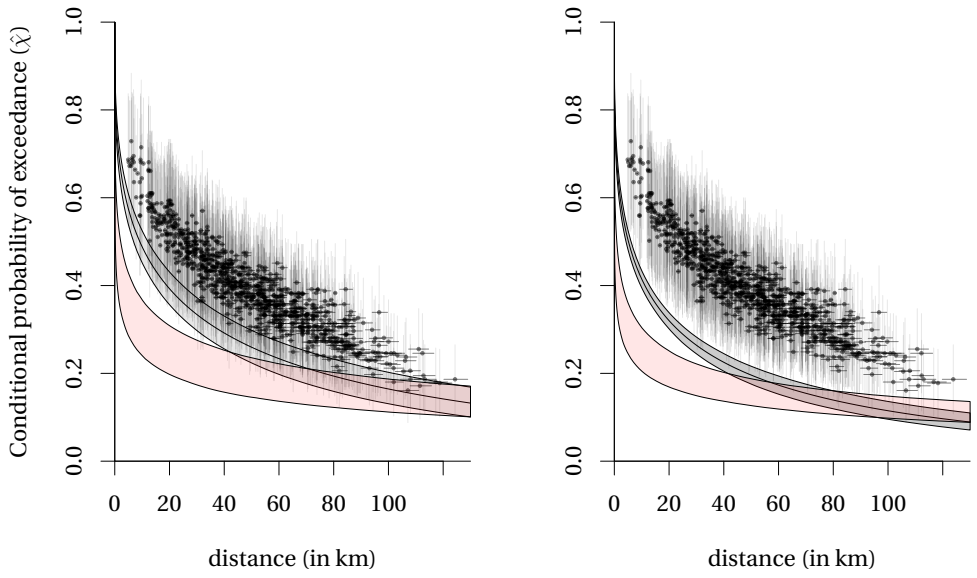

Figure 3.19 – *Left: conditional probability of exceedance above the 0.975 marginal percentile for the Zürich data, $\widehat{\chi}(h)$. Because of the geometric anisotropy, the distance between the sites is random and the points (segments) indicate the median distance (50% posterior predictive interval); vertical segments give approximate Wald 95% confidence intervals for the estimates of χ based on the delta-method. The fitted curve corresponds to the Brown–Resnick model with a power variogram (grey) and the extremal Student with a power exponential correlation function (red) with simultaneous 95% credible interval and median curve for the full likelihood fit (left) and the two-stage approach (right).*

method	$\mid \theta$	σ_1	ξ	α	λ	a	ϱ
full	5%	8.81	0.14	0.48	7.03	1.71	0.58
	median	9.99	0.18	0.53	9.55	1.95	0.69
	95%	11.32	0.23	0.57	12.7	2.22	0.79
	ESS(×0.01)	154	398	444	116	508	570
two-stage	5%	10.21	0.04	0.57	7.5	2.06	0.63
	median	10.97	0.06	0.58	7.54	2.18	0.67
	95%	11.82	0.08	0.6	7.68	2.3	0.71
	ESS(×0.01)	1380	746	234	312	216	266

Table 3.16 – *Posterior median, 90% marginal credible intervals, and effective sample size (ESS) for the parameters of the Brown–Resnick model with a power variogram for the Swiss rainfall dataset.*

model with power variogram to the Swiss rainfall data gives very different shape estimates; the right panel of Figure 3.20 shows density estimates of posterior draws for ξ, which do not overlap. There is a tradeoff between scale λ and ξ (Table 3.16) which results in an increase of the log-likelihood by nearly 3185 units if we evaluate the latter at the median a posteriori, suggesting that the draws for the two-stage approach are far from the maximum a posteriori.

$\mid \theta$	σ_1	ξ	β	α	λ	a	ϱ
5%	8.92	0.15	0.01	0.46	9.34	1.65	0.55
median	10.12	0.19	0.27	0.84	12.04	1.89	0.67
95%	11.46	0.23	0.62	1.17	15.02	2.17	0.77
ESS(×0.01)	154	370	198	202	146	366	534

Table 3.17 – *Posterior median, 90% marginal credible intervals, and effective sample size (ESS) for the parameters of the Brown–Resnick model with a Schlather variogram for the Swiss rainfall dataset.*

$\mid \theta$	σ_1	ξ	α	λ^α	a	ϱ	ν
5%	9.49	0.36	0.23	142.94	1.36	0.66	5.63
median	10.8	0.39	0.25	189.05	2.02	1	6.97
95%	12.31	0.41	0.29	253.61	2.95	1.28	8.82
ESS(×0.01)	90	530	208	180	324	378	238

Table 3.18 – *Posterior median, 90% marginal credible intervals, and effective sample size (ESS) for the parameters of the extremal Student model with a power exponential correlation function for the Swiss rainfall dataset.*

Such differences between marginal parameters are not uncommon even in the bivariate censored likelihood estimation, where marginal parameter estimates can compensate for what is potentially an incorrect dependence model. The marginal posterior estimates for ξ for the extremal Student (Table 3.18) are unreasonably high, potentially indicating a misspecified model. Such issues do not arise in simulation from the model; it would be interesting to compute the maximum composite likelihood estimator with a curvature adjustment to see if similar discrepancies between marginal and joint fit arise.

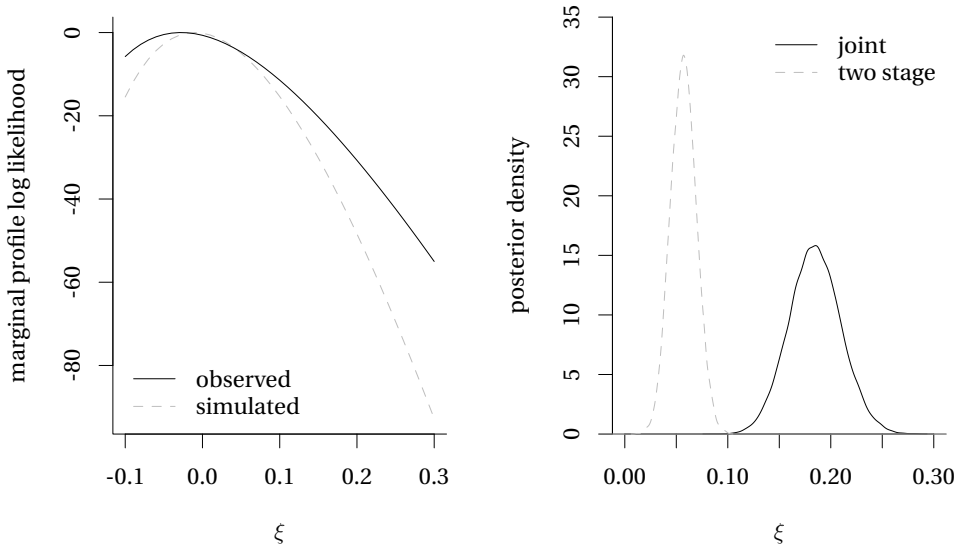

Figure 3.20 – *Left panel: profile log-likelihood for ξ for independent generalized Pareto distributions with common shape parameter (left) for the Swiss rainfall dataset and for simulated observations from the Brown–Resnick model with common shape parameter whose marginal parameters are obtained from the marginal fit. Right panel: posterior density estimates for ξ for a Brown–Resnick model with power variogram, using the two stage approach (dashed grey) and full likelihood (full).*

3.7.3 Concluding remarks

Bayesian modelling of functional peaks-over-threshold remains difficult and the simulations illustrate some of the challenges arising from censoring, which is responsible for the computational bottleneck of any frequentist or Bayesian estimation method. While efficiency gains could be achieved by porting routines to low level programming languages such as C++, the most expensive computations come from repeated matrix decompositions and calculation of Gaussian distribution functions, and there are limited alternatives. Data augmentation, which replaces the Monte Carlo average by a single draw, is costly to implement because of the need to compute the conditional distribution and we found that, despite faster computing time because we can use the full likelihood, use of auxiliary variables makes it extremely hard to design efficient proposals for the Markov chain Monte Carlo sampler. The loss of information for parameter estimation from going from full to censored likelihood depends on identifiability constraints, but overall remains moderately small as the dimension increases, in particular for dependence functions that have few parameters.

Adjusting a multivariate generalized Pareto model in large dimensions is difficult and real data, in particular in environmental settings, may be poorly approximated by the model. While an adaptive Metropolis-within-Gibbs Markov chain Monte Carlo algorithm worked in the simulation study and for the data applications, the method requires running long chains to get a good effective sample size and reliable estimate posterior quantiles of interest. The

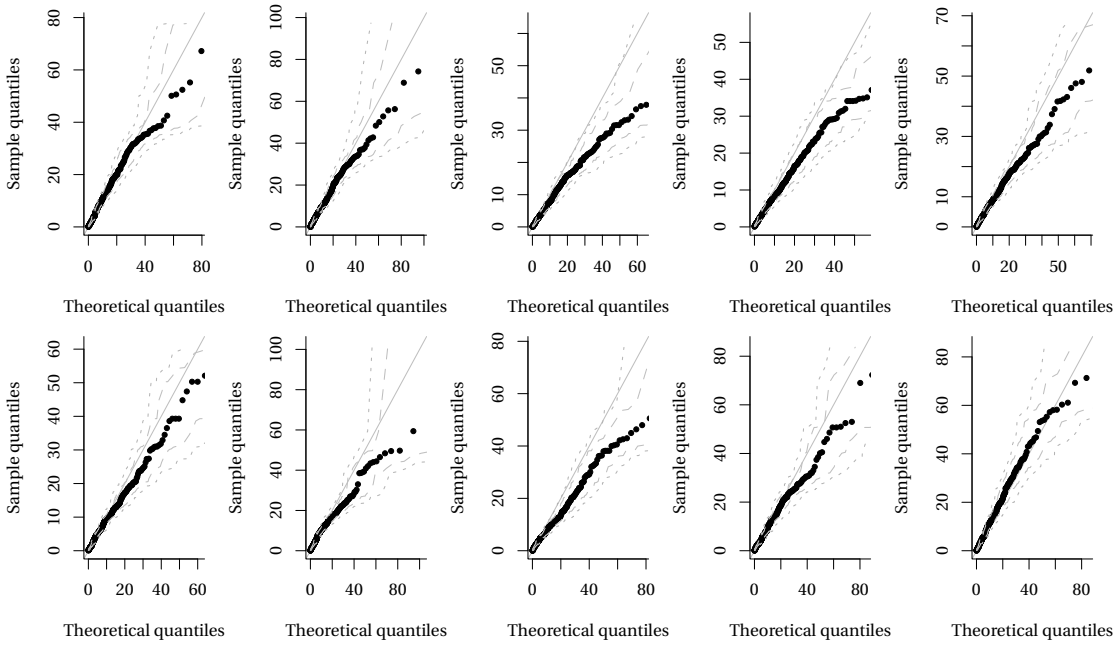

Figure 3.21 – *Quantile-quantile plots for marginal threshold exceedances above the 0.98%, based on the generalized Pareto distribution, using posterior samples from the fit of the Brown–Resnick model with a power variogram. The parameter values are posterior medians for* (σ_i, ξ) *for 10 sites used to fit the model, using all exceedances. The dashed (dotted) lines give approximate 95% pointwise (overall) credible intervals derived from the posterior predictive distribution.*

estimation is complex and special care must be taken to ensure that the proposals satisfy boundary constraints; use of truncated Gaussian proposals seems natural in many cases for sampling from the conditional posterior of a single parameter. However, the number of marginal parameters scales linearly in the dimension, the number of points exceeding their threshold decreases as D increases, leaving us with the unsatisfying option of having to evaluate numerically Gaussian distribution functions in dimension close to D for most of the observation vectors if we adopt a Brown–Resnick model. The computational burden increases linearly in the number of observations. As such, inference is and will remain limited to a small number of sites, $D \approx 20$, if we wish to obtain estimates in reasonable time. The computing for the application to the Zürich data took approximately 70 hours, but the time series are strongly spatially dependent. In contrast, some of the simulations in Section 3.6 with the extremal Student, the likelihood of which is more costly to evaluate, took a full 10 days of computing; while we can run multiple chains on a cluster, the sequential nature of the computations makes it difficult to exploit parallel computing architecture. In contrast, latent Gaussian models are very cheap to compute and the strategies outlined at the end of Section 3.4 and used in, e.g., Dyrrdal et al. (2014) can lead to efficient samplers without tuning the proposals. In dimension $D = 20$, it takes less than one hour to obtain about 40 000 posterior samples. However, the model falsely assumes that sites are independent, which means that the uncertainty is underestimated: nearby sites often share information about extreme events

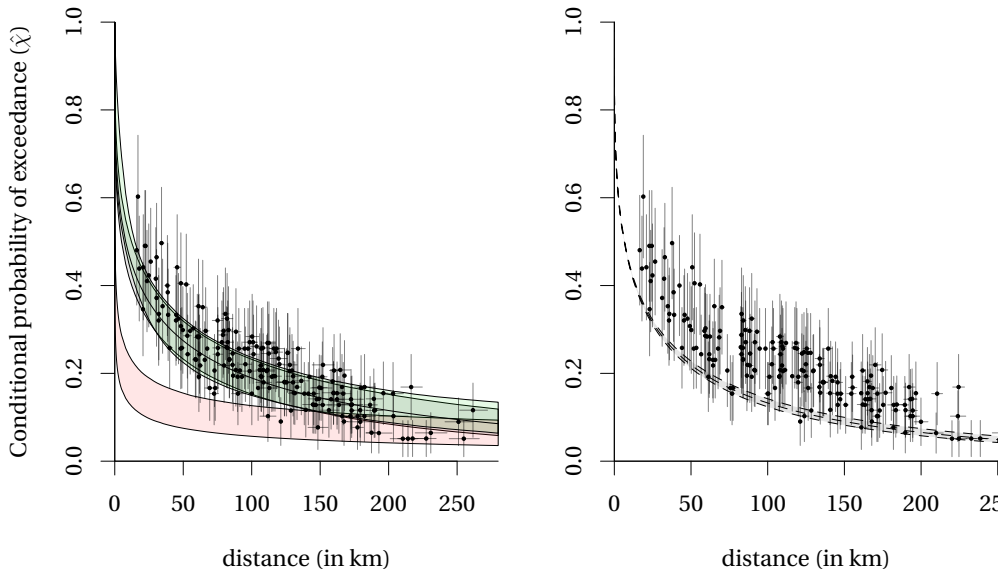

Figure 3.22 – *Left: conditional probability of exceedances above the 0.98 marginal percentile for the Swiss rainfall data, along with fitted curves obtained from posterior samples from the full likelihood fit (left) and the two-stage approach (right). Because of the geometric anisotropy, the distance between the sites is random; the points are median pairwise estimates of $\widehat{\chi}$ and the segments give 50% posterior predictive interval for the distances; vertical segments give approximate Wald 95% confidence intervals for the estimates of χ based on the delta-method. The fitted simultaneous 95% credible intervals curves correspond to a Brown–Resnick model with power variogram (grey) and with a Schlather variogram (green), and to the fit from the extremal Student with power exponential correlation (red).*

if the latter are spatially extended. Hewitt et al. (2019) suggest using the extremal coefficient to downweight observations, but this is *ad hoc*.

Conclusion and future work

The unifying theme of this thesis is likelihood-based inference. In Chapter 1, we considered inference for univariate extremes and derived closed-form expressions for cumulants of the generalized extreme value distribution. A key finding of the simulations is that confidence intervals based on the profile likelihood, while having adequate coverage, are typically shifted to the left and, as such, lead to higher error rate than confidence intervals derived using the tangent exponential model approximation. While derivation of higher-order cumulants is feasible, Bartlett and Cox–Snell corrections were shown to have poor finite-sample properties. It may be better avenue to rely on bootstrap methods, which while computationally expensive can more easily be generalized to arbitrary functionals. Often, the quality of the approximation provided by extreme value distributions is inadequate. On the other hand, these limiting models are the key tools for the extrapolation of risk. An avenue for future research is development of sub-asymptotic models for threshold exceedances, designed to be more flexible so that more observations from the bulk of the distribution can be used for inference. Automated threshold selection methods for high dimensional problems remain elusive, and likewise the problem for choosing the functional threshold for generalized ℓ-Pareto processes has not yet been addressed.

Chapter 2 comprises an extensive literature review, unifying the different characterizations of multivariate and functional extremes under a common umbrella. There have been many recent developments in the field, yet much remains to be done. While we have not covered temporal extremes, the models can be easily accommodated even if further complications, related to clustering, arise. In particular, the conditional spatial extremes model has not really been employed for modelling spatial extremes. Because it allows for modelling of possibly non-extreme data, the need for inferential censoring is less acute and extensions of the model, based on a mixture representation, could be envisioned. The main computational bottleneck is inferential censoring; while few models exist, extensions that account for skewness are likely to be even more difficult to fit. Inference appears feasible in high dimensions through gradient scoring, because the latter bypasses the estimation of the measure of the risk set and downweights rather than censors small observations. The gradient score method is limited to models that do not have mass on the boundary. Many open questions regarding estimation using gradient scoring also need to be addressed, including uncertainty quantification, the choice of weighting function and model selection.

Exploration of alternative hierarchical scale mixture model constructions to allow truly high-dimensional models to be fitted to extremal data is an obvious next step. Possible computational gains could come from approximating the Gaussian field with a basis expansion comprising stochastic Gaussian weights and compactly-supported basis functions. This corresponds to the finite element representation of the solution to a particular class of stochastic partial differential equations (SPDE) which give a Gaussian Markov random field (GMRF) approximation with Matérn covariance (Lindgren et al., 2011). The GMRF approximation results in sparse precision matrices for the weights and embedding the GMRF in a hierarchical model in place of the Gaussian process could drastically speed up computations. Use of weighted estimators rather than censoring could also allow better exploitation of the GMRF approximation; alternative estimators such as the gradient score have not yet been considered. Rather than starting from extreme value models, it may make sense to consider extremal properties of established models that are scalable, such as that presented in Wallin and Bolin (2015). An alternative to the SPDE approach for sparse modelling of high-dimensional Gaussian fields is the nearest-neighbour Gaussian process (Datta et al., 2016).

We estimated marginal and dependence parameters simultaneously for generalized max-Pareto processes in Chapter 3, finding that this is feasible in low dimensions and that it provides a large improvement in terms of overall fit compared to the two-step approach. There is no guarantee that the parameter estimates obtained from the two-stage procedure are optimal, and the application to the Zürich rainfall data shows this. The main issue is the fact that the posterior is doubly intractable, which means no information about the geometry of the posterior can be captured without gradient and hessian. Our attempts at bypassing this using data imputation were unsuccessful, and we had to rely on an adaptive Metropolis-within-Gibbs scheme. Pooling of spatial information can lead to better marginal estimates if the model is correctly specified, but censoring leads to a loss of information. There was some evidence of model misspecification in the application to the Zürich rainfall data and perhaps one could consider more complex models. However, only with additional spatial information can parameters of a more flexible model be estimated reliably and we are far from being able to tackle problems where the number of spatial sites of interest range in the hundreds. Even in bivariate models, incorrect specification of the joint model can lead to poor marginal adjustment, so it should perhaps not be too surprising that this carries over to higher dimensions.

A Supplementary material for Chapter 1

A.1 Cumulants of the generalized extreme value distribution

The moments of the derivatives log-likelihood μ for the generalized extreme value distribution are readily obtained after making the change of variable $y = \mu + \sigma(z^{-\xi} - 1)/\xi$. The moments of second order ν_{ij}, are defined for $\xi > -1/2$, while ν_{ijk} exists whenever $\xi > -1/3$, etc. Let $\zeta_* \approx 1.202$ denote the Riemann zeta function evaluated at 3, $\psi^{(0)}(x) = \Gamma'(x)/\Gamma(x)$ the digamma function, $\psi^{(1)}(x)$ the psigamma function $\mathrm{d}\psi^{(0)}(x)/\mathrm{d}x$ and, lastly, $\gamma_* \approx 0.5772$ the Euler–Mascheroni constant. Then

$$
\begin{aligned}
\nu_{\mu\mu} &= \frac{(1+\xi)\Gamma(3+2\xi)}{2\sigma^2(1+2\xi)},\\
\nu_{\mu\sigma} &= \frac{\Gamma(2+\xi) - (1+\xi)^2\Gamma(1+2\xi)}{\sigma^2\xi},\\
\nu_{\mu\xi} &= \frac{(1+\xi)^2\Gamma(1+2\xi) - \xi(1+\xi)\Gamma(1+\xi)\left\{1+\psi^{(0)}(2+\xi)\right\} + \Gamma(1+\xi)}{\sigma\xi^2},\\
\nu_{\sigma\sigma} &= \frac{(1+\xi)^2\Gamma(1+2\xi) - 2\Gamma(2+\xi) + 1}{\sigma^2\xi^2},\\
\nu_{\sigma\xi} &= \frac{(\gamma_*-1)\xi - (1+\xi)^2\Gamma(1+2\xi) + \Gamma(1+\xi)\{\xi(2+\xi) + \xi(1+\xi)\psi^{(0)}(2+\xi) + 2\} - 1}{\sigma\xi^3},\\
\nu_{\xi\xi} &= \frac{\left\{(\gamma_*-1)^2 + \pi^2/6\right\}\xi^2 - 2\Gamma(1+\xi)\left\{\xi^2 + \xi + (1+\xi)\xi\psi^{(0)}(2+\xi) + 1\right\}}{\xi^4}\\
&\quad - \frac{2(\gamma_*-1)\xi + (1+\xi)^2\Gamma(1+2\xi) + 1}{\xi^4},\\
\nu_{\mu\sigma\xi} &= \frac{(1+\xi)\left[\Gamma(1+2\xi)\left\{3\xi^2 + 7\xi + (1+2\xi)\xi\psi^{(0)}(2+2\xi) + 2\right\} - \left(4\xi^2 + 5\xi + 1\right)\Gamma(1+3\xi)\right]}{\sigma^2\xi^3}\\
&\quad + \frac{(1+\xi)(1+2\xi)\xi\Gamma(1+2\xi)\psi^{(0)}(2+2\xi) - \Gamma(2+\xi)\{2\xi + \xi\psi^{(0)}(2+\xi) + 1\}}{\sigma^2\xi^3}\\
&\quad + \frac{2(1+\xi)\Gamma(1+2\xi) - 6\xi\Gamma(1+3\xi) - \Gamma(2+\xi) - \Gamma(1+3\xi) - (1+\xi)\xi\Gamma(2+\xi)\psi^{(0)}(2+\xi)}{\sigma^2\xi^3},\\
\nu_{\mu\mu\mu} &= \frac{(1+\xi)^2(1+4\xi)\Gamma(1+3\xi)}{\sigma^3},
\end{aligned}
$$

$$
\begin{aligned}
\nu_{\mu\mu\sigma} &= \frac{(1+\xi)\{\Gamma(2+2\xi)-(1+\xi)(1+4\xi)\Gamma(1+3\xi)\}}{\sigma^3\xi},\\
\nu_{\mu\mu\xi} &= \frac{(1+\xi)\{(1+\xi)(1+4\xi)\Gamma(1+3\xi)-\Gamma(1+2\xi)(2\xi(2+\xi)+\xi(1+2\xi)\psi^{(0)}(2+2\xi)+1)\}}{\sigma^2\xi^2},\\
\nu_{\mu\sigma\sigma} &= \frac{(1+4\xi)(1+\xi)^2\Gamma(1+3\xi)+(1-\xi)\Gamma(2+\xi)-\Gamma(3+2\xi)}{\sigma^3\xi^2},\\
\nu_{\mu\xi\xi} &= \frac{\left(8\xi^2+6\xi+1\right)(1+\xi)^2\Gamma(1+3\xi)-\Gamma(3+2\xi)\left\{3\xi^2+5\xi+(1+2\xi)\xi\psi^{(0)}(2+2\xi)+1\right\}}{\sigma\xi^4(1+2\xi)},\\
&\quad+\frac{\Gamma(1+\xi)\left\{2\xi^3+5\xi^2+(1+\xi)\xi^2\psi^{(0)}(2+\xi)^2+(1+\xi)\xi^2\psi^{(1)}(2+\xi)\right\}}{\sigma\xi^4}\\
&\quad+\frac{\Gamma(1+\xi)\left\{2\left(2\xi^2+3\xi+1\right)\xi\psi^{(0)}(2+\xi)+6\xi+1\right\}}{\sigma\xi^4},\\
\nu_{\sigma\xi\xi} &= \frac{8\xi^3\Gamma(1+2\xi)-3(9+4\xi)\xi^3\Gamma(3\xi)+4\xi^3\Gamma(1+2\xi)\psi^{(0)}(2+2\xi)+4\xi^2+22\xi^2\Gamma(1+2\xi)}{\sigma\xi^5}\\
&\quad+\frac{6\xi^2\Gamma(1+2\xi)\psi^{(0)}(2+2\xi)+\gamma_*^2\xi^2+\pi^2\xi^2/6-6\gamma_*\xi^2-2\gamma_*\xi+7\xi-6\xi\Gamma(1+3\xi)}{\sigma\xi^5}\\
&\quad+\frac{17\xi\Gamma(1+2\xi)+3\Gamma(2\xi+1)-\Gamma(1+3\xi)+2\xi\Gamma(1+2\xi)\psi^{(0)}(2+2\xi)+1}{\sigma\xi^5}\\
&\quad-\frac{\Gamma(1+\xi)\left[\xi^2(1+\xi)\left\{\psi^{(0)}(2+\xi)^2+\psi^{(1)}(2+\xi)\right\}+2\left(3\xi^2+5\xi+2\right)\xi\psi^{(0)}(2+\xi)\right]}{\sigma\xi^5}\\
&\quad-\frac{\Gamma(1+\xi)(4\xi^3+13\xi^2+16\xi+3)}{\sigma\xi^5},\\
\nu_{\sigma\sigma\xi} &= \frac{\xi^3\Gamma(3\xi)(12\xi+27)-\gamma_*\xi^2+\xi^2-4(1+\xi)\xi^2\Gamma(1+2\xi)+\gamma_*\xi-2\xi+5\xi\Gamma(2+\xi)}{\sigma^2\xi^4}\\
&\quad+\frac{6\xi\Gamma(1+3\xi)-10(1+\xi)\xi\Gamma(2\xi+1)+3\Gamma(2+\xi)-3(1+\xi)\Gamma(1+2\xi)+\Gamma(1+3\xi)}{\sigma^2\xi^4}\\
&\quad+\frac{2\xi\Gamma(2+\xi)\psi^{(0)}(2+\xi)-(1+\xi)(1+2\xi)\xi\Gamma(1+2\xi)\psi^{(0)}(2+2\xi)-1}{\sigma^2\xi^4},\\
\nu_{\sigma\sigma\sigma} &= \frac{3(\xi-1)\Gamma(2+\xi)-3\xi-(1+4\xi)(1+\xi)^2\Gamma(1+3\xi)+3\Gamma(3+2\xi)/2+1}{\sigma^3\xi^3},\\
\nu_{\xi\xi\xi} &= -\frac{3\Gamma(3+2\xi)\left(4\xi^2+6\xi+(1+2\xi)\xi\psi^{(0)}(2+2\xi)+1\right)}{2\xi^6(1+2\xi)}\\
&\quad+\frac{(1+2\xi)\xi^3\left\{4\zeta_*+\gamma_*\left(24+\pi^2\right)+2\gamma_*^3-18\gamma_*^2-3\pi^2-8\right\}}{2\xi^6(1+2\xi)}\\
&\quad+\frac{(1+2\xi)\left\{6(\gamma_*-4)\xi+2(1+\xi)^2(4\xi+1)\Gamma(1+3\xi)-\left(30-42\gamma_*+6\gamma_*^2+\pi^2\right)\xi^2-2\right\}}{2\xi^6(1+2\xi)}\\
&\quad+\frac{3(1+2\xi)\Gamma(1+\xi)\left(4\xi^3+7\xi^2+8\xi+1+2\left(3\xi^2+4\xi+1\right)\xi\psi^{(0)}(2+\xi)\right)}{\xi^6(1+2\xi)}\\
&\quad+\frac{3(1+2\xi)\Gamma(1+\xi)\left[(1+\xi)\xi^2\left\{\psi^{(0)}(2+\xi)^2+\psi^{(1)}(2+\xi)\right\}\right]}{\xi^6(1+2\xi)}.
\end{aligned}
$$

Some entries of the Fisher information matrix of the generalized extreme value distribution, i.e., second-order cumulants, are undefined at $\xi = 0$ and can be obtained as $\lim_{\xi\to 0} \nu_{..}$, viz.

$$
\begin{aligned}
i_{\mu\mu} &= \frac{1}{\sigma^2}, \\
i_{\mu\sigma} &= \frac{1-\gamma_*}{\sigma^2}, \\
i_{\mu\xi} &= \frac{\pi^2/12 + \gamma_*^2/2 - \gamma_*}{\sigma}, \\
i_{\sigma\xi} &= \frac{\gamma_* + \gamma_*^3/2 + \gamma_*\pi^2/3 - \pi^2\left(\gamma_* - 1\right)/12 - 3\gamma_*^2/2 - \pi^2/3 + \zeta_*}{\sigma}, \\
i_{\sigma\sigma} &= \frac{\gamma_*^2 - 2\gamma_* + \pi^2/6 + 1}{\sigma^2}, \\
i_{\xi\xi} &= \frac{1}{4}\gamma_*^4 + \frac{1}{4}\gamma_*^2\pi^2 + \frac{3}{80}\pi^4 - \gamma_*^3 - \frac{1}{2}\gamma_*\pi^2 + \gamma_*^2 + \frac{1}{6}\pi^2 + 2\gamma_*\zeta_* - 2\zeta_*.
\end{aligned}
$$

A.2 Cumulants of the generalized Pareto distribution

Calculating the cumulants of the generalized Pareto distribution is a somewhat easier task; the nth order moment exist if and only if $\xi > -1/n$. The (mixed) moments of the derivatives of the log-likelihood up to fourth order are

$$
\begin{aligned}
\nu_{\xi\xi} &= -\frac{2}{(1+\xi)(1+2\xi)}, \\
\nu_{\sigma\sigma} &= -\frac{1}{\sigma^2(1+2\xi)}, \\
\nu_{\xi\sigma} &= -\frac{1}{\sigma(1+\xi)(1+2\xi)}, \\
\nu^{(\sigma)}_{\xi\xi} &= 0, \\
\nu^{(\sigma)}_{\xi\sigma} &= \frac{1}{\sigma^2(1+\xi)(1+2\xi)}, \\
\nu^{(\sigma)}_{\sigma\sigma} &= \frac{2}{\sigma^3(1+2\xi)}, \\
\nu^{(\xi)}_{\xi\xi} &= \frac{2\,(3+4\xi)}{(1+\xi)^2(1+2\xi)^2}, \\
\nu^{(\xi)}_{\xi\sigma} &= \frac{(3+4\xi)}{\sigma(1+\xi)^2(1+2\xi)^2}, \\
\nu^{(\xi)}_{\sigma\sigma} &= \frac{2}{\sigma^2(1+2\xi)^2}, \\
\nu_{\xi\xi\xi} &= \frac{24}{(1+\xi)(1+2\xi)(1+3\xi)}, \\
\nu_{\xi\xi\sigma} &= \frac{8}{\sigma(1+\xi)(1+2\xi)(1+3\xi)}, \\
\nu_{\xi\sigma\sigma} &= \frac{4}{\sigma^2(1+2\xi)(1+3\xi)}, \\
\nu_{\sigma\sigma\sigma} &= \frac{4}{\sigma^3(1+3\xi)}, \\
\nu_{\xi\xi,\xi} &= -\frac{2\left(12\xi^2+23\xi+9\right)}{(1+\xi)^2(1+2\xi)^2(1+3\xi)}, \\
\nu_{\xi\xi,\sigma} &= -\frac{8}{\sigma(1+\xi)(1+2\xi)(1+3\xi)}, \\
\nu_{\xi\sigma,\xi} &= -\frac{4\xi^2+11\xi+5}{\sigma(1+\xi)^2(1+2\xi)^2(1+3\xi)}, \\
\nu_{\xi\sigma,\sigma} &= -\frac{\xi+3}{\sigma^2(1+\xi)(1+2\xi)(1+3\xi)}, \\
\nu_{\sigma\sigma,\xi} &= -\frac{2(1+\xi)}{\sigma^2(1+2\xi)^2(1+3\xi)}, \\
\nu_{\sigma\sigma,\sigma} &= -\frac{2\,(1+\xi)}{\sigma^3(1+2\xi)(1+3\xi)}, \\
\nu_{\xi,\xi,\xi} &= \frac{6\left(4\xi^2+11\xi+5\right)}{(1+\xi)^2(1+2\xi)^2(1+3\xi)}, \\
\nu_{\xi,\xi,\sigma} &= \frac{2\left(4\xi^2+11\xi+5\right)}{\sigma(1+\xi)^2(1+2\xi)^2(1+3\xi)}, \\
\nu_{\xi,\sigma,\sigma} &= -\frac{2\left(\xi^2-3\xi-2\right)}{\sigma^2(1+\xi)(1+2\xi)^2(1+3\xi)}, \\
\nu_{\sigma,\sigma,\sigma} &= -\frac{2\,(\xi-1)}{\sigma^3(1+2\xi)(1+3\xi)},
\end{aligned}
$$

$$
\begin{aligned}
\nu_{\xi\xi\xi\xi} &= -\frac{18}{\sigma^4(1+4\xi)},\\
\nu_{\xi\xi\xi\sigma} &= -\frac{18}{\sigma^3(1+3\xi)(1+4\xi)},\\
\nu_{\xi\xi\sigma\sigma} &= -\frac{36}{\sigma^2(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi\sigma\sigma\sigma} &= -\frac{108}{\sigma(1+\xi)(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\sigma\sigma\sigma\sigma} &= -\frac{432}{(1+\xi)(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi,\xi,\xi,\xi} &= \frac{24\left(108\xi^5+464\xi^4+831\xi^3+686\xi^2+259\xi+36\right)}{(1+\xi)^3(1+2\xi)^3(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi,\xi,\xi,2} &= \frac{6\left(108\xi^5+464\xi^4+831\xi^3+686\xi^2+259\xi+36\right)}{\sigma(1+\xi)^3(1+2\xi)^3(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi,\xi,\sigma,\sigma} &= \frac{2\left(12\xi^5+96\xi^4+389\xi^3+453\xi^2+203\xi+31\right)}{\sigma^2(1+\xi)^2(1+2\xi)^3(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi,\sigma,\sigma,\sigma} &= \frac{3\left(4\xi^4-4\xi^3+27\xi^2+30\xi+7\right)}{\sigma^3(1+\xi)(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\sigma,\sigma,\sigma,\sigma} &= \frac{3\left(2\xi^2-\xi+3\right)}{\sigma^4(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi\xi,\xi\xi} &= \frac{8\left(30\xi^2+59\xi+23\right)}{(1+\xi)^2(1+2\xi)^2(1+3\xi)(1+4\xi)},\\
\nu_{\xi\xi,\xi\sigma} &= \frac{2\left(30\xi^2+59\xi+23\right)}{\sigma(1+\xi)^2(1+2\xi)^2(1+3\xi)(1+4\xi)},\\
\nu_{\xi\xi,\sigma\sigma} &= \frac{2\left(26\xi^2+29\xi+9\right)}{\sigma^2(1+\xi)(1+2\xi)^2(1+3\xi)(1+4\xi)},\\
\nu_{\xi\sigma,\xi\sigma} &= \frac{2\left(\xi+7\right)}{\sigma^2(1+\xi)(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi\sigma,\sigma\sigma} &= \frac{\xi+7}{\sigma^3(1+\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\sigma\sigma,\sigma\sigma} &= \frac{2\xi+5}{\sigma^4(1+2\xi)(1+4\xi)},\\
\nu_{\xi\xi,\xi,\xi} &= -\frac{4\left(360\xi^5+1596\xi^4+2624\xi^3+1973\xi^2+686\xi+89\right)}{(1+\xi)^3(1+2\xi)^3(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi\xi,\xi,\sigma} &= -\frac{2\left(126\xi^3+377\xi^2+268\xi+53\right)}{\sigma(1+\xi)^2(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi\xi,\sigma,\sigma} &= \frac{2\left(2\xi^2-35\xi-19\right)}{\sigma^2(1+\xi)(1+2\xi)^2(1+3\xi)(1+4\xi)},\\
\nu_{\xi\sigma,\xi,\xi} &= -\frac{2\left(108\xi^5+464\xi^4+831\xi^3+686\xi^2+259\xi+36\right)}{\sigma(1+\xi)^3(1+2\xi)^3(1+3\xi)^2(1+4\xi)},
\end{aligned}
$$

$$
\begin{aligned}
\nu_{\xi\sigma,\xi,\sigma} &= \frac{12\xi^4 - 46\xi^3 - 169\xi^2 - 124\xi - 25}{\sigma^2(1+\xi)^2(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi\sigma,\sigma,\sigma} &= \frac{2\xi^2 - 3\xi - 11}{\sigma^3(1+\xi)(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\sigma\sigma,\xi,\xi} &= -\frac{2\left(132\xi^4 + 252\xi^3 + 195\xi^2 + 74\xi + 11\right)}{\sigma^2(1+\xi)(1+2\xi)^3(1+3\xi)^2(1+4\xi)},\\
\nu_{\sigma\sigma,\xi,\sigma} &= \frac{12\xi^4 - 48\xi^3 - 89\xi^2 - 50\xi - 9}{\sigma^3(1+\xi)(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\sigma\sigma,\sigma,\sigma} &= \frac{6\xi^2 - 5\xi - 5}{\sigma^4(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi\xi\xi,\xi} &= \frac{48\left(18\xi^3 + 46\xi^2 + 31\xi + 6\right)}{(1+\xi)^2(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi\xi\xi,\sigma} &= \frac{108}{\sigma(1+\xi)(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi\xi\sigma,\xi} &= \frac{4\left(18\xi^3 + 85\xi^2 + 70\xi + 15\right)}{\sigma(1+\xi)^2(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi\xi\sigma,\sigma} &= \frac{4\left(\xi + 7\right)}{\sigma^2(1+\xi)(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\xi\sigma\sigma,\xi} &= \frac{4\left(6\xi^2 + 13\xi + 4\right)}{\sigma^2(1+2\xi)^2(1+3\xi)^2(1+4\xi)},\\
\nu_{\xi\sigma\sigma,\sigma} &= \frac{2\left(2\xi + 5\right)}{\sigma^3(1+2\xi)(1+3\xi)(1+4\xi)},\\
\nu_{\sigma\sigma\sigma,\xi} &= \frac{6\left(1+\xi\right)}{\sigma^3(1+3\xi)^2(1+4\xi)},\\
\nu_{\sigma\sigma\sigma,\sigma} &= \frac{6\left(1+\xi\right)}{\sigma^4(1+3\xi)(1+4\xi)}.
\end{aligned}
$$

Parameter	Method | F	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Quantile	MLE	−2	8	5	−3	11	−4	1	0	−2
	TEM	0	10	7	−1	12	−2	2	1	0
	Severini (TEM)	−2	6	3	−3	7	−5	−1	−1	−2
	Severini (cov.)	−1	7	4	−3	9	−4	0	−1	−1
N-obs. median	MLE	−2	10	6	−4	13	−5	1	0	−2
	TEM	0	12	9	−2	14	−3	3	2	0
	Severini (TEM)	−3	7	4	−4	9	−6	−1	−1	−2
	Severini (cov.)	−1	9	5	−3	11	−5	0	0	−1
N-obs. mean	MLE	−2	13	9	−4	21	−6	2	0	−2
	TEM	−1	15	12	−2	21	−4	4	2	0
	Severini (TEM)	−3	11	7	−4	16	−7	0	−1	−2
	Severini (cov.)	−1	13	8	−3	18	−6	1	0	−1

Table A.1 – *Truncated mean ($\alpha = 0.1$) of relative bias (in %), block maximum method with $m = 30$, $k = 60$. The largest standard error, obtained using a nonparametric bootstrap, is 0.44%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

Parameter	Method | F	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Quantile	MLE	−1	5	2	−3	9	−3	1	0	−2
	TEM	1	7	5	−1	10	−1	3	2	0
	Severini (TEM)	−2	3	0	−3	4	−4	−1	−1	−2
	Severini (cov.)	0	5	2	−2	7	−3	0	0	−1
N-obs. median	MLE	−1	7	4	−3	11	−4	1	0	−2
	TEM	1	9	7	−1	13	−1	3	2	1
	Severini (TEM)	−2	4	1	−4	6	−5	−1	−1	−2
	Severini (cov.)	1	6	3	−2	9	−3	1	0	−1
N-obs. mean	MLE	−1	10	7	−3	19	−4	2	0	−2
	TEM	1	13	10	−1	20	−1	5	3	1
	Severini (TEM)	−2	6	3	−4	12	−5	0	−1	−2
	Severini (cov.)	1	10	6	−2	17	−4	2	1	−1

Table A.2 – *Truncated mean ($\alpha = 0.1$) of relative bias (in %), block maximum method with $m = 45$, $k = 40$. The largest standard error, obtained using a nonparametric bootstrap, is 0.51%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

A.3 Simulation results for higher order methods

This section includes additional tables of simulation results, which are discussed on page 54.

Parameter	Method | F	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Quantile	MLE	0	3	1	−2	7	−2	1	0	−2
	TEM	2	6	5	1	10	2	4	3	1
	Severini (TEM)	−2	−1	−3	−3	0	−4	−3	−2	−3
	Severini (cov.)	3	4	2	0	8	1	2	2	0
N-obs. median	MLE	1	4	2	−2	10	−1	2	0	−2
	TEM	3	8	6	1	13	2	5	4	2
	Severini (TEM)	−2	−1	−2	−3	1	−4	−3	−2	−3
	Severini (cov.)	4	6	4	0	11	1	4	3	1
N-obs. mean	MLE	2	9	6	−2	21	−1	5	2	−1
	TEM	5	13	10	1	23	3	8	6	2
	Severini (TEM)	−1	2	0	−3	7	−4	−1	−1	−3
	Severini (cov.)	9	12	9	1	25	3	8	5	2

Table A.3 – *Truncated mean ($\alpha = 0.1$) of relative bias (in %), block maximum method with $m = 90, k = 20$. The largest standard error, obtained using a nonparametric bootstrap, is 0.81%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

Parameter	Method | F	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Quantile	MLE	−3	−4	−5	−2	−6	−4	−4	−3	−2
	TEM	2	3	2	1	3	2	3	3	1
	Severini (TEM)	−1	−1	−2	−1	−2	−1	−1	0	−1
	Severini (cov.)	−1	−1	−2	−1	−2	−1	−1	−1	−1
N-obs. median	MLE	−3	−4	−5	−3	−6	−4	−5	−3	−2
	TEM	2	4	3	1	5	2	3	3	2
	Severini (TEM)	0	0	−1	−1	−1	−1	−1	0	0
	Severini (cov.)	0	0	−1	−1	0	−1	0	0	0
N-obs. mean	MLE	−3	−3	−4	−3	−3	−4	−4	−2	−2
	TEM	4	8	7	2	14	4	7	5	3
	Severini (TEM)	1	3	2	−1	6	0	2	1	0
	Severini (cov.)	1	4	3	0	8	1	3	2	0

Table A.4 – *Truncated mean ($\alpha = 0.1$) of relative bias (in %), peaks-over-threshold method with $k = 20$. The largest standard error, obtained using a nonparametric bootstrap, is 0.77%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

Parameter	Method | F	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Quantile	MLE	−3	−4	−5	−3	−4	−5	−5	−3	−3
	TEM	−1	1	0	−1	3	−2	0	0	−1
	Severini (TEM)	−2	−2	−3	−3	0	−4	−3	−2	−2
	Severini (cov.)	−2	−2	−3	−3	0	−4	−3	−2	−2
N-obs. median	MLE	−4	−4	−5	−4	−3	−6	−5	−3	−4
	TEM	−1	2	0	−2	4	−2	0	0	−1
	Severini (TEM)	−2	−1	−3	−3	1	−4	−3	−2	−3
	Severini (cov.)	−2	−1	−2	−3	1	−4	−3	−2	−2
N-obs. mean	MLE	−4	−3	−4	−4	0	−6	−5	−3	−4
	TEM	0	4	2	−2	10	−2	1	1	−1
	Severini (TEM)	−2	1	−1	−3	6	−4	−2	−1	−2
	Severini (cov.)	−2	1	0	−3	6	−4	−2	−1	−2

Table A.5 – *Truncated mean ($\alpha = 0.1$) of relative bias (in %), peaks-over-threshold method with $k = 40$. The largest standard error, obtained using a nonparametric bootstrap, is 0.56%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

Parameter	Method | F	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Quantile	MLE	−3	−2	−4	−4	−1	−5	−3	−3	−3
	TEM	−1	2	0	−2	4	−3	0	0	−1
	Severini (TEM)	−2	0	−2	−3	2	−4	−2	−2	−2
	Severini (cov.)	−2	0	−2	−3	2	−4	−2	−2	−2
N-obs. median	MLE	−3	−1	−4	−4	0	−6	−4	−3	−3
	TEM	−1	3	0	−2	5	−3	0	0	−1
	Severini (TEM)	−2	1	−2	−4	3	−5	−2	−2	−3
	Severini (cov.)	−2	1	−2	−4	3	−5	−2	−2	−3
N-obs. mean	MLE	−4	0	−3	−5	4	−7	−3	−3	−3
	TEM	−1	5	2	−3	11	−4	1	0	−1
	Severini (TEM)	−2	3	0	−4	8	−5	−1	−1	−3
	Severini (cov.)	−2	3	0	−4	8	−5	−1	−1	−2

Table A.6 – *Truncated mean ($\alpha = 0.1$) of relative bias (in %), peaks-over-threshold method with $k = 60$. The largest standard error, obtained using a nonparametric bootstrap, is 0.49%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile						N-obs. mean					
F	Method | Error rate	0.5	2.5	5	5	2.5	0.5	0.5	2.5	5	5	2.5	0.5
F_1	Wald	0.0	0.0	0.5	18.0	14.0	8.5	0.0	0.0	0.0	19.5	15.5	10.0
	profile	0.5	2.0	3.5	7.5	3.5	0.5	0.5	1.5	3.5	8.5	4.0	0.5
	TEM	0.5	2.0	4.5	6.0	2.5	0.5	0.5	2.0	4.0	6.5	3.0	0.5
	Severini (TEM)	0.5	1.5	3.0	8.5	4.0	0.5	0.5	1.5	3.0	9.0	4.5	0.5
	Severini (cov.)	0.5	2.0	3.5	6.5	3.0	0.5	0.5	2.0	4.0	6.5	3.0	0.5
F_2	Wald	0.0	0.5	2.5	9.0	7.0	4.0	0.0	0.5	2.5	8.5	6.5	4.0
	profile	0.5	3.5	6.5	3.5	1.5	0.5	1.0	4.0	7.5	3.0	1.5	0.5
	TEM	1.0	4.0	7.0	2.5	1.0	0.0	1.0	4.5	8.5	2.5	1.0	0.0
	Severini (TEM)	0.5	2.5	5.5	4.0	1.5	0.5	1.0	3.5	7.0	3.0	1.5	0.5
	Severini (cov.)	0.5	3.0	6.0	3.5	1.5	0.5	1.0	4.0	7.5	3.0	1.5	0.5
F_3	Wald	0.0	0.5	1.5	11.0	8.5	5.0	0.0	0.0	1.5	10.5	8.0	5.0
	profile	0.5	2.0	4.5	4.5	2.0	0.5	0.5	2.5	5.0	4.0	2.0	0.5
	TEM	0.5	2.0	5.0	3.5	1.5	0.5	0.5	2.5	5.5	3.0	1.5	0.5
	Severini (TEM)	0.0	1.5	3.5	5.0	2.5	0.5	0.5	2.0	4.5	4.5	2.0	0.5
	Severini (cov.)	0.5	1.5	4.0	4.5	2.0	0.5	0.5	2.5	5.0	4.0	2.0	0.5
F_4	Wald	0.0	0.0	0.5	25.0	21.0	15.0	0.0	0.0	0.0	27.0	22.5	16.5
	profile	0.5	1.5	3.0	11.0	6.5	1.5	0.5	1.5	2.5	11.5	6.5	2.0
	TEM	0.5	2.5	4.5	8.0	4.0	1.0	0.5	2.0	4.0	8.5	4.5	1.0
	Severini (TEM)	0.5	1.5	2.5	11.5	6.5	1.5	0.0	1.0	2.5	12.5	7.0	2.0
	Severini (cov.)	0.5	1.5	3.0	9.5	5.0	1.5	0.5	1.5	2.5	10.5	5.5	1.5
F_5	Wald	0.0	1.0	3.0	8.0	6.0	3.5	0.0	0.5	2.5	7.5	5.5	3.0
	profile	0.5	3.0	6.5	3.5	1.5	0.5	1.0	4.0	8.5	3.0	1.5	0.5
	TEM	0.5	3.5	6.5	3.0	1.5	0.5	1.0	4.5	8.0	2.5	1.0	0.5
	Severini (TEM)	0.5	2.5	5.5	4.0	2.0	0.5	0.5	3.5	7.0	3.5	1.5	0.5
	Severini (cov.)	0.5	3.0	6.0	3.5	1.5	0.5	1.0	4.0	7.5	3.0	1.5	0.5
F_6	Wald	0.0	0.0	0.5	22.0	18.0	12.0	0.0	0.0	0.5	24.0	20.0	14.0
	profile	0.5	2.0	3.0	11.5	6.5	1.5	0.5	1.5	3.0	13.5	7.5	2.0
	TEM	0.5	2.0	4.0	9.0	5.0	1.0	0.5	2.0	3.5	10.5	6.0	1.5
	Severini (TEM)	0.5	1.5	3.0	12.5	7.0	2.0	0.5	1.5	2.5	13.5	8.0	2.5
	Severini (cov.)	0.5	1.5	3.0	11.5	6.5	1.5	0.5	1.5	3.0	12.5	7.0	2.0
F_7	Wald	0.0	0.5	1.5	14.5	11.0	6.5	0.0	0.0	1.0	15.0	11.5	7.0
	profile	0.5	2.0	4.5	6.5	3.5	1.0	0.5	2.0	5.0	7.0	3.5	1.0
	TEM	0.5	2.5	5.0	5.5	2.5	0.5	0.5	2.5	5.5	5.5	3.0	0.5
	Severini (TEM)	0.5	2.0	4.0	7.0	4.0	1.0	0.5	2.0	4.5	7.5	4.0	1.0
	Severini (cov.)	0.5	2.0	4.0	7.0	3.5	1.0	0.5	2.0	4.5	7.0	3.5	1.0
F_8	Wald	0.0	0.0	1.0	16.5	13.0	8.5	0.0	0.0	0.5	16.5	13.5	8.5
	profile	0.5	2.0	4.5	6.5	3.5	1.0	0.5	2.0	4.5	6.5	3.5	1.0
	TEM	0.5	2.5	5.5	5.0	2.5	0.5	0.5	2.5	5.5	5.0	2.5	0.5
	Severini (TEM)	0.5	2.0	4.0	7.5	4.0	1.0	0.5	2.0	4.0	7.0	4.0	1.0
	Severini (cov.)	0.5	2.0	4.0	6.5	3.5	0.5	0.5	2.0	4.5	6.5	3.5	0.5
F_9	Wald	0.0	0.0	0.5	19.5	16.0	11.0	0.0	0.0	0.5	20.0	16.5	11.5
	profile	0.5	1.5	3.5	7.5	4.0	1.0	0.5	1.5	3.5	7.5	4.0	1.0
	TEM	0.5	2.5	5.0	5.5	2.5	0.5	0.5	2.5	5.0	5.5	2.5	0.5
	Severini (TEM)	0.5	1.5	3.0	8.0	4.5	1.0	0.0	1.5	3.0	8.0	4.0	1.0
	Severini (cov.)	0.5	1.5	3.5	7.0	3.5	1.0	0.5	1.5	3.5	7.0	3.5	0.5

Table A.7 – *One-sided nominal error rate (in %) for lower (first to third columns) and upper (fourth to sixth columns) confidence intervals, block maximum method with* $m = 30, k = 60$. *The distributions (from top to bottom) are Burr* (F_1)*, Weibull* (F_2)*, generalized gamma* (F_3)*, normal* (F_4)*, lognormal* (F_5)*, Student t* (F_6)*,* $\mathsf{GEV}(\xi = 0.1)$ (F_7)*, Gumbel* (F_8) *and* $\mathsf{GEV}(\xi = -0.1)$ (F_9).

	Parameter	Quantile						N-obs. mean					
F	Method | Error rate	0.5	2.5	5	5	2.5	0.5	0.5	2.5	5	5	2.5	0.5
F_1	Wald	0.0	0.0	0.0	27.0	23.5	18.0	0.0	0.0	0.0	28.5	25.5	20.0
	profile	0.5	1.5	3.5	9.5	4.5	0.5	0.5	1.5	3.0	10.0	5.0	0.5
	TEM	0.5	3.0	5.5	3.5	1.0	0.0	0.5	3.0	5.5	3.5	1.0	0.0
	Severini (TEM)	0.5	1.5	3.5	7.5	3.5	0.5	0.5	1.5	3.5	8.0	3.5	0.5
	Severini (cov.)	0.5	1.5	3.5	7.5	3.5	0.5	0.0	1.5	3.5	7.5	3.5	0.5
F_2	Wald	0.0	0.0	0.0	24.0	21.0	16.0	0.0	0.0	0.0	24.5	21.5	17.0
	profile	0.5	1.5	3.0	8.5	4.5	0.5	0.5	1.5	3.0	8.0	4.0	0.5
	TEM	0.5	2.5	5.5	3.5	1.5	0.0	0.5	3.0	6.0	3.0	1.0	0.0
	Severini (TEM)	0.0	1.5	3.0	7.0	3.5	0.5	0.0	1.5	3.5	6.5	3.0	0.5
	Severini (cov.)	0.0	1.5	3.0	7.0	3.5	0.5	0.5	1.5	3.5	6.5	3.0	0.5
F_3	Wald	0.0	0.0	0.0	25.0	22.0	17.5	0.0	0.0	0.0	25.5	22.5	18.0
	profile	0.0	1.0	2.0	10.0	6.0	1.0	0.0	1.0	2.0	9.5	5.5	1.0
	TEM	0.0	1.5	4.0	4.5	2.0	0.0	0.0	2.0	5.0	4.5	1.5	0.0
	Severini (TEM)	0.0	0.5	2.0	8.5	4.5	0.5	0.0	1.0	2.5	8.0	4.5	0.5
	Severini (cov.)	0.0	1.0	2.0	8.0	4.5	0.5	0.0	1.0	2.5	8.0	4.0	0.5
F_4	Wald	0.0	0.0	0.0	31.0	27.0	21.5	0.0	0.0	0.0	32.5	29.0	23.5
	profile	0.5	1.5	3.0	8.5	3.5	0.0	0.5	1.0	2.5	8.5	3.5	0.0
	TEM	0.5	3.0	5.5	2.0	0.5	0.0	0.5	2.5	5.0	2.0	0.5	0.0
	Severini (TEM)	0.0	1.0	2.5	6.5	2.5	0.0	0.0	1.0	2.0	6.5	2.0	0.0
	Severini (cov.)	0.0	1.0	2.5	6.0	2.5	0.0	0.0	1.0	2.5	6.0	2.0	0.0
F_5	Wald	0.0	0.0	0.0	21.5	18.0	13.5	0.0	0.0	0.0	22.0	19.0	14.5
	profile	0.5	1.5	3.5	9.0	5.0	1.0	0.5	2.0	4.0	8.5	4.5	1.0
	TEM	0.5	3.0	5.5	4.5	2.5	0.5	0.5	3.5	7.0	4.5	2.0	0.0
	Severini (TEM)	0.5	2.0	4.0	8.0	4.5	1.0	0.5	2.0	4.5	7.5	4.0	0.5
	Severini (cov.)	0.5	2.0	4.0	8.0	4.5	1.0	0.5	2.5	4.5	7.0	4.0	0.5
F_6	Wald	0.0	0.0	0.0	29.0	25.5	20.0	0.0	0.0	0.0	31.0	27.5	22.0
	profile	0.5	2.0	3.5	11.5	6.0	1.0	0.5	1.5	3.0	12.0	6.5	1.0
	TEM	0.5	3.0	6.0	5.0	2.0	0.0	0.5	3.0	6.0	5.0	2.0	0.0
	Severini (TEM)	0.5	1.5	3.5	9.5	5.0	0.5	0.5	1.5	3.0	10.0	5.0	0.5
	Severini (cov.)	0.5	1.5	3.5	9.5	5.0	0.5	0.5	1.5	3.5	9.5	5.0	0.5
F_7	Wald	0.0	0.0	0.0	25.5	22.5	17.5	0.0	0.0	0.0	27.0	23.5	19.0
	profile	0.5	1.5	3.5	10.0	5.5	1.0	0.5	1.5	3.0	10.0	5.5	1.0
	TEM	0.5	3.0	6.0	4.5	2.0	0.0	0.5	3.0	6.0	4.5	2.0	0.0
	Severini (TEM)	0.5	1.5	3.5	8.5	4.5	1.0	0.5	1.5	3.5	8.5	4.5	0.5
	Severini (cov.)	0.5	1.5	3.5	8.5	4.5	0.5	0.5	2.0	4.0	8.0	4.5	0.5
F_8	Wald	0.0	0.0	0.0	26.0	23.0	18.0	0.0	0.0	0.0	27.0	24.0	19.5
	profile	0.5	1.5	3.5	8.5	4.0	0.5	0.5	1.5	3.0	8.5	4.0	0.5
	TEM	0.5	3.0	6.0	3.0	1.0	0.0	0.5	3.0	6.0	3.0	1.0	0.0
	Severini (TEM)	0.0	1.0	3.0	7.0	3.0	0.5	0.0	1.0	3.0	6.5	3.0	0.0
	Severini (cov.)	0.0	1.0	3.0	6.5	3.0	0.5	0.0	1.5	3.0	6.5	3.0	0.0
F_9	Wald	0.0	0.0	0.0	29.5	26.0	20.5	0.0	0.0	0.0	30.5	27.0	22.0
	profile	0.5	1.5	3.0	7.5	3.0	0.0	0.0	1.0	2.5	7.0	3.0	0.0
	TEM	0.5	3.0	5.5	1.5	0.5	0.0	0.5	2.5	5.5	1.5	0.5	0.0
	Severini (TEM)	0.0	1.0	2.0	5.5	2.0	0.0	0.0	1.0	2.0	5.0	2.0	0.0
	Severini (cov.)	0.0	1.0	2.0	5.5	2.0	0.0	0.0	1.0	2.5	5.0	1.5	0.0

Table A.8 – *One-sided nominal error rate (in %) for lower (first to third columns) and upper (fourth to sixth columns) confidence intervals, peaks-over-threshold method with* $k = 40$*. The distributions (from top to bottom) are Burr* (F_1)*, Weibull* (F_2)*, generalized gamma* (F_3)*, normal* (F_4)*, lognormal* (F_5)*, Student t* (F_6)*,* $\mathsf{GP}(\xi = 0.1)$ (F_7)*, exponential* (F_8) *and* $\mathsf{GP}(\xi = -0.1)$ (F_9)*.*

	Parameter	Quantile			N-obs. median			N-obs. mean		
F	Method | Conf. level (%)	90	95	99	90	95	99	90	95	99
F_1	Wald	78	71	57	77	70	56	73	65	49
	TEM	102	102	102	103	103	102	103	103	103
	Severini (TEM)	96	95	95	95	95	95	95	95	94
	Severini (cov.)	103	103	103	103	103	103	105	105	105
F_2	Wald	74	68	55	73	67	55	69	62	49
	TEM	101	101	101	101	101	101	101	101	101
	Severini (TEM)	94	94	94	94	94	94	94	94	94
	Severini (cov.)	96	96	96	97	96	96	96	96	96
F_3	Wald	74	68	55	73	67	54	69	61	48
	TEM	101	101	101	101	101	101	101	101	101
	Severini (TEM)	94	94	94	94	94	94	94	93	93
	Severini (cov.)	96	96	96	96	96	96	96	96	96
F_4	Wald	73	68	57	73	67	57	71	66	55
	TEM	105	105	105	106	106	105	106	106	106
	Severini (TEM)	98	98	98	98	98	98	98	97	97
	Severini (cov.)	102	102	101	102	101	101	102	102	102
F_5	Wald	77	69	54	76	68	53	67	57	41
	TEM	99	99	99	99	99	99	98	98	98
	Severini (TEM)	93	93	92	92	92	92	90	90	90
	Severini (cov.)	95	95	94	95	95	94	94	93	93
F_6	Wald	78	71	58	78	71	57	75	67	52
	TEM	103	103	103	104	104	103	104	104	104
	Severini (TEM)	96	96	96	96	96	96	96	96	95
	Severini (cov.)	99	99	98	99	98	98	99	98	98
F_7	Wald	76	70	56	76	69	55	71	63	48
	TEM	102	101	101	102	102	102	102	102	102
	Severini (TEM)	95	95	94	95	95	94	94	94	93
	Severini (cov.)	97	96	96	96	96	96	96	96	96
F_8	Wald	74	69	57	74	68	57	72	65	53
	TEM	103	103	103	103	103	103	104	104	104
	Severini (TEM)	96	96	96	96	96	96	96	96	96
	Severini (cov.)	98	98	98	98	98	98	98	98	98
F_9	Wald	72	68	58	72	68	58	71	66	57
	TEM	105	105	104	105	105	104	105	105	104
	Severini (TEM)	98	97	97	97	97	97	97	97	97
	Severini (cov.)	100	100	100	100	100	100	100	100	100

Table A.9 – *Truncated mean ($\alpha = 0.1$) of the ratio of the confidence interval width relative to the width of profile confidence interval (in %), block maximum method with $m = 30, k = 60$. The largest standard error, obtained using a nonparametric bootstrap, is 0.12%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile			N-obs. median			N-obs. mean		
F	Method | Conf. level (%)	90	95	99	90	95	99	90	95	99
F_1	Wald	68	59	44	67	58	42	61	52	35
	TEM	103	103	103	104	103	103	104	104	104
	Severini (TEM)	93	93	92	93	93	92	92	91	91
	Severini (cov.)	105	105	105	106	106	105	111	111	112
F_2	Wald	66	58	45	65	58	44	60	52	39
	TEM	102	102	102	102	102	102	102	102	102
	Severini (TEM)	92	92	92	92	92	92	91	91	91
	Severini (cov.)	97	96	96	97	97	96	97	97	97
F_3	Wald	66	58	44	65	57	44	59	51	37
	TEM	102	102	102	102	102	102	102	102	102
	Severini (TEM)	92	92	92	92	92	92	91	91	91
	Severini (cov.)	96	96	96	96	96	96	96	96	96
F_4	Wald	64	58	47	64	58	46	62	55	43
	TEM	108	108	107	108	108	107	108	108	108
	Severini (TEM)	97	96	96	96	96	96	96	96	96
	Severini (cov.)	104	104	104	104	104	103	105	105	104
F_5	Wald	67	58	42	66	57	40	55	44	28
	TEM	99	99	98	99	99	99	98	98	97
	Severini (TEM)	90	89	88	89	89	88	86	85	85
	Severini (cov.)	94	94	93	94	94	94	93	93	93
F_6	Wald	68	60	46	67	59	44	63	54	38
	TEM	104	104	104	105	105	105	105	105	105
	Severini (TEM)	94	94	93	94	94	93	93	93	92
	Severini (cov.)	100	99	99	100	99	99	100	100	99
F_7	Wald	67	59	44	66	58	43	60	51	35
	TEM	102	102	102	103	102	102	103	103	102
	Severini (TEM)	92	92	91	92	92	91	91	90	90
	Severini (cov.)	97	96	96	96	96	96	96	96	95
F_8	Wald	66	59	46	65	58	46	62	54	42
	TEM	105	105	104	105	105	105	105	105	105
	Severini (TEM)	94	94	94	94	94	94	93	93	94
	Severini (cov.)	99	99	98	99	99	98	99	99	99
F_9	Wald	64	59	49	64	59	49	62	57	46
	TEM	107	107	106	107	107	106	108	107	106
	Severini (TEM)	96	96	96	96	96	96	96	96	95
	Severini (cov.)	102	102	101	102	102	101	102	102	101

Table A.10 – *Truncated mean ($\alpha = 0.1$) of the ratio of the confidence interval width relative to the width of profile confidence interval (in %), block maximum method with $m = 45, k = 40$. The largest standard error, obtained using a nonparametric bootstrap, is 0.13%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile			N-obs. median			N-obs. mean		
F	Method | Conf. level (%)	90	95	99	90	95	99	90	95	99
F_1	Wald	46	37	22	45	36	22	37	28	15
	TEM	105	105	105	106	106	106	106	106	106
	Severini (TEM)	85	84	83	85	84	83	81	80	81
	Severini (cov.)	117	118	118	118	118	117	149	151	149
F_2	Wald	47	39	26	47	38	26	40	32	22
	TEM	104	105	104	105	105	104	105	105	104
	Severini (TEM)	85	85	85	85	85	86	82	83	85
	Severini (cov.)	102	101	100	102	102	101	103	102	101
F_3	Wald	47	39	25	46	38	25	39	31	20
	TEM	104	105	104	105	105	104	104	104	104
	Severini (TEM)	85	85	85	85	85	85	82	82	84
	Severini (cov.)	101	100	99	101	101	100	101	101	100
F_4	Wald	47	40	29	47	40	29	43	36	25
	TEM	114	114	112	115	114	111	115	114	111
	Severini (TEM)	92	92	91	92	91	91	90	89	89
	Severini (cov.)	116	116	113	116	115	112	122	121	118
F_5	Wald	46	36	21	44	34	20	31	22	12
	TEM	98	98	98	98	98	98	96	96	96
	Severini (TEM)	80	78	77	79	78	78	72	72	75
	Severini (cov.)	98	97	97	98	98	98	101	100	101
F_6	Wald	47	38	24	46	37	23	39	30	17
	TEM	107	107	107	108	108	108	108	108	108
	Severini (TEM)	87	86	84	86	85	85	83	82	83
	Severini (cov.)	108	107	107	107	107	106	113	113	112
F_7	Wald	47	37	23	45	36	22	37	28	16
	TEM	104	104	104	104	104	104	104	104	104
	Severini (TEM)	84	83	82	84	83	83	80	79	80
	Severini (cov.)	101	101	100	101	101	100	103	102	101
F_8	Wald	48	40	28	47	39	28	42	35	24
	TEM	109	109	108	109	109	108	110	109	107
	Severini (TEM)	89	88	88	88	88	88	86	86	87
	Severini (cov.)	106	105	104	106	105	103	108	107	105
F_9	Wald	49	42	32	48	42	32	45	39	30
	TEM	114	113	110	114	112	109	114	112	108
	Severini (TEM)	92	92	91	92	91	91	90	90	89
	Severini (cov.)	112	110	108	111	109	107	113	112	108

Table A.11 – *Truncated mean ($\alpha = 0.1$) of the ratio of the confidence interval width relative to the width of profile confidence interval (in %), block maximum method with $m = 90, k = 20$. The largest standard error, obtained using a nonparametric bootstrap, is 0.34%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GEV}(\xi = 0.1)$ (F_7), Gumbel (F_8) and $\mathsf{GEV}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile			N-obs. median			N-obs. mean		
F	Method | Conf. level (%)	90	95	99	90	95	99	90	95	99
F_1	Wald	34	24	11	32	22	11	22	14	6
	TEM	203	202	179	207	202	172	253	238	184
	Severini (TEM)	127	127	127	128	128	127	135	133	126
	Severini (cov.)	131	131	131	133	133	131	142	140	130
F_2	Wald	36	26	14	35	25	14	26	18	10
	TEM	192	186	158	193	184	152	221	205	168
	Severini (TEM)	126	126	125	126	126	124	129	127	121
	Severini (cov.)	130	130	128	131	131	128	135	133	125
F_3	Wald	36	26	14	34	25	14	24	17	9
	TEM	193	189	162	195	187	155	224	206	168
	Severini (TEM)	126	126	125	126	126	124	130	128	123
	Severini (cov.)	131	131	129	132	132	129	137	134	126
F_4	Wald	37	29	18	36	28	18	30	22	14
	TEM	182	171	144	182	168	140	201	187	157
	Severini (TEM)	124	123	120	124	122	118	122	119	112
	Severini (cov.)	128	126	121	129	127	121	128	125	116
F_5	Wald	34	23	10	32	21	9	18	10	4
	TEM	211	213	189	217	214	179	293	266	184
	Severini (TEM)	130	130	129	131	131	128	144	141	129
	Severini (cov.)	134	135	132	136	136	132	152	148	132
F_6	Wald	35	25	12	33	23	11	24	15	7
	TEM	203	202	179	207	202	172	248	235	185
	Severini (TEM)	126	127	127	127	128	127	133	132	126
	Severini (cov.)	130	131	130	132	133	131	140	138	129
F_7	Wald	35	24	12	33	23	11	22	14	6
	TEM	203	201	176	207	201	168	252	235	180
	Severini (TEM)	127	127	127	128	128	127	135	133	126
	Severini (cov.)	131	132	131	133	133	131	142	140	129
F_8	Wald	37	28	16	35	27	16	28	20	12
	TEM	186	178	151	187	175	146	209	195	162
	Severini (TEM)	124	124	122	125	124	121	125	122	115
	Severini (cov.)	129	129	125	130	129	125	131	129	120
F_9	Wald	39	31	21	37	30	21	33	26	18
	TEM	170	158	136	168	155	133	185	171	148
	Severini (TEM)	122	121	117	122	120	115	120	116	110
	Severini (cov.)	126	123	117	126	123	117	125	120	112

Table A.12 – *Truncated mean ($\alpha = 0.1$) of the ratio of the confidence interval width relative to the width of profile confidence interval (in %), peaks-over-threshold method with $k = 20$. The largest standard error, obtained using a nonparametric bootstrap, is 1.17%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile			N-obs. median			N-obs. mean		
F	Method | Conf. level (%)	90	95	99	90	95	99	90	95	99
F_1	Wald	53	42	25	51	40	23	42	31	16
	TEM	143	145	146	147	148	147	161	162	156
	Severini (TEM)	112	112	112	113	113	113	117	117	116
	Severini (cov.)	113	113	113	114	114	114	118	118	117
F_2	Wald	52	43	27	51	42	27	44	34	21
	TEM	140	141	141	142	143	140	152	151	144
	Severini (TEM)	111	112	112	112	112	112	115	114	113
	Severini (cov.)	112	113	113	113	113	113	116	116	114
F_3	Wald	52	42	27	51	41	26	43	33	20
	TEM	140	142	142	143	144	142	154	153	146
	Severini (TEM)	112	112	112	112	113	112	115	115	114
	Severini (cov.)	112	113	113	113	113	113	116	116	115
F_4	Wald	51	43	30	50	42	29	46	38	26
	TEM	140	141	138	143	142	137	147	146	139
	Severini (TEM)	111	111	110	112	111	110	112	112	110
	Severini (cov.)	112	112	111	113	113	111	114	114	112
F_5	Wald	53	41	23	51	39	22	37	25	12
	TEM	144	147	150	149	151	151	175	175	161
	Severini (TEM)	113	114	114	114	115	114	121	121	119
	Severini (cov.)	113	114	115	115	115	115	122	123	120
F_6	Wald	53	43	26	51	41	25	44	33	18
	TEM	143	145	146	147	148	148	159	160	156
	Severini (TEM)	112	112	112	113	113	113	116	116	115
	Severini (cov.)	113	113	113	114	114	113	117	117	116
F_7	Wald	53	42	25	51	40	24	42	31	17
	TEM	143	145	146	147	148	146	161	161	154
	Severini (TEM)	112	112	112	113	113	113	117	117	116
	Severini (cov.)	113	113	113	114	114	114	118	118	117
F_8	Wald	52	43	29	51	42	29	46	37	24
	TEM	140	141	139	142	142	138	148	148	141
	Severini (TEM)	111	111	111	112	112	111	113	113	112
	Severini (cov.)	112	112	112	112	112	112	114	114	113
F_9	Wald	52	44	33	51	44	33	48	41	30
	TEM	138	137	132	140	138	131	143	140	132
	Severini (TEM)	111	110	109	111	110	108	111	110	107
	Severini (cov.)	112	112	110	112	112	110	113	112	110

Table A.13 – *Truncated mean ($\alpha = 0.1$) of the ratio of the confidence interval width relative to the width of profile confidence interval (in %), peaks-over-threshold method with $k = 40$. The largest standard error, obtained using a nonparametric bootstrap, is 0.28%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

	Parameter	Quantile			N-obs. median			N-obs. mean		
F	Method | Conf. level (%)	90	95	99	90	95	99	90	95	99
F_1	Wald	64	53	36	62	52	34	55	44	26
	TEM	127	128	130	130	131	131	137	138	138
	Severini (TEM)	108	108	108	109	109	109	111	111	111
	Severini (cov.)	108	108	108	109	109	109	111	111	111
F_2	Wald	62	53	38	61	52	37	55	45	30
	TEM	126	127	127	128	128	128	133	133	131
	Severini (TEM)	108	108	108	108	108	108	110	110	109
	Severini (cov.)	108	108	108	108	109	109	110	110	110
F_3	Wald	62	53	37	61	52	36	55	44	29
	TEM	126	127	128	128	129	129	134	135	132
	Severini (TEM)	108	108	108	108	108	109	110	110	110
	Severini (cov.)	108	108	109	108	109	109	110	111	110
F_4	Wald	60	53	40	59	52	39	57	49	36
	TEM	126	127	126	128	128	127	131	131	129
	Severini (TEM)	108	108	107	108	108	107	109	109	108
	Severini (cov.)	108	108	108	108	108	108	109	109	108
F_5	Wald	64	53	34	63	51	32	50	37	20
	TEM	127	129	132	130	132	133	144	146	143
	Severini (TEM)	108	109	109	109	110	110	113	114	114
	Severini (cov.)	109	109	110	110	110	110	114	114	114
F_6	Wald	64	54	37	63	53	36	57	46	29
	TEM	127	128	130	130	131	131	137	137	137
	Severini (TEM)	108	108	108	109	109	109	111	110	110
	Severini (cov.)	108	108	108	109	109	109	111	111	111
F_7	Wald	63	53	36	62	52	35	55	43	26
	TEM	128	128	130	130	131	131	138	138	137
	Severini (TEM)	108	108	108	109	109	109	111	111	111
	Severini (cov.)	108	108	109	109	109	109	111	111	111
F_8	Wald	62	54	39	61	53	38	57	48	34
	TEM	126	126	126	127	128	127	131	131	129
	Severini (TEM)	107	108	108	108	108	108	109	109	109
	Severini (cov.)	108	108	108	108	108	108	109	109	109
F_9	Wald	60	54	42	60	53	42	58	51	39
	TEM	125	125	123	126	126	123	128	127	124
	Severini (TEM)	108	107	107	108	107	106	108	108	106
	Severini (cov.)	108	108	107	108	108	107	109	108	107

Table A.14 – *Truncated mean ($\alpha = 0.1$) of the ratio of the confidence interval width relative to the width of profile confidence interval (in %), peaks-over-threshold method with $k = 60$. The largest standard error, obtained using a nonparametric bootstrap, is 0.12%. The distributions (from top to bottom) are Burr (F_1), Weibull (F_2), generalized gamma (F_3), normal (F_4), lognormal (F_5), Student t (F_6), $\mathsf{GP}(\xi = 0.1)$ (F_7), exponential (F_8) and $\mathsf{GP}(\xi = -0.1)$ (F_9).*

B Variogram and covariance models

We begin with some commonly used terminology employed in spatial statistics. For bounded variograms, we define the sill to be the limiting value attained by the function, i.e. $\lim_{h\to\infty}\gamma(h)$. The smallest lag distance at which this value is attained is the range, which represents the spatial extent of the process. We speak of effective range as the smallest lag distance h at which 95% of the sill value is attained if the latter is never reached by the variogram.

We review here some of the most commonly used variogram models; see e.g., Table 2.2 of Banerjee, Carlin and Gelfand (2014, p.29) for a more comprehensive list, focusing on intrinsically stationary and isotropic models. For interpolation purposes, the behaviour of the variogram at the origin is the most important feature of the function, as it describes the smoothness of the process. The sum of variogram functions, termed "gigogne", also define valid models. Correlation functions can be obtained by taking $\exp\{-\gamma(\cdot)\}$ for a valid semi-variogram γ.

Nugget

The simplest variogram model is the nugget,

$$\gamma(h) = \tau^2\mathbf{I}_{h>0}, \qquad \tau^2 > 0,$$

a spatially-constant noise. The latter is often used in conjunction with other models and creates a discontinuity at the origin. Nugget processes are often associated to microscale variation that are best left out of the modelling. They are also justified by measurement errors, particularly if data sets include replicated measures.

Power exponential variogram

$$\gamma(h) = \sigma^2[1-\exp\{-(\phi h)^\alpha\}]\mathbf{I}_{h>0}, \qquad \sigma^2, \phi > 0, \alpha \in (0,2].$$

The classical exponential model is recovered by setting $\alpha = 1$. The random field that results from the powered exponential variogram is continuous, but not differentiable.

Power variogram

$$\gamma(h) = (h/\lambda)^{\alpha}\mathbf{I}_{h>0}, \qquad \lambda > 0, \alpha \in (0,2).$$

Oesting et al. (2017) utilize a variant designed to be smoother near the origin,

$$\gamma(h) = \sigma^2 \frac{\phi^2 h^2}{(1+\phi^2 h^2)^{\alpha}}.$$

Generalized Cauchy variogram

The model, due to Gneiting et al. (2010), has variogram

$$\gamma(h) = (1+\phi^{\alpha} h^{\alpha})^{-\beta/\alpha}, \qquad \alpha \in (0,2], \phi, \beta > 0.$$

Matérn variogram

The Matérn model is a bounded variogram of the form

$$\gamma(h) = \sigma^2 \mathcal{M}(h;\phi,\nu) := \sigma^2 \left\{1 - \frac{(\phi h)^{\nu}}{2^{\nu-1}\Gamma(\nu)} \mathrm{K}_{\nu}(\phi h)\right\} \mathbf{I}_{h>0}, \qquad \phi, \nu, \sigma^2 > 0,$$

where K_{ν} the modified Bessel function of the second kind. The parameter ν controls the smoothness of the process. The latter is m times differentiable provided $\nu > m$ (Handcock and Stein, 1993). The parameter ϕ gives the reciprocal of the range. The use of the Matérn family is advocated by Stein (1999) and is widely employed thanks to its inherent flexibility and its good performance for kriging. It includes as a special case the exponential model $\gamma(h) = \sigma^2 \exp(-\phi h)$, when $\nu = 1/2$, and the Gaussian model $\gamma(h) = \sigma^2 \exp(-\phi h^2)$, the latter being obtained as limiting case when $\nu \to \infty$. The range parameter of the Matérn is notoriously difficult to estimate, since if it often close to the region of interest over which the data collection took place. Moreover, while the product $\sigma^2\phi^{2\nu}$ is identifiable, none of the individual parameters is. Since kriging is only sensitive to the product (Zhang, 2004), the reciprocal range ϕ is often set to a fixed value, but simulation studies due to Kaufman and Shaby (2013) suggest that it may be best to estimate both in finite samples even if the estimators are not consistent. (Banerjee et al., 2014, p. 125), recommend setting a vague prior for σ^2 and a more informative

prior for ϕ. It is also common in Markov chain Monte Carlo algorithms to reparametrize the model as $\alpha = 2\nu^{1/2}\phi$ and $\eta = \sigma^2\phi^{2\nu}$ to improve mixing. For rough processes for which $\nu < 1$, it may be interesting to use instead an (unbounded) power-law variogram to avoid the need to estimate the range parameter.

Schlather's variogram

The choice of variogram in spatial extremes modelling has important implications when utilized within Gaussian processes which appear as building blocks in the de Haan's spectral representation. For example, the ergodicity of Brown–Resnick processes depends on the boundedness of the underlying variogram model. Schlather and Moreva (2017) details how large scale extreme events align with non-ergodic processes while localized events are captured by ergodic processes. Schlather and Moreva proposes a variogram family able to capture both states, namely

$$\gamma(h) = \frac{\left\{1 + h^{\alpha}\right\}^{\beta/\alpha} - 1}{2^{\beta/\alpha} - 1}, \qquad 0 < \alpha \leq 2, \beta \leq 2. \tag{B.1}$$

The function is obtained through a composition of complete Bernstein functions, i.e., infinitely differentiable non-negative functions f whose derivative is completely monotone so that $(-1)^{n+1} f^{(n)}(x)/dx^n \geq 0$, and parametrized so that $\gamma(1) = 1$. As for the power-law variogram family, the parameter α controls the smoothness of the process, while β affects the behaviour at large lags. The variogram is unbounded for non-negative values of β and bounded otherwise, the case $\beta = 0$ being understood as the limiting case as $\beta \to 0$, i.e., $\gamma(h) = \log(1 + h^{\alpha})/\log(2)$. The pair (α, β) is generally strongly negatively correlated, but reparametrization is not straightforward due to the support constraints.

C Properties and simulation of elliptical distributions

C.1 Bayesian linear model and properties of elliptical distributions

Proposition C.1 (Bayesian Gaussian linear model with conjugate location prior)
Consider a hierarchical linear regression model of the form

$$\boldsymbol{Y} \mid \boldsymbol{\beta} \sim \mathsf{No}_q(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad \boldsymbol{\beta} \sim \mathsf{No}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}),$$

The joint distribution of $(\boldsymbol{\beta}, \boldsymbol{Y})$ is

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{Y} \end{pmatrix} \sim \mathsf{No}_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{X}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & \boldsymbol{\Omega}\mathbf{X}^\top \\ \mathbf{X}\boldsymbol{\Omega} & \boldsymbol{\Sigma} + \mathbf{X}\boldsymbol{\Omega}\mathbf{X}^\top \end{pmatrix} \right).$$

The posterior distribution $\boldsymbol{\beta} \mid \boldsymbol{Y} = \boldsymbol{y}$ is Gaussian with

$$\begin{aligned} \mathsf{E}\left(\boldsymbol{\beta} \mid \boldsymbol{Y} = \boldsymbol{y}\right) &= \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{X}^\top(\boldsymbol{\Sigma} + \mathbf{X}\boldsymbol{\Omega}\mathbf{X}^\top)^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}) \\ &= \left(\boldsymbol{\Omega}^{-1} + \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} + \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right), \\ \mathsf{Var}\left(\boldsymbol{\beta} \mid \boldsymbol{Y} = \boldsymbol{y}\right) &= \boldsymbol{\Omega} - \boldsymbol{\Omega}\mathbf{X}^\top(\boldsymbol{\Sigma} + \mathbf{X}\boldsymbol{\Omega}\mathbf{X}^\top)^{-1}\mathbf{X}\boldsymbol{\Omega} \\ &= \left(\boldsymbol{\Omega}^{-1} + \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}, \end{aligned}$$

the second expression follows from an application of Woodbury's formula or by completing the square.

Proposition C.2 (Conditional distribution of Gaussian subvectors)
Suppose that $\boldsymbol{W} \sim \mathsf{No}_p(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ is a Gaussian random p-vector partitioned into $(\boldsymbol{W}_1, \boldsymbol{W}_2)$, respectively $p_1 \times 1$ and $p_2 \times 1$ dimensional, where $p_1 + p_2 = p$. We can decompose the model into a marginal component and a conditional one as $p(\boldsymbol{W}) = p(\boldsymbol{W}_1 \mid \boldsymbol{W}_2 = \boldsymbol{w}_2)p(\boldsymbol{W}_2)$, where $\boldsymbol{W}_2 \sim \mathsf{No}_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ and

$$\begin{aligned} \boldsymbol{W}_1 \mid \boldsymbol{W}_2 = \boldsymbol{w}_2 &\sim \mathsf{No}_{p_1}\left(\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\boldsymbol{w_2} - \boldsymbol{\mu}_2), \mathbf{Q}_{11}^{-1}\right) \\ &\sim \mathsf{No}_{p_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{w_2} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right). \end{aligned}$$

Proposition C.3 (Conditional distribution of Student subvectors)
Let $\boldsymbol{X} \sim \mathsf{St}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ be a p-dimensional vector and consider the partition $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, where $\boldsymbol{X}_1$ is $p_1 \times 1$ and $\boldsymbol{X}_2 = p_2 \times 1$ for $p_1 + p_2 = p$. The conditional distribution of $\boldsymbol{X}_1 \mid \boldsymbol{X}_2$ is again Student distributed, with

$$\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim \mathsf{St}_{p_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \frac{\nu + (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)}{\nu + p_2}\boldsymbol{\Sigma}_{1|2}, \nu + p_2\right).$$

where $\boldsymbol{\Sigma}_{1|2} \coloneqq \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ is the Schur complement of $\boldsymbol{\Sigma}_{11}$, while $\boldsymbol{X}_2 \sim \mathsf{St}_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu)$ (e.g. Ding, 2016).

C.2 Simulation of (truncated) Gaussian processes

Simulation of Gaussian processes

The most widely-used parametrization of the multivariate Gaussian distribution is in terms of the positive definite matrix covariance matrix $\boldsymbol{\Sigma}$ and samples can be obtained through Algorithm C.1. We can consider instead the precision matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ and its Cholesky decomposition is $\mathbf{Q} = \mathbf{T}\mathbf{T}^\top$, where $\mathbf{T}$ is a lower triangular matrix and the lower triangular matrix $\mathbf{T}^{-1}$, can be obtained by back-solving. Since $\mathbf{T}^{-\top}\mathbf{T}^{-1} = \boldsymbol{\Sigma}$, the matrix $\mathbf{L} \equiv \mathbf{T}^{-\top}$ can be used in Algorithm C.1. For Gaussian Markov random fields, Algorithm C.2 proposed by Rue (2001) can be used for sampling Gaussian models $\mathsf{No}_p(\mathbf{Q}^{-1}\mathbf{b}, \mathbf{Q}^{-1})$ without calculating explicitly $\mathbf{Q}^{-1}$.

If the precision or covariance matrix is sparse, its Cholesky root will also be sparse; there are dedicated algorithms to do these calculations efficiently using the sparse structure, though these may require reordering the variable beforehand to obtain a banded matrix. Given the Cholesky root, the log determinant of the precision matrix is readily obtained as $\log\det(\mathbf{Q}) = 2\sum_{j=1}^{n}\log\left([\mathbf{L}]_{jj}\right)$ since the matrix is upper triangular.

Simulation of truncated Gaussian distributions

The distribution function of an absolutely continuous random variable $X \sim F$, truncated on the interval (a, b), is $\{F(x) - F(a)\}/\{F(b) - F(a)\}$ for $-\infty \leq a < b \leq \infty$. The inverse distribution method of Devroye (1986) for this case is summarized in Algorithm C.3. This approach, albeit simple, requires the evaluation of the inverse distribution function. In the Gaussian case, evaluation of Φ^{-1} is numerically unstable for small (large) values of a (b) even if accurate approximations are used for Φ and Φ^{-1}; the method suffers from numerical overflow whenever $a \geq 8.3$. Whenever $a > 0$, the following equivalent representation of Algorithm C.3,

$$-\Phi^{-1}[\overline{\Phi}(a) - \{\overline{\Phi}(a) - \overline{\Phi}(b)\}U],$$

is more numerically accurate. For values of $a > 37$, different approximation routines must be employed. Botev and L'Écuyer (2017) show that the truncated series expansion of Mill's

Algorithm C.1 Gaussian samples from $\mathsf{No}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Require: mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$.
1: compute the Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$;
2: sample $\boldsymbol{y} \sim \mathsf{No}_D(\mathbf{0}_D, \mathbf{I}_D)$
3: **return** $\boldsymbol{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{y}$.

Algorithm C.2 Gaussian samples from $\mathsf{No}(\mathbf{Q}^{-1}\boldsymbol{b}, \mathbf{Q}^{-1})$ (Rue, 2001, Algorithm 3.1.2)

Require: mean $\mathbf{Q}^{-1}\boldsymbol{b}$, precision matrix $\boldsymbol{Q}$.
1: compute the Cholesky decomposition $\mathbf{Q} = \mathbf{L}\mathbf{L}^\top$;
2: solve $\mathbf{L}\mathbf{v} = \boldsymbol{b}$ for $\mathbf{v}$;
3: solve $\mathbf{L}^\top\boldsymbol{\nu} = \mathbf{v}$ for $\boldsymbol{\nu}$;
4: solve $\mathbf{L}^\top\boldsymbol{y} = \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathsf{No}_n(\mathbf{0}, \mathbf{I})$, for $\boldsymbol{y}$;
5: **return** $\boldsymbol{\nu} + \boldsymbol{y}$.

Algorithm C.3 Inverse distribution function sampler for two-sided truncated distributions

Require: distribution function F, lower bound a, upper bound b.
1: sample $U \sim \mathsf{U}(0,1)$;
2: **return** $F^{-1}[F(a) + U\{F(b) - F(a)\}]$.

ratio, $\{1-\Phi(x)\}/\phi(x)$, provides upper and lower bounds. The problem can be reduced to root finding for the equation

$$-\overline{\Phi}(x) + \overline{\Phi}(a) - \{\overline{\Phi}(a) - \overline{\Phi}(b)\}u = 0, \tag{C.1}$$

using the Rayleigh distribution with survival function $\exp(-x^2/2)$ as a substitute for $1-\Phi(x)$. The approximate root of eq. (C.1) is

$$x \approx [a^2 - 2\log\{1 - u + u\exp(a^2/2 - b^2/2)\}]^{1/2},$$

and this approximation can be refined if we employ Newton's method on eq. (C.1) as in Algorithm C.4.

The alternative to numerical inversion via the inverse distribution function is accept-reject methods. for unidimensional truncated Gaussian variables, Robert (1995) proposes an accept-reject algorithm for sampling from $\mathsf{TNo}(\mu, \sigma^2; -\infty, b)$ distribution with $b < \mu$ based on exponential proposals (Algorithm C.5) and suggests a MCMC scheme for higher dimensions that uses a series of univariate simulations at the cost of exactness. If the truncation point b lies above the mean, one can simply use an accept-reject by sampling from $\mathsf{No}(\mu, \sigma^2)$ and keep only draws that fall below b.

Chopin (2011) suggests a faster alternative for finite intervals or semi-finite intervals using an adaptation of the Ziggurat algorithm that is extended to three variables. Other proposals include Damien and Walker (2001) and Yu and Tian (2011).

Algorithm C.4 Numerical inversion for uth quantile of the truncated Gaussian $\mathsf{TNo}(0,1;a,b)$ (Botev and L'Écuyer, 2017)

Let $q(x) = \sum_{j=0}^{r}(-1)^j x^{-2j-1}(2j-1)!!$, where !! denotes the double factorial.

Require: $u \in (0,1)$, numerical tolerance δ, lower bound a, upper bound b.

$c \leftarrow q(\alpha)(1-u) + q(b)u\exp(a^2/2 - b^2/2)$;
$\delta_x \leftarrow \infty$;
$z \leftarrow 1 - u + u\exp(a^2/2 - b^2/2)$;
$x \leftarrow \{a^2 - 2\log(z)\}^{1/2}$;
while $\delta_x \leq \delta$ **do**
    $z \leftarrow z - x\{zq(x) - c\}$;
    $x_0 \leftarrow \{a^2 - 2\log(z)\}^{1/2}$;
    $\delta_x \leftarrow |x_0 - x|/x$;
    $x \leftarrow x_0$;
return x.

Algorithm C.5 Right-truncated univariate Gaussian samples using exponential accept-reject Robert (1995)

Let $\alpha^*(x) = \left\{x + (x^2+4)^{1/2}\right\}/2$.

Require: mean μ, upper bound $b < \mu$, variance σ^2.

set $b_s = (b-\mu)/\sigma$;
sample $e + b_s \sim \mathsf{E}(\alpha^*)$;
set $\rho(e) = \exp[-\{e - \alpha^*(b_s)\}^2/2]$;
sample $b \sim \mathsf{U}(0,1)$;
if $b \leq \rho(e)$ **then**
    return $-\sigma e + \mu$.
else
    return to step 2;

The region of interest for the truncated process need not be rectangular. Rodriguez-Yam et al. (2004) propose a Gibbs sampling algorithm for sampling from a truncated Gaussian distribution conditional on a set of linear constraints, which they efficiently transform. Pakman and Paninski (2014) propose a Hamiltonian Monte Carlo algorithm to sample from a truncated multivariate Gaussian distribution subject to linear or quadratic constraints. The samples are however only approximate and the sampler does not work in rare-event estimation, e.g., if the upper truncation level is much smaller than the mean of the process. Botev (2017) and Botev and L'Ecuyer (2015) propose a more efficient way to sample exactly from Gaussian and Student-t multivariate vectors subject to linear constraints.

D Functionalities of the R package mev

The manual of the R package mev (version 1.12) is available from the Comprehensive R Archive Network at https://cran.r-project.org/web/packages/mev/index.html.

Bibliography

1. Amestoy, P. R., Davis, T. A. and Duff, I. S. (1996) An approximate minimum degree ordering algorithm. SIAM Journal on Matrix Analysis and Applications **17**(4), 886–905.
2. Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics **37**(2), 697–725.
3. Andrieu, C. and Thoms, J. (2008) A tutorial on adaptive MCMC. Statistics and Computing **18**(4), 343–373.
4. Asadi, P., Davison, A. C. and Engelke, S. (2015) Extremes on river networks. The Annals of Applied Statistics **9**(4), 2023–2050.
5. Bader, B., Yan, J. and Zhang, X. (2017) Automated selection of *r* for the *r* largest order statistics approach with adjustment for sequential testing. Statistics and Computing **27**(6), 1435–1451.
6. Bader, B., Yan, J. and Zhang, X. (2018) Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. The Annals of Applied Statistics **12**(1), 310–329.
7. Banerjee, S., Carlin, B. and Gelfand, A. (2014) Hierarchical Modeling and Analysis for Spatial Data. Second edition. Boca Raton: CRC Press, 584 p.
8. Barbi, E., Lagona, F., Marsili, M., Vaupel, J. W. and Wachter, K. W. (2018) The plateau of human mortality: Demography of longevity pioneers. Science **360**(6396), 1459–1461.
9. Barlow, A. M., Sherlock, C. and Tawn, J. (2019) Inference for extreme values under threshold-based stopping rules. arXiv e-prints p. arXiv:1901.03151.
10. Barndorff-Nielsen, O. (1980) Conditionality resolutions. Biometrika **67**(2), 293–310.
11. Barndorff-Nielsen, O. (1983) On a formula for the distribution of the maximum likelihood estimator. Biometrika **70**(2), 343–365.
12. Barndorff-Nielsen, O. (1988) Parametric Statistical Models and Likelihood. Heidelberg: Springer–Verlag, 276 p.
13. Barndorff-Nielsen, O. E. (1986) Inference on full or partial parameters based on the standardized signed log likelihood ratio. Biometrika **73**(2), 307–322.
14. Barndorff-Nielsen, O. E. and Cox, D. R. (1994) Inference and Asymptotics. London: Taylor & Francis, 360 p.
15. Barnett, V. (1976) The ordering of multivariate data (with discussion). Journal of the Royal Statistical Society. Series A (General) **139**(3), 318–355.
16. Bartlett, M. S. (1953) Approximate confidence intervals. II. More than one unknown parameter. Biometrika **40**(3/4), 306–317.
17. Basrak, B., Davis, R. A. and Mikosch, T. (2002) A characterization of multivariate regular variation. The Annals of Applied Probability **12**(3), 908–920.
18. Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) Statistics of Extremes: Theory and Applications. Chichester: John Wiley & Sons, 490 p.

19. Belzile, L. R. and Nešlehová, J. G. (2017) Extremal attractors of Liouville copulas. Journal of Multivariate Analysis **160**, 68–92.
20. Beranger, B., Padoan, S. A. and Sisson, S. A. (2017) Models for extremal dependence derived from skew-symmetric families. Scandinavian Journal of Statistics **44**(1), 21–45.
21. Betancourt, M. (2017) A conceptual introduction to Hamiltonian Monte Carlo. arXiv e-prints arXiv:1701.02434.
22. Bie, O., Borgan, Ø. and Liestøl, K. (1987) Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. Scandinavian Journal of Statistics **14**(3), 221–233.
23. Boldi, M.-O. (2009) A note on the representation of parametric models for multivariate extremes. Extremes **12**(3), 211–218.
24. Bolin, D. and Lindgren, F. (2015) Excursion and contour uncertainty regions for latent Gaussian models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **77**(1), 85–106.
25. Botev, Z. and L'Écuyer, P. (2017) Simulation from the normal distribution truncated to an interval in the tail. In Proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools on 10th EAI International Conference on Performance Evaluation Methodologies and Tools, pp. 23–29.
26. Botev, Z. I. (2017) The normal law under linear restrictions: simulation and estimation via minimax tilting. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(1), 125–148.
27. Botev, Z. I. and L'Ecuyer, P. (2015) Efficient probability estimation and simulation of the truncated multivariate student-t distribution. In 2015 Winter Simulation Conference (WSC), pp. 380–391.
28. Brazzale, A. R., Davison, A. C. and Reid, N. (2007) Applied Asymptotics: Case Studies in Small-Sample Statistics. Cambridge: Cambridge University Press, 236 p.
29. Breiman, L. (1965) On some limit theorems similar to the arc-sin law. Theory of Probability & Its Applications **10**(2), 323–331.
30. Brooks, S., Gelman, A., Jones, G. and Meng, X. (eds) (2011) Handbook of Markov Chain Monte Carlo. Boca Raton: CRC Press, 619 p.
31. Brown, B. M. and Resnick, S. I. (1977) Extreme values of independent stochastic processes. Journal of Applied Probability **14**(4), 732–739.
32. Bücher, A. and Segers, J. (2017) On the maximum likelihood estimator for the generalized extreme-value distribution. Extremes **20**(4), 839–872.
33. Buishand, T. A. (1984) Bivariate extreme-value data and the station-year method. Journal of Hydrology **69**, 77–95.
34. de Carvalho, M., Oumow, B., Segers, J. and Warchoł, M. (2013) A euclidean likelihood estimator for bivariate tail dependence. Communications in Statistics - Theory and Methods **42**(7), 1176–1192.
35. Casson, E. and Coles, S. (1999) Spatial regression models for extremes. Extremes **1**(4), 449–468.
36. Castruccio, S., Huser, R. and Genton, M. G. (2016) High-order composite likelihood inference for max-stable distributions and processes. Journal of Computational and Graphical Statistics **25**(4), 1212–1229.
37. Chilès, J.-P. and Delfiner, P. (2012) Geostatistics: Modeling Spatial Uncertainty. Second edition. Hoboken: Wiley, 699 p.
38. Chopin, N. (2011) Fast simulation of truncated Gaussian distributions. Statistics and Computing **21**(2), 275–288.
39. Christensen, D. (2005) Fast algorithms for the calculation of Kendall's τ. Computational Statistics **20**(1), 51–62.
40. Coles, S. (2001) An Introduction to Statistical Modeling of Extreme Values. London: Springer–Verlag, 209 p.

41. Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence measures for extreme value analyses. Extremes **2**(4), 339–365.
42. Coles, S. and Pericchi, L. (2003) Anticipating catastrophes through extreme value modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics) **52**(4), 405–416.
43. Coles, S., Pericchi, L. R. and Sisson, S. (2003) A fully probabilistic approach to extreme rainfall modeling. Journal of Hydrology **273**(1–4), 35–50.
44. Coles, S. G. (1993) Regional modelling of extreme storms via max-stable processes. Journal of the Royal Statistical Society. Series B (Methodological) **55**(4), 797–816.
45. Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. Journal of the Royal Statistical Society. Series B (Methodological) **53**(2), 377–392.
46. Coles, S. G. and 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), 1–48.
47. Cooley, D., Naveau, P. and Poncet, P. (2006) Variograms for spatial max-stable random fields. In Dependence in Probability and Statistics, eds P. Bertail, P. Soulier and P. Doukhan, volume 187 of Lecture Notes in Statistics, pp. 373–390. New York: Springer.
48. Cooley, D., Nychka, D. and Naveau, P. (2007) Bayesian spatial modeling of extreme precipitation return levels. Journal of the American Statistical Association **102**(479), 824–840.
49. Cordeiro, G. and Cribari-Neto, F. (2014) An Introduction to Bartlett Correction and Bias Reduction. Berlin: Springer,
50. Cordeiro, G. M. and Klein, R. (1994) Bias correction in ARMA models. Statistics & Probability Letters **19**(3), 169–176.
51. Cowles, M. K. and Carlin, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Association **91**(434), 883–904.
52. Cox, D. R., Isham, V. S. and Northrop, P. J. (2002) Floods: some probabilistic and statistical approaches. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **360**(1796), 1389–1408.
53. Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) **49**(1), 1–39.
54. Cox, D. R. and Reid, N. (1993) A note on the calculation of adjusted profile likelihood. Journal of the Royal Statistical Society. Series B (Methodological) **55**(2), 467–471.
55. Cox, D. R. and Snell, E. J. (1968) A general definition of residuals. Journal of the Royal Statistical Society. Series B. (Methodological) **30**(2), 248–275.
56. Cressie, N. (1993) Statistics for Spatial Data. New York: John Wiley & Sons, Inc., 928 p.
57. Daley, D. and Vere-Jones, D. (2002) An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods. New York: Springer–Verlag, 471 p.
58. Damien, P. and Walker, S. G. (2001) Sampling truncated normal, beta, and gamma densities. Journal of Computational and Graphical Statistics **10**(2), 206–215.
59. Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association **111**(514), 800–812.
60. Davis, R. A. and Mikosch, T. (2009) The extremogram: A correlogram for extreme events. Bernoulli **15**(4), 977–1009.
61. Davis, R. A., Mikosch, T. and Cribben, I. (2012) Towards estimating extremal serial dependence via the bootstrapped extremogram. Journal of Econometrics **170**(1), 142–152.
62. Davison, A. C. (1984) Modelling excesses over high thresholds, with an application. In Statistical Extremes and Applications, ed. J. de Oliveira, volume 131 of NATO ASI Series, pp. 461–482. Springer Netherlands.

63. Davison, A. C. (2003) Statistical Models. Cambridge: Cambridge University Press, 736 p.
64. Davison, A. C. (2018) 'The life of man, solitary, poore, nasty, brutish, and short': Discussion of the paper by Rootzén and Zholud. Extremes **21**(3), 365–372.
65. Davison, A. C. and Gholamrezaee, M. M. (2011) Geostatistics of extremes. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **468**(2138), 581–608.
66. Davison, A. C. and Hinkley, D. V. (1997) Bootstrap Methods and Their Application. New York: Cambridge University Press, 582 p.
67. Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modeling of spatial extremes (with discussion). Statistical Science **27**(2), 161–186.
68. Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). Journal of the Royal Statistical Society. Series B. (Methodological) **52**(3), 393–442.
69. de Oliveira, V. (2005) Bayesian inference and prediction of Gaussian random fields based on censored data. Journal of Computational and Graphical Statistics **14**(1), 95–115.
70. Deheuvels, P. (1978) Caractérisation complète des lois extrêmes multivariées et de la convergence des types extrêmes. Publications de l'Institut de statistique de l'Université de Paris **XXIII**(3), 1–36.
71. Demarta, S. and McNeil, A. J. (2005) The *t* copula and related copulas. International Statistical Review **73**(1), 111–129.
72. Devroye, L. (1986) Non-Uniform Random Variate Generation. New York: Springer,
73. Dietrich, C. R. and Newsam, G. N. (1993) A fast and exact method for multidimensional Gaussian stochastic simulations. Water Resources Research **29**(8), 2861–2869.
74. Ding, P. (2016) On the conditional distribution of the multivariate *t* distribution. The American Statistician **70**(3), 293–295.
75. Dombry, C., Engelke, S. and Oesting, M. (2016) Exact simulation of max-stable processes. Biometrika **103**(2), 303–317.
76. Dombry, C. and Eyi-Minko, F. (2013) Regular conditional distributions of continuous max-infinitely divisible random fields. Electronic Journal of Probability **18**(7), 1–21.
77. Dombry, C. and Ferreira, A. (2019) Maximum likelihood estimators based on the block maxima method. Bernoulli **25**(3), 1690–1723.
78. Dombry, C., Éyi Minko, F. and Ribatet, M. (2013) Conditional simulation of max-stable processes. Biometrika **100**(1), 111–124.
79. Dombry, C., Oesting, M. and Ribatet, M. (2015) Conditional simulation of max-stable processes. In Extreme Value Modeling and Risk Analysis Methods and Applications, eds D. K. Dey and J. Yan, pp. 215–238. Boca Raton: Chapman and Hall/CRC. ISBN 978-1-4987-0129-7.
80. Dombry, C. and Ribatet, M. (2015) Functional regular variations, Pareto processes and peaks over threshold. Statistics and Its Interface **8**(1), 9–17.
81. Dombry, C., Ribatet, M. and Stoev, S. (2018) Probabilities of concurrent extremes. Journal of the American Statistical Association **113**(524), 1565–1582.
82. Doucet, A., Pitt, M. K., Deligiannidis, G. and Kohn, R. (2015) Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Biometrika **102**(2), 295–313.
83. Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. Physics Letters B **195**, 216–222.
84. Dupuis, D. (1999) Exceedances over high thresholds: A guide to threshold selection. Extremes **1**(3), 251–261.
85. Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L. and Stordal, F. (2014) Bayesian hierarchical modeling of extreme hourly precipitation in norway. Environmetrics **26**(2), 89–106.
86. Eastoe, E. F. and Tawn, J. A. (2009) Modelling non-stationary extremes with application to surface level ozone. Journal of the Royal Statistical Society: Series C (Applied Statistics) **58**(1), 25–45.

87. Efron, B. (1975) Defining the curvature of a statistical problem (with applications to second order efficiency). The Annals of Statistics **3**(6), 1189–1242.
88. Efron, B. and Tibshirani, R. J. (1993) An Introduction to the Bootstrap. Boca Raton: CRC Press, 456 p.
89. Einmahl, J. H. J. and Segers, J. (2009) Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. The Annals of Statistics **37**(5B), 2953–2989.
90. Einmahl, J. J., Einmahl, J. H. and de Haan, L. (2019) Limits to human life span through extreme value theory. Journal of the American Statistical Association **To appear**, 1–15.
91. Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) Modelling Extremal Events — For Insurance and Finance. Berlin: Springer–Verlag, 645 p.
92. Engelke, S., Malinowski, A., Kabluchko, Z. and Schlather, M. (2015) Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **77**(1), 239–265.
93. Erhardt, R. J. and Smith, R. L. (2012) Approximate bayesian computing for spatial extremes. Computational Statistics & Data Analysis **56**(6), 1468–1481.
94. Fang, K., Kotz, S. and Ng, K. (1990) Symmetric Multivariate and Related Distributions. London: Chapman and Hall, 220 p.
95. Ferreira, A. and de Haan, L. (2014) The generalized Pareto process; with a view towards application and simulation. Bernoulli **20**(4), 1717–1737.
96. Firth, D. (1993) Bias reduction of maximum likelihood estimates. Biometrika **80**(1), 27–38.
97. Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical Proceedings of the Cambridge Philosophical Society **24**, 180–190.
98. de Fondeville, R. (2018) Functional Peaks-Over-Threshold Analysis for Complex Extreme Events. Ph.D. thesis, EPFL, Lausanne.
99. de Fondeville, R. and Davison, A. C. (2018) High-dimensional peaks-over-threshold inference. Biometrika **105**(3), 575–592.
100. Fougères, A.-L. and Mercadier, C. (2012) Risk measures and multivariate extensions of Breiman's theorem. Journal of Applied Probability **49**(2), 364–384.
101. Fraser, D. A. S. (2003) Likelihood for component parameters. Biometrika **90**(2), 327–339.
102. Fraser, D. A. S. (2004) Ancillaries and conditional inference. Statistical Science **19**(2), 333–369.
103. Fraser, D. A. S., Reid, N. and Sartori, N. (2016) Accurate directional inference for vector parameters. Biometrika **103**(3), 625–639.
104. Fraser, D. A. S., Reid, N. and Wu, J. (1999) A simple general formula for tail probabilities for frequentist and Bayesian inference. Biometrika **86**(2), 249–264.
105. Fuentes, M., Henry, J. and Reich, B. (2013) Nonparametric spatial models for extremes: application to extreme temperature data. Extremes **16**(1), 75–101.
106. Fung, T. and Seneta, E. (2018) Quantile function expansion using regularly varying functions. Methodology and Computing in Applied Probability **20**(4), 1091–1103.
107. Gabda, D., Towe, R., Wadsworth, J. and Tawn, J. (2012) Discussion of "Statistical modeling of spatial extremes" by A. C. Davison, S. A. Padoan and M. Ribatet. Statistical Science **27**(2), 189–192.
108. Galambos, J. (1975) Order statistics of samples from multivariate distributions. Journal of the American Statistical Association **70**(351, part 1), 674–680.
109. Geirsson, Ó. P., Hrafnkelsson, B. and Simpson, D. (2015) Computationally efficient spatial modeling of annual maximum 24h precipitation on a fine grid. Environmetrics **26**(5), 339–353.
110. Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013) Bayesian Data Analysis. Third edition. Boca Raton: Taylor & Francis,

111. Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. Statistical Science **7**(4), 457–472.
112. Genz, A. and Bretz, F. (2009) Computation of Multivariate Normal and *t* Probabilities. Heidelberg: Springer, 126 p.
113. Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics 4, eds J. Bernardo, J. Berger, A. Dawid and A. Smith, volume 4, pp. 169–193. Oxford: Clarendon Press.
114. Gibson, G. J., Glasbey, C. A. and Elston, D A, R. (1994) Monte Carlo evaluation of multivariate normal integrals and sensitivity to variate ordering. In Advances in Numerical Methods and Applications: Proceedings of the Third International Conference, eds I. T. Dimov, B. Sendov and V. P. S, River Edge: World Scientific Publishing, pp. 120–126.
115. Giles, D. E., Feng, H. and Godwin, R. T. (2016) Bias-corrected maximum likelihood estimation of the parameters of the generalized Pareto distribution. Communications in Statistics - Theory and Methods **45**(8), 2465–2483.
116. Giné, E., Hahn, M. G. and Vatan, P. (1990) Max-infinitely divisible and max-stable sample continuous processes. Probability Theory and Related Fields **87**(2), 139–165.
117. Gnedenko, B. (1943) Sur la distribution limite du terme maximum d'une série aléatoire. Annals of Mathematics. Second Series **44**, 423–453.
118. Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**(2), 243–268.
119. Gneiting, T., Kleiber, W. and Schlather, M. (2010) Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association **105**(491), 1167–1177.
120. Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378.
121. Godwin, R. T. and Giles, D. E. (2019) Analytic bias correction for maximum likelihood estimators when the bias function is non-constant. Communications in Statistics - Simulation and Computation **48**(1), 15–26.
122. Grimshaw, S. D. (1993) Computing maximum likelihood estimates for the generalized Pareto distribution. Technometrics **35**(2), 185–191.
123. Gumbel, É. J. (1960) Distributions des valeurs extrêmes en plusieurs dimensions. Publications de l'Institut de statistique de l'Université de Paris **9**, 171–173.
124. Gyarmati-Szabó, J., Bogachev, L. V. and Chen, H. (2017) Nonstationary pot modelling of air pollution concentrations: Statistical analysis of the traffic and meteorological impact. Environmetrics **28**(5).
125. de Haan, L. (1984) A spectral representation for max-stable processes. The Annals of Probability **12**(4), 1194–1204.
126. de Haan, L., Drees, H. and Ferreira, A. (2004) On maximum likelihood estimation of the extreme value index. The Annals of Applied Probability **14**(3), 1179–1201.
127. de Haan, L. and Ferreira, A. (2006) Extreme Value Theory: An Introduction. New York: Springer, 418 p.
128. de Haan, L. and Lin, T. (2001) On convergence toward an extreme value distribution in $C[0,1]$. The Annals of Probability **29**(1), 467–483.
129. de Haan, L. and Resnick, S. (1996) Second-order regular variation and rates of convergence in extreme-value theory. The Annals of Probability **24**(1), 97–124.
130. de Haan, L. and Stadtmüller, U. (1996) Generalized regular variation of second order. Journal of the Australian Mathematical Society. Series A. Pure Mathematics and Statistics **61**(3), 381–395.
131. Haario, H., Saksman, E. and Tamminen, J. (2001) An adaptive Metropolis algorithm. Bernoulli **7**(2), 223–242.

132. Hanayama, N. and Sibuya, M. (2016) Estimating the upper limit of lifetime probability distribution, based on data of Japanese centenarians. The Journals of Gerontology: Series A **71**(8), 1014–1021.
133. Handcock, M. S. and Stein, M. L. (1993) A Bayesian analysis of kriging. Technometrics **35**(4), 403–410.
134. Heffernan, J. E. and Resnick, S. I. (2007) Limit laws for random vectors with an extreme component. The Annals of Applied Probability **17**(2), 537–571.
135. Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) **66**(3), 497–546.
136. Heinrich, L. (2013) Asymptotic methods in statistics of random point processes. In Stochastic Geometry, Spatial Statistics and Random Fields: Asymptotic Methods, ed. E. Spodarev, pp. 115–150. Berlin: Springer.
137. Hewitt, J., Fix, M. J., Hoeting, J. A. and Cooley, D. S. (2019) Improved return level estimation via a weighted likelihood, latent spatial extremes model. Journal of Agricultural, Biological and Environmental Statistics p. To appear.
138. Ho, Z. W. O. (2018) Contributions to stochastic algorithms for Big Data and multivariate extreme value theory. Ph.D. thesis, Université Bourgogne Franche-Comté, Besançon.
139. Hofert, M., Mächler, M. and McNeil, A. J. (2012) Likelihood inference for Archimedean copulas in high dimensions under known margins. Journal of Multivariate Analysis **110**, 133–150.
140. Homan, M. D. and Gelman, A. (2014) The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. The Journal of Machine Learning Research **15**(1), 1593–1623.
141. Hosking, J. R. M. and Wallis, J. R. (1987) Parameter and quantile estimation for the generalized Pareto distribution. Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences **29**(3), 339–349.
142. Hult, H. and Lindskog, F. (2005) Extremal behavior of regularly varying stochastic processes. Stochastic Processes and their Applications **115**(2), 249–274.
143. Hult, H. and Lindskog, F. (2006) Regular variation for measures on metric spaces. Publications de l'Institut Mathématique. Nouvelle Série **80**(94), 121–140.
144. Huser, R. and Davison, A. C. (2013) Composite likelihood estimation for the Brown–Resnick process. Biometrika **100**(2), 511–518.
145. Huser, R., Davison, A. C. and Genton, M. G. (2016) Likelihood estimators for multivariate extremes. Extremes **19**(1), 79–103.
146. Huser, R., Dombry, C., Ribatet, M. and Genton, M. G. (2019) Full likelihood inference for max-stable data. Stat **8**(1), 1–14.
147. Huser, R., Opitz, T. and Thibaud, E. (2017) Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. Spatial Statistics **21**(Part A), 166–186.
148. Huser, R. and Wadsworth, J. L. (2019) Modeling spatial processes with unknown extremal dependence class. Journal of the American Statistical Association **114**(525), 434–444.
149. Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**, 695–709.
150. Jalbert, J., Favre, A.-C., Bélisle, C. and Angers, J.-F. (2017) A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. Journal of the Royal Statistical Society: Series C (Applied Statistics) **66**(5), 941–962.
151. Joe, H. (1990) Families of min-stable multivariate exponential and multivariate extreme value distributions. Statistics & Probability Letters **9**(1), 75–81.
152. Kabluchko, Z. (2011) Extremes of independent Gaussian processes. Extremes **14**(3), 285–310.
153. Kabluchko, Z., Schlather, M. and de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. The Annals of Probability **37**(5), 2042–2065.

154. Kaufman, C. G. and Shaby, B. A. (2013) The role of the range parameter for estimation and prediction in geostatistics. Biometrika **100**(2), 473–484.
155. Keef, C., Tawn, J. A. and Lamb, R. (2013) Estimating the probability of widespread flood events. Environmetrics **24**(1), 13–21.
156. Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017) Median bias reduction of maximum likelihood estimates. Biometrika **104**(4), 923–938.
157. Kent, J. T. (1982) Robust properties of likelihood ratio test. Biometrika **69**(1), 19–27.
158. Kiriliouk, A., Rootzén, H., Segers, J. and Wadsworth, J. L. (2019) Peaks over thresholds modeling with multivariate generalized Pareto distributions. Technometrics **61**(1), 123–135.
159. Kotz, S. and Nadarajah, S. (2000) Extreme Value Distributions. London: Imperial College Press, 196 p.
160. Krüger, F., Lerch, S., Thorarinsdottir, T. L. and Gneiting, T. (2017) Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. arXiv e-prints arXiv:1608.06802.
161. Lawley, D. N. (1956) A general method for approximating to the distribution of likelihood ratio criteria. Biometrika **43**(3/4), 295–303.
162. Ledford, A. W. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. Biometrika **83**(1), 169–187.
163. Lee, S. M. and Young, G. A. (2005) Parametric bootstrapping with nuisance parameters. Statistics & Probability Letters **71**(2), 143–153.
164. Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017) Forecaster's dilemma: Extreme events and forecast evaluation. Statistical Science **32**(1), 106–127.
165. van Lieshout, M.-C. (2010) Spatial point process theory. In Handbook of Spatial Statistics, eds A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp, pp. 263–282. Chapman & Hall.
166. Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(4), 423–498.
167. Lindskog, F., Resnick, S. I. and Roy, J. (2014) Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. Probability Surveys **11**, 270–314.
168. Lindsten, F. and Doucet, A. (2016) Pseudo-Marginal Hamiltonian Monte Carlo. arXiv e-prints p. arXiv:1607.02516.
169. Liu, J. S., Wong, W. H. and Kong, A. (1995) Covariance structure and convergence rate of the Gibbs sampler with various scans. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 157–169.
170. Liu, Y. and Tawn, J. (2014) Self-consistent estimation of conditional multivariate extreme value distributions. Journal of Multivariate Analysis **127**, 19–35.
171. Liu, Z., Blanchet, J. H., Dieker, A. B. and Mikosch, T. (2016) Optimal exact simulation of max-stable and related random fields. ArXiv e-prints arXiv:1609.06001.
172. Lugrin, T. (2018) Semiparametric Bayesian Risk Estimation for Complex Extremes. Ph.D. thesis, EPFL, Lausanne.
173. Maire, F., Friel, N., Mira, A. and Raftery, A. E. (2019) Adaptive incremental mixture markov chain monte carlo. Journal of Computational and Graphical Statistics (To appear), 1–48.
174. Marani, M. and Ignaccolo, M. (2015) A metastatistical approach to rainfall extremes. Advances in Water Resources **79**, 121–126.
175. Marani, M. and Zanetti, S. (2015) Long-term oscillations in rainfall extremes in a 268 year daily time series. Water Resources Research **51**(1), 639–647.
176. Martins, E. S. and Stedinger, J. R. (2000) Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resources Research **36**(3), 737–744.
177. McCullagh, P. (2018) Tensor Methods in Statistics. Second edition. Mineola: Dover Publications, 304 p.

178. McCullagh, P. and Tibshirani, R. (1990) A simple method for the adjustment of profile likelihoods. Journal of the Royal Statistical Society. Series B (Methodological) **52**(2), 325–344.
179. Meinguet, T. and Segers, J. (2010) Regularly varying time series in Banach spaces. ArXiv e-prints arXiv:1001.3262.
180. Møller, J. and Waagepetersen, R. (2003) Statistical Inference and Simulation for Spatial Point Processes. Boca Raton: CRC Press, 320 p.
181. Naveau, P., Guillou, A., Cooley, D. and Diebolt, J. (2009) Modelling pairwise dependence of maxima in space. Biometrika **96**(1), 1–17.
182. Neal, R. M. (2011) Optimal proposal distributions and adaptive MCMC. In Handbook of Markov Chain Monte Carlo, eds S. Brooks, A. Gelman, G. Jones and X. Meng, pp. 113–162. Boca Raton: CRC Press.
183. Nikoloulopoulos, A. K., Joe, H. and Li, H. (2009) Extreme value properties of multivariate t copulas. Extremes **12**(2), 129–148.
184. Northrop, P. J., Attalides, N. and Jonathan, P. (2016) Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. Journal of the Royal Statistical Society: Series C (Applied Statistics) pp. 93–120.
185. Northrop, P. J. and Coleman, C. L. (2014) Improved threshold diagnostic plots for extreme value analyses. Extremes **17**(2), 289–303.
186. Nuyens, D. and Cools, R. (2006) Fast component-by-component construction, a reprise for different kernels. In Monte Carlo and Quasi-Monte Carlo Methods 2004, eds H. Niederreiter and D. Talay, Heidelberg: Springer, pp. 373–387.
187. O'Brien, G. L. (1987) Extreme values for stationary and Markov sequences. The Annals of Probability **15**(1), 281–291.
188. Oesting, M. (2018) Equivalent representations of max-stable processes via ℓ_p-norms. Journal of Applied Probability **55**(1), 54–68.
189. Oesting, M., Ribatet, M. and Dombry, C. (2015) Simulation of max-stable processes. In Extreme Value Modeling and Risk Analysis Methods and Applications, eds D. K. Dey and J. Yan, pp. 195–214. Boca Raton: Chapman and Hall/CRC. ISBN 978-1-4987-0129-7.
190. Oesting, M., Schlather, M. and Friederichs, P. (2017) Statistical post-processing of forecasts for extremes using bivariate Brown–Resnick processes with an application to wind gusts. Extremes **20**(2), 309–332.
191. Oesting, M., Schlather, M. and Schillings, C. (2019) Sampling sup-normalized spectral functions for Brown–Resnick processes. Stat **8**(1), e228.
192. Oesting, M., Schlather, M. and Zhou, C. (2018) Exact and fast simulation of max-stable processes on a compact set using the normalized spectral representation. Bernoulli **24**(2), 1497–1530.
193. Opitz, T. (2013a) Extremal t processes: Elliptical domain of attraction and a spectral representation. Journal of Multivariate Analysis **122**, 409–413.
194. Opitz, T. (2013b) Extrêmes multivariés et spatiaux : approches spectrales et modèles elliptiques. Ph.D. thesis, Université Montpellier 2.
195. Opitz, T. (2016) Modeling asymptotically independent spatial extremes based on laplace random fields. Spatial Statistics **16**, 1–18.
196. Owen, A. B. (2013) Self-concordance for empirical likelihood. Canadian Journal of Statistics **41**(3), 387–397.
197. Padoan, S. A., Ribatet, M. and Sisson, S. A. (2010) Likelihood-based inference for max-stable processes. Journal of the American Statistical Association **105**(489), 263–277.
198. Pakman, A. and Paninski, L. (2014) Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. Journal of Computational and Graphical Statistics **23**(2), 518–542.
199. Papalexiou, S. M. and Koutsoyiannis, D. (2013) Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resources Research **49**(1), 187–201.

200. Papastathopoulos, I. and Tawn, J. A. (2013) Extended generalised Pareto models for tail estimation. Journal of Statistical Planning and Inference **143**(1), 131–143.
201. Pickands, III, J. (1986) The continuous and differentiable domains of attraction of the extreme value distributions. The Annals of Probability **14**(3), 996–1004.
202. Pires, J. F., Cysneiros, A. H. M. A. and Cribari-Neto, F. (2018) Improved inference for the generalized Pareto distribution. Brazilian Journal of Probability and Statistics **32**(1), 69–85.
203. Prescott, P. and Walden, A. T. (1980) Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. Biometrika **67**(3), 723–724.
204. Reich, B. J. and Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. The Annals of Applied Statistics **6**(4), 1430–1451.
205. Resnick, S. I. (1987) Extreme values, regular variation, and point processes. New York: Springer–Verlag, xii+320 p.
206. Resnick, S. I. (2007) Heavy-Tail Phenomena. New York: Springer, 404 p.
207. Ribatet, M. (2013) Spatial extremes: max-stable processes at work. Journal de la Société Française de Statistique **154**(2), 156–177.
208. Ribatet, M., Cooley, D. and Davison, A. C. (2012) Bayesian inference from composite likelihoods, with an application to spatial extremes. Statistica Sinica **22**(2), 813–845.
209. Robert, C. P. (1995) Simulation of truncated normal variables. Statistics and Computing **5**(2), 121–125.
210. Robert, C. P. and Casella, G. (2005) Monte Carlo Statistical Methods (Springer Texts in Statistics). Berlin: Springer-Verlag,
211. Rocke, D. M. (1989) Bootstrap Bartlett adjustment in seemingly unrelated regression. Journal of the American Statistical Association **84**(406), 598–601.
212. Rodriguez-Yam, G., A Davis, R. and Scharf, L. (2004) Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. Technical report.
213. Roodman, D. (2018) Bias and size corrections in extreme value modeling. Communications in Statistics - Theory and Methods **47**(14), 3377–3391.
214. Rootzén, H., Segers, J. and Wadsworth, J. L. (2018) Multivariate peaks over thresholds models. Extremes **21**(1), 115–145.
215. Rootzén, H. and Zholud, D. (2017) Human life is unlimited — but short. Extremes **20**(4), 713–728.
216. Rootzén, H. and Katz, R. W. (2013) Design life level: Quantifying risk in a changing climate. Water Resources Research **49**(9), 5964–5972.
217. Rootzén, H., Segers, J. and Wadsworth, J. L. (2018) Multivariate generalized pareto distributions: Parametrizations, representations, and properties. Journal of Multivariate Analysis **165**, 117–131.
218. Rootzén, H. and Tajvidi, N. (2006) Multivariate generalized Pareto distributions. Bernoulli **12**(5), 917–930.
219. Rosenthal, J. S. (2011) Optimal proposal distributions and adaptive MCMC. In Handbook of Markov Chain Monte Carlo, eds S. Brooks, A. Gelman, G. Jones and X. Meng, pp. 93–111. Boca Raton: CRC Press.
220. Rue, H. (2001) Fast sampling of Gaussian Markov random fields. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **63**(2), 325–338.
221. Rue, H. and Held, L. (2005) Gaussian Markov Random Fields: Theory and Applications. Boca Raton: CRC Press, 280 p.
222. Sabourin, A. and Segers, J. (2017) Marginal standardization of upper semicontinuous processes. with application to max-stable processes. Journal of Applied Probability **54**(3), 773–796.
223. Sang, H. and Gelfand, A. E. (2010) Continuous spatial process models for spatial extreme values. Journal of Agricultural, Biological, and Environmental Statistics **15**(1), 49–65.
224. Sansó, B. and Guenni, L. (1999) Venezuelan rainfall data analysed by using a Bayesian space-time model. Journal of the Royal Statistical Society: Series C (Applied Statistics) **48**(3), 345–362.

225. Scarrott, C. and MacDonald, A. (2012) A review of extreme-value threshold estimation and uncertainty quantification. REVSTAT – Statistical Journal **10**(1), 33–60.
226. Schlather, M. (2002) Models for stationary max-stable random fields. Extremes **5**(1), 33–44.
227. Schlather, M. and Moreva, O. (2017) A parametric model bridging between bounded and unbounded variograms. Stat **6**(1), 47–52.
228. Schlather, M. and Tawn, J. A. (2003) A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika **90**(1), 139–156.
229. Sebille, Q. (2016) Modélisation spatiale de valeurs extrêmes: application à l'étude de précipitations en France. Ph.D. thesis, Université de Lyon.
230. Segers, J. (2012) Max-stable models for multivariate extremes. REVSTAT **10**(1), 61–82.
231. Segers, J., Sibuya, M. and Tsukahara, H. (2017) The empirical beta copula. Journal of Multivariate Analysis **155**, 35 – 51.
232. Serinaldi, F. and Kilsby, C. G. (2014) Rainfall extremes: Toward reconciliation after the battle of distributions. Water Resources Research **50**(1), 336–352.
233. Severini, T. (2000) Likelihood Methods in Statistics. New York: Oxford University Press, 392 p.
234. Severini, T. A. (1999) An empirical adjustment to the likelihood ratio statistic. Biometrika **86**(2), 235–247.
235. Shaby, B. A. (2014) The open-faced sandwich adjustment for MCMC using estimating functions. Journal of Computational and Graphical Statistics **23**(3), 853–876.
236. Shaby, B. A. and 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**(8), 638–648.
237. Sharkey, P. and Tawn, J. A. (2017) A Poisson process reparameterisation for Bayesian inference for extremes. Extremes **20**(2), 239–263.
238. Sharkey, P. and Winter, H. C. (2019) A Bayesian spatial hierarchical model for extreme precipitation in Great Britain. Environmetrics **30**(1).
239. Shi, D. (1995) Fisher information for a multivariate extreme value distribution. Biometrika **82**(3), 644–649.
240. Skovgaard, I. M. (1996) An explicit large-deviation approximation to one-parameter tests. Bernoulli **2**(2), 145–165.
241. Skovgaard, I. M. (2001) Likelihood asymptotics. Scandinavian Journal of Statistics **28**(1), 3–32.
242. Smith, R. L. (1985) Maximum likelihood estimation in a class of nonregular cases. Biometrika **72**(1), 67–90.
243. Smith, R. L. (1987a) Approximations in extreme value theory. Center for Stochastic Processes, University of North Carolina Chapel Hill, Technical Report 205.
244. Smith, R. L. (1987b) Estimating tails of probability distributions. The Annals of Statistics **15**(3), 1174–1207.
245. Smith, R. L. (1988) Bias and variance approximations for estimators of extreme quantiles. Center for Stochastic Processes, University of North Carolina Chapel Hill, Technical Report 249.
246. Smith, R. L. (1990) Max-stable processes and spatial extremes. Unpublished technical report.
247. Smith, R. L., Tawn, J. A. and Coles, S. G. (1997) Markov chain models for threshold exceedances. Biometrika **84**(2), 249–268.
248. Stacy, E. W. (1962) A generalization of the gamma distribution. Ann. Math. Stat. **33**(3), 1187–1192.
249. Stein, M. L. (1999) Interpolation of Spatial Data: Some Theory for Kriging. New York: Springer, 249 p.
250. Stephenson, A. (2003) Simulating multivariate extreme value distributions of logistic type. Extremes **6**(1), 49–59.
251. Stephenson, A. and Tawn, J. (2005) Exploiting occurrence times in likelihood inference for componentwise maxima. Biometrika **92**(1), 213–227.

252. Stephenson, A. G. (2009) High-dimensional parametric modelling of multivariate extreme events. Australian & New Zealand Journal of Statistics **51**(1), 77–88.
253. Süveges, M. (2007) Likelihood estimation of the extremal index. Extremes **10**(1), 41–55.
254. Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association **82**(398), 528–540.
255. Tawn, J., Shooter, R., Towe, R. and Lamb, R. (2018) Modelling spatial extreme events with environmental applications. Spatial Statistics **28**, 39 – 58.
256. Tawn, J. A. (1990) Modelling multivariate extreme value distributions. Biometrika **77**(2), 245–253.
257. Thibaud, E. (2014) Contributions to spatial statistics: species distributions and rare events. Ph.D. thesis, EPFL, Lausanne.
258. Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C. and Heikkinen, J. (2016) Bayesian inference for the Brown–Resnick process, with an application to extreme low temperatures. The Annals of Applied Statistics **10**(4), 2303–2324.
259. Thibaud, E. and Opitz, T. (2015) Efficient inference and simulation for elliptical Pareto processes. Biometrika **102**(4), 855–870.
260. Tsai, W.-Y., Jewell, N. P. and Wang, M.-C. (1987) A note on the product-limit estimator under right censoring and left truncation. Biometrika **74**(4), 883–886.
261. Varin, C. (2008) On composite marginal likelihoods. AStA - Advances in Statistical Analysis **92**(1), 1–28.
262. Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. Statistica Sinica **21**, 5–42.
263. Vatan, P. (1985) Max-infinite divisibility and max-stability in infinite dimensions. In Probability in Banach Spaces V: Proceedings of the International Conference held in Medford, USA, July 16–27, 1984, eds A. Beck, R. Dudley, M. Hahn, J. Kuelbs and M. Marcus, pp. 400–425. Berlin: Springer.
264. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.-C. (2019) Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. arXiv e-prints arXiv:1903.08008.
265. Wadsworth, J. and Tawn, J. (2013) A new representation for multivariate tail probabilities. Bernoulli **19**(5B), 2689–2714.
266. Wadsworth, J. L. (2015) On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. Biometrika **102**(3), 705–711.
267. Wadsworth, J. L. (2016) Exploiting structure of maximum likelihood estimators for extreme value threshold selection. Technometrics **58**(1), 116–126.
268. Wadsworth, J. L. and Tawn, J. A. (2012) Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **74**(3), 543–567.
269. Wadsworth, J. L. and Tawn, J. A. (2014) Efficient inference for spatial extreme value processes associated to log-gaussian random functions. Biometrika **101**(1), 1–15.
270. Wadsworth, J. L. and Tawn, J. A. (2018) Spatial conditional extremes. Technical report.
271. Wadsworth, J. L., Tawn, J. A., Davison, A. C. and Elton, D. M. (2017) Modelling across extremal dependence classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(1), 149–175.
272. Wadsworth, J. L., Tawn, J. A. and Jonathan, P. (2010) Accounting for choice of measurement scale in extreme value modeling. The Annals of Applied Statistics **4**(3), 1558–1578.
273. Wakefield, J. C., Gelfand, A. E. and Smith, A. F. M. (1991) Efficient generation of random variates via the ratio-of-uniforms method. Statistics and Computing **1**(2), 129–133.
274. Wallin, J. and Bolin, D. (2015) Geostatistical modelling using non-Gaussian Matérn fields. Scandinavian Journal of Statistics **42**(3), 872–890.

275. Wang, Y. and Stoev, S. A. (2011) Conditional sampling for spectrally discrete max-stable random fields. Advances in Applied Probability **43**(2), 461–483.
276. Wellner, J. A. and Hall, W. J. (1980) Confidence bands for a survival curve from censored data. Biometrika **67**(1), 133–143.
277. White, H. (1982) Maximum likelihood estimation of misspecified models. Econometrica **50**(1), 1–25.
278. Wieczorek, G., Larsen, M., Eaton, L., Morgan, B. and Blair, J. L. (2001) Debris-flow and flooding hazards associated with the December 1999 storm in coastal Venezuela and strategies for mitigation. Technical Report 01-0144, U.S. Geological Survey.
279. Wood, A. T. A. and Chan, G. (1994) Simulation of stationary Gaussian processes in $[0,1]^d$. Journal of Computational and Graphical Statistics **3**(4), 409–432.
280. Yu, J.-W. and Tian, G.-L. (2011) Efficient algorithms for generating truncated multivariate normal distributions. Acta Mathematicae Applicatae Sinica, English Series **27**(4), 601–612.
281. Yuen, R. and Guttorp, P. (2014) A hierarchical Gauss-Pareto model for spatial prediction of extreme precipitation. Technical Report 535, University of Michigan.
282. Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. Journal of the American Statistical Association **99**(465), 250–261.
283. Zorzetto, E., Botter, G. and Marani, M. (2016) On the emergence of rainfall extremes from ordinary events. Geophysical Research Letters **43**(15), 8076–8082.

R packages

284. Belzile, L., Wadsworth, J. L., Northrop, P. J., Grimshaw, S. D. and Huser, R. (2019) mev: Multivariate Extreme Value Distributions. R package version 1.12.
285. Botev, Z. and Belzile, L. (2019) TruncatedNormal: Truncated Multivariate Normal and Student Distributions. R package version 1.1.10.
286. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. Journal of Statistical Software **76**(1), 1–32.
287. de Fondeville, R. and Belzile, L. (2018) mvPot: Multivariate Peaks-over-Threshold Modelling for Spatial Extreme Events. R package version 0.1.5.
288. Hofert, M. and Mächler, M. (2016a) Parallel and other simulations in R made easy: An end-to-end study. Journal of Statistical Software **69**(4), 1–44.
289. Hofert, M. and Mächler, M. (2016b) Parallel and other simulations in R made easy: An end-to-end study. Journal of Statistical Software **69**(4), 1–44.
290. R Development Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
291. Northrop, P. J. (2019a) revdbayes: Ratio-of-Uniforms Sampling for Bayesian Extreme Value Analysis. R package version 1.3.3.
292. Northrop, P. J. (2019b) rust: Ratio-of-Uniforms Simulation with Transformation. R package version 1.3.6.
293. Ribatet, M. (2018) SpatialExtremes: Modelling Spatial Extremes. R package version 2.0-7.
294. Schlather, M., Malinowski, A., Menck, P. J., Oesting, M. and Strokorb, K. (2015) Analysis, simulation and prediction of multivariate random fields with package RandomFields. Journal of Statistical Software **63**(8), 1–25.
295. Stan Development Team (2018) RStan: the R interface to Stan. R package version 2.18.2.
296. Stephenson, A. G. (2002) evd: Extreme value distributions. R News **2**(2), 31–32.

Index

A

ancillarity 39, 40, 44

B

Barndorff-Nielsen's r^* approximation 40
Bartlett identities 27, 32
Bartlett's correction 37, 38
Bayesian hierarchical model 158, 174, 175, 188
Bayesian model averaging 175
bias 165
  asymptotic 13, 15
  Cox–Snell 32, 33
  Firth's modified score 34
bootstrap 47, 65
Breiman's lemma 81, 88, 100, 128, 156, 171
Brown–Resnick process . 87, 165, 178, 188, 206
  censored likelihood 128, 129
  exponent measure 90
  extremal coefficient 145
  gradient score 134
  intensity 88, 89
  parametrization 91
  simulation 110

C

Campbell's theorem 74
composite likelihood 131
  information criterion 132
cone 78
  bounded away 78
covariance 45
cumulant 26, 49
  generating function 26

D

data augmentation 160, 178
dataset
  Danish insurance 43
  Fremantle sea height 44
  Italian supercentenarian 63
  Maiquetia rainfall 9, 62
  Padova rainfall 23
  Swiss rainfall 142, 147, 149, 150
de Haan's representation 83
distribution
  Burr 16, 18
  Gaussian 12
  generalized extreme value . 4, 8, 11, 46, 49
  generalized gamma 16, 18
  generalized Pareto 4, 8, 11, 46
  lognormal 12
  normal 15
  Student t 17
  Weibull 15, 17, 18

E

effective sample size 163
ergodicity 76
exponent measure 85, 86
extremal attractor 4, 7, 8, 19
extremal coefficient 139
  F-madogram estimator 146
  λ-madogram estimator 147
  properties 145
  Schlather's estimator 145
  Smith's estimator 146
extremal concurrence probability 148
extremal function 107, 110
  conditional distribution 115
extremal Student process 88, 165, 178, 188, 206
  exponent measure 91
  gradient score 135
  intensity 89
  Pareto process
    unconditional simulation 117

simulation
unconditional 111
extremogram 142

G

Gaussian scale mixture 156, 171, 173
generalized Pareto
Bartlett's correction 37
Cox–Snell bias 33
profile likelihood 30
quantiles 42
generalized ℓ-Pareto process 102, 156
likelihood 126
simulation 120
geometric anisotropy 78
Godambe information 132
gradient score 134
Grimshaw's algorithm 31, 49

H

Heffernan–Tawn model 104, 106, 122, 136
hierarchical kernel extreme value process . 86, 176
hitting scenario 106, 115

I

information
Fisher 26–28
observed 26
observed (profile) 30

K

Kullback–Leibler divergence 18

L

latent Gaussian model 175
Ledford–Tawn likelihood 131
likelihood 26
Cox–Reid 42, 44
penalized profile 45, 53
profile 29, 46, 47, 65
likelihood ratio test 35, 47, 52, 65
linearization 35, 42

M

Markov chain Monte Carlo 115, 120, 158
diagnostics 163
Gibbs sampling 159, 174
Hamiltonian Monte Carlo 69, 180
Metropolis adjusted Langevin 159
Metropolis–Hastings 158, 160
Metropolis-within-Gibbs ... 159, 177, 188
proposal covariance 159, 164
max-stability 4, 9, 52
max-stable distribution
asymmetric logistic 93
asymmetric negative logistic 93
Brown–Resnick 87–91, 110, 128, 129, 134, 145
extremal Dirichlet 95
extremal Student 88, 89, 91, 111, 117, 135
Hüsler–Reiss 96
logistic 92, 180
multilogistic 94
negative logistic 93
scaled extremal Dirichlet 94, 95
max-stable process 81
angular measure 85
nonparametric estimation 136
change of measure 84
convergence to 82
likelihood 124
ℓ_p representation 87
simulation 106
conditional 113
unconditional 108
spectral representation 83, 89
unconditional simulation 109, 110
uniqueness 84
measure 2
mapping theorem 79
pushforward 80
weak-hash convergence 79
metastatistical model 17
minimax exponential tilting 167, 169, 178
multivariate extreme value 86

N

NHPP 5
likelihood 6, 27, 31
nuisance parameter 30

O

orthogonal 41

P

ℓ-Pareto process 99

change of measure 98
likelihood 128
simulation
accept-reject 117
composition sampling 118
MCMC methods 120
rejection sampling 116
penultimate approximation . 10–12, 15, 16, 19, 21
pivot 7, 40, 44
point process 73
homogeneous 74
intensity 74
locally finite 73
marked 76
moments 74
simple 74
Poisson process 74
simulation 75
transformation 75
posterior distribution 62, 158
projection 79
pseudo-likelihood 136
pseudo-marginal 161
pseudo-polar decomposition 79, 85, 99
p^* approximation 39

R

r-largest
information 28
likelihood 6
regular variation 7
functional 80
second order generalized 14
reparametrization 41
residuals 7
return level 9, 12, 17, 52
$\widehat{R}$ 164
risk functional 97

S

scaled extremal Dirichlet
angular density 96
intensity 95
spectral representation 94
score 26, 45
score test 22, 36
separation of variables 165, 169, 178
variable reordering 168
Slivnyak–Mecke theorem 76, 109
stationary
intrinsic 77, 105
strict 76
weak 77
Stephenson–Tawn likelihood 125
sub-extremal function 107

T

tangent exponential model 44, 45
threshold selection
changepoint tests 22
extended generalized Pareto 23
robust 20
stability plot 21
white noise plot 21
threshold stability 4

V

variance reduction factor 28
variogram 77, 89
generalized Cauchy 228
Matérn 228
nugget 227
power 187
power exponential 228
power law 228
Schlather's 229
von Mises conditions 8

W

Wald statistic 35, 52
weak-hash convergence 79

Léo Belzile

Canadian citizen
curriculum vitæ

Assistant professor
Département de sciences de la décision
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec), Canada
H3T 2A7
✉ leo.belzile@hec.ca

Education

2014–2019 **École doctorale en mathématiques**
Director: Pr. Anthony C. Davison
École polytechnique fédérale de Lausanne
Thesis: *Contributions to Likelihood-Based Modelling of Extreme Values*

2013–2014 **Master of Science**
Statistics, CGPA 4/4
Université McGill, Montréal

2010–2013 **Bachelor of Science**
Probability and Statistics
Minor in Economics
First Class Honours, CGPA 3.73/4
Université McGill, Montréal

Publications

ORCID: 0000-0002-9135-014X

- Belzile, Léo R. and Johanna G. Nešlehová; *Extremal attractors of Liouville copulas*, Journal of Multivariate Analysis (2017), 160C, pp. 68–92. doi.org/10.1016/j.jmva.2017.05.008
- Saarela, Olli, Belzile, Léo R. and David A. Stephens; *A Bayesian view of doubly robust causal inference* (2016), Biometrika, 103 (3): 667-681. doi:10.1093/biomet/asw025
- Raymond-Belzile, Léo and Johanna G. Nešlehová (supervisor) (2014); *Extremal and inferential properties of Liouville copulas*, master's thesis, 125 p.

Technical Skills

I mostly carry out my statistical analysis in R, partly using the C^{++} Armadillo library. Most of my code is available online on Github.

R packages:

- `lcopula` [CRAN] (author, maintainer).
- `mev` [CRAN] (author, maintainer)
- `TruncatedNormal` [CRAN] (author, maintainer).
- `BMAmevt` [CRAN] (maintainer)
- `mvPot` [CRAN] (author, contributor)

Language: French (native), English (fluent)

Course Experience (Graduate)

Extensive course experience (18 session long graduate courses), 2 specialized schools and 24 continuing education mini-courses offered mostly with the CUSO doctoral program (4.5 hours each) on specialized topics.

Graduate courses completed include:
Algorithmic Game Theory • Stochastic Processes • Introduction to Time Series Analysis • Seminar (Statistical Computation in R) • Mathematical Statistics 1 and 2 • Generalized Linear Models • Computation Intensive Statistics • Models for Financial Economics • Advanced Probability 1 • Analysis of Extreme Values with Applications to Financial Engineering • Reading course (Bayesian Statistics) • PASI Spatio-Temporal Statistics Winter School • Advanced Topics in Statistics of Extremes • Principles of Statistical Inference • Statistique multivariée • Extreme Value Modeling and Water Resources Summer School • Robust and nonparametric statistics • Introductory course about Spatial Statistics • Multivariate Extreme Value and Max-Stable Processes

Academic Work Experience

Teaching assistantships: At EPFL: Applied biostatistics, Linear Models (×2), Time series, Probabilités et statistique, Statistique multivariée (×2), Risk, rare events and extremes, Algèbre linéaire. At McGill: Calculus 2.
I have a keen interest in evidence-based learning methods. At EPFL, I followed training at the teaching support center (Teaching Toolkit, Effective Interactive Teaching and Learning).

Extracurricular Involvement

- Chair (July 2019– June 2020) and member (July 2018–June 2020) of the Committee on membership, Statistical Society of Canada,
- PhD Student Representative and delegate to the PhD student representative, EDMA, (October 2016–July 2019)
- Organizer, Recent Advances in Extremes Winter School (June 2016 - February 2017)
- Organizer of the CUSO Career Day 2015, (January 2015–November 2015)
- Chair (July 2015–March 2017) and member (August 2014–June 2017) of the Student and Recent Graduate Committee, Statistical Society of Canada
- Organizer of the 14th Graduate Colloquium of the Swiss Doctoral Program in Mathematics, (December 2014–February 2015)
- Cultural ambassador, Vidy Theatre (October 2014–October 2017)
- QED representative to SMA (November 2014–December 2015)
- Quebec representative on the Student Committee, Canadian Mathematical Society (CMS-Studc), (July 2013–July 2014)
- Organizer of the 2014 Seminars in Undergraduate Mathematics in Montreal (SUMM), (November 2013–May 2014)
- Editor-in-chief of *The Delta-Epsilon*, (January 2013–May 2014)
- President and organizer of the 2013 Seminars in Undergraduate Mathematics in Montreal (SUMM), (October 2012–November 2013)
- Organizer of the XVI Colloque panquébécois des étudiants de l'Institut des sciences mathématiques, (April 2013–May 2013)
- Student representative to the Committee of graduate affairs (CGA), (September 2013–April 2014)
- Student representative to the Committee of undergraduate affairs (CUA), (September 2012–April 2013)
- VP-Academics of the Society of Undergraduate Mathematics Students (SUMS), (May 2012–April 2013)

(a) I have experience organizing events involving large numbers of attendees (ranging from 40 to 100). Planning usually involves funding requests to sponsors, production of financial audits and reports, handling of logistics, coordinating a team of volunteers during the event, and overseeing meetings.

(b) My work of representative on the CMS and the SSC has focused on elaborating new communication policies, increasing involvement in the SSC and promoting accreditation of programs.

(c) As a student committee member, I have written reports identifying problems in programs, notably workload balance. At the undergraduate level, I was involved in the changes in the analysis syllabus, pushed for the creation of an undergraduate major in statistics as well as a rotation of the specialized courses offer. At EPFL, I pioneered voluntary teaching evaluations and spearheaded the response of PhD student representatives to a teaching workload reform.

Conference and presentations

- *Likelihood inference for univariate extremes: higher-order asymptotics* (contributed talk), Extreme Value Analysis, Zagreb, Croatia, 2019-07-04.
- *Multivariate extreme values in R: the* mev *package* (invited talk), Extreme Value Analysis, Zagreb, Croatia, 2019-06-30.
- SpatialExtremes package (invited talk), Extreme Value Analysis, Zagreb, Croatia, 2019-06-30.
- *Modèles hiérarchiques bayésiens pour excès de seuils de processus spatiaux* (invited talk), HEC Montréal, 2018-12-10; Université de Sherbrooke, 2018-12-14; Université Laval, 2019-01-08.
- *A Bayesian Hierarchical Model for Spatial Extremes* (contributed talk), SSC Annual Meeting, Montreal, 2018-06-04.
- *Extremal attractors of Liouville copulas* (contributed talk), Extreme Value Analysis, Delft, Netherlands, 2017-06-26.
- *Extreme values, from theory to practice* (invited talk), CM Stat, Sevilla, Spain, 2016-12-09.
- *A Bayesian View of Doubly Robust Causal Inference* (poster), SIAM Uncertainty Quantification, Lausanne. 2016-04-05.
- *Extremal properties of Liouville copulas* (poster), Extreme-Value Analysis 2015, Ann Arbor. 2015-06-15.

Awards

2014–2017*	**CGS D Alexander-Graham-Bell**
105 000$	NSERC
2014–2017	**Postgraduate scholarship (PGS D)**
63 000$	NSERC
2014–2017*	**Bourses de doctorat en recherche**
60 000$	FRQNT
2012*	**Undergraduate Summer Scholarship**
5000$	Institut des Sciences Mathématiques
2012*	**Arts Undergrad. Research Internship**
4000$	McGill University
2010–2011	**J. W. McConnell Scholarship**
3000$	McGill University
2009–2011	**Millenium Excellence Award Bursary**
11,500$	Canada Millenium Scholarship Foundation
2007	**Governor general bronze medal**
	for best academic average
	Lieutenant governor diploma
	for social implication

(*) Award declined

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**