Biometrika (2018), **105**, 3, *pp.* 575–592
Printed in Great Britain

doi: 10.1093/biomet/asy026
Advance Access publication 21 June 2018

High-dimensional peaks-over-threshold inference

BY R. DE FONDEVILLE AND A. C. DAVISON

Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Station 8,
1015 Lausanne, Switzerland

raphael.de-fondeville@epfl.ch anthony.davison@epfl.ch

SUMMARY

Max-stable processes are increasingly widely used for modelling complex extreme events, but existing fitting methods are computationally demanding, limiting applications to a few dozen variables. r-Pareto processes are mathematically simpler and have the potential advantage of incorporating all relevant extreme events, by generalizing the notion of a univariate exceedance. In this paper we investigate the use of proper scoring rules for high-dimensional peaks-over-threshold inference, focusing on extreme-value processes associated with log-Gaussian random functions, and compare gradient score estimators with the spectral and censored likelihood estimators for regularly varying distributions with normalized marginals, using data with several hundred locations. When simulating from the true model, the spectral estimator performs best, closely followed by the gradient score estimator, but censored likelihood estimation performs better with simulations from the domain of attraction, though it is outperformed by the gradient score in cases of weak extremal dependence. We illustrate the potential and flexibility of our ideas by modelling extreme rainfall on a grid with 3600 locations, based on exceedances for locally intense and for spatially accumulated rainfall, and discuss diagnostics of model fit. The differences between the two fitted models highlight how the definition of rare events affects the estimated dependence structure.

Some key words: Functional regular variation; Gradient score; Pareto process; Peaks-over-threshold analysis; Proper scoring rule; Statistics of extremes.

1. INTRODUCTION

Recent contributions to extreme value theory describe models capable of handling spatiotemporal phenomena (e.g., Kabluchko et al., 2009) and provide a flexible framework for modelling rare events, but their complexity makes inference difficult, if not intractable, for high-dimensional data. For instance, the number of terms in the block maximum likelihood for a Brown–Resnick process grows with the dimension D like the Bell numbers (Huser & Davison, 2013), so computationally cheaper methods such as composite likelihood (Padoan et al., 2010) or the inclusion of partition information (Stephenson & Tawn, 2005; Thibaud et al., 2016) have been advocated. The first is slow and the second, though more efficient, is prone to bias if the partition is incorrect (Wadsworth, 2015).

An attractive alternative to the use of block maxima is peaks-over-threshold analysis, which includes more information by focusing on single extreme events. In the multivariate case, specific definitions of exceedances have been used (e.g., Rootzén & Tajvidi, 2006; Ferreira & de Haan, 2014; Engelke et al., 2015), which can be unified within the framework of r-Pareto processes

(Dombry & Ribatet, 2015). For this approach, a full likelihood is often available in closed form, thus increasing the maximum number of variables that can be jointly modelled from a handful to a few dozen, but biased estimation may occur if nonextreme components are included. Censored likelihood, proposed in this context by Wadsworth & Tawn (2014), is more robust with regard to nonextreme observations, but it involves multivariate normal distribution functions, which can be computationally expensive. Nevertheless, inference is feasible for $D \approx 30$.

Nonparametric alternatives to full likelihood inference developed using the tail dependence coefficient (Davis & Mikosch, 2009; Davis et al., 2013) or the stable tail dependence function (Einmahl et al., 2016) rely on pairwise estimators and allow peaks-over-threshold inference for $D \approx 100$, but they are potentially inefficient and may be limited by combinatorial considerations.

Applications of max-stable processes (e.g., Asadi et al., 2015) or Pareto processes (Thibaud & Opitz, 2015) have focused on small regions and have used at most a few dozen locations with particular types of exceedance, but exploitation of much larger gridded datasets, along with more complex definitions of risk, is needed for a better understanding of extreme events and to reduce model uncertainties. The goals of this paper are to highlight the advantages of functional peaks-over-threshold modelling using r-Pareto processes, to show the feasibility of high-dimensional inference for the Brown–Resnick model with hundreds of locations, and to compare the robustness of different procedures with regard to finite thresholds. We develop an estimation method based on the gradient score (Hyvärinen, 2005) for a general notion of exceedances, for which the computation of multivariate normal probabilities is not needed and computational complexity is driven by matrix inversion, as with classical Gaussian likelihood inference. This method focuses on single extreme events and a general notion of exceedance, modelled by Pareto processes, instead of the max-stable approach.

2. Modelling exceedances over a high threshold

2·1. *Univariate model*

The statistical analysis of extremes was first developed for block maxima (Gumbel, 1958, § 5.1). This approach is widely used and can give good results, but the reduction of a complex dataset to maxima can lead to a significant loss of information (Madsen et al., 1997), so the modelling of exceedances over a threshold is often preferred (Davison & Smith, 1990). Let X be a random variable for which there exist sequences of constants $a_n > 0$ and b_n such that

$$n\,\text{pr}(X > b_n + a_n x) \to -\log G(x) \tag{1}$$

as $n \to \infty$, where G is a nondegenerate distribution function. Then X is said to belong to the max-domain of attraction of G and, for a large enough threshold $u < \inf\{x : F(x) = 1\}$, we can use the approximation

$$\text{pr}(X - u > x \mid X > u) \approx H_{(\xi,\sigma)}(x) = \begin{cases} (1 + \xi x/\sigma)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-x/\sigma), & \xi = 0, \end{cases} \tag{2}$$

where $\sigma = \sigma(u) > 0$ and $a_+ = \max(a, 0)$. If the shape parameter ξ is negative, then x must lie in the interval $[0, -\sigma/\xi]$, whereas x can take any positive value with positive or zero ξ. The implication is that the distribution over a high threshold u of any random variable X satisfying the conditions for (2) can be approximated by

$$G_{(\xi,\sigma,u)}(x) = 1 - \zeta_u H_{(\xi,\sigma)}(x - u), \quad x > u,$$

where ζ_u is the probability that X exceeds the threshold u. In its simplest form, this model for univariate exceedances applies to independent and identically distributed variables, but it has also been used for time series, nonstationary and spatial data.

The modelling of exceedances can be generalized to a multivariate setting (Rootzén & Tajvidi, 2006) and to continuous processes (Ferreira & de Haan, 2014; Dombry & Ribatet, 2015) within the functional regular variation framework.

2·2. *Functional regular variation*

Let S be a compact metric space, such as $[0, 1]^2$ for spatial applications. We write $\mathscr{F}_+ = C\{S, [0, \infty)\}$ for the closed subset of the Banach space of continuous functions $x : S \to \mathbb{R}$ endowed with the uniform norm $\|x\| = \sup_{s\in S} |x(s)|$, write $\mathscr{F}$ for $\mathscr{F}_+$ with the singleton $\{0\}$ excluded, and write $\mathscr{B}(\Xi)$ for the Borel σ-algebra associated with a metric space Ξ. Let $M_{\mathscr{F}}$ denote the class of Borel measures on $\mathscr{B}(\mathscr{F})$; we say that a set $A \in \mathscr{B}(\mathscr{F})$ is bounded away from $\{0\}$ if $d(A, \{0\}) = \inf_{x\in A} \|x\| > 0$. A sequence of measures $\{\Lambda_n\} \subset M_{\mathscr{F}}$ is said to converge to a limit $\Lambda \in M_{\mathscr{F}}$ (Hult & Lindskog, 2005) if $\lim_{n\to\infty} \Lambda_n(A) = \Lambda(A)$ for all $A \in \mathscr{B}(\mathscr{F})$ bounded away from $\{0\}$ with $\Lambda(\partial A) = 0$, where ∂A denotes the boundary of A. For equivalent definitions of this so-called $\hat{w}$-convergence, see Lindskog et al. (2014, Theorem 2.1).

Regular variation provides a flexible mathematical setting in which to characterize the tail behaviour of random processes in terms of $\hat{w}$-convergence of measures. A stochastic process X with sample paths in $\mathscr{F}$ is regularly varying (Hult & Lindskog, 2005) if there exists a sequence of positive real numbers $a_1, a_2, \ldots$ with $\lim_{n\to\infty} a_n = \infty$ and a measure $\Lambda \in M_{\mathscr{F}}$ such that

$$n\,\mathrm{pr}(a_n^{-1} X \in \cdot) \to \Lambda(\cdot) \tag{3}$$

as $n \to \infty$; then we write $X \in \mathrm{RV}(\mathscr{F}, a_n, \Lambda)$; note the link to (1). For a normalized process X^*, obtained for instance by standardizing the margins of X to unit Fréchet (Coles & Tawn, 1991, § 5) or unit Pareto (Klüppelberg & Resnick, 2008), regular variation is equivalent to the convergence of the renormalized pointwise maximum $n^{-1} \max_{i=1}^n X_i^*$ of independent replicates of X^* to a nondegenerate process Z^* with unit Fréchet margins and exponent measure Λ^* (de Haan & Lin, 2001). The process Z^* is called a simple max-stable process, and X^* is said to lie in the max-domain of attraction of Z^*.

Regular variation also affects the properties of exceedances over high thresholds. For any nonnegative measurable functional $r : \mathscr{F} \to [0, \infty)$ and stochastic process $\{X(s)\}_{s\in S}$, an r-exceedance is defined to be an event $\{r(X) > u_n\}$ where the threshold u_n is such that $\mathrm{pr}\{r(X) > u_n\} \to 0$ as $n \to \infty$. We further require that r be homogeneous, i.e., there exists $\alpha > 0$ such that $r(ax) = a^\alpha r(x)$ for $a > 0$ and $x \in \mathscr{F}$. As $r(\cdot)$ could be replaced by $r(\cdot)^{1/\alpha}$ without loss of generality, in what follows we assume that $\alpha = 1$. Dombry & Ribatet (2015) called r a cost functional, and in his 2013 Université de Montpellier II PhD thesis Thomas Opitz called it a radial aggregation function; but we prefer the term risk functional because r determines the type of extreme event whose risk is to be studied.

A natural formulation of subsequent results on r-exceedances uses a pseudo-polar decomposition. For a norm $\|\cdot\|_{\mathrm{ang}}$ on $\mathscr{F}$, called the angular norm, and a risk functional r, a pseudo-polar transformation T_r is a map such that

$$T_r : \mathscr{F} \to [0, \infty) \times \mathscr{S}_{\mathrm{ang}}, \quad T_r(x) = \left\{ v = r(x),\ w = \frac{x}{\|x\|_{\mathrm{ang}}} \right\},$$

where $\mathscr{S}_{\mathrm{ang}}$ is the unit sphere $\{x \in \mathscr{F} : \|x\|_{\mathrm{ang}} = 1\}$. If r is continuous and T is restricted to $\{x \in \mathscr{F} : r(x) > 0\}$, then T is a homeomorphism with inverse $T_r^{-1}(v, w) = v \times w/r(w)$.

Theorem 2.1 in Lindskog et al. (2014) provides an equivalent pseudo-polar formulation of (3). For any $X \in \textsc{rv}(\mathscr{F}, a_n, \Lambda)$ and any uniformly continuous risk functional r not vanishing Λ-almost everywhere, there exist $\beta > 0$ and a measure σ_r on $\mathscr{B}(\mathscr{S}_{\text{ang}})$ such that

$$n\,\mathrm{pr}\{T_r^{-1}(a_n^{-1}v, w) \in \cdot\} \to \Lambda \circ T_r^{-1}(\cdot) = \Lambda^{\beta} \times \sigma_r(\cdot) \tag{4}$$

as $n \to \infty$, where $\Lambda^{\beta}[v,\infty) = v^{-\beta}$ and $\sigma_r(\cdot) = \Lambda\{x \in \mathscr{F} : r(x) \geqslant 1,\ x/\|x\|_{\text{ang}} \in (\cdot)\}$ is called the angular measure. The converse holds if $\{x \in \mathscr{F} : r(x) = 0\} = \emptyset$ (Lindskog et al., 2014, Corollary 4.4).

The functional $r(x) = \sup_{s\in S} x(s)$, used by Rootzén & Tajvidi (2006) in a multivariate setting and by Ferreira & de Haan (2014) for continuous processes, implies that realizations of $X(s)$ exceeding the threshold at any location $s \in S$ will be labelled extreme, but this functional can only be used in applications where $X(s)$ is observed throughout S. Therefore it may be preferable to use functions such as $\max_{s\in S'} X(s)$ or $\max_{s\in S'} X(s)/u(s)$, where $S' \subset S$ is a finite set of gauged sites. Other risk functionals include $\int_S X(s)\,\mathrm{d}s$ for the study of areal rainfall (Coles & Tawn, 1996), $\min_{s\in S'} X(s)/u(s)$, and $X(s_0)$ for risks impacting a specific location s_0. Although the choice of risk functional allows one to focus on particular types of extreme event, the choice of the angular norm $\|\cdot\|_{\text{ang}}$ has no effect and is usually made for convenience.

Finally, for a common angular norm $\|\cdot\|_{\text{ang}}$, the angular measures of two risk functionals r_1 and r_2 that are strictly positive Λ-almost everywhere are linked by the expression

$$\sigma_{r_1}(\mathrm{d}w) = \left\{\frac{r_1(w)}{r_2(w)}\right\}^{\beta} \sigma_{r_2}(\mathrm{d}w), \quad \mathrm{d}w \in \mathscr{B}(\mathscr{S}_{\text{ang}}). \tag{5}$$

Equation (5) is useful for simulation and when we are interested in r_2-exceedances but inference has been performed based on r_1. All the previous definitions and results hold for finite dimensions, i.e., for D-dimensional random vectors, upon replacing $\hat{w}$-convergence by vague convergence (Resnick, 2007, § 3.3.5) on $M_{\mathbb{R}_+^D\setminus\{0\}}$, the class of Borel measures on $\mathscr{B}(\mathbb{R}_+^D \setminus \{0\})$ endowed with the $\|\cdot\|$ norm; see the PhD thesis of Thomas Opitz mentioned above.

2·3. *r-Pareto processes*

In this section, r denotes a functional that is nonnegative and homogeneous with $\alpha = 1$. The r-Pareto processes (Dombry & Ribatet, 2015) are important for modelling exceedances and may be constructed as

$$P = U\frac{Q}{r(Q)},$$

where U is a univariate Pareto random variable with $\mathrm{pr}(U > v) = 1/v^{\beta}$ $(v \geqslant 1)$ and Q is a random process with sample paths in $\mathscr{S}_{\text{ang}}^{r} = \{x \in \mathscr{F} : r(x) \geqslant 1, \|x\|_{\text{ang}} = 1\}$ and probability measure σ_{ang}. The process P is then called an r-Pareto process with tail index $\beta > 0$ and angular measure σ_{ang}; to distinguish different Pareto processes, below we use the notation P^{r}_{β,σ_r} for P.

An important property of r-Pareto processes is threshold invariance: for all $A \in \mathscr{B}(\{x \in \mathscr{F} : r(x) \geqslant 1\})$ and all $u \geqslant 1$ such that $\mathrm{pr}\{r(P) \geqslant u\} > 0$,

$$\mathrm{pr}\{u^{-1}P \in A \mid r(P) \geqslant u\} = \mathrm{pr}(P \in A).$$

Furthermore, for $X \in \textsc{rv}(\mathscr{F}, a_n, \Lambda)$ with index $\beta > 0$ and for a risk functional r that is continuous at the origin and does not vanish Λ-almost everywhere, the distribution of the r-exceedances converges weakly to that of an r-Pareto process, i.e.,

$$\mathrm{pr}\big\{u^{-1}X \in \cdot \mid r(X) \geqslant u\big\} \to \mathrm{pr}\big(P^{r}_{\beta,\sigma_r} \in \cdot\big)$$

as $n \to \infty$, with tail index β and probability measure σ_r as defined in (4) (Dombry & Ribatet, 2015, Theorem 2). When working with a normalized process X^*, the exponent measure Λ^* of the limiting max-stable process Z^* and the measure $\Lambda^1 \times \sigma_r$ of the Pareto process are the same up to a coordinate transform, as suggested by (4).

For two different risk functionals r_1 and r_2 and angular measures σ_{r_1} and σ_{r_2} for which there exists $\Lambda \in M_{\mathscr{F}}$ such that

$$\Lambda \circ T_{r_1}^{-1}(\cdot) = \Lambda \circ T_{r_2}^{-1}(\cdot),$$

the associated Pareto processes $P^{r_1}_{\beta,\sigma_{r_1}}$ and $P^{r_2}_{\beta,\sigma_{r_2}}$ are defined on different subsets of $\mathscr{F}$, but, as suggested by (5), if there exists a threshold $u_{\min}$ such that

$$\{x \in \mathscr{F} : r_1(x) \geqslant u_{\min}\} \subset \{x \in \mathscr{F} : r_2(x) \geqslant 1\},$$

then

$$\mathrm{pr}\left\{ \frac{P^{r_1}_{\beta,\sigma_{r_1}}}{u} \in \cdot \;\middle|\; r_2\big(P^{r_1}_{\beta,\sigma_{r_1}}\big) \geqslant u \right\} = \mathrm{pr}\big(P^{r_2}_{\beta,\sigma_{r_2}} \in \cdot\big), \quad u \geqslant u_{\min}. \tag{6}$$

Simulation of r-Pareto processes is feasible only for a few risk functionals, such as $r_1(x) = \|x\|_1$, but (6) can be used to obtain samples of one process from those of another: for independent replicates $x^1, \ldots, x^N$ from $P^{r_1}_{\beta,\sigma_{r_1}}$, $\{y^n = x^n/u_{\min} : r_2(y^n) \geqslant 1\}$ is a sample from $P^{r_2}_{\beta,\sigma_{r_2}}$.

Finally, let σ_r be a probability measure on $\mathscr{S}^{r}_{\mathrm{ang}}$, and define the process

$$M(s) = \max_{n\in\mathbb{N}} U^n \frac{Q^n(s)}{r(Q^n)}, \quad s \in S, \tag{7}$$

where $\{U^n : n \in \mathbb{N}\}$ is a Poisson process on $(0,\infty)$ with intensity $u^{-2}\,\mathrm{d}u$ and $Q^1, Q^2, \ldots$ are replicates of a process Q with probability measure σ_r. Then M is a max-stable process with exponent measure $\Lambda_\theta\{A_{\max}(x)\} = \Lambda^1 \times \sigma_r\{A_{\max}(x)\}$, where $A_{\max}(x) = \{y \in \mathscr{F} : \sup_{s\in S} y(s)/x(s) \geqslant 1\}$. Thus equation (7) connects an r-Pareto process and its max-stable counterpart.

2·4. *Extreme value processes associated with log-Gaussian random functions*

We focus on a class of r-Pareto processes based on log-Gaussian stochastic processes, whose max-stable counterparts are Brown–Resnick processes. This class is particularly useful, not only for its flexibility but also because it is based on Gaussian models widely used in applications. Chiles & Delfiner (1999, pp. 84–108) review these classical models.

Let Z be a zero-mean Gaussian process with stationary increments, i.e., the semivariogram $\gamma(s,s') = E[\{Z(s) - Z(s')\}^2]/2$ $(s,s' \in S)$ depends only on the difference $s - s'$ (Chiles & Delfiner, 1999, p. 30), and let $\sigma^2(s) = \mathrm{var}\{Z(s)\}$. If $Z^1, Z^2, \ldots$ are independent replicates of Z and $\{U^n : n \in \mathbb{N}\}$ is a Poisson process on $(0,\infty)$ with intensity $u^{-2}\,\mathrm{d}u$, independent of the Z^n, then

$$M(s) = \max_{n\in\mathbb{N}} U^n \exp\{Z^n(s) - \sigma^2(s)/2\}, \quad s \in S,$$

is a stationary Brown–Resnick process with standard Fréchet margins, whose distribution depends only on γ (Kabluchko et al., 2009); such processes are max-stable. Let γ_θ denote a parameterized semivariogram whose parameter θ lies in a compact set Θ, and let σ_θ^2 denote the corresponding variance function.

Let $s_1, \ldots, s_D$ be locations of interest in S. In the rest of the paper, x will denote an element of $\mathbb{R}_+^D$ with components $x_d \equiv x(s_d)$ $(d = 1, \ldots, D)$. The finite-dimensional exponent measure of a simple Brown–Resnick process with D variables is

$$\Lambda_\theta\{A_{\max}(x)\} = E\left[\max_{d=1,\ldots,D}\left\{\frac{Z(s_d) - \sigma_\theta^2(s_d)/2}{x_d}\right\}\right], \tag{8}$$

where $\Lambda_\theta(\cdot)$ is the finite-dimensional projection of the measure defined in (3). Then we can write (Huser & Davison, 2013)

$$\Lambda_\theta\{A_{\max}(x)\} = \sum_{d=1}^{D} \frac{1}{x_d}\,\Phi\{\eta_d(x), R_d\}, \tag{9}$$

where η_d is the $(D-1)$-dimensional vector with ith component $\eta_{d,i} = (\gamma_{d,i}/2)^{1/2} + \log(x_i/x_d)/(2\gamma_{d,i})^{1/2}$, $\gamma_{d,i}$ denotes $\gamma(s_d, s_i)$ $(s_d, s_i \in S)$, and $\Phi(\cdot, R_d)$ is the multivariate normal distribution function with mean zero and covariance matrix R_d having (i,j) entry $(\gamma_{d,i} + \gamma_{d,j} - \gamma_{i,j})/\{2(\gamma_{d,i}\gamma_{d,j})^{1/2}\}$, where $i,j \in \{1, \ldots, d-1, d+1, \ldots, D\}$.

The r-Pareto processes associated with log-Gaussian random functions are closely related to the intensity function λ_θ corresponding to the measure Λ_θ, which can be found by taking partial derivatives of $\Lambda_\theta(x)$ with respect to $x_1, \ldots, x_D$, yielding (Engelke et al., 2015)

$$\lambda_\theta(x) = \frac{|\Sigma_\theta|^{-1/2}}{x_1^2 x_2 \cdots x_D (2\pi)^{(D-1)/2}} \exp\left(-\frac{1}{2}\tilde{x}^{\mathrm{T}}\Sigma_\theta^{-1}\tilde{x}\right), \quad x \in \mathbb{R}_+^D, \tag{10}$$

where $\tilde{x}$ is the $(D-1)$-dimensional vector with components $\{\log(x_i/x_1) + \gamma_{i,1} : i = 2, \ldots, D\}$ and Σ_θ is the $(D-1) \times (D-1)$ matrix with elements $\{\gamma_{i,1} + \gamma_{j,1} - \gamma_{i,j}\}_{i,j\in\{2,\ldots,D\}}$. Wadsworth & Tawn (2014) derived an alternative symmetric expression for (10) that will be useful in § 3·3, but (10) is more readily interpreted. Similar expressions exist for extremal-t processes (Thibaud & Opitz, 2015).

3. Inference for r-Pareto processes

3·1. *Generalities*

In this section, $x^1, \ldots, x^N$ are independent replicates of a D-dimensional r-Pareto random vector P with tail index $\beta = 1$, and $y^1, \ldots, y^N$ are independent replicates of a regularly varying D-dimensional random vector Y^* with normalized margins.

As in the univariate setting, statistical inference based on block maxima and the max-stable framework discards information by focusing for maxima instead of single events. Models for maxima are difficult to fit not only due to the small number of replicates, but also because the likelihood is usually too complex to compute in high dimensions (Castruccio et al., 2016). For the Brown–Resnick process, the full likelihood cannot be computed for D greater than around 10 (Huser & Davison, 2013), except in special cases. When the occurrence times of maxima are available, inference is usually possible up to $D \approx 30$ (Stephenson & Tawn, 2005; Thibaud et al., 2016).

A useful alternative is composite likelihood inference (Padoan et al., 2010; Varin et al., 2011) based on subsets of observations of sizes smaller than D, which trades off a gain in computational efficiency against a loss of statistical efficiency. The number of possible subsets increases very rapidly with D, and their selection can be troublesome, though some statistical efficiency can be retrieved by taking higher-dimensional subsets. Castruccio et al. (2016) found higher-order composite likelihoods to be more robust than the spectral estimator, but in realistic cases these methods are limited to fairly small dimensions.

Estimation based on threshold exceedances and the Pareto process has the advantages that individual events are used, the likelihood function is usually simpler, and the choice of risk functional can be a means of tailoring the definition of an exceedance to the application. Equation (4) suggests that the choice of risk functional should not affect the estimates, but this is not entirely true, because the threshold cannot be taken arbitrarily high and the events selected depend on the risk functional, the choice of which enables the detection of mixtures in the extremes and can improve subasymptotic behaviour by fitting the model using only those observations closest to the chosen type of extreme event. For example, we might expect the extremal dependence of intense local rainfall events to differ from that of heavy large-scale precipitation, even in the same geographical region.

The probability density function of a Pareto process for r-exceedances over the threshold vector $u \in \mathbb{R}_+^D$ can be found by rescaling the intensity function λ_θ by $\Lambda_\theta\{A_r(u)\}$, yielding

$$\lambda^r_{\theta,u}(x) = \frac{\lambda_\theta(x)}{\Lambda_\theta\{A_r(u)\}}, \quad x \in A_r(u), \tag{11}$$

where

$$\Lambda_\theta\{A_r(u)\} = \int_{A_r(u)} \lambda_\theta(x)\,\mathrm{d}x \tag{12}$$

and $A_r(u)$ is the exceedance region $\{x \in \mathbb{R}_+^D : r(x/u) \geqslant 1\}$. Equation (11) yields the loglikelihood

$$\ell(\theta; x^1, \ldots, x^N) = \sum_{n=1}^{N} \mathbb{1}\left\{r\left(\frac{x^n}{u}\right) \geqslant 1\right\} \log\left[\frac{\lambda_\theta(x^n)}{\Lambda_\theta\{A_r(u)\}}\right], \tag{13}$$

where division of vectors is componentwise and $\mathbb{1}$ denotes the indicator function. Maximization of ℓ gives an estimator $\hat{\theta}_r(x^1, \ldots, x^N)$ that is consistent, asymptotically normal and efficient under mild conditions.

Numerical evaluation of the D-dimensional integral $\Lambda_\theta\{A_r(u)\}$ is generally intractable for large D, though it simplifies for certain risk functionals; an example is $r(x) = \max_d x_d$, for which the integral is a sum of multivariate probability functions; see (9). Similarly, $\Lambda_\theta\{A_r(u)\}$ does not depend upon θ when $r(x) = D^{-1}\sum_d x_d$ (Coles & Tawn, 1991); we call the corresponding version of (13) the spectral loglikelihood and its maximizer the spectral estimator.

In practice observations cannot be assumed to be exactly Pareto distributed; it is usually more plausible that they lie in the domain of attraction of some extremal process. As a consequence of Theorem 3.1 of de Haan & Resnick (1993), asymptotic properties of $\hat{\theta}_r(x^1, \ldots, x^N)$ hold for $\hat{\theta}_r(y^1, \ldots, y^N)$ as $N \to \infty$ and $u \to \infty$ with the number of exceedances $N_u = o(N) \to \infty$; see § 3·3. However, the threshold u is finite and so low components of $y^n \in A_r(u)$ may lead to biased estimation. As it is due to model misspecification, this bias is unavoidable; moreover, it grows with D, so these methods can perform poorly, especially if the extremal dependence is weak, because it is then more likely that at least one component of y^n will be small (Engelke et al., 2015; Thibaud & Opitz, 2015; Huser et al., 2016). The bias can be reduced by a censored likelihood proposed in the multivariate setting by Joe et al. (1992) and used for the Brown–Resnick model by Wadsworth & Tawn (2014) and for the extremal-t process by Thibaud & Opitz (2015). This method works well in practice but typically requires the computation of multivariate normal and t probabilities, which can be challenging in realistic cases if standard code is used. Some modest changes to the code to perform quasi-Monte Carlo maximum likelihood estimation with hundreds of locations are described in § 3·2.

For spatiotemporal applications, inference for r-Pareto processes must be performed using data from thousands of locations, and in § 3·3 we discuss an approach that applies to a wide range of risk functionals and is computationally fast, statistically efficient and robust with regard to finite thresholds.

3·2. *Efficient censored likelihood inference*

Censored likelihood estimation for extreme value processes associated with log-Gaussian random functions was developed by Wadsworth & Tawn (2014) and is based on (13) with $\max_d x_d/u_d$ as the risk functional and where any component lying below the threshold vector $(u_1, \ldots, u_D) > 0$ is treated as censored. The corresponding estimator has a higher variance but a lower bias than the spectral estimator. The censored likelihood density function for the Brown–Resnick process is (Asadi et al., 2015)

$$\lambda^{\text{cens}}_{\theta,u}(x) = \frac{1}{\Lambda_\theta\{A_{\max}(u)\}} \frac{1}{x_1^2 x_2 \cdots x_k} \phi_{k-1}(\tilde{x}_{2:k}; \Sigma_{2:k}) \Phi_{D-k}\{\mu_{\text{cens}}(x_{1:k}), \Sigma_{\text{cens}}(x_{1:k})\}, \quad x \in A_{\max}(u),$$

where k components exceed their thresholds, $\tilde{x}_{2:k}$ and $\Sigma_{2:k}$ are subsets of the variables $\tilde{x}$ and Σ_θ in equation (10), and ϕ_{k-1} and Φ_{D-k} are the multivariate Gaussian density and distribution functions with mean zero. The argument and covariance matrix for Φ_{D-k} are

$$\mu_{\text{cens}}(x_{1:k}) = \{\log(u_j/x_1) + \gamma_{j,1}\}_{j=k+1,\ldots,D} - \Sigma_{(k+1):D,2:k} \Sigma^{-1}_{2:k,2:k} \tilde{x}_{2:k},$$
$$\Sigma_{\text{cens}}(x_{1:k}) = \Sigma_{(k+1):D,(k+1):D} - \Sigma_{(k+1):D,2:k} \Sigma^{-1}_{2:k,2:k} \Sigma_{2:k,(k+1):D}.$$

Wadsworth & Tawn (2014) derived similar expressions. The estimator

$$\hat{\theta}_{\text{cens}}(y^1, \ldots, y^N) = \arg\max_{\theta \in \Theta} \sum_{n=1,\ldots,N} \mathbb{1}\left\{\max_d \left(\frac{y^n_d}{u_d}\right) \geqslant 1\right\} \log \lambda^{\text{cens}}_{\theta,u}(\mathrm{y}^n) \tag{14}$$

is also consistent and asymptotically normal as $u \to \infty$, $N \to \infty$ and $N_u \to \infty$ with $N_u = o(N)$. For finite thresholds, $\hat{\theta}_{\text{cens}}$ has been found to be more robust with respect to the presence of low components than the spectral estimator (Engelke et al., 2015; Huser et al., 2016), but it is awkward because of the potentially large number of multivariate normal integrals involved, thus far limiting its application to $D \lesssim 30$ (Wadsworth & Tawn, 2014; Thibaud et al., 2016).

When maximizing the right-hand side of (14), the normalizing constant $\Lambda_\theta\{A_{\max}(u)\}$ described in (8) and the multivariate normal distribution functions require the computation of multidimensional integrals. Theorem 7 of Geyer (1994) suggests that we approximate $\hat{\theta}_{\text{cens}}$ by maximizing

$$\ell^{p}_{\text{cens}}(\theta) \\ = \sum_{n=1}^{N} \mathbb{1}\left\{\max\left(\frac{x^n}{u}\right) \geqslant 1\right\}\left[\log\left\{\frac{\phi_{k-1}(\tilde{x}_{2:k};\, \Sigma_{2:k})}{(x_1^n)^2 x_2^n \cdots x_k^n}\right\} + \log\frac{\Phi^{p}_{D-k}\{\mu_{\text{cens}}(x^n_{1:k}),\, \Sigma_{\text{cens}}(x^n_{1:k})\}}{\Lambda^{p}_{\theta}\,\{A_{\max}(u)\}}\right],$$

where Φ^{p}_{D-k} and Λ^{p}_{θ} are Monte Carlo estimates of the corresponding integrals based on p simulated samples, yielding a maximizer $\hat{\theta}^{p}_{\text{cens}}$ that converges almost surely to $\hat{\theta}_{\text{cens}}$ as $p \to \infty$.

Classical Monte Carlo estimation for multivariate integrals yields a probabilistic error bound that is $O(\omega p^{-1/2})$, where $\omega = \omega(\phi)$ is the square root of the variance of the integrand ϕ. Quasi-Monte Carlo methods can achieve higher rates of convergence and thus improve computational efficiency while preserving the consistency of $\hat{\theta}^{p}_{\text{cens}}$. For estimation of multivariate normal distribution functions, Genz & Bretz (2009, § 4.2.2) advocate the use of randomly shifted deterministic lattice rules, which can achieve a convergence rate of order $O(p^{-2+\epsilon})$ for some $\epsilon > 0$. Lattice rules rely on regular sampling of the hypercube $[0,1]^D$, taking

$$\mathrm{z}_q = \left|2 \times \mathrm{frac}(qv' + \Delta) - 1\right| \quad (q = 1, \ldots, p),$$

where frac(z) denotes the componentwise fractional part of $\mathrm{z} \in \mathbb{R}^D$, p is a prime number of samples in the hypercube $[0,1]^D$, $v' \in \{1, \ldots, p\}^D$ is a carefully chosen generating vector, and $\Delta \in [0,1]^D$ is a uniform random shift. Fast construction rules have been developed to find an optimal v' for given numbers of dimensions D and samples p (Nuyens & Cools, 2004). The existence of generating vectors achieving a nearly optimal convergence rate, with integration error independent of the dimension, has been proved, and methods for their construction are available (Dick & Pillichshammer, 2010).

Our implementation of this approach applied to (14) and coupled with parallel computing is tractable for D of the order of a few hundred; see the Supplementary Material for details. Our algorithm can be extended to the extremal-t model by writing multivariate t probabilities in terms of the multivariate normal distribution function; see Genz & Bretz (2009) for more details.

3·3. *Score matching*

Classical likelihood inference requires either evaluation or simplification of the scaling constant $\Lambda_\theta\{A_r(u)\}$, whose complexity increases with the number of dimensions. Hence we seek alternatives that do not require its computation.

Let $\mathscr{A}$ be a sample space such as $\mathbb{R}^D_+$, $\mathscr{P}$ a convex class of probability measures on $\mathscr{A}$, and X a random variable with distribution $F \in \mathscr{A}$. A proper scoring rule (Gneiting & Raftery, 2007) is a functional $\delta : \mathscr{P} \times \mathscr{A} \to \mathbb{R}$ such that

$$\Delta_\delta(G, F) = E_X\{\delta(G, X)\} - E_X\{\delta(F, X)\} \geqslant 0, \quad G \in \mathscr{P}.$$

The scoring rule is said to be strictly proper if $\Delta_\delta(G, F) = 0$ if and only if $G = F$, and under this hypothesis Δ_δ defines a divergence measure on $\mathscr{P}$ (Thorarinsdottir et al., 2013).

Let δ denote a strictly proper scoring rule, let $\{F_\theta : \theta \in \Theta\} \subset \mathscr{A}$ be a parametric family of distributions, and let $X^1, \ldots, X^N$ be independent observations from F_{θ_0}. The first term of the divergence $\Delta_\delta(F_\theta, F_{\theta_0})$ can be estimated by

$$\delta(\theta) = \frac{1}{N}\sum_{i=1}^{N} \delta(F_\theta, X^i),$$

minimization of which defines an unbiased and asymptotically normal estimator of θ_0 under suitable regularity conditions (Dawid et al., 2016, Theorem 4.1). Consequently, for a risk functional r, the estimator

$$\hat{\theta}^r_{\delta,u}(X^1,\ldots,X^N) = \arg\min_{\theta\in\Theta} \sum_{n=1}^{N} \mathbb{1}\left\{r\left(\frac{X^n}{u}\right) > 1\right\} \delta\left(\lambda^r_{\theta,u}, \frac{X^n}{u}\right)$$

is also consistent and asymptotically normal. Owing to de Haan & Resnick (1993, Propositions 3.1 and 3.2), these asymptotic properties can be generalized to samples from a regularly varying random vector with normalized marginals; see the Supplementary Material for the proof.

PROPOSITION 1. *Let $Y^1,\ldots,Y^N$ be independent replicates of a regularly varying random vector Y^* with normalized marginals and limiting measure Λ_{θ_0}, and let δ be a strictly proper scoring rule satisfying the conditions of Theorem 4.1 of Dawid et al. (2016). If $N \to \infty$ and $N_u \to \infty$ in such a way that $N_u/N \to 0$ as $N \to \infty$, then*

$$N_u{}^{1/2}\{\hat{\theta}^r_{\delta,N/N_u}(Y^1,\ldots,Y^N) - \theta_0\} \to \mathcal{N}\{0, K^{-1}J(K^{-1})^{\mathrm{T}}\}$$

in distribution, where

$$J = E\left\{\frac{\partial\delta(\theta)}{\partial\theta}\frac{\partial\delta(\theta)}{\partial\theta^{\mathrm{T}}}\right\}\bigg|_{\theta=\theta_0}, \quad K = E\left\{\frac{\partial^2\delta(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\right\}\bigg|_{\theta=\theta_0}.$$

Estimates of the Godambe information matrix $KJ^{-1}K$ can be used for inference, and the scoring rule ratio statistic

$$W^\delta = 2\left\{\frac{\partial\delta(\theta_0)}{\partial\theta} - \frac{\partial\delta(\hat{\theta}^r_{\delta,N/N_u})}{\partial\theta}\right\},$$

properly calibrated, can be used to compare nested models (Dawid et al., 2016, § 4.1).

The loglikelihood is the proper scoring rule associated with the Kullback–Leibler divergence. Although efficient, it is not robust, which is problematic for fitting asymptotic models such as Pareto processes, and a closed form for the normalizing coefficient $\Lambda_\theta\{A_r(u)\}$ defined in (12) is available only in special cases. The gradient scoring rule (Hyvärinen, 2005) uses only the derivative $\nabla_x \log\lambda^r_{\theta,u}$ and thus does not require computation of $\Lambda_\theta\{A_r(u)\}$. Hyvärinen (2007) adapted this scoring rule for strictly positive variables, and we propose to extend it to any domain of the form $A_r(u) = \{x \in \mathbb{R}^D_+ : r(x/u) \geqslant 1\}$, using the divergence measure

$$\Delta_{\mathrm{grad}}(\theta,\theta_0) = \int_{A_r(u)} \left\|\nabla_x \log\lambda_\theta(x) \otimes w(x) - \nabla_x \log\lambda_{\theta_0}(x) \otimes w(x)\right\|_2^2 \lambda_{\theta_0}(x)\,\mathrm{d}x, \qquad (15)$$

where λ_θ is differentiable for all $\theta \in \Theta$ on $A_r(u) \setminus \partial A_r(u)$, with ∂A denoting the boundary of A, ∇_x is the gradient operator, $w : A_r(u) \to \mathbb{R}^D_+$ is a positive weight function, and $\otimes$ denotes the Hadamard product. If w is differentiable on $A_r(u)$ and satisfies certain boundary conditions discussed in the Supplementary Material, then the scoring rule

$$\delta_w(\lambda_\theta,x) = \sum_{d=1}^{D}\left(2w_d(x)\frac{\partial w_d(x)}{\partial x_d}\frac{\partial\log\lambda_\theta(x)}{\partial x_d} + w_d(x)^2\left[\frac{\partial^2\log\lambda_\theta(x)}{\partial x_d^2} + \frac{1}{2}\left\{\frac{\partial\log\lambda_\theta(x)}{\partial x_d}\right\}^2\right]\right)$$

for $x \in A_r(u)$ is strictly proper. The gradient score for a log-Gaussian Pareto process satisfies the regularity conditions of Theorem 4.1 in Dawid et al. (2016), so the resulting estimator $\hat{\theta}_w$ is asymptotically normal.

For the Brown–Resnick model, two possible weight functions are

$$\begin{aligned} w_d^1(x) &= x_d\big[1 - \exp\{1 - r(x/u)\}\big] \quad (d = 1, \ldots, D),\\ w_d^2(x) &= \big[1 - \exp\{-3(x_d - u_d)/u_d\}\big]\big[1 - \exp\{1 - r(x/u)\}\big] \quad (d = 1, \ldots, D), \end{aligned} \tag{16}$$

where r is a risk functional differentiable on $\mathbb{R}_+^D$ and the threshold vector u lies in $\mathbb{R}_+^D$. The weights w^1 are derived from Hyvärinen (2007), whereas w^2 is designed to approximate the effect of censoring by downweighting components of x^n near u. These weighting functions are well suited to extremes: a vector $x \in A_r(u)$ is penalized if $r(x/u)$ is close to 1, and low components of x induce low weights for the associated partial derivatives. For these reasons, inference using δ_w with the weighting functions in (16) should be more robust with respect to low components than is the spectral estimator. The estimator $\hat{\theta}_w$ is much cheaper to compute than $\hat{\theta}_{\text{cens}}$ and can be obtained for any risk functional differentiable on $\mathbb{R}_+^D$. Expressions for the gradient score for the Brown–Resnick model can be found in the Supplementary Material, and the performances of these inference procedures are compared in § 4.

The gradient score can be applied to any extremal model with a multivariate density function whose logarithm is twice differentiable away from the boundaries of its support, and if discontinuities are present on this support, then a carefully chosen weighting function w ensures the existence and the consistency of the score. Indeed, similar expressions can be derived for the extremal-t model, though choices for the weight functions are more restricted: w^2 satisfies the boundary conditions, but w^1 does not ensure that the score is proper.

4. Simulation study

4·1. *Exact simulation*

The inference procedures and simulation algorithms described herein are contained in an R package, mvPot (de Fondeville, 2017; R Development Core Team, 2018).

We first illustrate the feasibility of high-dimensional inference by simulating r-Pareto processes associated with log-normal random functions at D locations in $S = [0, 100]^2$. Details of the algorithm can be found in the Supplementary Material.

We use an isotropic power semivariogram, $\gamma(s, s') = (\|s - s'\|/\tau)^\kappa/2$, shape parameters $\kappa = 0{\cdot}5, 1, 1{\cdot}3, 1{\cdot}8$, and scale parameter $\tau = 2{\cdot}5$, chosen such that the dependence models defined on S cover strong to weak extremal dependence. For this simulation, the dependence model with $\kappa = 1{\cdot}8$ requires us to work on the log scale to avoid rounding errors. For each simulation, $N = 10\,000$ r-Pareto processes, with $r(x) = D^{-1}\sum_d x_d$, were simulated on regular 10 × 10, 20 × 10 and 20 × 15 grids. The grid size was restricted to at most 300 locations for ease of comparison with the second simulation study. For the gradient score, we use $r(x) = D^{-1}\sum_d x(s_d)$. The components of the threshold vector u are set equal to the empirical 0·99 quantile of $r(x^1), \ldots, r(x^N)$, giving $N_u = 100$. For censored likelihood inference, we use the approach described in the Supplementary Material with $\bar{p} = 10$ and threshold u equal to the empirical 0·99 quantile of $\max_d x_d^1, \ldots, \max_d x_d^N$, so that the conditions for (6) are satisfied. One hundred replicates are used in each case.

Table 1 displays the relative root mean squared error for estimation based on the censored loglikelihood and the gradient score with weights w^1 and w^2, compared to that based on the

Table 1. *Relative root mean squared error* (%) *for comparison of estimates based on censored loglikelihood, left, and the gradient score with weights* w^1, *middle, and* w^2, *right, relative to spectral estimates, for the parameters* κ *and* $\tau = 2{\cdot}5$. *Efficiency of* 100% *would correspond to the spectral estimator, and smaller values to less efficient estimators. Inference is performed using the top* 1% *of* 10 000 *Pareto processes with semivariogram* $\gamma(s, s') = (\|s - s'\|/\tau)^\kappa/2$ *simulated on regular* 10 × 10, 20 × 10 *and* 20 × 15 *grids*

Shape κ

Grid size	$\kappa = 0{\cdot}5$	$\kappa = 1$	$\kappa = 1{\cdot}3$	$\kappa = 1{\cdot}8$
10 × 10	53/46/44	10/32/33	4·7/39/39	1·0/51/52
20 × 10	67/51/52	10/25/24	5·4/34/35	1·0/54/55
20 × 15	67/47/47	11/30/31	4·1/25/25	1·4/49/49

Scale τ

Grid size	$\kappa = 0{\cdot}5$	$\kappa = 1$	$\kappa = 1{\cdot}3$	$\kappa = 1{\cdot}8$
10 × 10	52/58/57	19/60/59	10/63/66	1·7/53/53
20 × 10	41/80/79	17/70/70	9·2/71/70	3·3/52/51
20 × 15	38/68/69	17/82/81	7·1/62/61	3·9/51/52

spectral estimator. For all the methods and parameter combinations, bias is negligible and performance is driven mainly by the variance. As expected, efficiency is lower than 100% because when simulating and fitting from the true model, the spectral estimator performs best. The gradient score and censored likelihood estimators deteriorate as the extremal dependence weakens and the number of low components in the simulated vectors increases. The gradient score outperforms the censored likelihood except when censoring is low, i.e., when $\kappa = 0{\cdot}5$. The performance of censored likelihood estimation deteriorates as D increases, suggesting that the gradient score will be preferable in high dimensions. These results are not realistic, however, since the data are simulated from the model fitted, whereas in practice the model is used as a high-threshold approximation to the data distribution.

The optimization of the likelihood based on the spectral density and gradient score functions takes only a dozen seconds even for the finest grid. The same random starting point is used for each optimization to ensure fair comparison. Estimation using the censored approach takes several minutes and slows greatly as the dimension increases; see the Supplementary Material.

4·2. *Domain of attraction*

As the asymptotic regime is never reached in practice, we now compare the robustness of the different inference procedures for finite thresholds. The Brown–Resnick process belongs to its own max-domain of attraction, so its peaks-over-threshold distribution converges to a generalized Pareto process with log-Gaussian random function. We repeat the simulation study of § 4·1 with 10 000 Brown–Resnick processes and the same parameter values. This simulation uses the algorithm of Dombry et al. (2016) and is computationally expensive, so we used only 300 variables. It took around three hours with 16 cores to generate $N = 10\,000$ samples on the finest grid.

Table 2 shows the results. As expected when the model is misspecified, the root relative mean squared error is mainly driven by bias, which increases with the shape κ and the dimension D. Spectral estimation is on the whole outperformed by both of the other methods. For $\kappa = 0{\cdot}5$, the three methods show fairly similar overall performance, with the censored likelihood better

Table 2. *As Table 1 but with inference based on the top* 1% *of* 10 000 *simulated Brown–Resnick processes*

Shape κ

Grid size	$\kappa = 0{\cdot}5$	$\kappa = 1$	$\kappa = 1{\cdot}3$	$\kappa = 1{\cdot}8$
10×10	154/111/81	473/183/108	196/170/105	NC
20×10	172/122/95	413/150/114	309/181/137	144/168/122
20×15	142/119/99	369/133/110	314/170/140	163/173/137

Scale τ

Grid size	$\kappa = 0{\cdot}5$	$\kappa = 1$	$\kappa = 1{\cdot}3$	$\kappa = 1{\cdot}8$
10×10	107/127/116	263/38/35	109/231/452	NC
20×10	105/133/119	206/94/80	315/66/53	105/336/261
20×15	104/138/126	173/102/90	290/92/46	103/211/144

NC, optimization does not converge.

at capturing the shape parameter, while the gradient score does better for the scale parameter. The moderate extremal dependence cases, with $\kappa = 1$ and $1{\cdot}3$, are dominated by the censored likelihood, whereas for weak extremal dependence, $\kappa = 1{\cdot}8$, the gradient score performs best, because too much information is lost by censoring. For the 100-point grid, the optimization procedures do not converge when the extremal dependence is too weak. The choice of the weighting function w affects the robustness of the gradient score. Computation times are similar to those in § 4·1.

Quantile-quantile plots show that the score-matching estimators are very close to being normally distributed, but censored likelihood estimates can deviate somewhat from normality due to the quasi-Monte Carlo approximation; this can be remedied by increasing the value of p.

To summarize: for strong extremal dependence, the three types of estimator are roughly equivalent. For moderate extremal dependence, we recommend using the censored likelihood if the number of variables permits; this is $D \lesssim 500$ with our computational capabilities, although if extremal independence is reached at far distances and the grid is dense, the gradient score is an excellent substitute. Owing to its robustness and lack of dimensionality limitations, the gradient score appears to be the best choice for gridded applications with fine resolution. Empirical work suggests that it can be robustified by careful design of the weight function.

5. Extreme rainfall over Florida

5·1. *Real-data application*

We fit an r-Pareto process based on the Brown–Resnick model to radar measurements of rainfall taken every 15 minutes during the wet season, June–September, from 1999 to 2004 on a regular 2 km grid in a 120 km × 120 km region of east Florida; see Fig. 1. There are 3600 spatial observations in each radar image, and 70 272 images in all. The region was chosen to repeat the application of Buhl & Klüppelberg (2016), but in a spatial setting only; a spatiotemporal model is beyond the scope of the present paper. Buhl & Klüppelberg (2016) analysed daily maxima for 10 km × 10 km squares, but we use nonaggregated data to fit a nonseparable parametric model for spatial extremal dependence, using single extreme events instead of daily maxima.

The marginal distributions for each grid cell were first locally transformed to unit Pareto using their empirical distribution functions. For general application, where we wish to extrapolate the

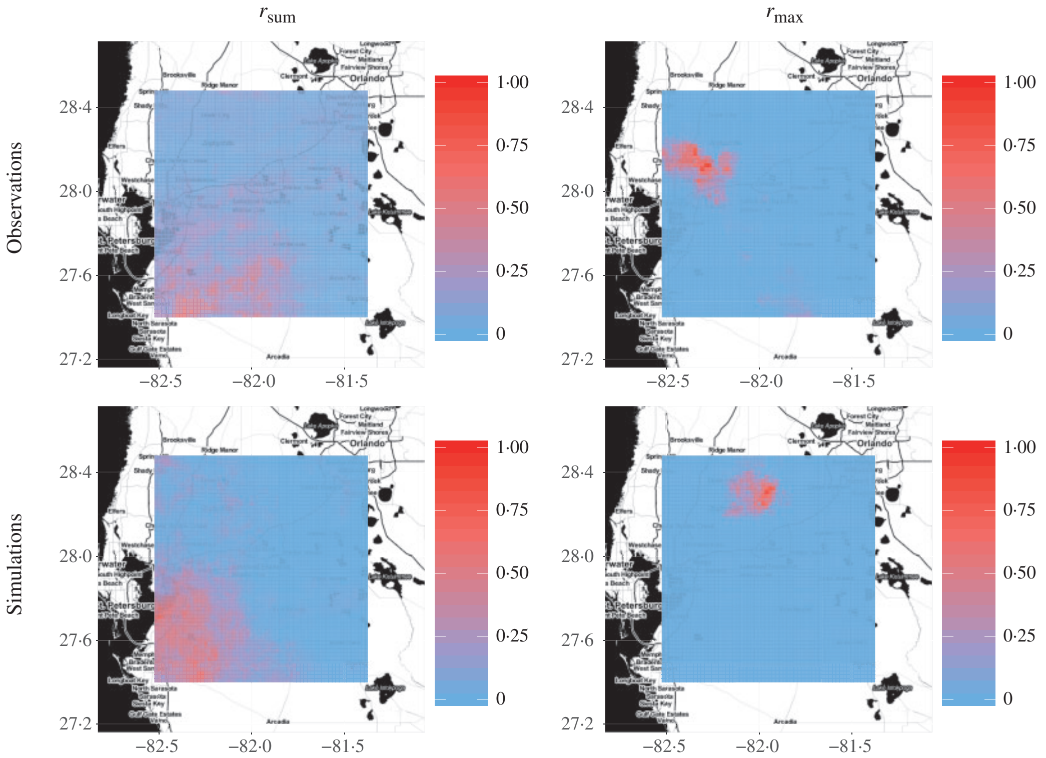

Fig. 1. Fifteen-minute accumulated rainfall in inches, observed (first row) and simulated (second row) for the risk functionals r_{sum} and r_{max} with an intensity equivalent to the 59th most intense event.

distribution above observed intensities, a model for the marginal distributions of exceedances is needed, but since our goal here is to illustrate the feasibility of dependence model estimation on dense grids, we regard marginal modelling as outside the scope of this study.

5·2. *Multivariate extremal dependence model*

The spatial model of Buhl & Klüppelberg (2016) is fully separable, i.e., it is a sum of two separate semivariograms. This has the advantage that inference for each direction can be performed separately, but it cannot capture anisotropy that does not follow the axis of the grid, i.e., is not in the South-North or East-West directions. Furthermore, their pairwise likelihood approach focuses on short-distance pairs, and so could mis-estimate dependence at longer distances. To better capture possible anisotropy, we use the nonseparable semivariogram

$$\gamma(s_i, s_j) = \left\| \frac{\Omega(s_i - s_j)}{\tau} \right\|^{\kappa}, \quad s_i, s_j \in [0, 120]^2, \quad i,j \in \{1, \ldots, 3600\}, \quad 0 < \kappa \leqslant 2, \quad \tau > 0,$$

and anisotropy matrix

$$\Omega = \begin{bmatrix} \cos\eta & -\sin\eta \\ a\sin\eta & a\cos\eta \end{bmatrix}, \quad \eta \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right], \quad a > 0.$$

The semivariogram γ achieves asymptotic extremal independence as the distance between sites tends to infinity, i.e., the pairwise extremal index increases to 2 as $\|s - s'\| \to \infty$.

Table 3. *Parameter estimates (with standard errors in parentheses) for an r-Pareto process derived from log-Gaussian random functions with the semivariogram* $\gamma(s, s') = \{\|\Omega(s - s')\|/\tau\}^{\kappa}$ *obtained by maximization of the gradient score for events corresponding to the* 59 *highest exceedances of the risk functionals* $r_{\rm sum}$ *and* $r_{\rm max}$ *for the Florida radar rainfall data*

Risk functional	κ	τ	η	a
$r_{\rm sum}$	0·814 (0·036)	25·63 (4·70)	−0·009 (0·458)	1·059 (0·031)
$r_{\rm max}$	0·955 (0·048)	3·540 (0·67)	−0·316 (0·410)	0·940 (0·029)

To apply the peaks-over-threshold approach, we must define exceedances by choosing risk functionals. We focus on two types of extremes: local very intense rainfall at any point of the region, and high cumulative rainfall over the whole region. We therefore take the risk functionals

$$r_{\rm max}(X^*) = \left\{\sum_{d=1}^{3600} X^*(s_d)^{20}\right\}^{1/20}, \quad r_{\rm sum}(X^*) = \left\{\sum_{d=1}^{3600} X^*(s_d)^{\xi_0}\right\}^{1/\xi_0}.$$

The function $r_{\rm max}$ is a differentiable approximation to $\max_d X(s_d)$, which cannot be used with the gradient score because of its nondifferentiability. Censored likelihood is computationally out of reach with so many locations. Directly summing normalized observations X^* makes no physical sense, so our function $r_{\rm sum}$, which selects extreme events with large spatial extent, attempts to transform the data back to the original scale; we take $\xi_0 = 0{\cdot}114$, which is the average of independent local estimates of a generalized Pareto distribution.

We fitted univariate generalized Pareto distributions to $r_{\rm sum}(x_n^*)$ and $r_{\rm max}(x_n^*)$ ($n = 1, \ldots, 70\,272$) with increasing thresholds. The estimated shape parameters are stable around the 99·9 percentile, which we used for event selection, giving 59 exceedances; just two events were found to be extreme relative to both risk functionals. This threshold may appear rather high, but it corresponds to around 10 events per year, which seems reasonable in light of the time-frame. Here we merely illustrate the feasibility of high-dimensional inference, so we treat them as independent, but in practice temporal declustering should be considered.

Optimization of the gradient score with the w^1 weighting function on a 16-core cluster took 1–6 hours, depending on the initial point. Different initial points must be considered because of the possibility of local maxima. Results are shown in Table 3, where standard deviations are obtained using a jackknife procedure with 20 blocks. Both the estimated bias and the variance are fairly low. For $r_{\rm sum}(x_n^*)$, we obtain a model similar to that of Buhl & Klüppelberg (2016).

The estimated parameters differ appreciably for the two risk functionals, suggesting the presence of a mixture of different types of extreme events. The structure for $r_{\rm max}$ is consistent with the database, in which the most intense events tend to be spatially concentrated. Our model suggests higher dependence for middle distances than was found by Buhl & Klüppelberg (2016), but they did note that their model underestimates dependence, especially for high quantiles. The estimated smoothness parameters are very close, and neither estimate of η differs significantly from zero, as imposed by Buhl & Klüppelberg (2016). For $r_{\rm sum}$, the estimated parameters show strong extremal dependence even at long distances, corresponding to exceedances of accumulated rainfall with large spatial cover. As $\hat{a} \approx 1$, there is no evidence of anistropy.

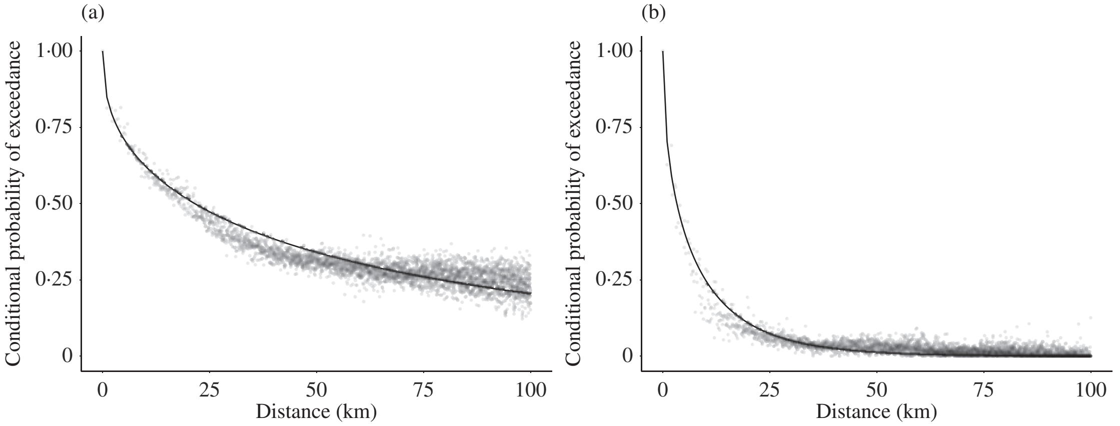

Fig. 2. Estimated conditional exceedance probabilities π_{ij} for the risk functionals (a) r_{sum} and (b) $r_{\max}$ as functions of the distance between locations s_i and s_j $(i,j = 1,\ldots,3600)$. In each panel the solid black curve represents the model fitted using gradient score estimation.

5·3. *Model checking and simulation*

For model checking, we propose to use the conditional exceedance probability,

$$\pi_{ij} = \text{pr}\big[X^*(s_j) \geqslant u' \mid \{X^*(s_i) \geqslant u'\} \cap \{r(X^*/u) \geqslant 1\}\big] = 2\left[1 - \Phi\left\{\left(\frac{\gamma_{ij}}{2}\right)^{1/2}\right\}\right],$$

where $\gamma_{i,j}$ is the semivariogram for sites s_i and s_j $(i,j = 1,\ldots,3600)$, as defined in (9), and $u' > 0$. An empirical estimator of π_{ij} is

$$\hat{\pi}_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}\big[\{r(x^{*n}/u) \geqslant 1\} \cap \{x_i^{*n} \geqslant u'\} \cap \{x_j^{*n} \geqslant u'\}\big]}{\sum_{n=1}^{N} \mathbb{1}\big[\{r(x^{*n}/u) \geqslant 1\} \cap \{x_i^{*n} \geqslant u'\}\big]},$$

whose asymptotic behaviour derives from Davis & Mikosch (2009). For both risk functionals, the fitted model, represented by the solid black lines in Fig. 2, follows the cloud of estimated conditional exceedance probabilities reasonably well and captures the general trend, but fails to represent some local variation, perhaps due to a lack of flexibility of the power model.

Finally, we use the models fitted in § 5·2 to simulate events with intensities equivalent to the weakest of the 59 events found by our risk functionals. Simulation is performed by generating the corresponding r-Pareto process with the fitted dependence structure, as in § 4·1. Figure 1 compares observations from the database and representative simulations, which seem to successfully reproduce both the spatial dependence and the intensity of the selected observations. A closer examination suggests that in both cases the models produce oversmooth rainfall fields. This could be addressed by improving event selection using risk functionals r that characterize special spatial structures or physical processes. Although we fail to detect anisotropy, more complex models for dependence that allow stochasticity of the spatial patterns might be worth considering.

These models can reproduce both spatial patterns and extreme intensity for spatially accumulated and local heavy rainfall. In both cases the fitted dependence model provides a reasonable fit and simulations seem broadly consistent with observations. However, the presence of two contrasting dependence structures highlights the complexity of extreme rainfall and suggests that a mixture model for both dependence and margins might be considered. Marginal and dependence parameters are often estimated separately, but with the presence of mixtures, which can be detected using different risk functionals, joint estimation is required, which is beyond the scope of this paper. For this reason and because we have neglected the temporal dependence, our model should be viewed as merely a first step towards a spatiotemporal rainfall generator.

Acknowledgement

We thank the Swiss National Science Foundation for financial support, Sebastian Engelke for helpful comments, and the reviewers for their constructive remarks. Chin Man Mok kindly provided the data, which were supplied by the Southwest Florida Water Management District.

Supplementary material

Supplementary material available at *Biometrika* online includes discussion of computational considerations, code to run the simulations, details of the gradient scoring rule, the proof of Proposition 1, and further numerical results.

References

Asadi, P., Davison, A. C. & Engelke, S. (2015). Extremes on river networks. *Ann. Appl. Statist.* **9**, 2023–50.

Buhl, S. & Klüppelberg, C. (2016). Anisotropic Brown–Resnick space-time processes: Estimation and model assessment. *Extremes* **19**, 627–60.

Castruccio, S., Huser, R. & Genton, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *J. Comp. Graph. Statist.* **25**, 1212–29.

Chiles, J.-P. & Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Coles, S. G. & Tawn, J. A. (1991). Modelling extreme multivariate events. *J. R. Statist. Soc.* B **53**, 377–92.

Coles, S. G. & Tawn, J. A. (1996). Modelling extremes of the areal rainfall process. *J. R. Statist. Soc.* B **58**, 329–47.

Davis, R. A., Klüppelberg, C. & Steinkohl, C. (2013). Max-stable processes for modelling extremes observed in space and time. *J. Korean Statist. Soc.* **42**, 399–414.

Davis, R. A. & Mikosch, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli* **15**, 977–1009.

Davison, A. C. & Smith, R. L. (1990). Models for exceedances over high thresholds (with Discussion). *J. R. Statist. Soc.* B **52**, 393–442.

Dawid, A. P., Musio, M. & Ventura, L. (2016). Minimum scoring rule inference. *Scand. J. Statist.* **43**, 123–38.

de Fondeville, R. (2017). *mvPot: Multivariate Peaks-over-Threshold Modelling for Spatial Extreme Events*. R package version 0.1.4, available at https://cran.r-project.org/package=mvPot.

de Haan, L. & Lin, T. (2001). On convergence toward an extreme value distribution in $C[0, 1]$. *Ann. Prob.* **29**, 467–83.

de Haan, L. & Resnick, S. I. (1993). Estimating the limit distribution of multivariate extremes. *Commun. Statist. Stoch. Mod.* **9**, 275–309.

Dick, J. & Pillichshammer, F. (2010). *Digital Nets and Sequences*. Cambridge: Cambridge University Press.

Dombry, C., Engelke, S. & Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika* **103**, 303–17.

Dombry, C. & Ribatet, M. (2015). Functional regular variations, Pareto processes and peaks over thresholds. *Statist. Interface* **8**, 9–17.

Einmahl, J. H. J., Kiriliouk, A., Krajina, A. & Segers, J. (2016). An M-estimator of spatial tail dependence. *J. R. Statist. Soc.* B **78**, 275–98.

Engelke, S., Malinowski, A., Kabluchko, Z. & Schlather, M. (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *J. R. Statist. Soc.* B **77**, 239–65.

Ferreira, A. & de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli* **20**, 1717–37.

Genz, A. & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Dordrecht: Springer.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. R. Statist. Soc.* B **56**, 261–74.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc.* **102**, 359–78.

Gumbel, E. J. (1958). *Statistics of Extremes*. New York: Columbia University Press.

Hult, H. & Lindskog, F. (2005). Extremal behavior of regularly varying stochastic processes. *Stoch. Proces. Appl.* **115**, 249–74.

HUSER, R. & DAVISON, A. C. (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika* **100**, 511–8.

HUSER, R., DAVISON, A. C. & GENTON, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes* **19**, 79–103.

HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6**, 695–708.

HYVÄRINEN, A. (2007). Some extensions of score matching. *Comp. Statist. Data Anal.* **51**, 2499–512.

JOE, H., SMITH, R. L. & WEISSMAN, I. (1992). Bivariate threshold methods for extremes. *J. R. Statist. Soc.* B **54**, 171–83.

KABLUCHKO, Z., SCHLATHER, M. & DE HAAN, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Prob.* **37**, 2042–65.

KLÜPPELBERG, C. & RESNICK, S. I. (2008). The Pareto copula, aggregation of risks, and the emperor's socks. *J. Appl. Prob.* **45**, 67–84.

LINDSKOG, F., RESNICK, S. I. & ROY, J. (2014). Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Prob. Surv.* **11**, 270–314.

MADSEN, H., RASMUSSEN, P. F. & ROSBJERG, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events. *Water Resources Res.* **33**, 747–57.

NUYENS, D. & COOLS, R. (2004). Fast component-by-component construction, a reprise for different kernels. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, H. Niederreiter & D. Talay, eds. Berlin: Springer, pp. 373–87.

PADOAN, S. A., RIBATET, M. & SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Am. Statist. Assoc.* **105**, 263–77.

R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

RESNICK, S. I. (2007). *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer.

ROOTZÉN, H. & TAJVIDI, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli* **12**, 917–30.

STEPHENSON, A. G. & TAWN, J. A. (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika* **92**, 213–27.

THIBAUD, E., AALTO, J., COOLEY, D. S., DAVISON, A. C. & HEIKKINEN, J. (2016). Bayesian inference for the Brown–Resnick process, with an application to extreme low temperatures. *Ann. Appl. Statist.* **10**, 2303–24.

THIBAUD, E. & OPITZ, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102**, 855–70.

THORARINSDOTTIR, T. L., GNEITING, T. & GISSIBL, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncert. Quant.* **1**, 522–34.

VARIN, C., REID, N. & FIRTH, D. (2011). An overview of composite marginal likelihoods. *Statist. Sinica* **21**, 5–42.

WADSWORTH, J. L. (2015). On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. *Biometrika* **102**, 705–11.

WADSWORTH, J. L. & TAWN, J. A. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101**, 1–15.

[*Received on* 7 *April* 2017. *Editorial decision on* 21 *February* 2018]