ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 4, *pp.* 871–932

Graphical models for extremes

Sebastian Engelke

University of Geneva, Switzerland

and Adrien S. Hitz

University of Oxford and Materialize.X, London, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 12th, 2020*, Professor G. P. Nason *in the Chair*]

Summary. Conditional independence, graphical models and sparsity are key notions for parsimonious statistical models and for understanding the structural relationships in the data. The theory of multivariate and spatial extremes describes the risk of rare events through asymptotically justified limit models such as max-stable and multivariate Pareto distributions. Statistical modelling in this field has been limited to moderate dimensions so far, partly owing to complicated likelihoods and a lack of understanding of the underlying probabilistic structures. We introduce a general theory of conditional independence for multivariate Pareto distributions that enables the definition of graphical models and sparsity for extremes. A Hammersley–Clifford theorem links this new notion to the factorization of densities of extreme value models on graphs. For the popular class of Hüsler–Reiss distributions we show that, similarly to the Gaussian case, the sparsity pattern of a general extremal graphical model can be read off from suitable inverse covariance matrices. New parametric models can be built in a modular way and statistical inference can be simplified to lower dimensional marginals. We discuss learning of minimum spanning trees and model selection for extremal graph structures, and we illustrate their use with an application to flood risk assessment on the Danube river.

Keywords: Conditional independence; Extreme value theory; Graphical models; Multivariate Pareto distribution; Sparsity

1. Introduction

Evaluation of the risk that is related to heat waves, extreme flooding, financial crises or other rare events requires the quantification of their small occurrence probabilities. Empirical estimates are unreliable since the regions of interest are in the tail of the distribution and typically contain few or no data points. Extreme value theory provides the theoretical foundation for extrapolations to the distributional tail of a d-dimensional random vector $\mathbf{X}$. The univariate case $d = 1$ is well studied and the generalized extreme value and Pareto distributions are widely applied in areas such as hydrology (Katz *et al.*, 2002), climate science (Min *et al.*, 2011) and finance (McNeil *et al.*, 2015); see also Embrechts *et al.* (1997) and Beirlant *et al.* (2004).

In the multivariate setting, $d \geqslant 2$, the result of the extrapolation strongly depends on the strength of extremal dependence between the components of $\mathbf{X}$. Most current statistical models assume multivariate regular variation for $\mathbf{X}$ (Resnick, 2008) since this entails mathematically elegant descriptions of the asymptotic tail distribution. Similarly to the univariate setting, two different but closely related approaches exist. Max-stable distributions arise as limits of nor-

Address for correspondence: Sebastian Engelke, Research Center for Statistics, University of Geneva, Boulevard du Pont d'Arve 40, 1205 Geneva, Switzerland.
E-mail: sebastian.engelke@unige.ch

 1369–7412/20/82871

malized maxima of independent copies of $\mathbf{X}$ and have been extensively studied and applied in multivariate and spatial risk problems (see de Haan (1984), Gudendorf and Segers (2010) and Davison *et al.* (2012)). In contrast, multivariate Pareto distributions describe the random vector $\mathbf{X}$ conditioned on the event that at least one component exceeds a high threshold; see Rootzén and Tajvidi (2006), Rootzén *et al.* (2018) and Kiriliouk *et al.* (2018a) for their construction, stability properties and statistical inference.

A drawback of the current multivariate models is their limitation to rather moderate dimensions d, and the construction of tractable parametric models in higher dimensions is challenging, both for max-stable and for multivariate Pareto distributions. Sparse multivariate models require the notion of conditional independence (Dawid, 1979), which is not easy to define for tail distributions. In fact, Papastathopoulos and Strokorb (2016) have shown that, if (Z_1, Z_2, Z_3) is a max-stable random vector with positive continuous density, then the conditional independence of $Z_1 \perp\!\!\!\perp Z_3 | Z_2$ already implies the independence $Z_1 \perp\!\!\!\perp Z_3$; see also Dombry and Éyi-Minko (2014). Meaningful conditional independence structures can thus only be obtained for max-stable distributions with discrete spectral measure (Gissibl and Klüppelberg, 2018). Since these models do not admit densities, this excludes most of the currently used parametric families.

In this work we take the perspective of threshold exceedances and introduce a new notion of conditional independence for a multivariate Pareto distribution $\mathbf{Y} = (Y_1, \ldots, Y_d)$, which we denote by '$\perp_{\mathrm{e}}$' to stress that it is designed for extremes. It is different from classical conditional independence since the support of $\mathbf{Y}$ is not a product space, but the homogeneity property of $\mathbf{Y}$ can be used to show that it is well defined. Conditional independence is tightly linked to graphical models. For an undirected graph $\mathcal{G} = (V, E)$ with nodes $V = \{1, \ldots, d\}$ and edge set E, we say that $\mathbf{Y}$ is an extremal graphical model if it satisfies the pairwise Markov property

$$Y_i \perp_{\mathrm{e}} Y_j | \mathbf{Y}_{\setminus \{i,j\}}, \qquad (i, j) \notin E. \tag{1}$$

The main advantage of conditional independence and graphical models is that they imply a simple probabilistic structure and possibly sparse patterns in multivariate random vectors (Lauritzen, 1996; Wainwright and Jordan, 2008). For extremal graphical models on decomposable graphs, we prove a Hammersley–Clifford-type theorem stating that property (1) is equivalent to the factorization of the density $f_{\mathbf{Y}}$ of $\mathbf{Y}$ into lower dimensional marginals. This underlines that our notion of conditional independence is in fact natural for multivariate Pareto distributions.

Applications of this result are manifold. From a probabilistic perspective, we analyse models in the literature regarding their graphical properties in the sense of our definition (1). Extremal graphical models whose underlying graph is a tree have a particularly simple multiplicative stochastic representation in terms of extremal functions: a notion that is known from the simulation of max-stable processes (Dombry *et al.*, 2016). In multivariate extremes, one may argue that the family of Hüsler–Reiss distributions (Hüsler and Reiss, 1989) takes a similar role to that of Gaussian distributions in the non-extreme world. Instead of covariance matrices, they are parameterized by a variogram matrix Γ. We show that the extremal graphical structure of a Hüsler–Reiss distribution can be identified by zero patterns on matrices derived from Γ.

Extremal graphical models enable the construction of parsimonious models for multivariate Pareto distributions $\mathbf{Y}$, which further enjoy the advantage of interpretability in terms of the underlying graph. Thanks to the factorization of the densities, statistical inference can be efficiently carried out on lower dimensional marginals. For decomposable graphs with singleton separator sets, so-called block graphs, this allows the use of multivariate Pareto models in fairly high dimensions. In many cases the underlying graphical structure is unknown and must be learned from data. We discuss how a maximum likelihood tree can be obtained by using

standard algorithms by Kruskal (1956) or Prim (1957), and how the best model can be selected among different extremal graphical models.

There is previous work on the construction of parsimonious extreme value models. A large body of literature studies spatial max-stable random fields (Schlather, 2002; Kabluchko *et al.*, 2009; Opitz, 2013). Such models have small parameter dimension but they rely on strong assumptions on stationarity and cannot be applied to multivariate, non-spatial data without information on an underlying space. Other approaches include constructions through factor copulas (Lee and Joe, 2018), ensembles of trees combining bivariate copulas (Yu *et al.*, 2017), graphical models for large censored observations (Hitz and Evans, 2016) and eigendecompositions (Cooley and Thibaud, 2019). Closely related to our concept of conditional independence is the work of Coles and Tawn (1991) and Smith *et al.* (1997) who proposed a Markov chain model where all bivariate marginals are extreme value distributions. This can be seen as a special case of our approach when the graph has the simple structure of a chain. Similar limiting objects also arise as the tail chains in the theory of extremes of stationary Markov chains with regularly varying marginals (Smith, 1992; Basrak and Segers, 2009; Janssen and Segers, 2014). This theory has recently been extended to regularly varying Markov trees (Segers, 2019). Gissibl and Klüppelberg (2018) and Gissibl *et al.* (2018) studied the causal structure of directed acyclic graphs for max-linear models, and they developed methods for model identification based on tail dependence coefficients. Their work is in some sense complementary to ours, since their models do not have densities whereas we shall explicitly assume the existence of densities.

To the best of our knowledge, our work is the first principled attempt to define conditional independence for general multivariate extreme value models that naturally extends to the factorization of densities, sparsity and graphical models. Section 2 introduces background on extreme value theory and graphical models that is needed throughout the paper. The new notion of conditional independence is defined in Section 3 and equivalent properties are derived. Section 4 contains the main probabilistic results on extremal graphical models, the representation of trees and the characterization for Hüsler–Reiss distributions. Statistical models on block graphs and their estimation, simulation and model selection are discussed in Section 5. In these graphical models the dependence is modelled directly between lower dimensional subsets of variables, whereas the global dependence is implicitly implied by the conditional independence structure of the graph. There are many potential applications of extremal graphical models. In Section 6, we illustrate the advantages of this structured approach compared with classical extreme value models on a data set related to flooding on a river network in the upper Danube basin (see Asadi *et al.* (2015)). The interpretation of the graphical structures that are obtained in this application is particularly interesting since there is a seemingly natural underlying tree associated with the flow connections. Our conditional independence is formulated for multivariate Pareto distributions, but the results in this paper have implications for max-stable distributions. This point and further research directions will be addressed in the discussion in Section 7. Appendices A–F contain proofs and some additional results.

The package `graphicalExtremes` (Engelke *et al.*, 2019) is an implementation for R (R Core Team, 2019). The code for the simulation study and application is available from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`.

2. Background

2.1. Notation

We introduce some standard notation that is used throughout the paper. Symbols in bold such

as $\mathbf{x} \in \mathbb{R}^d$ are column vectors with components denoted by x_i, $i \in \{1, \ldots, d\}$, and operations and relationships involving such vectors are meant componentwise. The vectors $\mathbf{0} = (0, \ldots, 0)$ and $\mathbf{1} = (1, \ldots, 1)$ are used as generic vectors with suitable dimension. Denoting the index set by $V = \{1, \ldots, d\}$, for a non-empty subset $I \subset V$, we write for the subvectors $\mathbf{x}_I = (x_i)_{i \in I}$ and $\mathbf{x}_{\setminus I} = (x_i)_{i \in V \setminus I}$. Similar notation is used for random vectors $\mathbf{X} = (X_i)_{i \in V}$ with values in $\mathbb{R}^d$. For a matrix $A = (A_{ij})_{i,j \in V} \in \mathbb{R}^{d \times d}$ with entries indexed by V, and subsets $I, J \subset V$, we let $A_{IJ} = (A_{ij})_{i \in I, j \in J}$ denote the $|I| \times |J|$ submatrix of A, and we abbreviate $A_I = A_{II}$. For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ with $\mathbf{a} \leqslant \mathbf{b}$, a multivariate interval is denoted by $[\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \ldots \times [a_d, b_d]$. The l_p-norm of a vector $\mathbf{x} \in \mathbb{R}^d$ for $p \geqslant 1$ is $\|\mathbf{x}\|_p = (\Sigma_{i \in V} |x_i|^p)^{1/p}$, and its l_∞-norm is $\|\mathbf{x}\|_\infty = \max_{i \in V} |x_i|$. The density of a random vector $\mathbf{X}$, if it exists, is denoted by $f_{\mathbf{X}}$. The density of the marginal $\mathbf{X}_I$ for a non-empty $I \subset V$ is denoted by f_I, if there is no ambiguity regarding the random vector.

2.2. Multivariate extreme value theory

The tail behaviour of the random vector $\mathbf{X} = (X_1, \ldots, X_d)$ can be described through two different approaches: one based on componentwise maxima and the other on threshold exceedances. We briefly discuss both approaches and the close link between them.

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$, be independent copies of $\mathbf{X}$ and denote the componentwise maximum by $\mathbf{M}_n = (M_{1n}, \ldots, M_{dn}) = (\max_{i=1}^n X_{i1}, \ldots, \max_{i=1}^n X_{id})$. Under mild conditions on the marginal distribution of X_j there are sequences of normalizing constants $b_{jn} \in \mathbb{R}$, $a_{jn} > 0$, $j = 1, \ldots, d$, such that

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{M_{jn} - b_{jn}}{a_{jn}} \leqslant x \right) = G_j(x) = \exp\{-(1 + \xi_j x)_+^{-1/\xi_j}\}, \qquad x \in \mathbb{R}, \tag{2}$$

where $z_+ = \max(z, 0)$, and G_j is the generalized extreme value distribution whose shape parameter $\xi_j \in \mathbb{R}$ determines the heaviness of the tail of X_j; see de Haan and Ferreira (2006), Embrechts *et al.* (1997) and Beirlant *et al.* (2004) for details. For analysis of the dependence structure, the marginal distributions F_j of X_j are typically estimated first to normalize the data by $1/\{1 - F_j(X_j)\}$ to standard Pareto distributions. For simplicity, we assume in what follows that the F_j are known and the vector $\mathbf{X}$ has been normalized to standard Pareto marginals. Joint estimation of marginals and dependence is discussed in Section 5.2.

The standardized vector $\mathbf{X}$ is said to be in the max-domain of attraction of the random vector $\mathbf{Z} = (Z_1, \ldots, Z_d)$ if for any $\mathbf{z} = (z_1, \ldots, z_d)$

$$\lim_{n \to \infty} \mathbb{P}\Big(\max_{i=1,\ldots,n} X_{i1} \leqslant n z_1, \ldots, \max_{i=1,\ldots,n} X_{id} \leqslant n z_d \Big) = \mathbb{P}(\mathbf{Z} \leqslant \mathbf{z}). \tag{3}$$

In this case, $\mathbf{Z}$ is max-stable with standard Fréchet marginals $\mathbb{P}(Z_j \leqslant z) = \exp(-1/z)$, $z \geqslant 0$, and we may write

$$\mathbb{P}(\mathbf{Z} \leqslant \mathbf{z}) = \exp\{-\Lambda(\mathbf{z})\}, \qquad \mathbf{z} \in \mathcal{E}, \tag{4}$$

where the exponent measure Λ is a Radon measure on the cone $\mathcal{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$, and $\Lambda(\mathbf{z})$ is shorthand for $\Lambda(\mathcal{E} \setminus [\mathbf{0}, \mathbf{z}])$. If Λ is absolutely continuous with respect to Lebesgue measure on $\mathcal{E}$, its Radon–Nikodym derivative, denoted by λ, has the following properties:

(a) homogeneity of order $-(d+1)$, i.e. $\lambda(t\mathbf{y}) = t^{-(d+1)} \lambda(\mathbf{y})$ for any $t > 0$ and $\mathbf{y} \in \mathcal{E}$;
(b) normalized marginals, i.e. for any $i = 1, \ldots, d$,

$$\int_{\mathbf{y} \in \mathcal{E} : y_i > 1} \lambda(\mathbf{y}) \mathrm{d}\mathbf{y} = 1.$$

The two properties follow from the max-stability and the standard Fréchet marginals of $\mathbf{Z}$ respectively. For a non-empty subset $I \subset \{1, \ldots, d\}$, we define the marginal of λ by

$$\lambda_I(\mathbf{y}_I) = \int_{[0,\infty)^{d-|I|}} \lambda(\mathbf{y}) \mathrm{d}\mathbf{y}_{\setminus I}, \tag{5}$$

and note that it is homogeneous of order $-(|I|+1)$. In particular, if $I = \{i\}$ for some $i = 1, \ldots, d$, then $\lambda_{\{i\}}(y_i) = 1/y_i^2$ as a consequence of properties (a) and (b). Conversely, any positive and continuous function λ satisfying properties (a) and (b) defines a valid density of an exponent measure $\Lambda(\mathbf{z})$ by integration over $\mathcal{E} \setminus [\mathbf{0}, \mathbf{z}]$, $\mathbf{z} \in \mathcal{E}$, that satisfies similar homogeneity and normalization properties as λ. By expression (4) this also defines a max-stable distribution.

Another perspective on multivariate extremes is through threshold exceedances. By proposition 5.17 in Resnick (2008), the convergence in equation (3) is equivalent to

$$\lim_{u \to \infty} u\{1 - \mathbb{P}(\mathbf{X} \leqslant u\mathbf{z})\} = \Lambda(\mathbf{z}), \qquad \mathbf{z} \in \mathcal{E}.$$

Consequently, the multivariate distribution of the threshold exceedances of $\mathbf{X}$ satisfies

$$\mathbb{P}(\mathbf{Y} \leqslant \mathbf{z}) = \lim_{u \to \infty} \mathbb{P}\left(\frac{\mathbf{X}}{u} \leqslant \mathbf{z} \,\middle|\, \|\mathbf{X}\|_\infty > u \right) = \frac{\Lambda(\mathbf{z} \wedge \mathbf{1}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})}, \qquad \mathbf{z} \in \mathcal{E}. \tag{6}$$

The distribution of the limiting random vector $\mathbf{Y}$ is called a multivariate Pareto distribution (see Rootzén and Tajvidi (2006)). It is defined through the exponent measure Λ of the max-stable distribution $\mathbf{Z}$, with support on the L-shaped space $\mathcal{L} = \{\mathbf{x} \in \mathcal{E} : \|\mathbf{x}\|_\infty > 1\}$. We say that $\mathbf{Z}$ and $\mathbf{Y}$ are associated, since their distributions mutually determine each other.

Multivariate Pareto distributions are the only possible limits in expression (6) and, owing to the homogeneity of the exponent measure, they enjoy certain stability properties (see Rootzén *et al.* (2018)). The exponent measure Λ, and hence the distribution of $\mathbf{Y}$, may place mass on some lower dimensional faces of the space $\mathcal{E}$. For the remainder of this paper we exclude this case to avoid technical difficulties. We further assume that the distribution of $\mathbf{Y}$ admits a positive and continuous density $f_{\mathbf{Y}}$ on $\mathcal{L}$, which is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\partial^d}{\partial y_1 \ldots \partial y_d} \mathbb{P}(\mathbf{Y} \leqslant \mathbf{y}) = \frac{\lambda(\mathbf{y})}{\Lambda(\mathbf{1})}, \qquad \mathbf{y} \in \mathcal{L},$$

since $\Lambda(\mathbf{y} \wedge \mathbf{1})$ is always constant along at least one co-ordinate for $\mathbf{y} \in \mathcal{L}$. The density $f_{\mathbf{Y}}$ is thus proportional to the density λ of the exponent measure Λ. By the homogeneity of λ, $f_{\mathbf{Y}}$ is also homogeneous of order $-(d+1)$. The normalization constant $\Lambda(\mathbf{1}) \in [1, d]$ is known as the d-variate extremal coefficient (see Schlather and Tawn (2003)). The assumption of a positive and continuous density $f_{\mathbf{Y}}$ implies that the multivariate Pareto distributions that we study are models for asymptotic extremal dependence, and all p-variate extremal coefficients, $1 \leqslant p \leqslant d$, are strictly smaller than their upper limit p.

For some non-empty subset $I \subset \{1, \ldots, d\}$, the subvector $\mathbf{X}_I = (X_j)_{j \in I}$, properly normalized, given that its l_∞-norm is large converges in the sense of expression (6) to the marginal $\mathbf{Y}_I = (Y_j)_{j \in I}$ of $\mathbf{Y}$ conditioned on $\{\|\mathbf{Y}_I\|_\infty > 1\}$, defined on $\mathcal{L}_I = \{\mathbf{x}_I \in [0, \infty)^{|I|} \setminus \{\mathbf{0}\} : \|\mathbf{x}_I\|_\infty > 1\}$ with homogeneous density of order $-(|I|+1)$ given by

$$f_I(\mathbf{y}_I) = \frac{\Lambda(\mathbf{1})}{\Lambda_I(\mathbf{1})} \int_{[0,\infty)^{d-|I|}} f_{\mathbf{Y}}(\mathbf{y}) \mathrm{d}\mathbf{y}_{\setminus I} = \frac{\lambda_I(\mathbf{y}_I)}{\Lambda_I(\mathbf{1})}, \qquad \mathbf{y}_I \in \mathcal{L}_I, \tag{7}$$

where Λ_I is the exponent measure of $\mathbf{Z}_I$, and λ_I is the density of Λ_I.

2.2.1. Example 1 (logistic distribution)

The extremal logistic distribution with parameter $\theta \in (0, 1)$ induces a multivariate Pareto distribution with density

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{d^{\theta}}(y_1^{-1/\theta} + \ldots + y_d^{-1/\theta})^{\theta - d} \prod_{i=1}^{d-1} \left(\frac{i}{\theta} - 1\right) \prod_{i=1}^{d} y_i^{-1/\theta - 1}, \qquad \mathbf{y} \in \mathcal{L}. \tag{8}$$

2.2.2. Example 2 (Hüsler–Reiss distribution)

The Hüsler–Reiss distribution (Hüsler and Reiss, 1989) is parameterized by a symmetric, strictly conditionally negative definite matrix $\Gamma = \{\Gamma_{ij}\}_{1 \leqslant i, j \leqslant d}$ with $\mathrm{diag}(\Gamma) = \mathbf{0}$ and non-negative entries, i.e. $\mathbf{a}^{\mathrm{T}}\Gamma\mathbf{a} < 0$ for all non-zero vectors $\mathbf{a} \in \mathbb{R}^d$ with $\Sigma_{i=1}^{d} a_i = 0$. The corresponding density of the exponent measure can be written for any $k \in \{1, \ldots, d\}$ as (see Engelke *et al.* (2015))

$$\lambda(\mathbf{y}) = y_k^{-2} \prod_{i \neq k} y_i^{-1} \phi_{d-1}(\tilde{\mathbf{y}}_{\setminus k}; \Sigma^{(k)}), \qquad \mathbf{y} \in \mathcal{E}, \tag{9}$$

where $\phi_p(\cdot; \Sigma)$ is the density of a centred p-dimensional normal distribution with covariance matrix Σ, $\tilde{\mathbf{y}} = \{\log(y_i/y_k) + \Gamma_{ik}/2\}_{i=1,\ldots,d}$ and

$$\Sigma^{(k)} = \tfrac{1}{2}\{\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}\}_{i,j \neq k} \in \mathbb{R}^{(d-1)\times(d-1)}. \tag{10}$$

The matrix $\Sigma^{(k)}$ is strictly positive definite; see Appendix B for details. The representation of the density in expression (9) seems to depend on the choice of k, but, in fact, the value of the right-hand side of this equation is independent of k. The Hüsler–Reiss multivariate Pareto distribution has density $f_{\mathbf{Y}}(\mathbf{y}) = \lambda(\mathbf{y})/\Lambda(\mathbf{1})$ and the strength of dependence between the ith and jth component is parameterized by Γ_{ij}, ranging from complete dependence for $\Gamma_{ij} = 0$ and independence for $\Gamma_{ij} = \infty$. In the bivariate case $d = 2$ we have

$$\lambda(y_1, y_2) = \frac{y_1^{-2} y_2^{-1}}{\sqrt{(2\pi\Gamma_{12})}} \exp\left[-\frac{\{\log(y_2/y_1) + \Gamma_{12}/2\}^2}{2\Gamma_{12}}\right], \qquad (y_1, y_2) \in \mathcal{E}, \tag{11}$$

and $\Lambda(1, 1) = 2\Phi(\sqrt{\Gamma_{12}}/2)$, where Φ is the standard normal distribution function. The extension of Hüsler–Reiss distributions to random fields are Brown–Resnick processes (Brown and Resnick, 1977; Kabluchko *et al.*, 2009), which are widely used models for spatial extremes.

2.2.3. Example 3 (bivariate Pareto distribution)

In the general bivariate case $d = 2$, because of homogeneity, the density λ of the exponent measure can be characterized by a univariate distribution. Indeed, for any positive random variable U_2^1 with density $f_{U_2^1}$ and $\mathbb{E}(U_2^1) = 1$,

$$\lambda(y_1, y_2) = y_1^{-3} f_{U_2^1}(y_2/y_1), \qquad (y_1, y_2) \in \mathcal{E}, \tag{12}$$

satisfies conditions (a) and (b) in Section 2.2 and thus defines a valid exponent measure density. We call U_2^1 the extremal function at co-ordinate 2, relative to co-ordinate 1 (see Dombry *et al.* (2013, 2016)). Equivalently, we can write the density in terms of the extremal function U_1^2 at co-ordinate 1, relative to co-ordinate 2, as $\lambda(y_1, y_2) = y_2^{-3} f_{U_1^2}(y_1/y_2)$, $(y_1, y_2) \in \mathcal{E}$, and U_1^2 is related to U_2^1 via the measure change $\mathbb{P}(U_1^2 \leqslant z) = \mathbb{E}(\mathbf{1}\{1/U_2^1 \leqslant z\} U_2^1)$, $z > 0$.

The above principle is a general construction principle, since every valid exponent measure density can be obtained in this way. The bivariate Hüsler–Reiss distribution in expression

(11) corresponds to the case of log-normal U_2^1 and U_1^2, but many other parametric and non-parametric examples are available (e.g. Boldi and Davison (2007), Cooley *et al.* (2010), Ballani and Schlather (2011) and de Carvalho and Davison (2014)).

2.3. Graphical models

A graph $\mathcal{G} = (V, E)$ is defined as a set of nodes $V = \{1, \ldots, d\}$, also called vertices, together with a set of edges $E \subset V \times V$ of pairs of distinct nodes. The graph is called undirected if, for two nodes $i, j \in V$, $(i, j) \in E$ if and only if $(j, i) \in E$. For notational convenience, for undirected graphs we sometimes represent edges as unordered pairs $\{i, j\} \in E$. When counting the number of edges, we count $\{i, j\} \in E$ such that each edge is considered only once. A subset $C \subset V$ of nodes is called complete if it is fully connected in the sense that $(i, j) \in E$ for all $i, j \in C$. We denote by $\mathcal{C}$ the set of all cliques, i.e. the complete subsets that are not properly contained within any other complete subset.

To each node $i \in V$ we associate a random variable X_i with continuous state space $\mathcal{X}_i \subset \mathbb{R}$. The resulting random vector $\mathbf{X} = (X_i)_{i \in V}$ takes values in the Cartesian product $\mathcal{X} = \times_{i \in V} \mathcal{X}_i$. Suppose that $\mathbf{X}$ has a positive and continuous Lebesgue density $f_{\mathbf{X}}$ on $\mathcal{X}$. For three disjoint subsets $A, B, C \subset V$ whose union is V, we say that $\mathbf{X}_A$ is conditionally independent of $\mathbf{X}_C$ given $\mathbf{X}_B$ if the density factorizes as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{f_{A \cup B}(\mathbf{x}_{A \cup B}) f_{B \cup C}(\mathbf{x}_{B \cup C})}{f_B(\mathbf{x}_B)}, \tag{13}$$

and we write $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_C | \mathbf{X}_B$. If $B = \emptyset$, then equation (13) amounts to independence of $\mathbf{X}_A$ and $\mathbf{X}_C$.

The random vector $\mathbf{X}$ is said to be a probabilistic graphical model on the graph $\mathcal{G} = (V, E)$ if its distribution satisfies the pairwise Markov property relative to $\mathcal{G}$, i.e. $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus \{i,j\}}$ for all $(i, j) \notin E$. If in addition, for any disjoint subsets $A, B, C \subset V$ such that B separates A from C in $\mathcal{G}$, $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_C | \mathbf{X}_B$, then $\mathbf{X}$ is said to obey the global Markov property relative to $\mathcal{G}$. Since $f_{\mathbf{X}}$ is positive and continuous, it follows from the Hammersley–Clifford theorem (see Lauritzen (1996), theorem 3.9) that the two Markov properties are equivalent, and they are further equivalent to the factorization of the density

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C), \qquad \mathbf{x} \in \mathcal{X}, \tag{14}$$

for suitable functions ψ_C on $\times_{i \in C} \mathcal{X}_i$. If the graph $\mathcal{G}$ is decomposable, then this factorization can be rewritten in terms of marginal densities

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} f_C(\mathbf{x}_C)}{\prod_{D \in \mathcal{D}} f_D(\mathbf{x}_D)}, \qquad \mathbf{x} \in \mathcal{X}, \tag{15}$$

where $\mathcal{D}$ is a multiset containing intersections between the cliques called separator sets; see Lauritzen (1996) and Appendix A for the definition of decomposability and separator sets.

2.3.1. Example 4

We recall that, for a normal distribution $\mathbf{W} = (W_i)_{i \in V}$ with invertible covariance matrix Σ, the precision matrix Σ^{-1} contains the conditional independences, or equivalently the graph structure, since, for $i, j \in V$,

$$W_i \perp\!\!\!\perp W_j | \mathbf{W}_{\setminus \{i,j\}} \Leftrightarrow \Sigma^{-1}_{ij} = 0.$$

3. Conditional independence for threshold exceedances

The notion of conditional independence has not been exploited in extreme value theory. In fact, for max-stable distributions it leads only to trivial probabilistic structures (Papastathopoulos and Strokorb, 2016). An exception is the max-linear model on a directed acyclic graph studied in Gissibl and Klüppelberg (2018) which does, however, not admit densities.

We therefore approach the problem from the perspective of threshold exceedances. Since the notion of independence is defined only on product spaces, the meaning of conditional independence is not straightforward for a multivariate Pareto distribution $\mathbf{Y} = (Y_i)_{i \in V}$, $V = \{1, \ldots, d\}$, with support on the L-shaped space $\mathcal{L} = \{\mathbf{x} \in \mathcal{E} : \|\mathbf{x}\|_\infty > 1\}$. In this section we show that there is nevertheless a natural definition of conditional independence for $\mathbf{Y}$. For this, we restrict $\mathbf{Y}$ to product spaces. For any $k \in V$, we consider the random vector $\mathbf{Y}^k$ defined as $\mathbf{Y}$ conditioned on the event that $\{Y_k > 1\}$. Clearly, $\mathbf{Y}^k$ has support on the product space $\mathcal{L}^k = \{\mathbf{x} \in \mathcal{L} : x_k > 1\}$ (Fig. 1) and it admits the density

$$f^k(\mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y})}{\int_{\mathcal{L}^k} f_{\mathbf{Y}}(\mathbf{y}) \mathrm{d}\mathbf{y}} = \lambda(\mathbf{y}), \qquad \mathbf{y} \in \mathcal{L}^k, \tag{16}$$

since $\int_{\mathcal{L}^k} f_{\mathbf{Y}}(\mathbf{y}) \mathrm{d}\mathbf{y} = 1/\Lambda(\mathbf{1})$ because of property (b) in Section 2.2. From expression (16) we see that the densities $f^1, \ldots, f^d$ coincide with λ on the intersection of their supports. Therefore there are no problems with lack of self-consistency as for instance in the model of Heffernan and Tawn (2004).

For any set $I \subset V$ with $k \in I$, the marginal $\mathbf{Y}_I^k$ has density

$$f_I^k(\mathbf{y}_I) = \int_{[0,\infty)^{d-|I|}} \lambda(\mathbf{y}) \mathrm{d}\mathbf{y}_{\setminus I} = \lambda_I(\mathbf{y}_I), \qquad \mathbf{y}_I \in \mathcal{L}_I^k,$$

which is homogeneous of order $-(|I|+1)$ on $\mathcal{L}_I^k = \{\mathbf{x}_I \in \mathcal{L}_I : x_k > 1\}$; see expression (5). This is, however, not so if $k \notin I$, since integration over $\mathbf{y}_{\setminus I}$ then includes y_k whose domain is $(1, \infty)$ in $\mathcal{L}^k$, and thus in general $f_I^k(\mathbf{y}_I) \neq \lambda_I(\mathbf{y}_I)$, $\mathbf{y}_I \in [0, \infty)^{|I|}$.

Definition 1. Suppose that $\mathbf{Y}$ is multivariate Pareto and admits a positive and continuous density $f_{\mathbf{Y}}$ on $\mathcal{L}$, and let $A, B, C \subset V$ be non-empty disjoint subsets whose union is $V = \{1, \ldots, d\}$. We say that $\mathbf{Y}_A$ is conditionally independent of $\mathbf{Y}_C$ given $\mathbf{Y}_B$ if

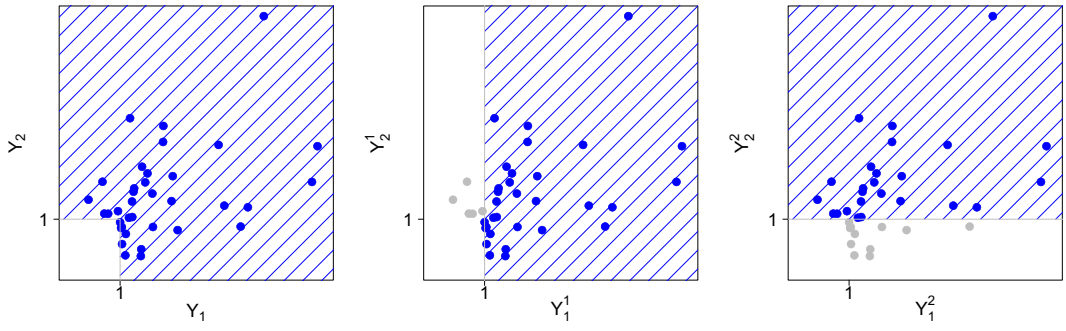


Fig. 1. The blue shaded areas are (a) the support $\mathcal{L}$ of $\mathbf{Y}$, and the supports (b) $\mathcal{L}^1$ of $\mathbf{Y}^1$ and (c) $\mathcal{L}^2$ of $\mathbf{Y}^2$: ●, samples of $\mathbf{Y}$

$$\forall\, k \in \{1, \ldots, d\} : \mathbf{Y}_A^k \perp\!\!\!\perp \mathbf{Y}_C^k | \mathbf{Y}_B^k. \tag{17}$$

In this case we write $\mathbf{Y}_A \perp_e \mathbf{Y}_C | \mathbf{Y}_B$.

In fact, this definition can be shown to be equivalent to a slightly weaker condition, and to a factorization of the exponent measure density λ.

Proposition 1. Let $f_{\mathbf{Y}}$ and the sets A, B and C be as in the above definition; then $\mathbf{Y}_A \perp_e \mathbf{Y}_C | \mathbf{Y}_B$ is equivalent to any of the following two conditions.

(a)

$$\exists\, k \in B : \quad \mathbf{Y}_A^k \perp\!\!\!\perp \mathbf{Y}_C^k | \mathbf{Y}_B^k. \tag{18}$$

(b) The density of the exponent measure factorizes as

$$\lambda(\mathbf{y}) = \frac{\lambda_{A \cup B}(\mathbf{y}_{A \cup B}) \lambda_{B \cup C}(\mathbf{y}_{B \cup C})}{\lambda_B(\mathbf{y}_B)}, \qquad \mathbf{y} \in \mathcal{L}. \tag{19}$$

A natural question is whether we can extend the definition of $\mathbf{Y}_A \perp_e \mathbf{Y}_C | \mathbf{Y}_B$ to the case where $B = \emptyset$, meaning that $\mathbf{Y}_A$ and $\mathbf{Y}_C$ are independent on $\mathcal{L}$. In terms of the original definition, that would mean that, for any $k \in V$, $f^k(\mathbf{y}) = f_A^k(\mathbf{y}_A) f_C^k(\mathbf{y}_C)$ for all $\mathbf{y} \in \mathcal{L}^k$. Without losing generality, suppose that $k \in A$; then $f_C^k(\mathbf{y}_C) = \lambda(\mathbf{y}_A, \mathbf{y}_C)/\lambda_A(\mathbf{y}_A)$ for any $\mathbf{y}_A \in \mathcal{L}_A^k$ and $\mathbf{y}_C \in [0, \infty)^{|C|}$. Therefore f_C^k would be homogeneous of order $-|C|$ and thus not integrable on $[0, \infty)^{|C|}$, which is a contradiction. In the next section we show that this property implies that all graphical models that are defined in terms of the conditional independence $\perp_e$ must be connected.

4. Graphical models for threshold exceedances

The notion of conditional independence enables us to define graphical models for threshold exceedances. As before, let $f_{\mathbf{Y}}$ be the positive and continuous density on $\mathcal{L}$ of a multivariate Pareto distribution $\mathbf{Y}$, proportional to the density λ of the exponent measure Λ, and homogeneous of order $-(d+1)$. Let $\mathcal{G} = (V, E)$ be an undirected graph with nodes $V = \{1, \ldots, d\}$ and edge set E. Similarly to the classical probabilistic graphical models that were described in Section 2.3, we say that $\mathbf{Y}$ satisfies the pairwise Markov property on $\mathcal{L}$ relative to $\mathcal{G}$ if

$$Y_i \perp_e Y_j | \mathbf{Y}_{\setminus \{i,j\}}, \qquad (i, j) \notin E, \tag{20}$$

i.e. Y_i and Y_j are conditionally independent in the sense of definition 1 given all other nodes whenever there is no edge between i and j in $\mathcal{G}$. By definition, this is equivalent to saying that $\mathbf{Y}^k$ satisfies the usual pairwise Markov property on $\mathcal{L}^k$ relative to $\mathcal{G}$ for all $k \in V$. The global Markov property for $\mathbf{Y}$ is defined similarly.

Definition 2. Let $\mathcal{G} = (V, E)$ be an undirected graph. If the multivariate Pareto distribution $\mathbf{Y}$ with positive and continuous density $f_{\mathbf{Y}}$ satisfies the pairwise Markov property (20) relative to $\mathcal{G}$, we call the distribution of $\mathbf{Y}$ an extremal graphical model with respect to $\mathcal{G}$.

For a decomposable graph $\mathcal{G}$ we obtain a factorization of the density $f_{\mathbf{Y}}$ that is similar to the classical Hammersley–Clifford theorem, showing that definition 1 of conditional independence is natural for multivariate Pareto distributions. Let $\mathcal{C}$ and $\mathcal{D}$ be the sequences of cliques and separators of $\mathcal{G}$ respectively, satisfying the running intersection property (44) in Appendix A.

Theorem 1. Let $\mathcal{G} = (V, E)$ be a decomposable graph and suppose that $\mathbf{Y}$ has a multivariate Pareto distribution with positive and continuous density $f_{\mathbf{Y}}$ on $\mathcal{L}$. Denote the corresponding

exponent measure and its density by Λ and λ respectively. Then the following results are equivalent.

(a) The distribution of $\mathbf{Y}$ satisfies the pairwise Markov property relative to $\mathcal{G}$.
(b) The distribution of $\mathbf{Y}$ satisfies the global Markov property relative to $\mathcal{G}$.
(c) The density $f_{\mathbf{Y}}$ factorizes according to $\mathcal{G}$, i.e.

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\Lambda(\mathbf{1})} \frac{\prod_{C \in \mathcal{C}} \lambda_C(\mathbf{y}_C)}{\prod_{D \in \mathcal{D}} \lambda_D(\mathbf{y}_D)}, \qquad \mathbf{y} \in \mathcal{L}, \tag{21}$$

where the marginals λ_I are positive, continuous and homogeneous of order $-(|I|+1)$ for any $I \subset V$.

In all cases, the graph $\mathcal{G}$ is necessarily connected.

Remark 1. Theorem 1 shows that only connected extremal graphical models can arise. This is related to the assumption of multivariate regular variation in expression (3) that implies asymptotic dependence between all components. Loosely speaking, unconnected components would correspond to asymptotically independent components.

Remark 2. If the graph $\mathcal{G}$ in theorem 1 is non-decomposable, it is expected that the density $f_{\mathbf{Y}}$ still factorizes into factors on the cliques of the graph. These factors can, however, no longer be identified with marginal densities, and it is an open problem whether they can be chosen to be homogeneous.

Since $\mathcal{L}$ is not a product space, unlike in the classical Hammersley–Clifford theorem for decomposable graphs in expression (15), the factors in the factorization of the density $f_{\mathbf{Y}}$ in expression (21) are not the marginals f_I but the marginals of the exponent measure density λ_I. It holds, however, that $f_I(\mathbf{y}_I) = \lambda_I(\mathbf{y}_I)/\Lambda_I(\mathbf{1})$ for all $\mathbf{y}_I \in \mathcal{L}_I \subset \{\mathbf{x}_I : \mathbf{x} \in \mathcal{L}\}$.

As a first application, theorem 1 enables us to analyse formally the conditional independences and graphical structures of models in the multivariate extreme value literature.

4.1. Examples

4.1.1. Example 5

One of the simplest examples of a graph is a chain, i.e.

$$E = \{\{1,2\},\{2,3\},\ldots,\{d-1,d\}\}.$$

Coles and Tawn (1991) proposed a model that factorizes with respect to this chain where all bivariate marginals are logistic (see example 1). This was extended to general bivariate marginals in Smith *et al.* (1997). More generally, in the study of extremes of stationary Markov chains the limiting objects are so-called tail chains. The tail chains induce multivariate Pareto distributions that can readily be seen to factorize with respect to a chain; see Smith (1992), Basrak and Segers (2009) and Janssen and Segers (2014).

4.1.2. Example 6

It turns out that many of the multivariate models in the literature do not have any conditional independences, i.e. their underlying graph is fully connected. For instance, this holds for the logistic multivariate Pareto distribution in example 1, the Dirichlet mixture model in Boldi and Davison (2007) and the pairwise beta distribution in Cooley *et al.* (2010).

4.1.3. Example 7

Similarly to Gaussian distributions, an appealing property of the Hüsler–Reiss model is its stability under taking marginals. Indeed, for any $I \subset V$ and $k \in I$ the marginal density of the exponent measure is

$$\lambda_I(\mathbf{y}_I) = \int_{[0,\infty)^{d-|I|}} \lambda_I(\mathbf{y}) \mathrm{d}\mathbf{y}_{\setminus I} = y_k^{-2} \prod_{i \in I \setminus \{k\}} y_i^{-1} \phi_{|I|-1}\{\tilde{\mathbf{y}}_{I\setminus\{k\}}; \Sigma_I^{(k)}\},$$

with the notation of example 2, where $\Sigma_I^{(k)}$ is the matrix in expression (10) induced by the submatrix Γ_I. Thus, $f_I(\mathbf{y}_I) = \lambda_I(\mathbf{y}_I)/\Lambda_I(\mathbf{1})$ is the density of the $|I|$-dimensional Hüsler–Reiss Pareto distribution with parameter matrix Γ_I.

By theorem 1, the density of a Hüsler–Reiss distribution that satisfies the pairwise Markov property relative to some decomposable graph $\mathcal{G}$ factorizes into lower dimensional Hüsler–Reiss distributions. The explicit formula is given in corollary 2 in Appendix C.

Theorem 1 can also be seen as a construction principle to build new classes of extreme value distributions in a modular way by combining lower dimensional marginals. The following corollary shows how a multivariate Pareto distribution can be defined that factorizes according to a desired underlying graphical structure. This is particularly useful in high dimensional problems to ensure model sparsity.

Corollary 1. Let $\mathcal{G}$ be a decomposable and connected graph and suppose that $\{\lambda_I : I \in \mathcal{C} \cup \mathcal{D}\}$ is a set of valid, positive and continuous exponent measure densities in the sense of properties (a) and (b) in Section 2.2. For $D \subset C$, $D \in \mathcal{D}$ and $C \in \mathcal{C}$, assume that they satisfy the consistency constraint

$$\lambda_D(\mathbf{y}_D) = \int_{[0,\infty)^{|C\setminus D|}} \lambda_C(\mathbf{y}_C) \mathrm{d}\mathbf{y}_{C\setminus D}. \tag{22}$$

The density of a valid d-dimensional exponent measure Λ is then given by

$$\lambda(\mathbf{y}) = \frac{\prod_{C \in \mathcal{C}} \lambda_C(\mathbf{y}_C)}{\prod_{D \in \mathcal{D}} \lambda_D(\mathbf{y}_D)}, \qquad \mathbf{y} \in \mathcal{L},$$

and the function $f_{\mathbf{Y}}(\mathbf{y}) = \lambda(\mathbf{y})/\Lambda(\mathbf{1})$, $\mathbf{y} \in \mathcal{L}$, is the density of a multivariate Pareto distribution satisfying the pairwise Markov property relative to $\mathcal{G}$.

4.2. Tree graphical models

A tree is a special case of a decomposable graphical model that is connected and has no cycles. All cliques are then of size 2 and the separators contain only one node. Let $\mathcal{T} = (V, E)$ be an undirected tree with nodes $V = \{1, \ldots, d\}$ and edge set $E \subset V \times V$. Suppose that $\mathbf{Y} = (Y_i)_{i \in V}$ follows a multivariate Pareto distribution on $\mathcal{L}$ with density $f_{\mathbf{Y}}$ that is an extremal graphical model with respect to the tree $\mathcal{T}$. Theorem 1 yields the factorization

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\Lambda(\mathbf{1})} \prod_{\{i,j\} \in E} \frac{\lambda_{ij}(y_i, y_j)}{y_i^{-2} y_j^{-2}} \prod_{i \in V} y_i^{-2}, \tag{23}$$

where $\lambda_{ij} = \lambda_{\{i,j\}}$ are the bivariate marginals of the exponent measure density λ corresponding to $\mathbf{Y}$. Formula (23) enables the extension of the modelling approach by Smith *et al.* (1997) that is

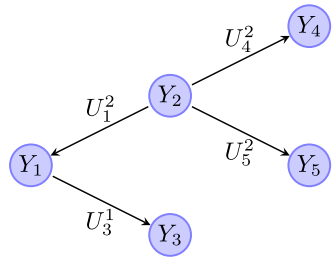


Fig. 2. A tree $\mathcal{T}^2$ rooted at node $k=2$ with the extremal functions in proposition 2 on the edges

described in example 5 from time series to general tree structures. Such tree models can represent more complex dependences and, moreover, are suitable beyond temporal data for multivariate or spatial applications.

Thanks to the relatively simple structure of a tree, more explicit results can be derived than for general graphical models. For this, we define a new directed tree $\mathcal{T}^k = (V, E^k)$ rooted at an arbitrary but fixed node $k \in V$. The edge set E^k consist of all edges $e \in E$ of the tree $\mathcal{T}$ pointing away from node k; see Fig. 2 for an example with $k=2$. For the resulting directed tree we define a set $(U_e)_{e\in E^k}$ of independent random variables, where for $e=(i,j)$, the distribution of $U_e = U_j^i$ is the extremal function of λ_{ij} at co-ordinate j, relative to co-ordinate i; see expression (12) in example 3 for the definition of extremal functions. The following stochastic representation of the random vectors $\mathbf{Y}^k$, $k \in V$, provides a better understanding of the stochastic structure of multivariate Pareto distributions factorizing on a tree, and it is the main ingredient for efficient simulation in Section 5.4.

Proposition 2. Let $\mathbf{Y}$ be a multivariate Pareto distribution with positive and continuous density on $\mathcal{L}$ that factorizes with respect to the tree $\mathcal{T}$. With the notation above, and for a standard Pareto distribution P, we have the joint stochastic representation for $\mathbf{Y}^k$ on $\mathcal{L}^k$

$$Y_i^k \overset{\mathrm{d}}{=} \begin{cases} P, & \text{for } i=k, \\ P \times \prod_{e\in \mathrm{ph}(ki)} U_e, & \text{for } i \in V\backslash\{k\}, \end{cases} \tag{24}$$

where $\mathrm{ph}(ki)$ denotes the set of edges on the unique path from node k to node i on the tree $\mathcal{T}^k$.

Remark 3. The same object as in representation (24) has been obtained in Segers (2019) as the limit of regularly varying random vectors that satisfy a Markov condition on a tree. In analogy to the tail chains in example 5, they term it a tail tree.

4.2.1. Example 8

Suppose that all bivariate marginals λ_{ij} for $\{i,j\} \in E$ of a tree Pareto model are of logistic type with parameter $\boldsymbol{\theta}_{ij} \in (0,1)$ as defined in example 1. This tree logistic model is a generalization of the chain logistic model that was considered in Coles and Tawn (1991). In this symmetric case, the extremal functions U_j^i and U_i^j have the same distribution with stochastic representation F/G, where F follows a Fréchet$(1/\boldsymbol{\theta}, c_{\boldsymbol{\theta}})$ distribution with scale parameter $c_\theta = \Gamma(1-\theta)^{-1}$ and $(G/c_\theta)^{-1/\theta}$ follows a gamma$(1-\boldsymbol{\theta}, 1)$ distribution, where we abbreviated $\boldsymbol{\theta} = \boldsymbol{\theta}_{ij}$ and Γ is the gamma function.

Similarly we can define a Hüsler–Reiss tree model, or use asymmetric models for λ_{ij} such as the Dirichlet model in Boldi and Davison (2007) for some or all of the edges $\{i,j\} \in E$. In asymmetric models, the extremal functions U_j^i and U_i^j in general have different distributions. We

refer to Section 4 in Dombry *et al.* (2016) for explicit formulae for extremal function distributions of commonly used model classes.

4.3. Hüsler–Reiss graphical models

In many ways, the class of Hüsler–Reiss distributions that were introduced in example 2 can be seen as the natural analogue of Gaussian distributions in the world of asymptotically dependent extremes. They are parameterized by the variogram of Gaussian distributions, and their statistical inference (Wadsworth and Tawn, 2014; Engelke *et al.*, 2015) and exact simulation (Dombry *et al.*, 2016) involves tools that are closely related to the corresponding methods for normal models.

Despite the similarities to Gaussian distributions, there are subtle but important differences that render analysis and statistical inference of Hüsler–Reiss distributions more difficult. To characterize conditional independence and graphical structures in these models, we first recall some results that are related to the original construction. The max-stable Hüsler–Reiss distribution has a stochastic representation as componentwise maxima

$$\mathbf{Z} = \max_{l \in \mathbb{N}} U_l \exp\{\mathbf{W}_l - \operatorname{diag}(\Sigma)/2\}, \tag{25}$$

where $\{U_l : l \in \mathbb{N}\}$ is a Poisson point process on $[0, \infty)$ with intensity $u^{-2}\mathrm{d}u$, and $\mathbf{W}_l$ are independent copies of a d-dimensional normal distribution $\mathbf{W}$ with zero mean and covariance matrix Σ. Subtracting $\mathbb{E}\{\exp(\mathbf{W})\} = \operatorname{diag}(\Sigma)/2$ in the exponent normalizes the marginals of $\mathbf{Z}$ to be standard Fréchet. Kabluchko *et al.* (2009) showed that the distribution of $\mathbf{Z}$ depends on only the strictly conditionally negative definite variogram matrix of $\mathbf{W}$,

$$\Gamma_{ij} = \mathbb{E}(W_i - W_j)^2, \qquad i, j \in V.$$

Importantly, this implies that the representation in equation (25) is not unique since any centred, possibly degenerate normal distribution $\mathbf{W}$ with variogram matrix Γ leads to the same max-stable Hüsler–Reiss distribution. Let

$$\mathcal{S}_\Gamma = \{\Sigma \in \mathbb{R}^{d \times d} \text{ covariance matrix } : \mathbf{1}\operatorname{diag}(\Sigma)^{\mathrm{T}} + \operatorname{diag}(\Sigma)\mathbf{1}^{\mathrm{T}} - 2\Sigma = \Gamma\} \tag{26}$$

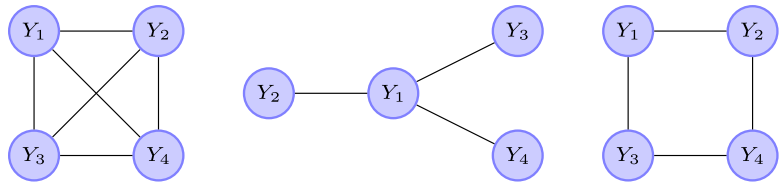
be the set of all covariance matrices that correspond to the same variogram matrix Γ; see Appendix B. The Hüsler–Reiss Pareto distribution $\mathbf{Y}$ that is associated with $\mathbf{Z}$ is defined by its density in example 2, which is also parameterized by Γ. We recall that, for a normal distribution $\mathbf{W}$ with invertible covariance matrix Σ, the precision matrix Σ^{-1} contains the conditional independences; see example 4. A first naive guess would be that the graph structure of $\mathbf{W}$ that is used in the construction of $\mathbf{Z}$ directly translates into the extremal graph structure of the Hüsler–Reiss Pareto distribution $\mathbf{Y}$. This is, however, not so.

4.3.1. Example 9

We consider three examples for $\mathbf{W}$ in the representation (25) with $d = 4$.

(a) Let W_i, $i = 1, \ldots, 4$, be independent standard normal distributions; then $\Sigma^{-1} = \operatorname{diag}(1, \ldots, 1)$ and Γ_{ij} equals 2 if $i \neq j$ and 0 otherwise. The graph underlying the distribution of $\mathbf{W}$ is the graph with four unconnected nodes. The graph of the corresponding Hüsler–Reiss Pareto distribution $\mathbf{Y}$ turns out to be the fully connected graph in Fig. 3(a).

(b) Consider the centred normal distribution $\mathbf{W}$ with precision matrix and variogram matrix



(a) (b) (c)

Fig. 3. Hüsler–Reiss graphical models described in example 9

$$\Sigma^{-1} = \begin{pmatrix} 12 & -4 & -4 & -1 \\ -4 & 2 & 1 & 0 \\ -4 & 1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 2 \\ 1 & 2 & 2 & 0 \end{pmatrix}$$

respectively. The Gaussian graphical model is the graph in Fig. 3(b) with an additional edge between nodes 2 and 3. In contrast, the corresponding Hüsler–Reiss model factorizes according to the graph in Fig. 3(b).

(c) Consider the centred normal distribution $\mathbf{W}$ with precision matrix and variogram matrix

$$\Sigma^{-1} = \begin{pmatrix} 2 & -0.5 & -0.5 & 0 \\ -0.5 & 1 & 0 & -0.5 \\ -0.5 & 0 & 1 & -0.5 \\ 0 & -0.5 & -0.5 & 1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 0 & 1.5 & 1.5 & 2 \\ 1.5 & 0 & 2 & 1.5 \\ 1.5 & 2 & 0 & 1.5 \\ 2 & 1.5 & 1.5 & 0 \end{pmatrix}$$

respectively. It can be checked that both the Gaussian and the Hüsler–Reiss graphical model are as in Fig. 3(c). Also note that this graph is not decomposable.

The above examples show that it is not possible simply to transfer the Gaussian graphical model of the covariance matrix Σ in the construction (25) to the extremal graphical structure of the corresponding Hüsler–Reiss Pareto distribution. This is not surprising since the covariance matrices in the set $\mathcal{S}_\Gamma$ can have very different graph structures, but all lead to the same Hüsler–Reiss graphical model. There is, however, a set of particular matrices that enable us to identify conditional independences and thus the graphical structure of a Hüsler–Reiss Pareto distribution. Recall the definition of $\Sigma^{(k)} \in \mathbb{R}^{(d-1)\times(d-1)}$ in expression (10). The same matrix including the kth row and column

$$\tilde{\Sigma}^{(k)} = \tfrac{1}{2}\{\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}\}_{i,j\in V} \in \mathbb{R}^{d\times d}, \tag{27}$$

is degenerate since the kth component has zero variance, but it is a valid choice in the construction (25), i.e. $\tilde{\Sigma}^{(k)} \in \mathcal{S}_\Gamma$, for any $k \in V$. Let $\mathbf{W}^k$ be a centred normal distribution with covariance matrix $\tilde{\Sigma}^{(k)}$ and note that $W^k_k = 0$ almost surely. For a random variable P with standard Pareto distribution, independent of $\mathbf{W}^k$, it can be seen that

$$\mathbf{Y}^k \overset{\mathrm{d}}{=} P \exp(\mathbf{W}^k - \Gamma_{\cdot k}/2), \tag{28}$$

by comparing the density of the right-hand side with expression (9) and noting that $\mathrm{diag}(\tilde{\Sigma}^{(k)}) = \Gamma_{\cdot k}$. This together with the original definition of conditional independence in expression (17) suggests that the matrices $\Sigma^{(k)}$ contain the graphical structure of a Hüsler–Reiss Pareto distribution.

We denote the precision matrix of $\Sigma^{(k)}$ by $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$. For notational convenience, the indices of the matrices $\Sigma^{(k)}$ and $\Theta^{(k)}$ range in $\{1, \ldots, d\}\backslash\{k\}$ instead of $\{1, \ldots, d-1\}$.

Lemma 1. For $k, k' \in V$, $k \neq k'$, the precision matrices $\Theta^{(k)}$ and $\Theta^{(k')}$ satisfy for $i, j \in V\backslash\{k'\}$

$$\begin{aligned} \Theta^{(k')}_{ij} &= \Theta^{(k)}_{ij}, && \text{if } i, j \neq k, \\ \Theta^{(k')}_{ik} &= -\sum_{l \neq k} \Theta^{(k)}_{il}, && \text{if } i \neq k, j = k, \\ \Theta^{(k')}_{kk} &= \sum_{l,m \neq k} \Theta^{(k)}_{lm}, && \text{if } i, j = k. \end{aligned}$$

Lemma 1 is of independent interest since it explains the link between the precision matrices $\Theta^{(k)}$ for different $k \in V$. The proof uses blockwise inversion of the precision matrices. This result is the crucial ingredient to characterize conditional independence in Hüsler–Reiss models.

Proposition 3. For a Hüsler–Reiss Pareto distribution $\mathbf{Y}$ with parameter matrix Γ, it holds for $i, j \in V$ with $i \neq j$, and for any $k \in V$, that

$$Y_i \perp_{\mathrm{e}} Y_j | \mathbf{Y}_{\backslash\{i,j\}} \Leftrightarrow \begin{cases} \Theta^{(k)}_{ij} = 0, & \text{if } i, j \neq k, \\ \sum_{l \neq k} \Theta^{(k)}_{lj} = 0, & \text{if } i = k, j \neq k, \\ \sum_{l \neq k} \Theta^{(k)}_{il} = 0, & \text{if } j = k, i \neq k. \end{cases} \tag{29}$$

For any $k \in V$, the single matrix $\Theta^{(k)}$ contains all information on conditional independence of $\mathbf{Y}$. Conditional independence concerning the kth component is encoded in the row and column sums of $\Theta^{(k)}$, and it might sometimes be easier to switch to another representation $\Theta^{(k')}$, $k' \neq k$, where it simply figures as a zero entry. In example 9 we can now easily determine the graphical model $\mathcal{G} = (V, E)$ for each of the three Hüsler–Reiss Pareto distributions. For a given Σ we first compute the matrix Γ as in expression (26), then we transform it by expression (10) to obtain $\Sigma^{(k)}$ for any $k \in V$ and use proposition 3 to decide whether $(i, j) \in E$ for all $i, j \in V$. These transformations are implemented in our R package `graphicalExtremes` (Engelke *et al.*, 2019).

4.3.2. Example 10

In spatial extreme value statistics, the finite dimensional distributions of the Brown–Resnick process (Kabluchko *et al.*, 2009) at locations $t_1, \ldots, t_d \in \mathbb{R}^D$ are Hüsler–Reiss distributed with variogram matrix $\Gamma_{ij} = \gamma(t_i - t_j)$, $i, j \in \{1, \ldots, d\}$, where γ is a variogram function on $\mathbb{R}^D$. The most commonly used model is the fractal variogram family $\gamma_\alpha(h) = \|h\|_2^\alpha$, for some $\alpha \in (0, 2]$. The corresponding d-variate Hüsler–Reiss distribution does not have conditional independences and its graph is thus fully connected. The only exception is the case of the original Brown–Resnick process in Brown and Resnick (1977) with $\alpha = 1$ and $D = 1$, where the corresponding graph is a chain as in example 5.

In this section, we have so far not required that the underlying graph $\mathcal{G}$ is decomposable. If this is so then, as shown in example 7, theorem 1 implies that the density of the Hüsler–Reiss

graphical model factorizes into lower dimensional Hüsler–Reiss densities; see corollary 2 in Appendix C.

5. Statistical inference for block graphs

5.1. Model construction

The notion of conditional independence and graphical models for multivariate Pareto distributions enables the construction of new statistical models with two major advantages. First, sparsity can be imposed on the model, which is a crucial ingredient for tractable and parsimonious models in higher dimensions. Second, under certain graphical structures, the model parameters can be estimated separately on lower dimensional subsets of the data.

We consider here, and throughout the rest of the paper, decomposable and connected graphs $\mathcal{G} = (V, E)$ with clique set $\mathcal{C}$ and separator set $\mathcal{D}$, where all separators in $\mathcal{D}$ are single nodes. Such graph structures with singleton separator sets are known as block graphs (see Harary (1963)) and have already been seen to have appealing properties for discrete data (Loh and Wainwright, 2013). In our case, they are a convenient way of restricting the model complexity to obtain a tractable class of extremal graphical models. In fact, corollary 1 provides a simple construction principle for multivariate Pareto distributions that factorize with respect to the block graph $\mathcal{G}$.

(a) For each clique $C \in \mathcal{C}$, choose possibly different parametric families of valid exponent measure densities $\{\lambda_C(\cdot; \theta_C) : \theta_C \in \Omega_C\}$ for suitable parameter spaces Ω_C. If $\mathcal{G}$ is a tree $\mathcal{T}$, then this reduces to choosing $d-1$ bivariate exponent measure densities λ_{ij}, for each $\{i, j\} \in E$; see example 3 for a general representation of such densities.
(b) Since all separator sets consist of a single node, the consistency constraint (22) is trivially fulfilled as a consequence of properties (a) and (b) in Section 2.2 and the fact that $\lambda_D(y_D) = y_D^{-2}$ for all $D \in \mathcal{D}$.
(c) For any fixed combination of parameters $\theta = (\theta_C)_{C \in \mathcal{C}} \in \Omega = \times_{C \in \mathcal{C}} \Omega_C$, the product of the normalized lower dimensional exponent measure densities,

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = \frac{1}{\Lambda(\mathbf{1}; \theta)} \prod_{C \in \mathcal{C}} \frac{\lambda_C(\mathbf{y}_C; \theta_C)}{\prod_{j \in C} y_j^{-2}} \prod_{i \in V} y_i^{-2}, \qquad \mathbf{y} \in \mathcal{L}, \tag{30}$$

defines a valid d-variate Pareto distribution factorizing according to the graph $\mathcal{G}$, which is a member of the parametric family parameterized by $\theta \in \Omega$. For a tree $\mathcal{T}$, this reduces to the density in equation (23).

Concrete examples for this construction are tree logistic or tree Hüsler–Reiss models as described in example 8, where all cliques have the same type of distributions. The above construction is much more flexible, as it allows us to use different distribution families for the different cliques. Moreover, some, or even all, of the cliques may be modelled by non-parametric methods; see Lafferty *et al.* (2012) for non-parametric tree models in the non-extreme case. In this direction, there is a line of research on kernel-based estimation of exponent measure densities (see de Carvalho and Davison (2014), Marcon *et al.* (2017) and Kiriliouk *et al.* (2018b)) that could be used as clique models. We shall not follow this approach here.

In the graphical models above, the dependence inside each clique is modelled directly, whereas dependence between components from different cliques is implicitly implied by the conditional independence structure of the graph. Even if all cliques are modelled with the same type of parametric family, the joint distribution (30) is typically not of this distribution type. For a tree logistic distribution, for instance, this can easily be seen by comparing its density (23) with that

of the d-variate logistic distribution in example 1. The latter has only one parameter governing the whole d-dimensional dependence structure, whereas the tree has $d-1$ logistic parameters $\{\theta_{ij}; \{i, j\} \in E\}$ and thus much higher flexibility.

An important exception is the family of Hüsler–Reiss distributions, which is stable under taking marginal distributions; see example 7. The following proposition shows that for a given graphical structure as above, if all cliques have Hüsler–Reiss distributions, then so has the full d-dimensional model. This is the converse of corollary 2 in Appendix C.

Proposition 4. Let $\mathcal{G} = (V, E)$ be a block graph as above, and suppose that, on each clique $C \in \mathcal{C}$, $\mathbf{Y}$ has a $|C|$-variate Hüsler–Reiss distribution with exponent measure density $\lambda_C(\cdot; \Gamma^{(C)})$ parameterized by a $(|C| \times |C|)$-dimensional variogram matrix $\Gamma^{(C)}$. Then there is a unique solution to the problem

$$\text{find a } (d \times d)\text{-dimensional variogram matrix } \Gamma,$$
$$\text{subject to} \begin{cases} \Gamma_{ij} = \Gamma^{(C)}_{ij}, & \text{for } i, j \in C \text{ and all } C \in \mathcal{C}, \\ \Theta^{(k)}_{ij} = 0, & \text{for all } k \in V, i, j \neq k \text{ and } (i, j) \notin E, \end{cases} \tag{31}$$

with the notation from proposition 3. The corresponding d-variate Hüsler–Reiss distribution factorizes according to the graph $\mathcal{G}$ into the lower dimensional Hüsler–Reiss densities on the cliques.

This is a matrix completion problem for variograms similar to what Dempster (1972) introduced for covariance matrices. In our case, the graph is decomposable and the above result relates to the marginal problem that was studied in Kellerer (1964) and Dawid and Lauritzen (1993). For Hüsler–Reiss marginals on block graphs we even see that the implied d-dimensional distribution is again Hüsler–Reiss. We give a direct constructive proof in Appendix F. This provides a method to construct high dimensional Hüsler–Reiss distributions out of many low dimensional distributions. The full d-variate Hüsler–Reiss model without any conditional independences has $d(d-1)/2$ parameters. A Hüsler–Reiss distribution as in proposition 4 that factorizes on a block graph with clique set $\mathcal{C}$ has only

$$\tfrac{1}{2} \textstyle\sum_{C \in \mathcal{C}} |C|(|C| - 1)$$

parameters, which can be much smaller than $d(d-1)/2$.

5.2. Estimation

Extremal graphical models can be used to build parsimonious statistical models for the tail of a multivariate random vector. In this section we discuss how the model parameters can be estimated efficiently by considering each clique distribution separately.

Let $\mathbf{X} = (X_j)_{j \in V}$, $V = \{1, \ldots, d\}$, be a random vector in the max-domain of attraction of the max-stable random vector $\mathbf{Z}$ as in expression (3), with marginal distribution X_j in the max-domain of attraction of a generalized extreme value distribution with shape parameter ξ_j, $j \in V$. Equivalently, there are a sequence of high thresholds $\mathbf{t}_u = (t_{u1}, \ldots, t_{ud})$ with t_{uj} tending to the upper end point of X_j as $u \to \infty$, and positive normalizing functions $\sigma_u = (\sigma_{u1}, \ldots, \sigma_{ud})$, such that the distribution of exceedances converges weakly:

$$\frac{\mathbf{X} - \mathbf{t}_u}{\sigma_u} \left| \left\| \frac{\mathbf{X}}{\mathbf{t}_u} \right\|_\infty > 1 \to \frac{\mathbf{Y}^\xi - 1}{\xi}, \qquad u \to \infty, \right. \tag{32}$$

where $\mathbf{Y}$ is the multivariate Pareto distribution that is associated with $\mathbf{Z}$. We assume that $\mathbf{Y}$ is in the model class of the previous section with density (30), and for now we suppose that the underlying graph $\mathcal{G} = (V, E)$ is known and fixed. The conditional density of $\mathbf{X} - \mathbf{t}_u$ given that $\|\mathbf{X}/\mathbf{t}_u\|_\infty > 1$ is then approximated by

$$f_{\mathbf{Y}}\left\{\left(1+\xi\frac{\mathbf{x}}{\sigma_u}\right)^{1/\xi};\theta\right\}\prod_{j\in V}\frac{1}{\sigma_{uj}}\left(1+\xi_j\frac{x_j}{\sigma_{uj}}\right)^{1/\xi_j-1}. \tag{33}$$

This density can be used to estimate jointly the marginal parameters (σ_{uj}, ξ_j), $j \in V$, and the dependence parameter vector $\theta = (\theta_C)_{C\in\mathcal{C}}$ of $f_{\mathbf{Y}}$.

In what follows we concentrate on estimation of the dependence, and we therefore assume that the marginal parameters are known or have been estimated separately. As described in Section 2.2, we can then normalize $\mathbf{X}$ to standard Pareto marginals, in which case $\xi_j = 1$, $t_{uj} = u$ and $\sigma_{uj} = u$ for all $j \in V$. We recover the standardized setting of expression (6) considered throughout the paper, where $\mathbf{X}/u$ given that $\|\mathbf{X}\|_\infty > u$ converges to $\mathbf{Y}$, whose likelihood is proportional as a function of θ to

$$f_{\mathbf{Y}}(\mathbf{y};\theta) \propto \frac{1}{Z_\theta}\prod_{C\in\mathcal{C}}\frac{\lambda_C(\mathbf{y}_C;\theta_C)}{\Lambda_C(\mathbf{1};\theta_C)}, \qquad Z_\theta = \frac{\Lambda(\mathbf{1};\theta)}{\prod_{C\in\mathcal{C}}\Lambda_C(\mathbf{1};\theta_C)}. \tag{34}$$

Direct maximization of the likelihood with contributions (34) for each data point is tedious since the normalizing constant Z_θ contains all parameters and does not factorize. Fortunately the class of block graphs has the property that we can estimate the parameters θ_C of each λ_C separately, without having to enforce the consistency constraints at the separator sets. In fact, we use the following observation. If $\mathbf{X}$ is in the domain of attraction of the family of multivariate Pareto distributions $\{f_{\mathbf{Y}}(\cdot;\theta) : \theta \in \Omega\}$, then, for a fixed clique $C \in \mathcal{C}$, the subvector $\mathbf{X}_C$ is in the domain of attraction of $\{f_C(\cdot;\theta_C) : \theta_C \in \Omega_C\}$, and the distribution of the normalized exceedance $\mathbf{X}_C/u \,|\, \|\mathbf{X}_C\|_\infty > u$ is approximated for large u by $\mathbf{Y}_C$ with density

$$f_C(\mathbf{y}_C;\theta_C) = \frac{\lambda_C(\mathbf{y}_C;\theta_C)}{\Lambda_C(\mathbf{1};\theta_C)}, \qquad \mathbf{y}_C \in \mathcal{L}_C; \tag{35}$$

see expression (7) in Section 2.2. We can therefore obtain an estimate of θ_C based only on data of the components in C, whose dimension is typically much smaller than the dimension d of the full graph. Estimating the cliques separately might in principle result in a loss of estimation efficiency compared with using the joint likelihood (34). The normalizing constant Z_θ does, however, not contain much information on the parameter θ and the maximum likelihood estimate by using $f_{\mathbf{Y}}(\mathbf{y};\theta)$ is generally very close to the estimate that is obtained by maximizing separate likelihoods based on expression (35). We discuss this point in the simulation study in Section 5.5.

In practice, some components of $\mathbf{X}$ might not have converged to the limiting distribution $\mathbf{Y}$. To avoid biased estimates of the dependence parameters θ_C, it has become a standard approach to apply censoring to the data; see Ledford and Tawn (1997) and Smith *et al.* (1997). For a data point $\mathbf{X}_C$ with $\|\mathbf{X}_C\|_\infty > u$ for a high threshold $u > 0$, define J to be the set of indices $j \in C$ such that $Y_j < 1$, i.e. $X_j < u$. For this data point we use the censored likelihood contribution

$$f_C^{\text{cens}}(\mathbf{y}_C;\theta_C) = \int_{[0,1]^{|J|}} f_C(\mathbf{y}_C;\theta_C)\mathrm{d}\mathbf{y}_J, \qquad \mathbf{y}_C \in \mathcal{L}_C, \tag{36}$$

which uses for all $j \in J$ only the information that this component of $\mathbf{Y}_C$ is smaller than 1, but not its exact value. For explicit forms of the censored likelihoods for many parametric models see Dombry *et al.* (2017) and Kiriliouk *et al.* (2018a).

For n independent data $\mathbf{y}^{(h)} \in \mathcal{L}$, $h = 1, \ldots, n$, of $\mathbf{X}/u | \|\mathbf{X}\|_\infty > u$, for each clique C we define $\hat{\theta}_C$ as the maximizer of the censored log-likelihood

$$L(\theta_C; \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}) = \sum_{\mathbf{y}^{(h)} \in \mathcal{L}_C} \log\{f_C^{\text{cens}}(\mathbf{y}_C^{(h)}; \theta_C)\}, \tag{37}$$

where $\mathcal{L}_C = \{\mathbf{y} \in \mathcal{L} : \exists\, j \in C \text{ such that } y_j > 1\}$, and each $\mathbf{y}_C^{(h)}$ has its own censoring set $J^{(h)} \subset C$.

Maximum likelihood estimation is only one possibility to infer the parameters θ_C on the basis of exceedances of $\mathbf{X}_C$ and the limiting distribution (35). Alternative methods use M-estimators (Einmahl *et al.*, 2012, 2016) or proper scoring rules (de Fondeville and Davison, 2018).

5.3. Model selection

Up to now we have assumed that a graphical structure $\mathcal{G}$ was *a priori* given and we analysed models that factorize with respect to this structure. In many applications the underlying graph structure is unknown and should be learned in a data-driven way. Theorem 1 implies that all extremal graphical structures are connected, and a simple and flexible class of connected graphs is the class of trees; see Section 4.1. It is thus natural first to build a suitable tree as a baseline model, and then to extend the tree by adding additional edges to obtain more complex graphs.

Since trees are a special case of general graphical models, there are specific methods to learn these simpler structures. The notion of a minimum spanning tree is crucial (Kruskal, 1956). Let $\mathcal{G}_0 = (V, E_0)$ be the fully connected graph on $V = \{1, \ldots, d\}$ with edge set $E_0 = \{(i, j) : i, j \in V\}$. Suppose that a positive weight $w_{ij} > 0$ is attached to each edge $(i, j) \in E_0$ of $\mathcal{G}_0$. This number can be seen as the length of the edge (i, j) or the distance between nodes i and j, and it is assumed that $w_{ij} = w_{ji}$ and $w_{ii} = 0$, $i, j \in V$. The minimum spanning tree is the tree $\mathcal{T}_{\text{mst}} = (V, E_{\text{mst}})$, with $E_{\text{mst}} \subset E_0$, that minimizes the sum of weights on that tree, i.e.

$$\mathcal{T}_{\text{mst}} = \arg\min_{\mathcal{T} = (V, E)} \sum_{(i,j) \in E} w_{ij}. \tag{38}$$

If all edges of $\mathcal{G}_0$ have distinct lengths, then $\mathcal{T}_{\text{mst}}$ is unique. This minimization problem can be solved efficiently by the greedy algorithms that were proposed in Kruskal (1956) or Prim (1957).

The weights w_{ij} determine the tree structure and should be chosen carefully. A common approach in graphical modelling is to search the conditional independence structure that maximizes the likelihood (see Cowell *et al.* (2006), chapter 11). Such a tree is also called a Chow–Liu tree (Chow and Liu, 1968). We fix a parametric family of bivariate Pareto distributions that is used for all pairs of nodes $\{f(\cdot; \theta_{ij}) : \theta_{ij} \in \Omega\}$. For n independent data $\mathbf{y}^{(h)}$, $h = 1, \ldots, n$, the maximal log-likelihood of a fixed tree within this parametric class is essentially the sum over the maximized clique log-likelihoods in equation (37) over all edges of this tree. To find the tree that maximizes the log-likelihood over all trees and all distributions in this parametric family, we therefore find the minimum spanning tree in equation (38) with weights

$$w_{ij} = -L(\hat{\theta}_{ij}; \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}) - 2 \sum_{y_i^{(h)} > 1} \log(y_i^{(h)}) - 2 \sum_{y_j^{(h)} > 1} \log(y_j^{(h)}), \tag{39}$$

where we include the censored marginal densities y_i^{-2} and y_j^{-2} in density (30) for the clique $\{i, j\}$, since now the edges are no longer fixed but parameters of the optimization. The resulting tree $\mathcal{T}_{\text{mst}}$ is the baseline model for the data. If the model fit is not satisfactory, it is possible to extend this tree to graphs with more complex structures by adding additional edges. The family of Hüsler–Reiss distributions is particularly appealing since the bivariate marginals

remain in the same class. We illustrate this model extension through a greedy forward selection in Section 5.5.

The different multivariate Pareto models can then be compared by the Akaike information criterion (Kiriliouk *et al.*, 2018a):

$$\mathrm{AIC} = 2p - 2L(\hat{\theta}; \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}), \tag{40}$$

where p is the number of parameters in the respective model, and the second term is twice the negative log-likelihood based on the censored version of expression (34), evaluated at the optimized parameters of each clique.

5.4. Exact simulation

Exact simulation of a max-stable random vector $\mathbf{Z}$ relies on the notion of extremal functions (Dombry and Éyi-Minko, 2013). The extremal function of $\mathbf{Z}$, or of its associated multivariate Pareto distribution $\mathbf{Y}$, relative to co-ordinate $k \in V$ is the d-dimensional random vector $\mathbf{U}^k$ with $U_k^k = 1$ such that the exponent measure density of $\mathbf{Z}$ can be written as

$$\lambda(\mathbf{y}) = y_k^{-(d+1)} f_{\mathbf{U}^k_{\setminus k}}(\mathbf{y}_{\setminus k}/y_k). \tag{41}$$

The distributions of the extremal functions $\mathbf{U}^k$, $k \in V$, for most commonly used models have explicit forms and are derived in section 4 of Dombry *et al.* (2016). Theorem 2 in Dombry *et al.* (2016) relates the distribution of the so-called spectral measure to these extremal functions. Together with the following representation of $\mathbf{Y}$, this enables simulation of multivariate Pareto distributions by rejection sampling. Recall that, for any $k \in V$, the random vector $\mathbf{Y}^k$ is defined as $\mathbf{Y}$ conditioned on the event that $\{Y_k > 1\}$.

Lemma 2. The distribution of the extremal function $\mathbf{U}^k$ of $\mathbf{Y}$ relative to co-ordinate $k \in V$ is given by the distribution of $\mathbf{Y}^k/Y_k^k$. Independently, let P be a standard Pareto random variable and T uniformly distributed on $\{1, \ldots, d\}$. We then have for any Borel set $A \subset \mathcal{L}$

$$\mathbb{P}(\mathbf{Y} \in A) = \mathbb{P}\left(\frac{P\mathbf{Y}^T}{\|\mathbf{Y}^T\|_1} \in A \,\middle|\, \frac{P\|\mathbf{Y}^T\|_\infty}{\|\mathbf{Y}^T\|_1} > 1 \right). \tag{42}$$

This representation yields the following simple algorithm for exact simulation of a multivariate Pareto distribution $\mathbf{Y}$ (*algorithm 1*); see also de Fondeville and Davison (2018).

Step 1: simulate a standard Pareto random variable P.
Step 2: simulate T uniformly on $\{1, \ldots, d\}$ and sample a realization of the extremal function $\mathbf{U}^{\mathrm{T}}$ relative to co-ordinate T.
Step 3: if $P\|\mathbf{U}^{\mathrm{T}}\|_\infty/\|\mathbf{U}^{\mathrm{T}}\|_1 > 1$, return $\mathbf{Y} = P\mathbf{U}^{\mathrm{T}}/\|\mathbf{U}^{\mathrm{T}}\|_1$ as a realization of the multivariate Pareto distribution.
Step 4: otherwise, reject the simulation and go back to step 1.

The complexity of this simulation algorithm as a function of the dimension d of the vector $\mathbf{Y}$ is driven by the number of times that we must sample from one of the extremal functions $\mathbf{U}^1, \ldots, \mathbf{U}^d$, since simulation of the variables P and T requires much less computational effort. Let $C_{\mathbf{Y}}(d)$ denote the number of extremal functions that must be simulated in algorithm 1. The random variable $C_{\mathbf{Y}}(d)$ follows a geometric distribution and from equation (50) in the proof of lemma 2 in Appendix F its expectation is

$$\mathbb{E}\{C_{\mathbf{Y}}(d)\} = d/\Lambda(\mathbf{1}) \in [1, d].$$

The expected complexity therefore depends on both the dimension and the strength of extremal dependence in $\mathbf{Y}$. Weak dependence implies a large coefficient $\Lambda(\mathbf{1})$ closer to d and therefore reduces the computational effort that is required for exact simulation. The simulation of multivariate Pareto distributions is in general computationally easier than for the associated max-stable distribution $\mathbf{Z}$. Indeed, exact simulation of the max-stable distribution is also based on samples from a mixture of the $\mathbf{U}^1, \ldots, \mathbf{U}^d$, and the fastest algorithm in Dombry *et al.* (2016) has expected complexity $\mathbb{E}\{C_{\mathbf{Z}}(d)\} = d$; see also Dieker and Mikosch (2015) and Oesting *et al.* (2018) for other exact simulation methods.

The complexity measures $C_{\mathbf{Y}}(d)$ and $C_{\mathbf{Z}}(d)$ consider only the number of extremal functions that are required for one exact simulation of $\mathbf{Y}$ and $\mathbf{Z}$ respectively. The computational effort of sampling $\mathbf{U}^k$ can, however, be significantly lower if $\mathbf{Y}$ has a sparse structure. If $\mathbf{Y}$ factorizes according to a graph, then, by the definition 1 of conditional independence, the $\mathbf{Y}^1, \ldots, \mathbf{Y}^d$ inherit the sparsity of this graph structure. This is particularly important in the case of trees and for Hüsler–Reiss distributions, as shown in the examples below. It is important to note that more efficient simulation of the extremal functions speeds up exact simulation of the multivariate Pareto distribution $\mathbf{Y}$, but also of the max-stable distribution $\mathbf{Z}$.

5.4.1. *Example 11*

Suppose that $\mathbf{Y}$ factorizes according to a tree $\mathcal{T} = (V, E)$. It follows from proposition 2 and lemma 2 that the extremal function $\mathbf{U}^k$ relative to co-ordinate $k \in V$ is

$$\mathbf{Y}^k / Y_k^k \stackrel{\mathrm{d}}{=} \left(\prod_{e \in \mathrm{ph}(ki)} U_e \right)_{i \in V}.$$

For exact simulation of $\mathbf{Y}$ it therefore suffices to simulate the univariate random variables U_e. This is feasible even in very large dimensions.

5.4.2. *Example 12*

If $\mathbf{Y}$ has a Hüsler–Reiss distribution that factorizes on the graph $\mathcal{G} = (V, E)$, then it follows from expression (28) that the extremal function $\mathbf{U}^k$ relative to co-ordinate $k \in V$ is

$$\mathbf{Y}^k / Y_k^k \stackrel{\mathrm{d}}{=} \exp(\mathbf{W}^k - \Gamma_{\cdot k}/2),$$

where $\mathbf{W}^k$ is a centred normal distribution with covariance matrix $\tilde{\Sigma}^{(k)}$ in expression (27); see also proposition 4 in Dombry *et al.* (2016). The normal distribution $\mathbf{W}^k_{\setminus k}$ factorizes in the classical sense on the subgraph $\mathcal{G}_{\setminus k}$, and efficient simulation algorithms exist if the graph is sparse (e.g. Rue and Held (2005)).

The exact simulation algorithms for both multivariate Pareto and max-stable distributions are implemented in our R package `graphicalExtremes` (Engelke *et al.*, 2019).

5.5. Simulation study

We assess the efficiency of parameter estimation and model selection in the framework of graphical models for extremes described in the previous sections. We fix a dimension d of variables or nodes $V = \{1, \ldots, d\}$ and a block graph $\mathcal{G} = (V, E)$ as in Section 5.1. In this study we simulate samples directly from the limiting distribution $\mathbf{Y}$ by using the exact algorithm 1, but we use the censored estimator since this is common practice in applications.

We first choose $d = 5$ and let $\mathcal{G}$ be the undirected version of the tree in Fig. 2. We simulate $n \in \{100, 200\}$ samples $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$ of a Hüsler–Reiss distribution with parameter matrix Γ that

factorizes according to $\mathcal{G}$. The entries of Γ need to be specified only on the submatrices $\Gamma^{(C)}$ for all cliques $C \in \mathcal{C}$ of $\mathcal{G}$, since the solution to the matrix completion problem in proposition 4 then yields the unique variogram matrix Γ. In this simulation we set

$$\Gamma = \begin{pmatrix} 0 & \Gamma_{12} & \Gamma_{13} & \Gamma_{12}+\Gamma_{24} & \Gamma_{12}+\Gamma_{25} \\ & 0 & \Gamma_{12}+\Gamma_{13} & \Gamma_{24} & \Gamma_{25} \\ & & 0 & \Gamma_{13}+\Gamma_{12}+\Gamma_{24} & \Gamma_{13}+\Gamma_{12}+\Gamma_{25} \\ & & & 0 & \Gamma_{24}+\Gamma_{25} \\ & & & & 0 \end{pmatrix}, \tag{43}$$

where we show only the upper triangular part of Γ because of symmetry. We specified the four parameters Γ_{ij} for $(i, j) \in E$, $i < j$, to the values $\Gamma_{12} = 1, \Gamma_{13} = 2, \Gamma_{24} = 1$ and $\Gamma_{25} = 2$, and the rest of the matrix is implied by the graph structure.

In this dimension we can still maximize the censored version of the joint likelihood (34) to obtain an estimate $\hat{\Gamma}_{ij}^{\text{joint}}$, $\{i, j\} \in E$, of the parameters corresponding to the four edges of the tree. We also obtain estimates $\hat{\Gamma}_{ij}$, $\{i, j\} \in E$, of the parameters of each clique separately by maximizing the censored clique likelihood (37). In both cases, the four estimated parameters yield estimates $\hat{\Gamma}^{\text{joint}}$ and $\hat{\Gamma}$ of the whole variogram matrix Γ through the graph structure. We repeat the simulation and estimation 200 times and compare the efficiency of both approaches in Fig. 4, displaying only the four free parameters that have actually been estimated.

The difference in efficiency between the joint and clique likelihoods seems to be small or even negligible. This is due to two reasons. For non-censored points the two likelihoods differ by only the normalizing constant Z_θ. Since this constant measures only the global strength of dependence and does not depend on the data, it seems not very sensitive to changes in the parameter θ. The second difference between the two approaches is that they use slightly different data. Consider a clique $C \in \mathcal{C}$ and the corresponding model parameter θ_C. The joint likelihood uses all data $\mathbf{Y}$ in the space $\mathcal{L} = \{\mathbf{y} \in \mathcal{E} : \exists\, j \in V \text{ such that } y_j > 1\}$, but censors all components with $y_j \leqslant 1$. In contrast, the clique likelihood uses the marginals $\mathbf{Y}_C$ of all data $\mathbf{Y}$ in $\mathcal{L}_C = \{\mathbf{y} \in \mathcal{L} : \exists\, j \in C \text{ such that } y_j > 1\}$. Consequently, the additional data that are used in the joint likelihood are in $\mathcal{L} \setminus \mathcal{L}_C = \{\mathbf{y} \in \mathcal{L} : y_j \leqslant 1 \text{ for all } j \in C\}$. But the contribution to the joint likelihood of data in this set with regard to the parameter θ_C is completely censored and does therefore not add significant additional information. These two considerations underline that estimating the parameters for each clique separately does not result in significant losses of efficiency. This is one of the main

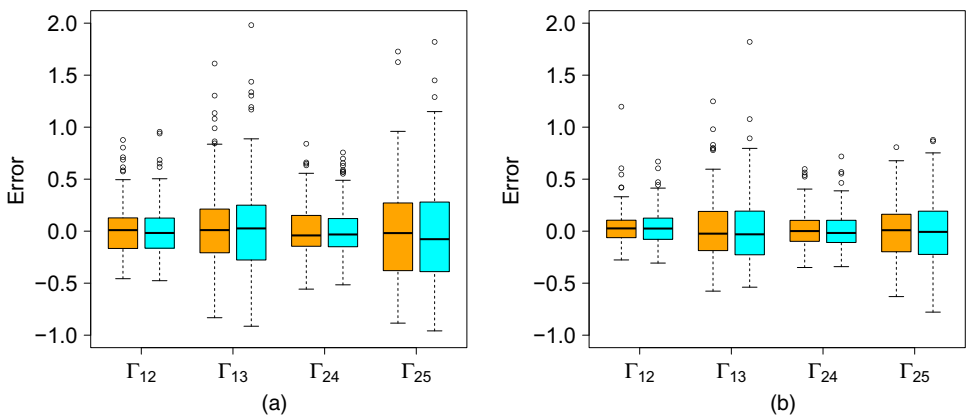


Fig. 4. Boxplots of errors of the four parameter estimates of the Hüsler–Reiss tree model in expression (43) based on joint (■) and clique likelihood (■) with sample size (a) $n = 100$ and (b) $n = 200$

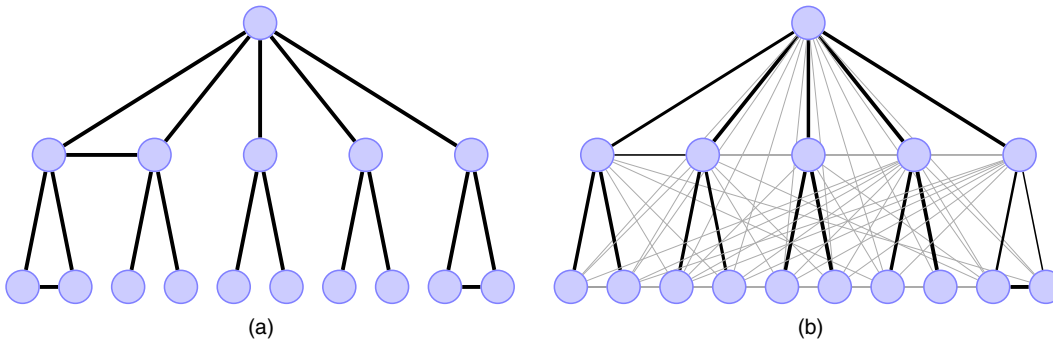


Fig. 5. (a) True underlying graphical structure and (b) the estimated structure in the second experiment, where the line thickness indicates the number of times that the edge has been selected

advantages of graphical models, namely that the distribution is defined locally by the cliques and extends globally by the conditional independence structure. In terms of computational aspects, the joint likelihood becomes infeasible even in moderate dimensions, whereas the clique likelihood is applicable in high dimensions as long as the cliques have sufficiently small sizes. Moreover, the computations for different cliques can be easily parallelized.

For the second experiment we take $d = 16$ and let $\mathcal{G}$ be the graph in Fig. 5(a), which is not a tree. We simulate $n = 100$ samples of a Hüsler–Reiss distribution with parameter matrix Γ that factorizes according to $\mathcal{G}$. The parameters of the $p = 18$ edges are independently sampled from a uniform distribution on $(0.5, 1)$, under the constraint that Γ is conditionally negative definite on cliques with three nodes. We illustrate how we can choose the best graphical model, where we restrict to block graphs as in Section 5.1 with cliques of sizes 2 and 3. We first construct the minimum spanning tree as described in Section 5.3 within the class of Hüsler–Reiss distributions. The estimated edge set of this tree is denoted by E_1. The 15 parameter estimates $\hat{\Gamma}_{ij}$, $\{i, j\} \in E_1$ that are obtained by fitting the clique likelihoods of each clique of the tree yield a unique estimate $\hat{\Gamma}$ of the $(d \times d)$-dimensional variogram matrix; see proposition 4. This tree model does not contain all edges of the true underlying graph. We therefore perform a greedy forward selection to add additional edges and to improve the model. In each step, we define an enlarged edge set $E_{m+1} = E_m \cup \{i, j\}, m = 1, 2, \ldots$, restricting to those edges $\{i, j\}, i, j \in V$, that still yield a block graph with cliques of maximal size 3. We continue this process until no more edges can be added in this way. For the same parameter matrix Γ, we repeat the simulation and model selection 100 times. Fig. 5(b) shows the graph with the selected edges, where the line width of each edge indicates the number of times that it has been selected among the first 18 edges. It can be seen that the graph structure is generally very well identified. For each model and each repetition we also compute the resulting AIC according to expression (40). The proportion of times that the model with $\{15, \ldots, 20\}$ edges has the smallest AIC are $\{0.01, 0.11, 0.23, 0.39, 0.23, 0.03\}$. Even though AIC is a criterion that was built for model estimation and not for identification (see Arlot and Celisse (2010)), it seems to be well suited to select the correct degree of sparsity for this extremal graphical model.

6. Application

We illustrate the applicability of extremal graphical models by the example of river discharges in the upper Danube basin: a region that is prone to serious flooding. The data are provided by the Bavarian Environmental Agency (`http://www.gkd.bayern.de`) and we use $d = 31$ gauging stations with 50 years of common daily data from 1960 to 2009. The tree that is induced by the

physical flow connections at these stations is shown in Fig. 6(a), where the path $10 \to 9 \to \ldots \to 1$ is on the Danube and the other branches are tributaries. The spatial extremal dependence structure of this data set has been studied in Asadi *et al.* (2015) and we follow their preprocessing steps to make the results comparable. Out of all daily data only the three months June, July and August are considered since the most severe floods occur in this period and are caused by heavy summer rain (Böhm and Wetzel, 2006). The $50 \times 92 = 4600$ observations in these months are declustered in time to remove temporal dependence and to match slightly shifted peak flows at different locations. We refer to Asadi *et al.* (2015) for more details on the data, the declustering method and exploratory analysis concerning stationarity and asymptotic dependence; see also Keef *et al.* (2009, 2013) for other approaches to flood risk assessment.

The declustering yields $N = 428$ supposedly independent events $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^d$. The univariate marginal distributions of these data were estimated in Asadi *et al.* (2015) by a regionalized extreme value model. We focus on estimation of the extremal dependence and normalize the data empirically to standard Pareto marginals. This still guarantees consistent inference of the dependence parameters (e.g. Genest *et al.* (1995) and Joe (2015)). We obtain $n = 117$ approximate samples of $\mathbf{Y}$ by $\mathbf{y}^{(h)} = \mathbf{x}^{(h)}/u$ for all observations with $\|\mathbf{x}^{(h)}\|_\infty > u$, where we choose the threshold u as the 90% quantile of the marginal Pareto distribution.

The max-stable Brown–Resnick model in Asadi *et al.* (2015) corresponds to a parametric family of Hüsler–Reiss Pareto distributions $\{f_{\mathbf{Y}}(\cdot;\theta) : \theta \in \Omega\}$ at the 31 gauging stations. The dependence model is tailor made for this particular application to river extremes and uses several covariates such as distance on the river network, catchment sizes and altitudes. In terms of our new notion of extremal graphical models it is readily checked by using the results of proposition 3 that for any parameter value $\theta \in \Omega$ their model does not exhibit conditional independences.

We propose a different Hüsler–Reiss model that factorizes according to a sparse graph and does not require any domain knowledge or additional covariates. In fact, we propose a sequence of models

$$M^{(l)} = \{f_{\mathbf{Y}}(\cdot;\theta^{(l)}) : \theta^{(l)} \in \Omega^{(l)}\}, \qquad l = 1, \ldots, L,$$

where $\theta^{(l)} = (\theta_C^{(l)})_{C \in \mathcal{C}^{(l)}}$, and $\mathcal{C}^{(l)}$ is the set of all cliques of the lth extremal graphical model $\mathcal{G}^{(l)}$ according to which the model family $M^{(l)}$ factorizes. As the simplest model we take $\mathcal{G}^{(1)}$ to be the minimum spanning tree within the family of Hüsler–Reiss models as described in Section 5.3. Similarly to the simulation study in Section 5.5, we obtain $\mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(L)}$ by successively adding edges to the tree $\mathcal{G}^{(1)}$ in a greedy way while restricting the model class to block graphs with cliques of size at most 3. The estimated tree $\mathcal{G}^{(1)}$ is shown in Fig. 9(a) in Appendix D. It is very similar to the tree in Fig. 6 that corresponds to the tree that is induced by the flow connections of the river network. There are, however, differences, and it is important to note that the flow connection tree is not necessarily the optimal tree structure in terms of extreme river discharges. Appendix D also contains a sensitivity analysis of the tree structure for various thresholds u, and a comparison with a Gaussian tree model fitted to non-extremal data.

Fig. 7 shows the AIC-values for the models $M^{(1)}, \ldots, M^{(L)}$. Forward selection is a greedy approach and it does not guarantee finding the optimal graph. We therefore also initialize the forward selection with the simplest model $\mathcal{G}^{(1)}$ being the flow connection tree in Fig. 6(a). This tree must have a larger AIC than the minimum spanning tree but, interestingly, Fig. 7(a) shows that by adding additional edges the optimal AIC is better than the previous optimal AIC. In this particular case, we thus choose the graph that is initiated with the flow connection tree with nine additional edges. In general, a tree structure appears to be too simple for this application. The reason is that only part of the extremal dependence of discharges at different locations can be explained by flow connections. Additional dependence may arise even between

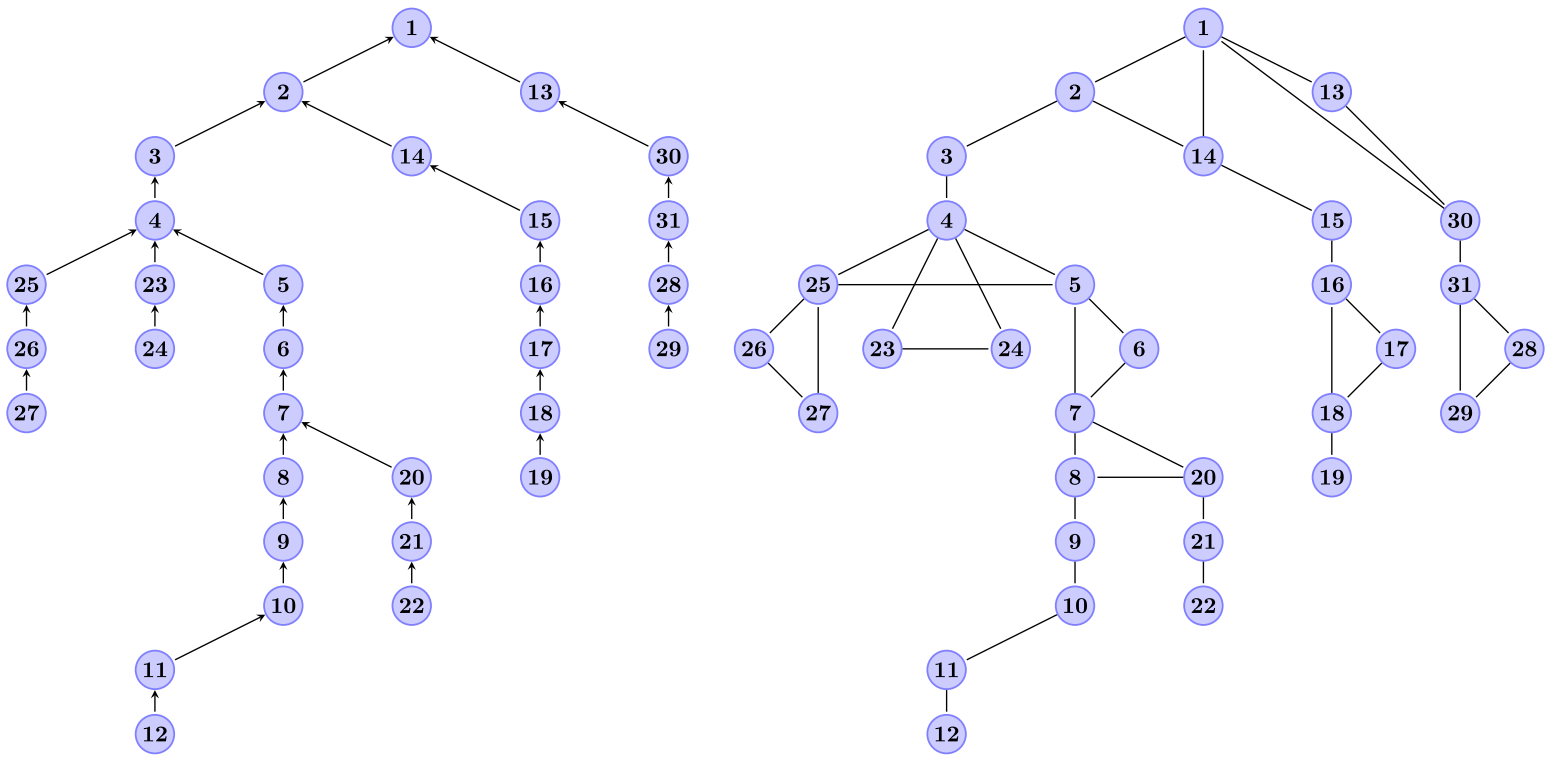


Fig. 6. (a) Tree induced by flow connections for the 31 stations in the upper Danube basin and (b) the estimated graph with the optimal AIC

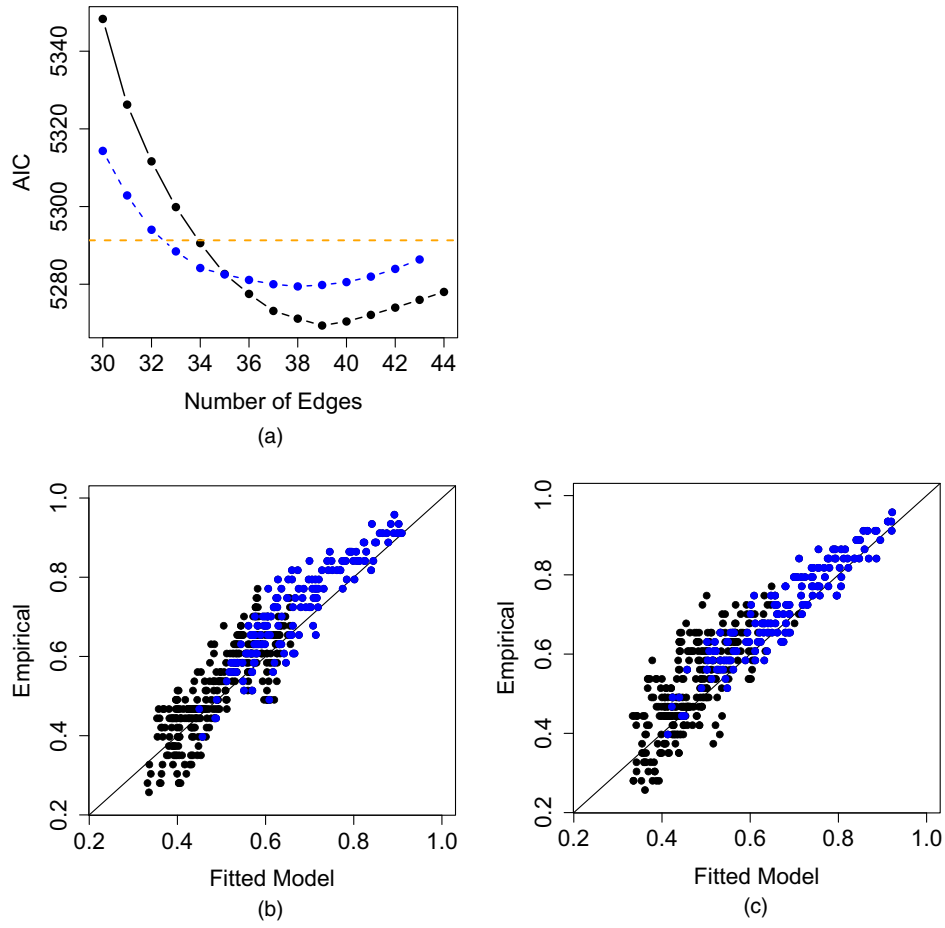


Fig. 7. (a) AIC-values for extremal graphical models for the Danube data set with an increasing number of edges, starting from the minimum spanning tree (●) and the flow connection tree (●) (- - - - -, AIC of the spatial model in Asadi *et al.* (2015)), and empirically estimated χ-coefficients against those implied by (b) the fitted spatial and (c) graphical model minimizing the AIC: ●, flow connected stations

flow-unconnected locations due to proximity of their catchments that are affected by the same spatial precipitation events. Asadi *et al.* (2015) modelled this explicitly through a variogram with two parts: one for the dependence on the river network and one for the spatial meteorological dependence. The nine additional edges of the graphical model in Fig. 6(b), which minimizes AIC, partly improve the model in terms of this spatial dependence between flow-unconnected stations but also strengthen it between some flow-connected locations. This best graphical model has 39 edges and an AIC of 5269.43. It significantly outperforms the simpler tree models with 30 edges and the spatial model of Asadi *et al.* (2015), which has only six parameters but an AIC of 5291.34, which is indicated by the broken orange line in Fig. 7(a).

A popular summary statistic for extremal dependence between Y_i and Y_j, $i, j \in V$, is the tail correlation (see Coles *et al.* (1999)), which can be expressed as $\chi_{ij} = 2 - \Lambda_{ij}(1, 1)$. Figs 7(b) and 7(c) compare empirical estimates of these statistics for all pairs of stations with those implied by the fitted models. In terms of this bivariate summary, both models seem to fit the data well, even though the graphical model seems to be slightly less biased than the spatial model. There are also

versions of χ that assess how a model captures the higher order extremal dependence structure. In Fig. 11 in Appendix E we compare the trivariate empirical χ-coefficients with those implied from the fitted spatial and graphical model. Both models fit well the trivariate dependence, again with a slightly lower bias of the graphical model.

In this application we have considered only block graphs, which are particularly convenient in terms of statistical inference as seen in the previous sections. In general it should be assessed whether this sparse model class is justified for the data. In our case, the bivariate and trivariate χ-coefficients indicate that block graphs are sufficiently flexible to capture the extremal dependence structure of the river data. This is further supported by the fact that the AIC-curve in Fig. 7(a) attains its minimum even before the maximal number of edges has been added in this model class. It is an important question for future research how extremal graphical models with more complicated structures can be estimated.

7. Discussion

The conditional independence relationship $\perp_e$ that is introduced in this paper is natural for a multivariate Pareto distribution $\mathbf{Y}$ as it explains the factorization of its density $f_{\mathbf{Y}}$ into lower dimensional marginals (see theorem 1). This establishes a link of extreme value statistics to the broad field of graphical models, and it opens the door to define sparsity and to perform structure learning for tail distributions. In this work we have studied the probabilistic structure and statistical inference for some important models, with the main purpose of modelling the extremal dependence structure. Many subsequent research directions are possible. Directed acyclic graphs as in Gissibl and Klüppelberg (2018) for max-linear models may be formulated in our setting and would yield factorizations that are different from those for undirected graphs, and this would form the basis for extending work on causal inference for extremes (Naveau *et al.*, 2018; Mhalla *et al.*, 2020; Gnecco *et al.*, 2019) to continuous extreme value distributions. The models in this paper are well suited for asymptotic dependence. Another line of research focuses on multivariate tail models under asymptotic independence (Ledford and Tawn, 1997; Heffernan and Tawn, 2004; Wadsworth *et al.*, 2017). Conditional independence and graphical models have not been studied in this framework, except for the special case of Markov chains (Kulik and Soulier, 2015; Papastathopoulos *et al.*, 2017).

Conditional independence for $\mathbf{Y}$ does not carry over to factorization of the density of the associated max-stable distribution $\mathbf{Z}$. By proposition 1, the conditional independence relationship $\perp_e$ does, however, imply the factorization of the exponent measure density λ of $\mathbf{Z}$, which is the key object in simulation (Dombry *et al.*, 2016) and full likelihood estimation (Thibaud *et al.*, 2016; Dombry *et al.*, 2017; Huser *et al.*, 2019) of max-stable processes. Thus, sparsity for multivariate Pareto distributions also facilitates inferential tasks for max-stable distributions: a fact that has been briefly discussed for simulation in Section 5.4 but which deserves further investigation.

The application to flood risk assessment is just one illustrative example. Unlike spatial models, extremal graphical models can be applied to multivariate problems without domain knowledge, as for instance in financial or insurance applications. The ability to learn underlying structures in a data-driven way has also great practical potential for exploratory analysis and data visualization. In on-going research we investigate efficient learning of extremal tree structures and, in the case of Hüsler–Reiss distributions, of more general graphs based on l_1-regularization.

Acknowledgements

We thank Robin J. Evans and Nicola Gnecco for helpful discussions. We are grateful to the

editorial team and the referees for knowledgeable comments that improved the paper. Financial support by the Swiss National Science Foundation (S. Engelke) and by the Berrow Foundation (A. S. Hitz) is gratefully acknowledged. The paper was completed while S. Engelke was a visitor at the Department of Statistical Sciences, University of Toronto.

Appendix A: Definitions for graphical models

Let $\mathcal{G} = (V, E)$ be an undirected graph with node set $V = \{1, \ldots, d\}$ and edge set $E \subset V \times V$; see Section 2.3. We define the notion decompositions and decomposability for the graph $\mathcal{G}$ (see Lauritzen (1996), definition 2.1).

Definition 3. A triplet (A, B, C) of disjoints subsets of V is said to form a decomposition of $\mathcal{G}$ into the components $\mathcal{G}_{A\cup B}$ and $\mathcal{G}_{B\cup C}$ if $V = A \cup B \cup C$ and

(a) B separates A from C (i.e. every path from A to C intersects B) and
(b) B is a complete subset.

The decomposition is called proper if A and C are both non-empty. A graph $\mathcal{G}$ is decomposable if it is complete or if there is a proper decomposition (A, B, C) into decomposable subgraphs $\mathcal{G}_{A\cup B}$ and $\mathcal{G}_{B\cup C}$. Decomposable graphs are also known as triangulated or chordal graphs.

For instance, $(\{1, 2, 3, 4, 5\}, \{4, 5\}, \{4, 5, 6\})$ is a proper decomposition of the decomposable graph in Fig. 8.

For a connected, decomposable graph $\mathcal{G}$, we can order the set of the cliques $\mathcal{C} = \{C_1, \ldots, C_m\}$ such that, for all $i = 2, \ldots, m$,

$$D_i := C_i \cap \bigcup_{j=1}^{i-1} C_j \subset C_k \qquad \text{for some } k < i, \tag{44}$$

which is a condition called the running intersection property; see Lauritzen (1996), chapter 2, and Green and Thomas (2013). The sets D_i, $i = 2, \ldots, m$, are called separators of the graph, and both $\mathcal{C}$ and the collection of separators $\mathcal{D} = \{D_2, \ldots, D_m\}$ are uniquely determined up to different orderings. The separators may not all be distinct, and we say that $\mathcal{D}$ is a multiset. A possible enumeration of cliques and separators for the graph in Fig. 8 that satisfies the running intersection property is

$$\mathcal{C} = (\{1, 2\}, \{2, 3, 4, 5\}, \{4, 5, 6\}), \quad \mathcal{D} = (\{2\}, \{4, 5\}).$$

From expression (44) we note that the clique C_m intersects the other cliques only in D_m. Consider the connected decomposable subgraph $\mathcal{G}_{m-1}$ of $\mathcal{G}$ with node set $V_{m-1} = V \backslash (C_m \backslash D_m)$ and corresponding induced edge set. Property (44) then holds for $\mathcal{G}_{m-1}$, which has one clique fewer. Continuing this process, we note that each C_j intersects the subgraph $\mathcal{G}_j$ only in D_j, $j = 2, \ldots, m$, and $\mathcal{G}_1$ with nodes $V_1 = C_1$ is complete.

Appendix B: Link between variogram and covariance matrices

For $k \in V = \{1, \ldots, d\}$, we denote by $\mathcal{P}_{d-1}^k$ the set of all strictly positive definite covariance matrices $\Sigma^{(k)} \subset \mathbb{R}^{(d-1)\times(d-1)}$ indexed by $V \backslash \{k\}$. In contrast, the space of strictly conditionally negative definite $d \times d$ matrices is denoted by

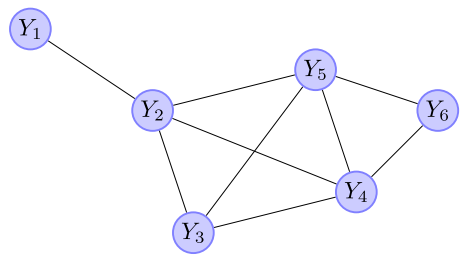


Fig. 8. Decomposable graph with set of nodes $V = \{1, \ldots, 6\}$: the cliques of the graph are $\{1, 2\}$, $\{2, 3, 4, 5\}$ and $\{4, 5, 6\}$; the separators are $\{2\}$ and $\{4, 5\}$

$$\mathcal{D}_d = \left\{ \Gamma \in [0,\infty)^{d\times d} : \mathbf{a}^{\mathrm{T}}\Gamma\mathbf{a} < 0 \text{ for all } \mathbf{a} \in \mathbb{R}^d \backslash \{0\} \text{ with } \sum_{i\in V} a_i = 0, \Gamma_{ii} = 0, \Gamma_{ij} = \Gamma_{ji} \text{ for all } i, j \in V \right\}.$$

Lemma 3. For any $k \in V$, there is a bijection $\varphi_k : \mathcal{D}_d \to \mathcal{P}^k_{d-1}$ given by

$$\begin{aligned} \varphi_k &: \Gamma \mapsto \tfrac{1}{2}\{\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}\}_{i,j\neq k}, \\ \varphi_k^{-1} &: \Sigma^{(k)} \mapsto \mathbf{1}\mathrm{diag}(\tilde{\Sigma}^{(k)})^{\mathrm{T}} + \mathrm{diag}(\tilde{\Sigma}^{(k)})\mathbf{1}^{\mathrm{T}} - 2\tilde{\Sigma}^{(k)}, \end{aligned} \tag{45}$$

where $\tilde{\Sigma}^{(k)}$ is the $d \times d$ matrix that coincides with $\Sigma^{(k)}$ for $i, j \neq k$ and that has 0s in the kth column and row.

Proof. It is easy to check that the mappings are their mutual inverses. To see that the strict positive definiteness of $\Sigma^{(k)}$ is equivalent to the strict conditionally negative definiteness of Γ, we observe for any $\mathbf{a}_{\backslash k} \in \mathbb{R}^{d-1}\backslash\{\mathbf{0}\}$ and $a_k = -\Sigma_{i\neq k} a_i$

$$\mathbf{a}_{\backslash k}^{\mathrm{T}} \Sigma^{(k)} \mathbf{a}_{\backslash k} = \tfrac{1}{2} \sum_{i,j\neq k} a_i a_j (\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}) = -\sum_{i\neq k} a_i a_k \Gamma_{ik} - \tfrac{1}{2} \sum_{i,j\neq k} a_i a_j \Gamma_{ij} = -\mathbf{a}^{\mathrm{T}}\Gamma\mathbf{a},$$

using the fact that Γ is symmetric and $\Gamma_{ii} = 0$ for all $i \in V$. The assertion then follows; see also the proof of lemma 3.2.1 in Berg *et al.* (1984).

Appendix C: Hüsler–Reiss densities on decomposable graphs

Corollary 2. Let $\mathcal{G} = (V, E)$ be a decomposable and connected graph, and suppose that $\mathbf{Y}$ is a Hüsler–Reiss Pareto distribution that satisfies the pairwise Markov property

$$Y_i \perp_{\mathrm{e}} Y_j | \mathbf{Y}_{\backslash\{i,j\}} \qquad \text{if } (i, j) \notin E.$$

Then the density of $\mathbf{Y}$ factorizes according to $\mathcal{G}$ into lower dimensional Hüsler–Reiss densities, i.e.

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{y_{k_1}^{-2} \prod_{j\neq k_1} y_j^{-1}}{\Lambda(\mathbf{1})} \frac{\prod_{i=1}^{m} \phi_{|C_i|-1}\{\log(\mathbf{y}_{C_i\backslash\{k_i\}}/y_{k_i}) + \Gamma_{C_i\backslash\{k_i\},k_i}; \Sigma^{(k_i)}_{C_i}\}}{\prod_{i=1}^{m-1} \phi_{|D_i|-1}\{\log(\mathbf{y}_{D_i\backslash\{k_i\}}/y_{k_i}) + \Gamma_{D_i\backslash\{k_i\},k_i}; \Sigma^{(k_i)}_{D_i}\}}, \qquad \mathbf{y} \in \mathcal{L},$$

where the sequences of cliques $\{C_1, \ldots, C_m\}$ and separator sets $\{D_2, \ldots, D_m\}$ have the running intersection property (44), and $k_i \in D_i$, $i = 2, \ldots, m$, $k_1 \in C_1$.

Proof. Theorem 1 and proposition 3 yield the factorization. It remains to show that the factors in front of the normal densities simplify to $y_{k_{m-1}}^{-2} \Pi_{i\neq k_{m-1}} y_i^{-1}$. Indeed, since we choose $k_i \in D_i \subset C_i$, $i = 2, \ldots, m$, the ratio $\lambda_{C_i}(\mathbf{y}_{C_i})/\lambda_{D_i}(\mathbf{y}_{D_i})$ contributes the factor y_j^{-1} for all $j \in C_i \backslash D_i$, and each such j appears exactly once. For $i = 1$, the contribution of $\lambda_{C_1}(\mathbf{y}_{C_1})$ is $y_{k_1}^{-2} \Pi_{i\in C_1\backslash\{k_1\}} y_i^{-1}$.

Appendix D: Minimum spanning tree for the Danube river data

Fig. 9(a) shows the estimated Hüsler–Reiss minimum spanning tree for the Danube data in Section 6 for a threshold u chosen as the 90% quantile of the marginal Pareto distribution. To assess the sensitivity of the tree structure with respect to the choice of threshold, we estimate the minimum spanning tree for thresholds u corresponding to a range of quantiles. The similarities of these trees in terms of the number of identical edges compared with the 90% quantile tree are shown in Fig. 10. We can see that there is some variation of the tree structure for different thresholds, but that most of the 30 edges are fairly stable throughout a wide range of thresholds. As a comparison, Fig. 9(b) shows the Gaussian minimum spanning tree fitted to all log-transformed data, using $\log(1 - \rho_{ij}^2)$ as distances in expression (38), where ρ_{ij} is the correlation coefficient between nodes $i, j \in V$. The Gaussian tree, which is a model for non-extremal data, is similar to the Hüsler–Reiss tree, which is a model for extreme flooding, but there are also some differences. For instance, for the extremal data the ordering of stations 16–19 seems to be less important since large discharges affect all at the same time. This is confirmed by the fact that, when the Hüsler–Reiss tree is extended to a block graph, then additional edges are introduced between these stations.

Fig. 9. (a) Estimated Hüsler–Reiss minimum spanning tree for the Danube data with 90% quantile threshold and (b) Gaussian minimum spanning tree using all log-transformed data

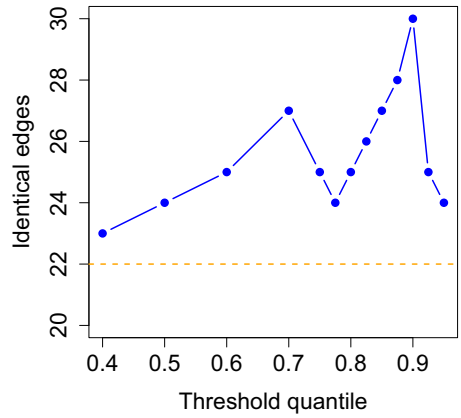


Fig. 10. For estimated minimum spanning trees corresponding to different threshold quantiles, the number of edges that are identical to the 90% quantile tree (●) and the number of identical edges for a Gaussian minimum spanning tree by using all log-transformed data (- - - - -)

Appendix E: Trivariate χ-coefficients

Fig. 11 shows the empirical estimates of the trivariate coefficients

$$\chi_{ijk} = 3 - \Lambda_{ij}(1,1) - \Lambda_{ik}(1,1) - \Lambda_{jk}(1,1) + \Lambda_{ijk}(1,1,1), \qquad i, j, k \in V,$$

against those implied by the fitted spatial model in Asadi *et al.* (2015) and our graphical model minimizing AIC.

Appendix F: Proofs

F.1. Proof (of proposition 1)

The implication that condition (17) implies part (a) is trivial. To show that part (a) implies part (b) let $k \in B$ and suppose that condition (18) holds, i.e.

$$f^k(\mathbf{y}) = \frac{f^k_{A\cup B}(\mathbf{y}_{A\cup B}) f^k_{B\cup C}(\mathbf{y}_{B\cup C})}{f^k_B(\mathbf{y}_B)}, \qquad \mathbf{y} \in \mathcal{L}^k.$$

For any $\mathbf{y} \in \mathcal{L}$ choose $0 < t < \min(y_k, 1)$, i.e. $\mathbf{y}/t \in \mathcal{L}^k$, and observe that

$$\begin{aligned}
\lambda(\mathbf{y}) &= t^{-(d+1)} f^k(\mathbf{y}/t) \\
&= t^{-(d+1)} \frac{f^k_{A\cup B}(\mathbf{y}_{A\cup B}/t) f^k_{B\cup C}(\mathbf{y}_{B\cup C}/t)}{f^k_B(\mathbf{y}_B/t)} \\
&= t^{-(d+1)} \frac{\lambda_{A\cup B}(\mathbf{y}_{A\cup B}/t)\lambda_{B\cup C}(\mathbf{y}_{B\cup C}/t)}{\lambda_B(\mathbf{y}_B/t)} \\
&= \frac{\lambda_{A\cup B}(\mathbf{y}_{A\cup B})\lambda_{B\cup C}(\mathbf{y}_{B\cup C})}{\lambda_B(\mathbf{y}_B)},
\end{aligned}$$

using the homogeneity of the λ_I, and the fact that $f^k_I(\mathbf{y}_I/t) = \lambda_I(\mathbf{y}_I/t)$ for any $I \subset V$ with $k \in I$. For this argument it is crucial that k is in an element of all three sets B, $A \cup B$ and $B \cup C$.

To show that part (b) implies condition (17) suppose that the factorization (19) of λ holds, and let $k \in V$. For all $\mathbf{y} \in \mathcal{L}^k$

$$f^k(\mathbf{y}) = \frac{\lambda_{A\cup B}(\mathbf{y}_{A\cup B})\lambda_{B\cup C}(\mathbf{y}_{B\cup C})}{\lambda_B(\mathbf{y}_B)} = g(\mathbf{y}_{A\cup B})h(\mathbf{y}_{B\cup C}),$$

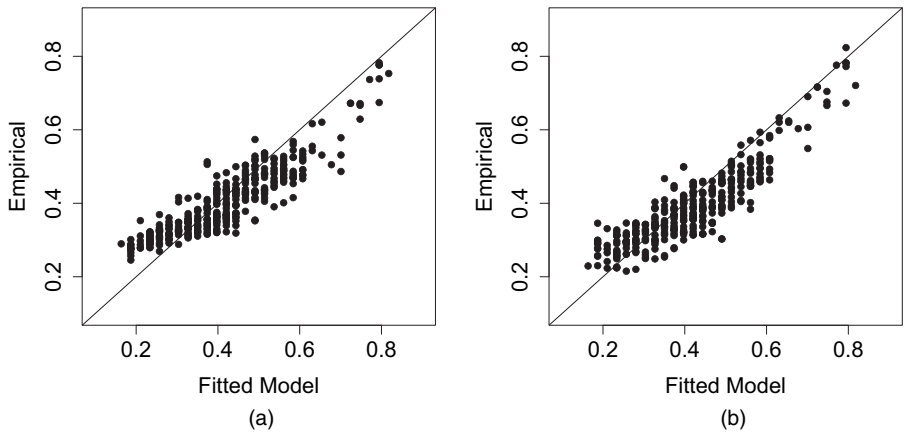


Fig. 11. Empirical estimates of the trivariate coefficients χ_{ijk}, $i, j, k \in V$, against those implied by (a) the fitted spatial model in Asadi *et al.* (2015) and (b) our graphical model minimizing AIC: only coefficients for 400 randomly selected triplets are shown

for suitable functions g and h, implying the required conditional independence of f^k (see Lauritzen (1996), chapter 3). This shows that condition (17) indeed holds and thus $\mathbf{Y}_A \perp_e \mathbf{Y}_C | \mathbf{Y}_B$.

F.2. Proof (of theorem 1)

We start by proving that, if $\mathbf{Y}$ satisfies the pairwise Markov property relative to $\mathcal{G}$, then the graph $\mathcal{G}$ is necessarily connected. Indeed, suppose that V can be split into non-empty disjoint subsets $V_1, V_2 \subset V$ such that for $(i, j) \in E$ it holds that either $i, j \in V_1$ or $i, j \in V_2$. For an arbitrary $k \in V$, by assumption, the pairwise Markov property relative to $\mathcal{G}$ is satisfied for f^k on $\mathcal{L}^k$ and the classical Hammersley–Clifford theorem implies the global Markov property for f^k, and in particular

$$f^k(\mathbf{y}) = f^k_{V_1}(\mathbf{y}_{V_1}) f^k_{V_2}(\mathbf{y}_{V_2}), \qquad \mathbf{y} \in \mathcal{L}^k.$$

The discussion after proposition 1 shows that such a factorization contradicts integrability of the multivariate Pareto density, and therefore the graph must be connected.

We now show that part (a) implies part (c). The pairwise Markov property of f^k relative to $\mathcal{G}$ implies by the classical Hammersley–Clifford theorem that

$$f^k(\mathbf{y}) = \frac{\prod_{C \in \mathcal{C}} f^k_C(\mathbf{y}_C)}{\prod_{D \in \mathcal{D}} f^k_D(\mathbf{y}_D)}, \qquad \mathbf{y} \in \mathcal{L}^k.$$

This representation is not of direct use since it cannot be extended to $f_{\mathbf{Y}}$ on the whole space $\mathcal{L}$, since all f^k_I with $k \notin I$ are not homogeneous. The result, however, tells us that $\mathbf{Y}^k$ also satisfies the global Markov property on $\mathcal{L}^k$ relative to $\mathcal{G}$, as defined in Section 2.3. The running intersection property implies that D_m separates $C_m \setminus D_m$ from $(C_1 \cup \ldots \cup C_{m-1}) \setminus D_m$. Choose $k \in D_m$; then the global Markov property for $\mathbf{Y}^k$ yields

$$f^k(\mathbf{y}) = \frac{f^k_{C_m}(\mathbf{y}_{C_m}) f^k_{C_1 \cup \ldots \cup C_{m-1}}(\mathbf{y}_{C_1 \cup \ldots \cup C_{m-1}})}{f^k_{D_m}(\mathbf{y}_{D_m})} = \frac{\lambda_{C_m}(\mathbf{y}_{C_m}) \lambda_{C_1 \cup \ldots \cup C_{m-1}}(\mathbf{y}_{C_1 \cup \ldots \cup C_{m-1}})}{\lambda_{D_m}(\mathbf{y}_{D_m})}, \qquad \mathbf{y} \in \mathcal{L}^k,$$

where the second equality holds since $k \in D_m$, and D_m is a subset of both C_m and $C_1 \cup \ldots \cup C_{m-1}$. By a homogeneity argument similar to the proof of proposition 1, this factorization extends to λ on the whole space $\mathcal{L}$, i.e.

$$\lambda(\mathbf{y}) = \frac{\lambda_{C_m}(\mathbf{y}_{C_m}) \lambda_{C_1 \cup \ldots \cup C_{m-1}}(\mathbf{y}_{C_1 \cup \ldots \cup C_{m-1}})}{\lambda_{D_m}(\mathbf{y}_{D_m})}, \qquad \mathbf{y} \in \mathcal{L}.$$

It remains to decompose $\lambda_{C_1\cup\ldots\cup C_{m-1}}$ in the same manner. For this, choose a new $k \in D_{m-1}$ and note that

$$f^k_{C_1\cup\ldots\cup C_{m-1}}(\mathbf{y}_{C_1\cup\ldots\cup C_{m-1}}) = \int_{[0,\infty)^{|C_m\setminus D_m|}} \frac{\prod_{C\in\mathcal{C}} f^k_C(\mathbf{y}_C)}{\prod_{D\in\mathcal{D}} f^k_D(\mathbf{y}_D)} \mathrm{d}\mathbf{y}_{C_m\setminus D_m} = \frac{\prod_{C\neq C_m} f^k_C(\mathbf{y}_C)}{\prod_{D\neq D_m} f^k_D(\mathbf{y}_D)},$$

and therefore satisfies the global Markov property relative to the subgraph induced on $C_1\cup\ldots\cup C_{m-1}$. Since $f^k_{C_1\cup\ldots\cup C_{m-1}} = \lambda_{C_1\cup\ldots\cup C_{m-1}}$ on $\mathcal{L}^k$, applying successively the same reasoning as before yields the factorization of λ that directly implies representation (21) for $f_{\mathbf{Y}}$.

To show that part (c) implies part (b), we need only to verify that $\mathbf{Y}^k$ satisfies the global Markov property on $\mathcal{L}^k$ for any $k \in V$. For disjoint sets $A, B, C \subset V$ such that B separates A from C, factorization (21) entails that

$$f^k(\mathbf{y}) = \Lambda(\mathbf{1}) f_{\mathbf{Y}}(\mathbf{y}) = g(\mathbf{y}_{A\cup B}) h(\mathbf{y}_{B\cup C}),$$

for suitable functions g and h, and thus $\mathbf{Y}^k_A \perp\!\!\!\perp \mathbf{Y}^k_C | \mathbf{Y}^k_B$.

The implication that part (b) implies part (a) holds trivially.

F.3. Proof (of corollary 1)

It is easy to check that λ and $f_{\mathbf{Y}}$ are homogeneous of order $-(d+1)$ on $\mathcal{L}$. Let $\{C_1,\ldots,C_m\}$ and $\{D_2,\ldots,D_m\}$ be the sequences of cliques and separators with the running intersection property (44). Sequential integration of the function $f_{\mathbf{Y}}$ on $C_m\setminus D_m,\ldots,C_2\setminus D_2$, together with the consistency constraint, yields that it defines in fact a probability density. Theorem 1 implies that the corresponding distribution on $\mathcal{L}$ satisfies the Markov property relative to $\mathcal{G}$.

F.4. Proof (of proposition 2)

The density of the random vector on the right-hand side of expression (24) is

$$y_k^{-2} \prod_{e=(i,j)\in E^k} y_i^{-1} f_{U_e}(y_j/y_i) = y_k^{-2} \frac{\prod_{(i,j)\in E^k} \lambda_{ij}(y_i, y_j)}{\prod_{(i,j)\in E^k} y_i^{-2}} = \prod_{\{i,j\}\in E} \frac{\lambda_{ij}(y_i, y_j)}{y_i^{-2} y_j^{-2}} \prod_{i\in V} y_i^{-2},$$

where we used expression (12) for the first equation, and the fact that each node $i \in V\setminus\{k\}$ has exactly one incoming arrow, and the kth node has no incoming arrows. However, we recall that the density of $\mathbf{Y}^k$ is $\lambda(\mathbf{y}) = \Lambda(\mathbf{1}) f_{\mathbf{Y}}(\mathbf{y})$, which factorizes with respect to the tree $\mathcal{T}$. Comparing the above density with expression (23) yields the result.

F.5. Proof (of lemma 1)

Without losing generality, we may and do assume that $k' = 1$ and $k = 2$. Let the vector $\mathbf{W}^1 = (0, W^1_2, \ldots, W^1_d)$ have a centred normal distribution with covariance matrix $\Sigma = \{\sigma_{ij}\} = \tilde{\Sigma}^{(1)}$, such that

$$\Sigma^{(1)} = \Sigma_{\setminus\{1\}} = \begin{pmatrix} \sigma_{22} & \Sigma_{2,\setminus\{1,2\}} \\ \Sigma_{\setminus\{1,2\},2} & \Sigma_{\setminus\{1,2\}} \end{pmatrix}.$$

The precision matrix is obtained by blockwise inversion as

$$\Theta^{(1)} = \begin{pmatrix} \sigma_{22}^{-1} + \sigma_{22}^{-2}\Sigma_{2,\setminus\{1,2\}} S^{-1} \Sigma_{\setminus\{1,2\},2} & -\sigma_{22}^{-1}\Sigma_{2,\setminus\{1,2\}} S^{-1} \\ -\sigma_{22}^{-1} S^{-1} \Sigma_{\setminus\{1,2\},2} & S^{-1} \end{pmatrix},$$

where $S = \Sigma_{\setminus\{1,2\}} - \sigma_{22}^{-1}\Sigma_{\setminus\{1,2\},2}\Sigma_{2,\setminus\{1,2\}}$ is the Schur complement of upper left-hand block σ_{22} in the matrix $\Sigma^{(1)}$. The random vector $\mathbf{W}^1$ can be transformed into

$$\mathbf{W}^2 = (-W_2^1, 0, W_3^1 - W_2^1, \ldots, W_d^1 - W_2^1),$$

which is readily verified to have a centred normal distribution with covariance matrix $\tilde{\Sigma}^{(2)}$. In contrast, we may write the covariance matrix $\Sigma^{(2)}$ of $(-W_2^1, W_3^1 - W_2^1, \ldots, W_d^1 - W_2^1)$ in terms of Σ as

$$\Sigma^{(2)} = \begin{pmatrix} \sigma_{22} & \sigma_{22}\mathbf{1}^{\mathrm{T}} - \Sigma_{2,\setminus\{1,2\}} \\ \sigma_{22}\mathbf{1} - \Sigma_{\setminus\{1,2\},2} & \Sigma_{\setminus\{1,2\}} + \sigma_{22}\mathbf{1}\mathbf{1}^{\mathrm{T}} - \Sigma_{\setminus\{1,2\},2}\mathbf{1}^{\mathrm{T}} - \mathbf{1}\Sigma_{2,\setminus\{1,2\}} \end{pmatrix}.$$

It can be checked that the Schur complement of the upper left-hand block σ_{22} in the matrix $\Sigma^{(2)}$ is again S. Thus, blockwise inversion yields

$$\Theta^{(2)} = \begin{pmatrix} \sigma_{22}^{-1} + \sigma_{22}^{-2}(\sigma_{22}\mathbf{1}^{\mathrm{T}} - \Sigma_{2,\setminus\{1,2\}})S^{-1}(\sigma_{22}\mathbf{1} - \Sigma_{\setminus\{1,2\},2}) & -\sigma_{22}^{-1}(\sigma_{22}\mathbf{1}^{\mathrm{T}} - \Sigma_{2,\setminus\{1,2\}})S^{-1} \\ -\sigma_{22}^{-1}S^{-1}(\sigma_{22}\mathbf{1} - \Sigma_{\setminus\{1,2\},2}) & S^{-1} \end{pmatrix}.$$

Comparing these representations of $\Theta^{(1)}$ and $\Theta^{(2)}$ yields the assertion for $i, j \in V\setminus\{1,2\}$. For $i \neq 2$ and $j = 2$, we observe that

$$\sum_{l\neq 2} \Theta_{il}^{(2)} = -\sum_{m\neq 1,2} S_{im}^{-1} + \sigma_{22}^{-1} \sum_{m\neq 1,2} S_{im}^{-1}\sigma_{m2} + \sum_{m\neq 1,2} S_{im}^{-1} = -\Theta_{i2}^{(1)}.$$

The case $i, j = 2$ follows similarly.

F.6. Proof (of proposition 3)

Let $i, j \in V$ with $i \neq j$ be fixed and choose a $k \neq i, j$. Let P and $\mathbf{W}$ be as in representation (28). Since $Y_k^k = P$ and because of the independence of P and $\mathbf{W}$ we obtain

$$\begin{aligned} Y_i^k \perp\!\!\!\perp Y_j^k | \mathbf{Y}_{\setminus\{i,j\}}^k &\Leftrightarrow P\exp(W_i^k - \Gamma_{ik}/2) \perp\!\!\!\perp P\exp(W_j^k - \Gamma_{jk}/2) | P, \mathbf{W}_{\setminus\{i,j,k\}}^k \\ &\Leftrightarrow W_i^k \perp\!\!\!\perp W_j^k | \mathbf{W}_{\setminus\{i,j,k\}}^k \\ &\Leftrightarrow \Theta_{ij}^{(k)} = 0, \end{aligned}$$

where the variable W_k^k can be deleted from the conditioning since it is deterministic given P, and therefore the reduced precision matrix $\Theta^{(k)}$ of the vector $\mathbf{W}_{\setminus k}^k$ appears. The last equivalence follows from the well-known fact that conditional independence in multivariate normal models corresponds to 0s in the precision matrix (see example 4).

Let now $k = i \neq j$ and choose a $k' \notin \{i, j\}$. Lemma 1 implies that

$$-\sum_{l\neq k} \Theta_{jl}^{(k)} = \Theta_{jk}^{(k')}. \tag{46}$$

Since $k' \in V\setminus\{i, j\}$, by proposition 1, $Y_i \perp_{\mathrm{e}} Y_j | \mathbf{Y}_{\setminus\{i,j\}}$ is equivalent to $Y_k^{k'} \perp\!\!\!\perp Y_j^{k'} | \mathbf{Y}_{\setminus\{k,j\}}^{k'}$. The latter, by the first part of the proof, is then equivalent to $\Theta_{jk}^{(k')} = 0$, which, together with equation (46), yields the assertion. The case $k = j \neq i$ is analogous by symmetry.

F.7. Proof (of proposition 4)

Let $C_1, \ldots, C_m$ be an enumeration of the cliques of the decomposable connected graph $\mathcal{G} = (V, E)$. Recall that, by assumption, all intersections between pairs of cliques are either empty or contain a single node. We show how to obtain the unique $(d \times d)$-dimensional variogram matrix Γ that solves the completion problem (31) by adding one clique after another. We first set

$$\Gamma_{ij} = \Gamma_{ij}^{(C_1)}, \qquad \text{for } i, j \in C_1. \tag{47}$$

Let $I_{p-1} = C_1 \cup \ldots \cup C_{p-1}$ be the union of the first $p-1$ cliques, $2 \leqslant p \leqslant m$, that have been chosen in an order such that $\mathcal{G}$ restricted to I_{p-1} forms a connected graph. Suppose that we have already constructed a unique $(|I_{p-1}| \times |I_{p-1}|)$-dimensional variogram matrix $\Gamma^{(I_{p-1})}$ that satisfies

$$\begin{aligned} \Gamma_{ij}^{(I_{p-1})} &= \Gamma_{ij}^{(C_l)}, && \text{for } i, j \in C_l \text{ and all } l = 1, \ldots, p-1, \\ \Theta_{ij}^{(I_{p-1},k)} &= 0, && \text{for all } i, j, k \in I_{p-1}, i, j \neq k \text{ and } (i, j) \notin E, \end{aligned} \tag{48}$$

where here and in what follows we use the notation $\Theta^{(J,k)}$ as the inverse of $\Sigma^{(J,k)} = \varphi_k(\Gamma^{(J)})$ for a variogram matrix $\Gamma^{(J)}$ on some index set $J \subset V$ and $k \in J$. We next choose a clique, say C_p, that intersects I_{p-1}, and this intersection must be a single node, say $k_0 \in V$. Let $I_p = I_{p-1} \cup C_p$ and define the matrix

$$\Theta^{(I_p,k_0)} = \begin{pmatrix} \Theta^{(I_{p-1},k_0)} & 0 \\ 0 & \Theta^{(C_p,k_0)} \end{pmatrix}. \tag{49}$$

This matrix is an invertible covariance matrix since its blocks are invertible covariance matrices, and its inverse $\Sigma^{(I_p,k_0)}$ has the same property with blocks $\Sigma^{(I_{p-1},k_0)}$ and $\Sigma^{(C_p,k_0)}$. This yields an $(|I_p| \times |I_p|)$-dimensional variogram matrix $\Gamma^{(I_p)}$ through the mapping $\varphi_{k_0}^{-1}$, which has the form

$$\Gamma_{ij}^{(I_p)} = \begin{cases} \Gamma_{ij}^{(I_{p-1})}, & \text{for } i, j \in I_{p-1}, \\ \Gamma_{ij}^{(C_p)}, & \text{for } i, j \in C_p, \\ \Gamma_{ik_0}^{(I_{p-1})} + \Gamma_{jk_0}^{(I_{p-1})} & \text{for } i \in I_{p-1}, j \in C_p \text{ or } j \in I_{p-1}, i \in C_p. \end{cases}$$

This variogram matrix clearly solves problem (48) with I_{p-1} replaced by I_p. It is unique by construction and the fact that φ_{k_0} and $\varphi_{k_0}^{-1}$ are bijections.

Starting with expression (47) and then adding all cliques for $p = 2, \ldots, m$ according to the above procedure, we obtain a unique $(d \times d)$-dimensional variogram $\Gamma = \Gamma^{(I_m)}$ matrix that satisfies all constraints in expression (31). Comparing with corollary 2 it follows that the corresponding density in expression (30) is d-variate Hüsler–Reiss with parameter matrix Γ.

F.8. Proof (of lemma 2)

The general formula for extremal functions in proposition 1 in Dombry *et al.* (2016) can be written in terms of the exponent measure density λ as

$$\begin{aligned} \mathbb{P}(\mathbf{U}^k \in A) &= \int_{\mathcal{E}} \mathbf{1}\{\mathbf{y}/y_k \in A\} \mathbf{1}\{y_k > 1\} \lambda(\mathbf{y}) \mathrm{d}\mathbf{y} \\ &= \int_{\mathcal{L}^k} \mathbf{1}\{\mathbf{y}/y_k \in A\} f^k(\mathbf{y}) \mathrm{d}\mathbf{y} \\ &= \mathbb{P}(\mathbf{Y}^k / Y_k^k \in A). \end{aligned}$$

Since the density of $\mathbf{U}_{\setminus k}^k = \mathbf{Y}_{\setminus k}^k / Y_k^k$ is readily seen to be $\lambda(\mathbf{y})$ for $\mathbf{y}_{\setminus k} \in [0, \infty)^{d-1}$ and $y_k = 1$, it follows with

$$\lambda(\mathbf{y}) = y_k^{-(d+1)} \lambda(\mathbf{y}/y_k) = y_k^{-(d+1)} f_{\mathbf{U}_{\setminus k}^k}(\mathbf{y}_{\setminus k}/y_k), \qquad \mathbf{y} \in \mathcal{E},$$

that expression (41) is an equivalent definition of extremal functions.

It follows from theorem 2 in Dombry *et al.* (2016) that, for a uniform distribution T on $\{1, \ldots, d\}$, the random vector $\mathbf{Y}^{\mathrm{T}} / \|\mathbf{Y}^{\mathrm{T}}\|_1$ follows the distribution of the spectral measure H on $S_{d-1} = \{\mathbf{x} \in \mathcal{E} : \|x\|_1 = 1\}$ associated with the max-stable distribution $\mathbf{Z}$, i.e.

$$\Lambda(A) = d \int_{S_{d-1}} \int_0^\infty u^{-2} \mathbf{1}\{u\mathbf{w} \in A\} \mathrm{d}u H(\mathrm{d}\mathbf{w}), \qquad A \subset \mathcal{E}.$$

If $A \subset \mathcal{L}$, then $u\mathbf{w} \in A$ implies that $u \geqslant 1$, and therefore

$$\begin{aligned} \mathbb{P}\left(\frac{P\mathbf{Y}^{\mathrm{T}}}{\|\mathbf{Y}^{\mathrm{T}}\|_1} \in A \right) &= \int_{S_{d-1}} \int_1^\infty f_P(u) \mathbf{1}\{u\mathbf{w} \in A\} \mathrm{d}u H(\mathrm{d}\mathbf{w}) \\ &= \frac{1}{d} \int_A \lambda(\mathbf{y}) \mathrm{d}\mathbf{y}, \end{aligned}$$

since $f_P(u) = 1/u^2, u \geqslant 1$. For $A = \mathcal{L} = \mathcal{E} \setminus [\mathbf{0}, \mathbf{1}]$ this yields for the conditioning event in expression (42)

$$\mathbb{P}\left(\frac{P\|\mathbf{Y}^{\mathrm{T}}\|_\infty}{\|\mathbf{Y}^{\mathrm{T}}\|_1} > 1 \right) = \frac{\Lambda(\mathcal{L})}{d} = \frac{\Lambda(\mathbf{1})}{d}. \tag{50}$$

Since $\mathbf{Y}$ has density $\lambda(\mathbf{y})/\Lambda(\mathbf{1})$, this concludes the proof.

References

Arlot, S. and Celisse, A. (2010) A survey of cross-validation procedures for model selection. *Statist. Surv.*, **4**, 40–79.

Asadi, P., Davison, A. C. and Engelke, S. (2015) Extremes on river networks. *Ann. Appl. Statist.*, **9**, 2023–2050.

Ballani, F. and Schlather, M. (2011) A construction principle for multivariate extreme value distributions. *Biometrika*, **98**, 633–645.

Basrak, B. and Segers, J. (2009) Regularly varying multivariate time series. *Stoch. Processes Appl.*, **119**, 1055–1080.

Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004) *Statistics of Extremes*. Chichester: Wiley.

Berg, C., Christensen, J. P. R. and Ressel, P. (eds) (1984) Theory of positive definite and related functions. In *Harmonic Analysis on Semigroups*. New York: Springer.

Böhm, O. and Wetzel, K.-F. (2006) Flood history of the Danube tributaries Lech and Isar in the alpine foreland of Germany. *Hydrol. Sci. J.*, **51**, 784–798.

Boldi, M.-O. and Davison, A. C. (2007) A mixture model for multivariate extremes. *J. R. Statist. Soc.* B, **69**, 217–229.

Brown, B. M. and Resnick, S. I. (1977) Extreme values of independent stochastic processes. *J. Appl. Probab.*, **14**, 732–739.

de Carvalho, M. and Davison, A. C. (2014) Spectral density ratio models for multivariate extremes. *J. Am. Statist. Ass.*, **109**, 764–776.

Chow, C. and Liu, C. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory*, **14**, 462–467.

Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence measures for extreme value analyses. *Extremes*, **2**, 339–365.

Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *J. R. Statist. Soc.* B, **53**, 377–392.

Cooley, D., Davis, R. A. and Naveau, P. (2010) The pairwise beta distribution: a flexible parametric multivariate model for extremes. *J. Multiv. Anal.*, **101**, 2103–2117.

Cooley, D. and Thibaud, E. (2019) Decompositions of dependence for high-dimensional extremes. *Biometrika*, **106**, 587–604.

Cowell, R. G., Dawid, P., Lauritzen, S. L. and Spiegelhalter, D. J. (2006) *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Berlin: Springer.

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modeling of spatial extremes. *Statist.Sci.*, **27**, 161–186.

Dawid, A. P. (1979) Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc.* B, **41**, 1–31.

Dawid, A. P. and Lauritzen, S. L. (1993) Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, **21**, 1272–1317.

Dempster, A. P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Dieker, A. B. and Mikosch, T. (2015) Exact simulation of Brown–Resnick random fields at a finite number of locations. *Extremes*, **18**, 301–314.

Dombry, C., Engelke, S. and Oesting, M. (2016) Exact simulation of max-stable processes. *Biometrika*, **103**, 303–317.

Dombry, C., Engelke, S. and Oesting, M. (2017) Bayesian inference for multivariate extreme value distributions. *Electron. J. Statist.*, **11**, 4813–4844.

Dombry, C. and Éyi-Minko, F. (2013) Regular conditional distributions of continuous max-infinitely divisible random fields. *Electron. J. Probab.*, **18**, no. 7, article 21.

Dombry, C. and Éyi-Minko, F. (2014) Stationary max-stable processes with the Markov property. *Stoch. Processes Appl.*, **124**, 2266–2279.

Dombry, C., Éyi-Minko, F. and Ribatet, M. (2013) Conditional simulation of max-stable processes. *Biometrika*, **100**, 111–124.

Einmahl, J. H. J., Kiriliouk, A., Krajina, A. and Segers, J. (2016) An M-estimator of spatial tail dependence. *J. R. Statist. Soc.* B, **78**, 275–298.

Einmahl, J. H. J., Krajina, A. and Segers, J. (2012) An M-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.*, **40**, 1764–1793.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events: for Insurance and Finance*. London: Springer.

Engelke, S., Hitz, S. A. and Gnecco, N. (2019) graphicalExtremes: statistical methodology for graphical extreme value models. *R Package Version 0.1.0*. (Available from `https://CRAN.R-project.org/package=graphicalExtremes`.)

Engelke, S., Malinowski, A., Kabluchko, Z. and Schlather, M. (2015) Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *J. R. Statist. Soc.* B, **77**, 239–265.

de Fondeville, R. and Davison, A. C. (2018) High-dimensional peaks-over-threshold inference. *Biometrika*, **105**, 575–592.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82**, 543–552.

Gissibl, N. and Klüppelberg, C. (2018) Max-linear models on directed acyclic graphs. *Bernoulli*, **24**, 2693–2720.
Gissibl, N., Klüppelberg, C. and Otto, M. (2018) Tail dependence of recursive max-linear models with regularly varying noise variables. *Econmetr. Statist.*, **6**, 149–167.
Gnecco, N., Meinshausen, N., Peters, J. and Engelke, S. (2019) Causal discovery in heavy-tailed models. *Preprint*. University of Geneva, Geneva. (Available from `https://arxiv.org/abs/1908.05097`.)
Green, P. J. and Thomas, A. (2013) Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, **100**, 91–110.
Gudendorf, G. and Segers, J. (2010) Extreme-value copulas. In *Copula Theory and Its Applications* (eds P. Jaworski, F. Durante, W. Härdle and T. Rychlik), pp. 127–145. Berlin: Springer.
de Haan, L. (1984) A spectral representation for max-stable processes. *Ann. Probab.*, **12**, 1194–1204.
de Haan, L. and Ferreira, A. (2006) *Extreme Value Theory*. New York: Springer.
Harary, F. (1963) A characterization of block-graphs. *Can. Math. Bull.*, **6**, 1–6.
Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *J. R. Statist. Soc.* B, **66**, 497–546.
Hitz, A. S. and Evans, J. R. (2016) One-component regular variation and graphical modeling of extremes. *J. Appl. Probab.*, **53**, 733–746.
Huser, R., Dombry, C., Ribatet, M. and Genton, M. G. (2019) Full likelihood inference for max-stable data. *Stat*, **8**, no. 1, article e218.
Hüsler, J. and Reiss, R.-D. (1989) Maxima of normal random vectors: between independence and complete dependence. *Statist. Probab. Lett.*, **7**, 283–286.
Janssen, A. and Segers, J. (2014) Markov tail chains. *J. Appl. Probab.*, **51**, 1133–1153.
Joe, H. (2015) *Dependence Modeling with Copulas*. Boca Raton: CRC Press.
Kabluchko, Z., Schlather, M. and de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. *Ann. Probab.*, **37**, 2042–2065.
Katz, R. W., Parlange, M. B. and Naveau, P. (2002) Statistics of extremes in hydrology. *Adv. Wat. Resour.*, **25**, 1287–1304.
Keef, C., Tawn, J. and Svensson, C. (2009) Spatial risk assessment for extreme river flows. *Appl. Statist.*, **58**, 601–618.
Keef, C., Tawn, J. A. and Lamb, R. (2013) Estimating the probability of widespread flood events. *Environmetrics*, **24**, 13–21.
Kellerer, H. G. (1964) Verteilungsfunktionen mit gegebenen Marginalverteilungen. *Z. Wahrsch. Ver. Geb.*, **3**, 247–270.
Kiriliouk, A., Rootzén, H., Segers, J. and Wadsworth, J. L. (2018a) Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, **61**, 123–135.
Kiriliouk, A., Segers, J. and Tafakori, L. (2018b) An estimator of the stable tail dependence function based on the empirical beta copula. *Extremes*, **21**, 581–600.
Kruskal, Jr, J. B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.*, **7**, 48–50.
Kulik, R. and Soulier, P. (2015) Heavy tailed time series with extremal independence. *Extremes*, **18**, 273–299.
Lafferty, J., Liu, H. and Wasserman, L. (2012) Sparse nonparametric graphical models. *Statist. Sci.*, **27**, 519–537.
Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Oxford University Press.
Ledford, A. W. and Tawn, J. A. (1997) Modelling dependence within joint tail regions. *J. R. Statist. Soc.* B, **59**, 475–499.
Lee, D. and Joe, H. (2018) Multivariate extreme value copulas with factor and tree dependence structures. *Extremes*, **21**, 147–176.
Loh, P.-L. and Wainwright, M. (2013) Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. *Ann. Statist.*, **41**, 3022–3049.
Marcon, G., Padoan, S., Naveau, P., Muliere, P. and Segers, J. (2017) Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials. *J. Statist. Planng Inf.*, **183**, 1–17.
McNeil, A. J., Frey, R. and Embrechts, P. (2015) *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton: Princeton University Press.
Mhalla, L., Chavez-Demoulin, V. and Dupuis, D. J. (2020) Causal mechanism of extreme river discharges in the Danube basin network. *Appl. Statist.*, **69**, in the press.
Min, S.-K., Zhang, X., Zwiers, F. and Hegerl, G. (2011) Human contribution to more-intense precipitation extremes. *Nature*, **470**, 378–381.
Naveau, P., Ribes, A., Zwiers, F., Hannart, A., Tuel, A. and Yiou, P. (2018) Revising return periods for record events in a climate event attribution context. *J. Clim.*, **31**, 3411–3422.
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**, 1497–1530.
Opitz, T. (2013) Extremal t processes: elliptical domain of attraction and a spectral representation. *J. Multiv. Anal.*, **122**, 409–413.
Papastathopoulos, I. and Strokorb, K. (2016) Conditional independence among max-stable laws. *Statist. Probab. Lett.*, **108**, 9–15.

Papastathopoulos, I., Strokorb, K., Tawn, J. A. and Butler, A. (2017) Extreme events of Markov chains. *Adv. Appl. Probab.*, **49**, 134–161.
Prim, R. C. (1957) Shortest connection networks and some generalizations. *Bell Syst. Tech. J.*, **36**, 1389–1401.
R Core Team (2019) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Resnick, S. I. (2008) *Extreme Values, Regular Variation and Point Processes*. New York: Springer.
Rootzén, H., Segers, J. and Wadsworth, J. L. (2018) Multivariate peaks over thresholds models. *Extremes*, **21**, 115–145.
Rootzén, H. and Tajvidi, N. (2006) Multivariate generalized Pareto distributions. *Bernoulli*, **12**, 917–930.
Rue, H. and Held, L. (eds) (2005) Theory and applications. In *Gaussian Markov Random Fields*. Boca Raton: Chapman and Hall–CRC.
Schlather, M. (2002) Models for stationary max-stable random fields. *Extremes*, **5**, 33–44.
Schlather, M. and Tawn, J. A. (2003) A dependence measure for multivariate and spatial extreme values: properties and inference. *Biometrika*, **90**, 139–156.
Segers, J. (2019) One- versus multi-component regular variation and extremes of Markov trees. *Preprint*. Université catholique de Louvain, Louvain-la-Neuve. (Available from `https://arxiv.org/abs/1902.02226`.)
Smith, R., Tawn, J. and Coles, S. (1997) Markov chain models for threshold exceedances. *Biometrika*, **84**, 249–268.
Smith, R. L. (1992) The extremal index for a Markov chain. *J. Appl. Probab.*, **29**, 37–45.
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. *Ann. Appl. Statist.*, **10**, 2303–2324.
Wadsworth, J. L. and Tawn, J. A. (2014) Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika*, **101**, 1–15.
Wadsworth, J. L., Tawn, J. A., Davison, A. C. and Elton, D. M. (2017) Modelling across extremal dependence classes. *J. R. Statist. Soc.* B, **79**, 149–175.
Wainwright, M. J. and Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Foundns Trends Mach. Learn.*, **1**, 1–305.
Yu, H., Uy, W. and Dauwels, J. (2017) Modeling spatial extremes via ensemble-of-trees of pairwise copulas. *IEEE Trans. Signl Process.*, **65**, 571–586.

Discussion on the paper by Engelke and Hitz

Jennifer L. Wadsworth (*Lancaster University*)

I congratulate Engelke and Hitz on an excellent paper, which is rich in both new ideas and elegant mathematical detail.

The setting of the paper is multivariate regular variation, which is a broadly applicable regularity assumption on the extremal dependence structure of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$. Assuming standard Pareto margins, multivariate regular variation implies convergenc of normalized 'threshold exceedances' to a multivariate Pareto distribution, with support $\mathcal{L}$, as in equation (6). It is further assumed that

(a) the d-dimensional joint density $\lambda(\mathbf{y})/\Lambda(1) > 0$, and
(b) the full support is on the interior of $\mathcal{L}$, i.e. no mass lies on regions of the form $\{\mathbf{x} \in \mathcal{L} : \min(x_1, \ldots, x_d) = 0\}$.

A loosely described consequence of these settings is that all variables will tend to take their largest values simultaneously, with no possibility that some groups of variables will tend to be large while others are small.

Assumptions (a) and (b) are reasonably common in the literature and in this case facilitate the vast progress achieved on the notions of conditional independence and graphical structure for extremes. In particular, proposition 1 and theorem 1 expose very neatly how ideas from the world of graphical modelling pass through to extreme value theory via the density of the exponent measure, $\lambda(\mathbf{y})$. For decomposable graphs, and more particularly block graphs, this leads to new ideas for high dimensional model construction, and inference on coherent high dimensional models via lower dimensional subgroups. Several interesting results are obtained for the Hüsler–Reiss model, and its parameterization is shown to reveal extremal conditional independence properties in a very natural way.

The new properties are explored via classical notions of conditional independence for the multivariate Pareto random vector $\mathbf{Y}$ with support restricted to $\mathcal{L}^k$, i.e. $\mathbf{Y}^k = \mathbf{Y} | Y_k > 1$. This presents an intriguing connection with the last Royal Statistical Society discussion paper on extremes, namely the so-called conditional model that was introduced by Heffernan and Tawn (2004) and Heffernan and Resnick (2007). In the setting of the paper with $\mathbf{X}$ standard Pareto,

$$\mathbf{X}/u \,|\, \|\mathbf{X}\|_\infty > u \xrightarrow{\mathrm{d}} \mathbf{Y}, \qquad \mathbf{X}/u \,|\, X_k > u = \mathbf{X}/u \,|\, \{\|\mathbf{X}\|_\infty > u, X_k > u\} \xrightarrow{\mathrm{d}} \mathbf{Y}^k, \qquad u \to \infty,$$

and

$$\mathbf{X}/X_k|X_k > u = \mathbf{X}/X_k|\{\|\mathbf{X}\|_\infty > u, X_k > u\} \xrightarrow{\mathrm{d}} \mathbf{Y}^k/Y_k^k = \mathbf{U}^k, \qquad u \to \infty, \tag{51}$$

with $\mathbf{U}^k$ the extremal function relative to co-ordinate k. Working in exponential-tailed margins. Heffernan and Tawn (2004) made the assumption that there exist $\mathbf{a}^k : \mathbb{R} \to \mathbb{R}^d$ and $\mathbf{b}^k : \mathbb{R} \to (0, \infty)^d$ with $a_k^k\{\log(X_k)\} = \log(X_k)$ and $b_k^k\{\log(X_k)\} = 1$ such that

$$\left.\frac{\log(\mathbf{X}) - \mathbf{a}^k\{\log(X_k)\}}{\mathbf{b}^k\{\log(X_k)\}}\right| \log(X_k) > u \xrightarrow{\mathrm{d}} \mathbf{Z}^k, \tag{52}$$

where $\mathbf{Z}_{\setminus k}^k$ is non-degenerate with no mass at ∞, and $\log(X_k) - u | \log(X_k) > u \xrightarrow{\mathrm{d}} E \sim \mathrm{Exp}(1)$ is independent of $\mathbf{Z}_{\setminus k}^k$. (The formulation with random normalization came later with Heffernan and Resnick (2007) but is equivalent to the Heffernan and Tawn (2004) case under the existence of densities.) Convergence (51) is recovered from expression (52) with $\mathbf{a}^k\{\log(X_k)\} = \log(X_k)\mathbf{1}$, $\mathbf{b}^k\{\log(X_k)\} = \mathbf{1}$, and $\mathbf{Z}^k = \log(\mathbf{U}^k)$.

Under assumptions (a) and (b), the exponent measure density $\lambda(\mathbf{y})$ provides the 'glue' linking the extremal functions together: $\lambda(\mathbf{y})$ yields the distribution of each $\mathbf{U}^k$ via equation (41). As a key example, the extremal functions for the Hüsler–Reiss model are log-Gaussian, and the paper shows how the conditional independence patterns in these log-Gaussian distributions yield the overall graphical structure of $\mathbf{Y}$.

The conditional model viewpoint seems to provide a possibility for alternative methodology to that outlined in the paper, as well as a first suggestion of how one might approach extending these ideas to cases where (a) and/or (b) may not hold. The latter failure becomes more likely as d grows, yielding a natural tension between the beautiful theory of this work and the messy reality of data.

Firstly, assuming a Hüsler–Reiss model, $\log(\mathbf{X}_{\setminus k}) - \log(X_k) | \log(X_k) > u \overset{\mathrm{d}}{\approx} \mathbf{Z}_{\setminus k}^k \sim N\{-\mathrm{diag}(\Sigma^{(k)})/2, \Sigma^{(k)}\}$ for a high threshold u (Engelke *et al.*, 2015). Applying a graphical lasso technique to $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ may give a sparse precision matrix, which in principle leads to Γ and the complete Hüsler–Reiss parameterization. A practical hurdle is the likelihood of inferring different graphs from different k, and that connections to the kth node are encoded in row or column sums of $\Theta^{(k)}$, which are not shrunk towards 0 in a standard implementation. However, this could provide a starting point to explore beyond trees and block graphs, should these appear inadequate.

The potential of the conditional formulation is particularly apparent in the case where mass of the multivariate Pareto distribution lies on regions of the form $\{\mathbf{x} \in \mathcal{L} : \min(x_1, \ldots, x_d) = 0\}$. In this case, the normalizations in expression (52) allow differing strengths of extremal dependence between the components of $\mathbf{X}$ and X_k, such that components of $\mathbf{Z}_{\setminus k}^k$ may not have mass at $-\infty$ where those of $\log(\mathbf{U}_{\setminus k}^k)$ do. As such, the representation provides more detail about the extremal dependence, increasing its utility for statistical modelling. To exploit assumption (52) in practice, one poses parametric forms for $\mathbf{a}^k$, $\mathbf{b}^k$ and the distribution of $\mathbf{Z}^k$. Suppose that we take $\mathbf{a}^k\{\log(X_k)\} = \boldsymbol{\alpha}^k \log(X_k)$, $\boldsymbol{\alpha}_{\setminus k}^k \in [0, 1]^{d-1}$, $\alpha_k^k = 1$ and $\mathbf{b}^k\{\log(X_k)\} = \mathbf{1}$ and, similarly to the Hüsler–Reiss model, assume that $\mathbf{Z}_{\setminus k}^k \sim N(\boldsymbol{\mu}^k, \Sigma^{(k)})$. Then, above a high threshold u,

$$\log(\mathbf{X}) | \log(X_k) > u \overset{\mathrm{d}}{\approx} \boldsymbol{\alpha}^k(E + u) + \mathbf{Z}^k, \qquad E \sim \mathrm{Exp}(1), \quad E \perp\!\!\!\perp \mathbf{Z}^k; \tag{53}$$

the Hüsler–Reiss model is a special case with, $\boldsymbol{\alpha}^k = \mathbf{1}$, and $\boldsymbol{\mu}^k = -\mathrm{diag}(\Sigma^{(k)})/2$ for all k. For illustration, model (53) was fitted to the Danube river data both with $\boldsymbol{\alpha}^k = \mathbf{1}$ fixed and estimated. The threshold u was taken as the 0.85 marginal quantile; higher thresholds produced some errors in sparse precision matrix estimation. In each case the components of $\mathbf{Z}_{\setminus k}^k$ were estimated individually and a graphical lasso applied to $\Theta^{(k)}$ by using `EBICglasso` in the R library `qgraph` (Epskamp *et al.*, 2012). To give an impression of results across all k, Fig. 12 displays connections selected at least half of the time. Notably, although most estimates $\hat{\alpha}_j^k < 1$, the set of connections is fairly similar. As a diagnostic, Fig. 13 displays $\chi_C(q) = \Pr\{F_i(X_i) > q \; \forall i \in C\}/(1 - q)$ for three sets with $|C| = 2$, and $C = \{1, \ldots, 31\}$. For a multivariate Pareto distribution $\chi_C(q) \equiv \chi_C$ for q sufficiently large (Rootzén *et al.*, 2018a), and the bivariate estimates from the fitted model in the paper are displayed. For model (53) with $k \in C$ and $\min_{j \in C}\{\alpha_j^k\} < 1$, $\chi_C(q) \searrow 0$ as $q \to 1$. This is often realistic for environmental data sets, though the Danube data display a high degree of extremal dependence.

Certain conditional independences could be established for model (53), but an interpretation along the lines of definition 1 in the paper is desirable. This seems to require a device like proposition 1, where the fact that $\lambda(\mathbf{y})$ does not depend on k is crucial, and it remains to be seen whether useful and coherent notions of graphical structure can be established in this case. The ideas that are presented in the paper nonetheless

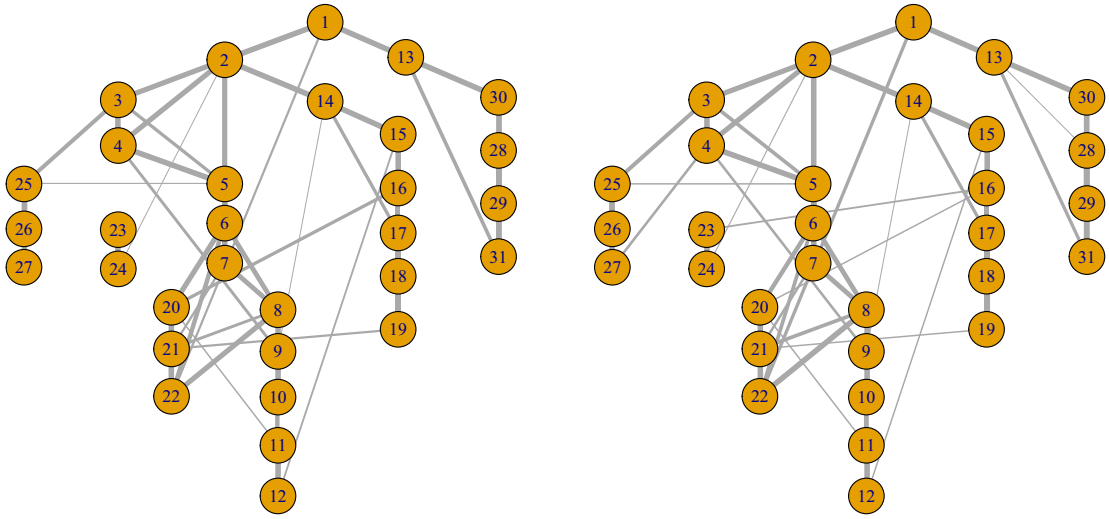


Fig. 12. (a) Connections estimated under model (53) with $\alpha^k = \mathbf{1}$ and (b) connections estimated under model (53) with α^k estimated: the line thickness is proportional to the number of times that connections were included in the graph, with only those selected at least half of the time displayed; some connections may not be visible because of the graph layout

form great inspiration for consideration of structured estimation and interpretation in the case of weaker extremal dependence.

I am very pleased to propose the vote of thanks for this thought-provoking work.

Ioannis Papastathopoulos (*University of Edinburgh*)

I congratulate Engelke and Hitz for this important contribution to an important problem. Graphical models are widely used for encoding the dependence properties of a multi-dimensional distribution and are an attractive means of dimension reduction. The key consequence of a graphical model is a reduction in the dimension of the parameter set that determines the law of the vector under study through a set of conditional independence properties induced by the graph. The authors introduce a natural concept for conditional independence in multivariate threshold exceedances and harness its implications by constructing flexible statistical models for extremes of fully asymptotically dependent random vectors. At first glance, graphical models are often criticized for being messy, yielding complicated conditions on parameters requiring plenty of skill to manipulate. There is, however, a special subclass of graphical models that brings *order* into chaos—the class of decomposable graphical models. The authors provide a construction principle for multivariate Pareto distributions on decomposable graphs and distil flexible models for block graphs. The paper is full of interesting ideas and provides fundamental insights into the class of Hüsler–Reiss Pareto distributions which is presented as an analogue of the class of multivariate Gaussian distributions for asymptotically dependent extremes.

But, let me ask, where did the Gaussian graph go? To be more specific, the methodology proposed can address only the situation where all variables in $\mathbf{X}$ grow large at the same rate, thereby excluding the possibility for modelling extremes of Gaussian random vectors. More generally, the methodology cannot encompass the modelling of asymptotically independent extremes and this brings into consideration the question of whether we can truly talk about flexible high dimensional inference when such key structure that is often exhibited in the data is not taken into account. I feel that this concern is exacerbated by the fact that the graph needs to be fully connected; an assumption that is likely to be violated when the number of variables d is large. Dr Kirstin Strokorb clarifies the interpretation of disconnected graphs under the new concept of extremal conditional independence but ramifications are still bewildering. The elegant exposition of conditional independence under full asymptotic dependence is illuminating and highly welcome, but is there not a gap in the theory that, potentially, requires refined concepts of extremal conditional

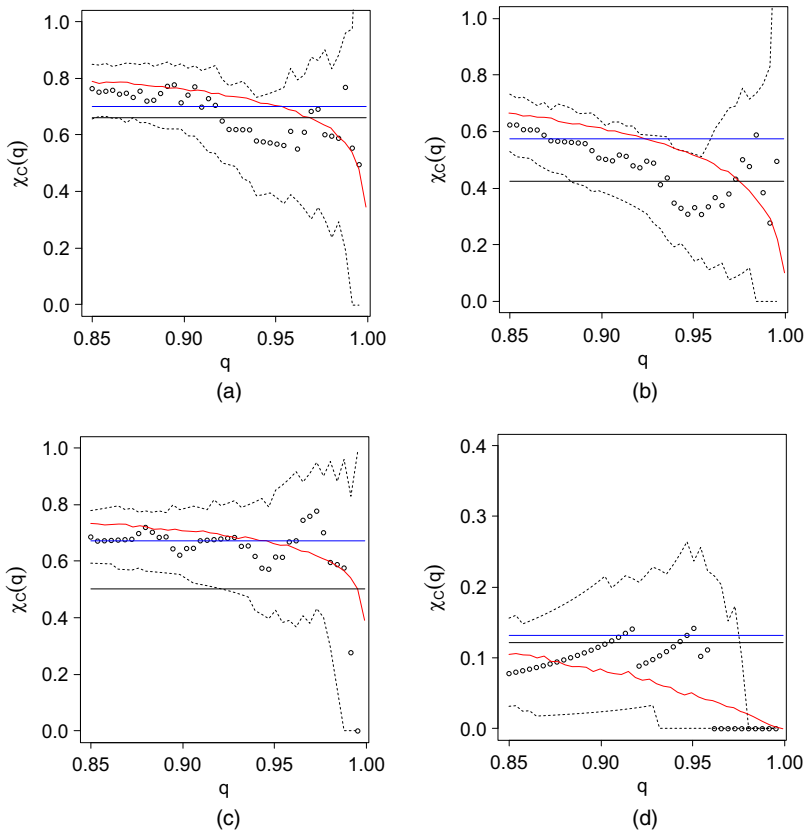


Fig. 13. Estimates of $\chi_C(q)$ for C equal to (a) $\{1,2\}$, (b) $\{1,12\}$, (c) $\{1,22\}$ and (d) $\{1,\ldots,31\}$: ○, empirical estimates: ———, estimate from model (3) with $\boldsymbol{\alpha}^k$ estimated produced conditioning on $k=1$; ———, estimate from model (3) with $\boldsymbol{\alpha}^k=\mathbf{1}$ fixed produced conditioning on $k=1$; ———, estimate from the fitted model in the paper; - - - - -, approximate 95% confidence intervals for empirical estimates obtained by using the non-parametric bootstrap

independence? Given the strong links with the conditional approach to multivariate extremes (Heffernan and Tawn, 2004) which provides a more general framework than the setting in the paper and where graphical structures are preserved in limiting conditional distributions, as for example in Papastathopoulos and Tawn (2019), I wonder, could an overarching concept of extremal conditional independence both for asymptotically dependent and asymptotically independent extremes be established? It should be emphasized that the perspective in Papastathopoulos and Tawn (2019) is different from the perspective that is adopted in the paper since a graphical structure is assumed on $\mathbf{X}$ whereas the authors rightly bypass such a restriction. Could valuable insight be gained by characterizing the domain of attraction of multivariate Pareto distributions that factorize on graphs?

Lastly, I would like to address some concerns that are usually raised about the model of Heffernan and Tawn (2004) with regard to lack of self-consistency. Assuming that $\mathbf{W}=h(\mathbf{X})$ is in standard Laplace margins, a slightly modified version of the assumption made by Heffernan and Tawn (2004) (see Keef *et al.* (2013)) is that, for any $k \in V$,

$$\left(W_k - u > x_k, \ \frac{\mathbf{W}-\mathbf{a}^k(W_k)}{\mathbf{b}^k(W_k)} \,\middle|\, W_k > u \right) \xrightarrow{\mathrm{d}} (E_k, \mathbf{Z}^k), \tag{54}$$

where E_k is unit exponential, $\mathbf{Z}^k_{\setminus k}$ is non-degenerate with no mass at ∞ and $E_k \perp\!\!\!\perp \mathbf{Z}^k_{\setminus k}$. Under asymptotic independence, it is possible to bind the limiting conditional distributions on any fixed ray (e.g. on $\{(x,y): y=x\}$) but this unfortunately restricts the dependence class. Furthermore, it turns out that it is not

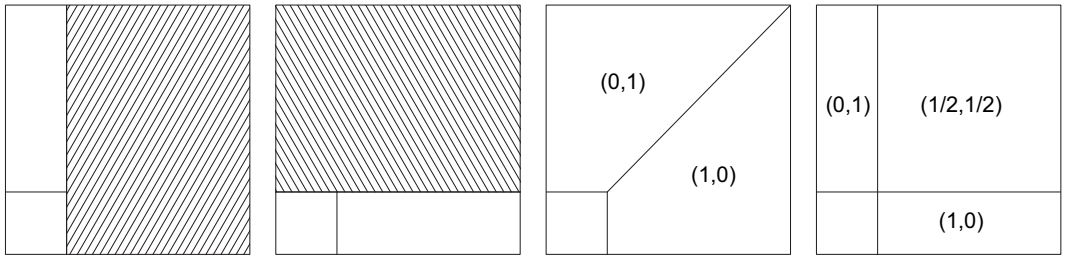


Fig. 14. (a), (b) Highlighted areas show $\{\mathbf{x} \in \mathbf{R}^2 : x_1 > u\}$ and $\{\mathbf{x} \in \mathbf{R}^2 : x_2 > u\}$; (c), (d) illustration of $(\pi_{2|1}(x_1, x_2), \pi_{1|2}(x_1, x_2))$ with $(x_1, x_2) \in \mathcal{L}^u$ under the original Heffernan and Tawn (2004) method and that in Wadsworth and Tawn (2019) respectively

possible to combine the limiting conditional distributions throughout the region where both variables are simultaneously large (Liu and Tawn, 2014). The authors allude to the limiting conditional distributions not being self-consistent under asymptotic independence which is true, yet it is my belief that this criticism often leads to a misconception about the model of Heffernan and Tawn, which, in fact, provides a valid probability measure on $\mathcal{L}^u = \{\mathbf{x} \in \mathbf{R}^d : \|\mathbf{x}\|_\infty > u\}$. For any random vector $\mathbf{W}$ admitting a positive density on $\mathbf{R}^d$, it is trivial to see that

$$f_{\mathbf{W}|\,\|\mathbf{W}\|_\infty > u}(\mathbf{x}) \propto \sum_{k \in V} \pi_{\setminus k|k}(\mathbf{x})\, f_{W_k}(x_k)\, f_{\mathbf{W}_{\setminus k}|W_k}(\mathbf{x}_{\setminus k}|x_k),$$

on $\mathcal{L}^u$, for any $\{\pi_{\setminus k|k}(\mathbf{x})\}_{k \in V}$ with $\Sigma_k \pi_{\setminus k|k}(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbf{R}^d$. This representation provides a construction principle which, for large u, permits the approximation of the joint distribution on any region $\mathcal{L}^u_k = \{\mathbf{x} : x_k > u\}$, $k \in V$, by the joint distribution of $(E, \mathbf{Z}^k_{\setminus k})$ from expression (54). In doing so, however, it is essential to avoid using the distribution of $(E_k, \mathbf{Z}^k_{\setminus k})$ on any subset of $\{\mathbf{x} : \|\mathbf{x}\|_\infty > u, x_k < u\}$ and this is accomplished by specifying the sequence of weight functions appropriately. Fig. 14 shows two possible choices for $\pi_{\setminus k|k}(\mathbf{x})$ in the bivariate case. Putting all together, this motivates the model

$$f^{\mathrm{HT}}_u(\mathbf{x}) = \sum_{k \in V} \pi_{\setminus k|k}(\mathbf{x}) \exp\{-(x_k - u)\}\, g^k_{\setminus k} \left\{ \frac{\mathbf{x}_{\setminus k} - \mathbf{a}^k_{\setminus k}(x_k)}{\mathbf{b}^k_{\setminus k}(x_k)} \right\} \Bigg/ \prod_{j \neq k} b^k_j(x_k) \qquad \text{on } \mathcal{L}^u,$$

where $g^k_{\setminus k}$ denotes the density of $\mathbf{Z}^k_{\setminus k}$ and can be thought of as a dynamic mixture model (Frigessi *et al.*, 2002). It follows that the model has a pure mixture representation and therefore defines a *valid probability measure* on $\mathcal{L}^u$. This model ensures consistency at any subasymptotic level, and any inconsistency in the limiting conditionals is entirely compatible with theory.

In summary, I found this to be an extremely stimulating paper. It is with great pleasure that I second the vote of thanks for what will be, I am certain, a very influential paper.

The vote of thanks was passed by acclamation.

Kirstin Strokorb (*Cardiff University*)

Engelke and Hitz are to be congratulated on their truly original contribution, by which they are laying the first solid foundations for graphical modelling for extreme values. I would like to reflect here briefly on a theoretical issue that was hinted at by the authors in remark 1. In the setting of the paper all extremal graphical models must be *connected* as the exponent measure Λ has been chosen to have a Lebesgue density on the non-negative orthant $\mathcal{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$. Indeed, as has been pointed out by the authors, the latter choice implies extremal dependence between all components of the random vector in question $\mathbf{X}$ and that the definition of conditional extremal independence $\mathbf{Y}_A \perp_e \mathbf{Y}_C | \mathbf{Y}_B$ cannot be extended to the case where $B = \emptyset$. For completeness, we revisit the broader framework in which the exponent measure Λ is allowed to place mass on all faces

$$\mathcal{E}^I = \{\mathbf{x} \in \mathcal{E} : \mathbf{x}_I > \mathbf{0}, \mathbf{x}_{\setminus I} = \mathbf{0}\} \qquad \text{of } \mathcal{E} = \bigcup_{\emptyset \neq I \subset \{1,\ldots,d\}} \mathcal{E}^I$$

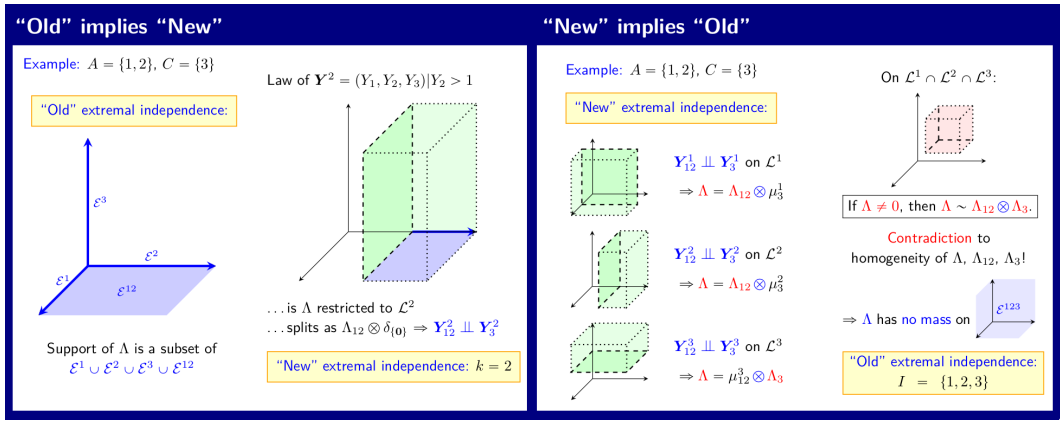


Fig. 15. Illustration of the key arguments to prove theorem 2

and let $\{1, \ldots, d\} = A \cup C$ with $A \cap C = \emptyset$. In this setting, the analogue of conditional extremal independence (17) for $B = \emptyset$ (the unconditional case) is

$$\mathbf{Y}_A^k \perp\!\!\!\perp \mathbf{Y}_C^k \qquad \text{for all } k = 1, \ldots, d.$$

It defines a *new notion of extremal independence* that we may abbreviate by $\mathbf{Y}_A \perp_e \mathbf{Y}_C$. Naturally, the question arises how it is related to the corresponding *traditional notion of extremal independence*, which can be expressed in terms of the support set of the exponent measure Λ. That is, $\mathbf{X}_A$ and $\mathbf{X}_C$ are said to be extremally independent (in the traditional sense), when the exponent measure Λ places mass only on the faces $\mathcal{E}^I$ for which $I \subset A$ or $I \subset C$. The answer to this question is as follows.

Theorem 2. The new notion of extremal independence and the traditional notion of extremal independence are equivalent.

Its proof, which does not require any considerations on densities, can be found in Strokorb (2020). The main idea is illustrated in Fig. 15.

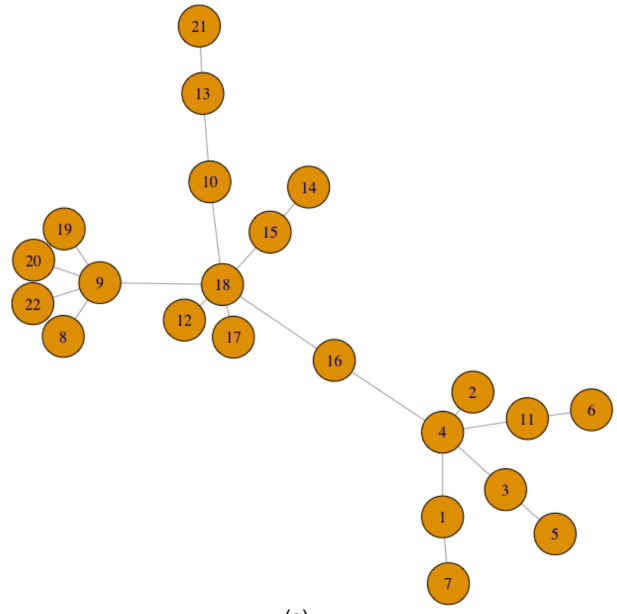
Charlotte Darné and Anthony C. Davison (*Ecole Polytechnique Fédérale de Lausanne*)
We congratulate Engelke and Hitz on this major addition to the burgeoning literature on graphical models and extreme value statistics. In other contexts graphs play a central role in model construction, because their Markov properties enable complex models to be built of simple parts, but until recently it appeared that the rigid dependence structures that are imposed by requiring extremal models to be max- or threshold stable strongly limited the possibilities for extensions to extremes. To have seen a way out of this impasse is a valuable step forward that should lead to many further developments, particularly since the construction involves the use of threshold exceedances, which allow both more detailed modelling and simpler inferences than analysis of maxima, through links to recent developments in multivariate threshold modelling (Rootzén and Tajvidi, 2006; Rootzén *et al.*, 2018a, b; Kiriliouk *et al.*, 2019). Causal modelling for extreme events should also benefit from this work.

The focus on maximum likelihood estimation is somewhat limiting, since other methods allow inference in very high dimensions (de Fondeville and Davison, 2018).

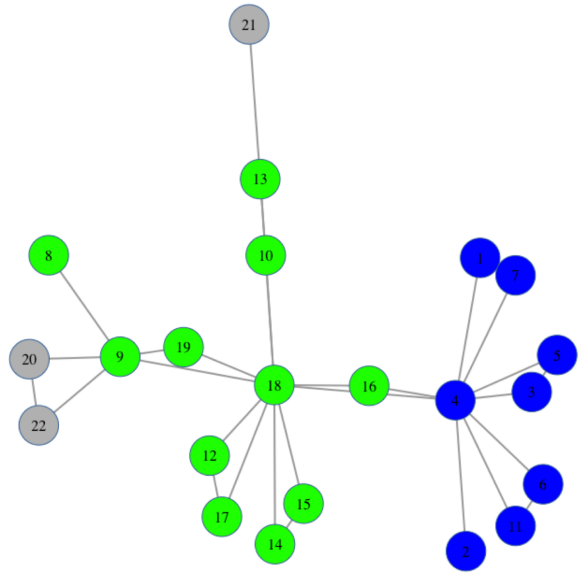
We applied the ideas to large negative daily returns for 22 banks from 2001 to 2018; 4780 time steps in all. There are US, European and three other banks: two Asian and one Canadian. After standard univariate analyses of the negative returns, the application of the methods that are suggested in the paper led to the graphs shown in Fig. 16, which shows the minimal spanning tree and the graph resulting from adding further links. The method seems to separate the European and US banks nicely, though those shown in grey seem less coherent; in particular the two Japanese banks are quite far apart.

We then split the data into the periods before and after the banking crisis of 2008. The corresponding networks, shown in Fig. 17, appear rather different, though the previous main groupings still appear, and some national links within Europe (e.g. British, French, Spanish and Swiss) persist.

This rather sketchy analysis suggests that use of the ideas in the paper gives what appear to be sensible results, but also that the restrictions of having a fully connected graph and allowing only tiny cliques



(a)



(b)

Fig. 16. Extremal analysis of the bank data—(a) minimum spanning tree estimated from days with extreme negative returns, 2001–2018, and (b) network after greedy forward selection based on AIC: banks 9, 16, 18 and 4 are respectively Santander, Deutsche Bank, Crédit Suisse and JP Morgan; in (b), the right-hand nodes are US banks, the middle nodes are European banks, and the left-hand nodes correspond to the rest of the world; the banks are the Bank of America (1), Mellon Bank of New York (2), Goldman Sachs (3), JP Morgan (4), Morgan Stanley (5), PNC Financial Services (6), Wells Fargo (7), Banco Bilbao (8), Santander (9), Barclays (10), Citigroup (11), HSBC (12), Lloyds Group (13), BNP Paribas (14), Société Générale (15), Crédit Suisse (16), UBS (17), Deutsche Bank (18), ING Bank (19), Bank of Montréal (20), DBS Group (21) and Mitsubishi Financial Group (22)

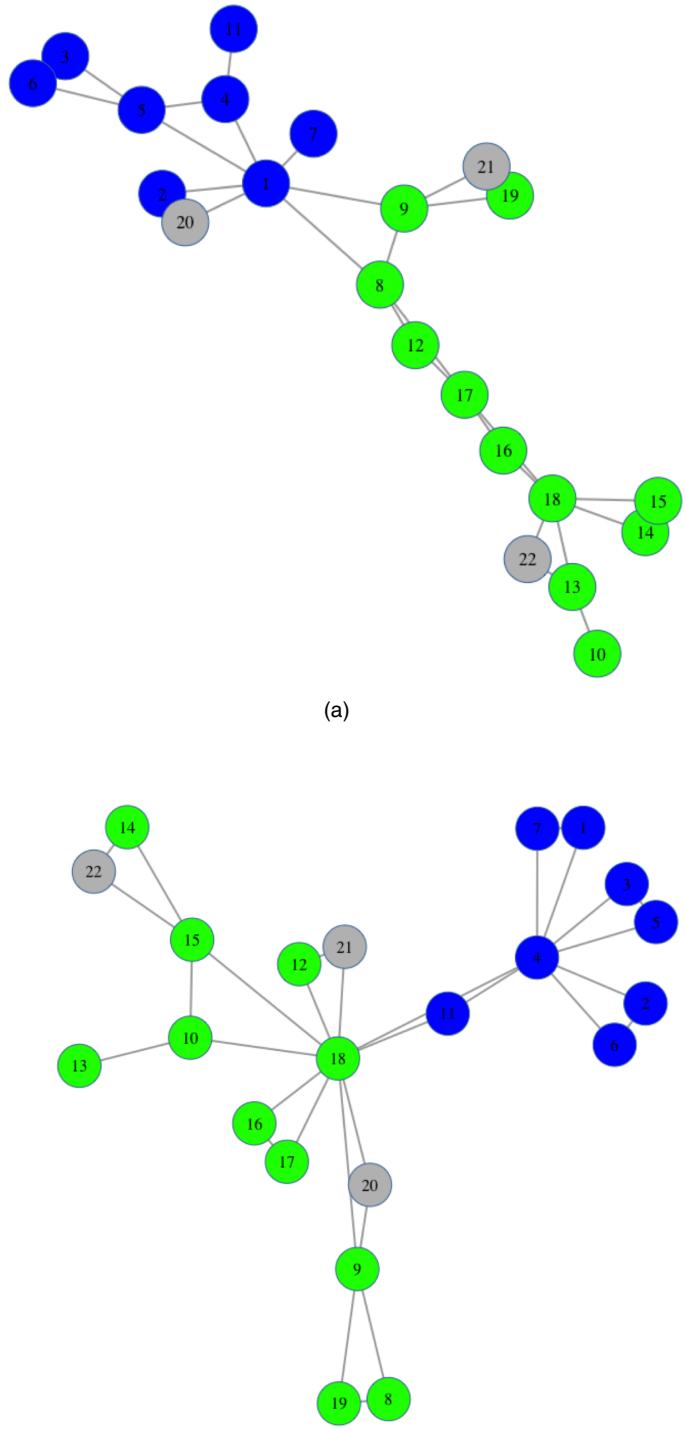


Fig. 17. Estimated extremal network for bank data for (a) 2001–2007 and (b) 2009–2018

need to be overcome; one would at least like to be able to test for more complex graphs, rather than stay close to trees. Moreover the imposition of asymptotic dependence on the variables that are represented by the nodes is a serious drawback, at least for environmental problems in which asymptotic independence of extremes seems to be quite common. Nevertheless this paper represents a large step forward and we applaud the authors for their elegant work.

Steffen Lauritzen (*University of Copenhagen*)

First I congratulate Engelke and Hitz for this inspiring paper which raises a large number of questions; this paper is destined to spawn many children. I shall focus on the properties of the independence model associated with threshold exceedances.

Recall that an abstract *independence model* $\perp_\sigma$ is a ternary relation on subsets of V. It is *semigraphoid* if for disjoint subsets A, B, C and D the following conditions hold.

Condition 1. If $A \perp_\sigma B|C$ then $B \perp_\sigma A|C$ (symmetry).

Condition 2. If $A \perp_\sigma (B \cup D)|C$ then $A \perp_\sigma B|C$ and $A \perp_\sigma D|C$ (decomposition).

Condition 3. If $A \perp_\sigma (B \cup C)|D$ then $A \perp_\sigma B|(C \cup D)$ (weak union).

Condition 4. If $A \perp_\sigma B|C$ and $A \perp_\sigma D|(B \cup C)$, then $A \perp_\sigma (B \cup D)|C$ (contraction).

It is a *graphoid* if conditions 1–4 and the following condition hold.

Condition 5. If $A \perp_\sigma B|(C \cup D)$ and $A \perp_\sigma C|(B \cup D)$ then $A \perp_\sigma (B \cup C)|D$ (intersection).

It appears that the *threshold independence* $\perp_e$ as defined in the paper obviously satisfies the semigraphoid axioms whereas the status of the intersection condition is unclear, and in particular it would be nice to see an extension of the definition which could encompass independence as well.

However the semigraphoid properties on their own indicate that sensible models along the lines in this paper may exist also for *directed acyclic graphs* (DAGs), e.g. defining multivariate Pareto or Hüsler–Reiss 'regressions' via conditional distributions $P_A^B(\cdot|x_B)$ of X_A given $X_B = x_B$ when $X_{A\cup B}$ follows a Hüsler–Reiss model. The joint distribution can then be defined recursively by suitable combination of Markov kernels $P_v^{\mathrm{pa}(v)}$, $v \in V$, along the DAG.

Since $\perp_e$ is a semigraphoid, equivalence of directed Markov properties is ensured whether or not densities are well defined (Lauritzen *et al.*, 1990) and it should be expected that, for example, if a DAG is *perfect*, i.e. all parents are married, the DAG Hüsler–Reiss model will most likely be equivalent to the decomposable version in the paper. In a non-perfect DAG, new and possibly interesting models may appear.

Rajendra Bhansali (*Imperial College London and University of Liverpool*)

This is a ground breaking paper in a mature subject area. I noted, however, that Engelke and Hitz use the Akaike information criterion AIC for comparing different multivariate Pareto models. But, this criterion is known to select 'overparameterized' models in many different situations, and its use in the present context needs further justification and investigation. Thus, for a standard auto-regressive process of order m, Shibata (1976) has demonstrated that AIC does not provide a consistent estimator of m. Moreover, following Bhansali and Downham (1977), it is possible to consider an extended AIC-criterion, AIC_α, say, in which the constant of 2 occurring in the first term on the right of their equation (40) is replaced by an arbitrary constant, α, with $\alpha > 1$. This extended criterion should have the advantage of assigning unequal weights to the 'variance' term that is represented by the first term on the right of equation (40) and the 'bias' term represented by the second term. Bhansali and Downham (1977) showed that the asymptotic probability of correctly selecting the unknown auto-regressive order m increases as α increases, but still remains fixed, and a consistent estimator of m may be obtained if $\alpha \to \infty$ simultaneously, but sufficiently slowly, with n. It is also relevant to note that a choice of $\alpha = \log(n)$ corresponds to the Bayesian information criterion of Schwarz (1980). Moreover, Shibata (1980, 1981) has demonstrated that although AIC does not provide a consistent estimator of m, it is asymptotically efficient for one-step prediction when the generating process admits an infinite auto-regressive representation which does not degenerate to a finite auto-regression; see also Bhansali (1996). It would be pertinent to examine whether similar results hold also for the model selection problem that was considered by the authors. Secondly, the authors recommend reading off the sparsity pattern of an extremal graphical model from a suitable inverse covariance matrix. For a multivariate normal distribution, however, the various elements in the inverse of the covariance matrix are known to admit a physical interpretation; see, for example, Rao (1973), page 524, Besag (1975)

and Bhansali (1990). It would be useful to investigate whether a similar interpretation also holds in the situation that was considered by the authors. Such an interpretation should enable a more systematic procedure to be developed for the problem mentioned above; the approach that is recommended by the authors, by contrast, seems rather *ad hoc*.

The following contributions were received in writing after the meeting.

Léo R. Belzile and Debbie J. Dupuis (*HEC Montréal*)

The main computational benefit of the Gaussian graphical model comes from using exclusively the precision matrix in the likelihood and employing dedicated algorithms for sparse matrices. This contrasts with the max-Pareto Hüsler–Reiss graphical model proposed where the covariance $\Sigma^{(k)}$ changes according to the component that exceeds the threshold: a different one must be used if the kth component is censored. The computation of the normalizing constant also requires numerical integration of D integrals of dimension $D-1$. The main benefit of the graphical model is that the conditional density of censored observations depends on fewer other components, reducing the effective dimension of the integrals.

To make inference feasible, Engelke and Hitz focus on block graphs and the numerical implementation in the accompanying R package restricts the maximum clique size to 3. These choices are pragmatic, but we question the robustness of the inference to this computational trick, which may yield structures that are too rigid for real data, and goodness-of-fit tools would be useful. Another aspect with large implications for inference is the standardization of the margins to unit Pareto. This standardization makes all variables equally important, possibly distorting relationships of interest. When considering the extremal dependence between world market indices, many observations that are flagged as extreme may correspond to more idiosyncratic events in smaller markets and estimated dependences between the major markets will then be weaker than we might obtain when analysing some subset of the indices.

It is thus tempting to choose a different risk functional to change the observations that are retained. Extremes are at present defined on the L-shaped space defined by $\mathcal{L} = \{\mathbf{Y} : \max_{j=1}^{D} Y_j > u\}$. Changing the risk region in the point process formulation would require the density of an observation to be scaled by the measure of the risk region, which could then be different from the exponent measure. Considering extremes in indices of major markets and the induced codependence in those of smaller markets, which is the maximum over fewer components, seems natural. The dimension of the problem is also reduced. Many risk functionals, such as $r(\mathbf{Y}) = \max_{j\in\mathcal{D}} Y_j$ for $\mathcal{D} \subset \{1, \ldots, D\}$ or $r(\mathbf{Y}) = \Sigma_{j=1}^{D} Y_j$ can be expressed as mixtures of extremal functions and explicit expressions for the normalizing constant (albeit possibly functions of high dimensional Gaussian integrals) exist. The conditional independence argument holds on these spaces by virtue of the homogeneity and rejection sampling can be employed to simulate from these models.

Adrian Casey (*University of Edinburgh*)

I congratulate Engelke and Hitz on this most important work that will no doubt inspire further exploration of Markov structures in extreme value models.

My comment is one such exploration and concerns an aspect of graphical structures in conditional extreme value models (Heffernan and Tawn, 2004). The assumption underlying these models is that, for a component X_k of a multivariate random variable $\mathbf{X}$ of dimension d with unit Laplace marginal distributions, there are normalizing functions $\mathbf{a} : \mathbb{R} \to \mathbb{R}^{d-1}$ and $\mathbf{b} : \mathbb{R} \to \mathbb{R}_{+}^{d-1}$ such that

$$\lim_{u\to\infty} \Pr\left\{ X_k - u > y, \frac{\mathbf{X}_{\setminus k} - \mathbf{a}^k(X_k)}{\mathbf{b}^k(X_k)} \leqslant \mathbf{z} | X_k > u \right\} = \exp(-y) G^k(\mathbf{z}) \qquad y > 0, \quad \mathbf{z} \in \mathbb{R}^{d-1}. \tag{55}$$

Here $G^k(\mathbf{z})$ has non-degenerate marginals and zero mass at ∞.

One problem is the identification of the residual distribution G^k; this distribution has no established family. For schemes with a Markov-type dependence, this problem may be treated by fitting low dimensional models for each clique that can then be combined to simulate from the conditional extremes distribution in equation (55) by applying a Gibbs sampling algorithm. We illustrate this idea with the simple block graph in Fig. 18.

Assuming that the model admits a density, conditioning on site 2 gives a factorized conditional density,

$$f_{\mathbf{X}_{\setminus 2}|X_2}(x_1, x_3, x_4, x_5 | x_2) = f_{X_1|X_2}(x_1|x_2) f_{X_3,X_4|X_2}(x_3, x_4|x_2) f_{X_5|X_4}(x_5|x_4). \tag{56}$$

Let G_A^k, $A \subset V \setminus \{k\}$ be the distribution that is associated with the residual random variable $\mathbf{Z}_A^k = \{\mathbf{X}_A -$

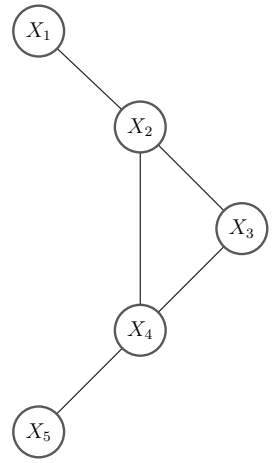


Fig. 18. Block graphical model, $\mathcal{G} = (V, E)$, $V = \{1, 2, 3, 4, 5\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{4, 5\}\}$

Table 1. Algorithm 2: a Gibbs sampler for a conditional extreme value model based on $\mathcal{G}$

1	*while samples* $< N$ *do*
2	draw $E \sim \exp(1)$ and set $X_2 = E + u$,
3	draw $Z_1^2 \sim G_1^2$ and set $X_1 = a_1^2(x_2) + b_1^2(x_2) Z_1^2$,
4	draw $\mathbf{Z}_{34} \sim G_{34}^2$ and set (X_3, X_4) as in the previous step,
5	*if* $X_4 > u$ *then*
6	draw $Z_5^4 \sim G_5^4$ and set $X_5 = a_5^4(x_4) + b_5^4(x_4) Z_5^4$;
7	*else*
8	draw a new value of X_5 from a model for the bulk conditional distribution $f_{X_5 \mid X_4}(x_5 \mid x_4)$;
9	after one sweep of the model, save as a sample

$\mathbf{a}_A(x_k)\}/\mathbf{b}_A(\mathbf{x}_k)$ for $X_k > u$ where u is a threshold that is sufficiently high for the asymptotic model equation (55) to apply.

For $x_2 > u$, the two densities $f_{X_1 \mid X_2}(x_1 \mid x_2)$ and $f_{X_3, X_4 \mid X_2}(x_3, x_4 \mid x_2)$ can be approximated by $g_1^2[\{x_1 - a_1^2(x_2)\}/b_1^2(x_2)]/b_1^2(x_2)$ and $g_{34}^2[\{\mathbf{x}_{34} - \mathbf{a}_{34}^2(x_2)\}/\mathbf{b}_{34}^2(x_2)]\{b_3^2(x_2) b_4^2(x_2)\}$. We can model $f_{X_5 \mid X_4}(x_5 \mid x_4)$ similarly, but only if $x_4 > u$. When x_4 is not in the asymptotic region, another model is required for the bulk distribution $f_{X_5 \mid X_4}(x_5 \mid x_4)$. Using these conditional distributions we can sample from $\mathbf{X}_{1345} \mid X_2 > u$ via the Gibbs sampling algorithm in Table 1.

This method has the advantage of breaking down the conditional extremes model into a series of univariate or bivariate models, at the cost of the requirement to model a bulk conditional distribution. The method is flexible in that it allows the use of differing low dimensional residual distributions G_A^k and the inclusion of covariates at each clique. Work is on going to apply these ideas to Markov random fields.

Valérie Chavez-Demoulin (*Université de Lausanne*)

I congratulate Engelke and Hitz on a very interesting paper introducing a general theory of graphical models for extremes. This discussion focuses on Section 5.3 on 'Model selection' and points towards an alternative way to present model selection for decomposable (block) graphs. In the paper, the proposed model selection consists first of building the minimum spanning tree as a baseline model for the data and 'If the model fit is not satisfactory, it is possible to extend this tree to graphs with more complex structures by adding additional edges'. The final models can be compared by using the Akaike information criterion AIC and a practical example is provided in Section 5.5. An alternative way to select models beyond the

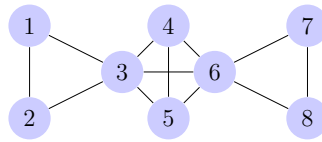


Fig. 19. Example of a block graph

family of trees may be to consider a greedy procedure that evaluates the benefit of adding or removing an edge through local computations. The approach, described in Drton and Maathuis (2017), is based on clique sum decompositions (Lauritzen (1996), chapter 3), as illustrated in the block graph of Fig. 19.

According to the clique sum of the graph of Fig. 19, the censored density satisfies

$$f^{\text{cens}}(\mathbf{y};\theta) = \frac{f^{\text{cens}}_{123}(y_1, y_2, y_3; \theta_{123}) f^{\text{cens}}_{3456}(y_3, y_4, y_5, y_6; \theta_{3456}) f^{\text{cens}}_{678}(y_6, y_7, y_8; \theta_{678})}{y_3^{-2} y_6^{-2}}$$

where the numerator corresponds to the clique prime components and the denominator to the censored marginal densities of the separators that are single nodes. 'Nested' graphs can be compared through likelihood ratio statistics. Suppose, for instance, that we want to compare this graph with the graph without the edge (3,4). Using proposition 4.30 and proposition 4.32 of Lauritzen (1996), the likelihood ratio statistics can be computed by considering only the clique (3,4,5,6), i.e. the clique to which the edge (3,4) belongs. Therefore, a formal likelihood ratio test within the clique of the clique sum decomposition seems convenient for block graphs and may be considered in practice as model selection.

D. R. Cox (*Nuffield College, Oxford*)

This impressive paper deals with extremal problems in a graph theoretical context. Some emphasis lies on the graphical structure itself. In the wide-ranging discussion of extreme values in a hydrological setting the system of observational points is usually to be regarded as given and the interest lies in the extremes of a stochastic process defined over that set of points. An exception was connected with the setting up of a rain gauge network for a river basin in South West England (Moore *et al.*, 2000) where the area was divided into smallish squares, all of which had at least one gauge and some had many to enable the exploration of dependences on different spatial scales. What design issues are implicit in the present paper?

Extremal problems in a hydrological context may involve minima, as in the study of arid and semiarid river systems or maxima as in the study of floods. Reservoir systems need study of both extremes.

Some of the properties of large systems, extreme cases being, for instance, the Amazon and the Ganges, have relationships described by Horton's (1945) laws, initially empirical relationships between position in the graphical system and flow. The connection with fundamental issues of statistical mechanics was explored by Rodriguez-Iturbe and Rinaldo (1997).

Extremal problems and graphical structures underpin these hydrological situations although with quite a different emphasis from that of the paper. Do the authors see fruitful links?

Robin J. Evans (*University of Oxford*)

I congratulate Engelke and Hitz on a thoroughly exceptional piece of work.

I would like to draw attention to the possibility of using the graphical structures they have derived for probabilistic expert systems involving extremes. We mostly use the same notation as the paper, but write $\tilde{Y}_i$ for Y_i/Y_2.

Example 1

Consider the example graph in their Fig. 3(b), reproduced here as Fig. 20, and the corresponding Hüsler–Reiss distribution used in example 9(b). Suppose that we know that $Y_2 \geqslant 1$ (so it meets a threshold for being 'extreme'); what then is the new conditional distribution of Y_1/Y_2 if we know that $\{Y_3 = 10Y_2\}$?

Recall from Section 5.4.2 in the paper that, for a Hüsler–Reiss distribution with variogram $\boldsymbol{\Gamma}$ and standard Pareto margins, the conditional distribution given $\{\log(Y_2) \geqslant y\}$ is

$$\begin{pmatrix} \log(\tilde{Y}_1) \\ \log(\tilde{Y}_3) \\ \log(\tilde{Y}_4) \end{pmatrix} = \begin{pmatrix} \log(Y_1) \\ \log(Y_3) \\ \log(Y_4) \end{pmatrix} - \begin{pmatrix} \log(Y_2) \\ \log(Y_2) \\ \log(Y_2) \end{pmatrix} \Bigg| \{\log(Y_2) \geqslant y\} \sim N_3\left(-\frac{1}{2}\Gamma_{\cdot 2}, \tilde{\Sigma}^{(2)} \right). \tag{57}$$

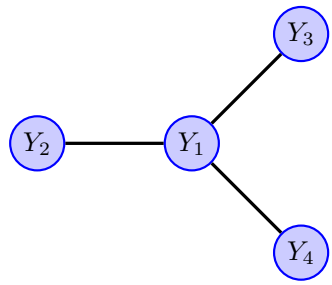


Fig. 20. Undirected graphical model used in example 1

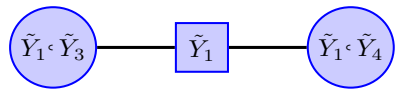


Fig. 21. Junction tree used in example 2

From this we can deduce immediately that

$$\log(\tilde{Y}_1)|\{\log(Y_2) \geqslant y, \log(\tilde{Y}_3) = x\} \sim N\left\{ -\frac{1}{2}\Gamma_{12} + \frac{\tilde{\sigma}^{(2)}_{13}}{\tilde{\sigma}^{(2)}_{33}}\left(x + \frac{\Gamma_{32}}{2}\right), \tilde{\sigma}^{(2)}_{11.3} \right\}.$$

For $x = \log(10)/y = 1$ this suggests that $\log(\tilde{Y}_1)$ will be normally distributed with mean $\log(10)/2$ and variance $\frac{1}{2}$. This matches what we observe empirically.

Example 2

The advantage of the graphical structure is that we can also use a junction tree to perform the same calculation. A junction tree for this graph is shown in Fig. 21. Suppose that we wish to estimate $\mathbb{P}(\tilde{Y}_4 \geqslant 1 | Y_2 \geqslant 1, \tilde{Y}_3 = 1)$. We can first set up the probability tables so that $(\tilde{Y}_1, \tilde{Y}_4)$ has their original distribution in equation (57) (see Lauritzen (1996) for details), i.e.

$$\begin{pmatrix} \log(\tilde{Y}_1) \\ \log(\tilde{Y}_4) \end{pmatrix} \Bigg| \{\log(Y_2) \geqslant y\} \sim N_2\left(-\frac{1}{2}\Gamma_{14,2}, \tilde{\Sigma}^{(2)}_{14,14} \right),$$

and similarly for $(\tilde{Y}_1, \tilde{Y}_3)$ and the separator $\tilde{Y}_1$. If we condition on $\{\tilde{Y}_3 = 1\}$ then

$$\log(\tilde{Y}_1)|\{\log(Y_2) \geqslant y, \log(\tilde{Y}_3) = 0\} \sim N\left(-\frac{1}{2}\Gamma_{12} + \frac{\tilde{\sigma}^{(2)}_{13}}{2\tilde{\sigma}^{(2)}_{33}}\Gamma_{32}, \tilde{\sigma}^{(2)}_{11.3} \right).$$

Passing a message through the separator set means we end up with

$$\begin{pmatrix} \log(\tilde{Y}_1) \\ \log(\tilde{Y}_4) \end{pmatrix} \Bigg| \{\log(Y_2) \geqslant y, \tilde{Y}_3 = 1\} \sim N_2\left(-\frac{1}{2}\Gamma_{14,2} + \frac{1}{2\tilde{\sigma}^{(2)}_{33}}\tilde{\Sigma}^{(2)}_{14,3}\Gamma_{32}, \tilde{\Sigma}^{(2)}_{14,14.3} \right).$$

In our case this is

$$\log(\tilde{Y}_4)|\{\log(Y_2) \geqslant y, \tilde{Y}_3 = 1\} \sim N\left(-\frac{1}{2}, \frac{3}{2} \right).$$

A Kolmogorov–Smirnov test starting with 2×10^7 independent samples (and leaving 44402 observations after conditioning on $Y_2 \geqslant 1$ and $\tilde{Y}_3 \in [0.99, 1.01]$) suggests that this is indeed the correct distribution ($p = 0.34$).

The primary benefit of the derivation above is that we never have to consider the full joint distribution, which may be prohibitively large. In particular, $\tilde{Y}_3$ and $\tilde{Y}_4$ never need to be considered together. For larger graphs with relatively small cliques, this may be a crucial advantage.

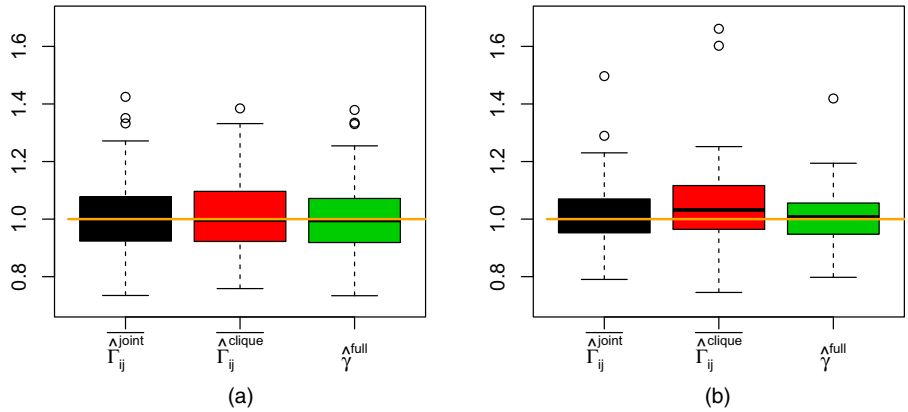


Fig. 22. Boxplots of estimates of γ based on the three estimators mentioned in the text, for trees in dimensional (a) $d = 5$ and (b) $d = 16$: ——, true value $\gamma = 1$

Raphaël Huser and Daniela Cisneros (*King Abdullah University of Science and Technology, Thuwal*)
We congratulate Engelke and Hitz on this impressive and timely paper, which makes a giant step forward in modelling and inference for high dimensional extremes, opening avenues for future research in causal inference and machine learning for extremes. State of the art methods in statistics of extremes have so far been limited to very low dimensions when modelling maxima (Castruccio *et al.*, 2016; Huser *et al.*, 2019) and moderately high dimensions when modelling high threshold exceedances (de Fondeville and Davison, 2018). By defining notions such as 'conditional independence for multivariate Pareto distributions' and 'extremal graphical models', Engelke and Hitz have developed a new framework for building and learning sparse models for multivariate extremes in potentially very high dimensions.

Our comments mostly revolve around the extension of the proposed methodology to the spatial extremes context. A crucial assumption in theorem 1 is graph decomposability. Although this assumption includes important classes (e.g. trees and block graphs), it is fairly restrictive for spatial data. For example, when spatial locations form a regular lattice in $\mathbb{R}^2$, any non-trivial tree structure violates spatial stationarity. The most natural spatially structured Markov random field on a regular lattice would assign the same distribution to all first-order neighbours and assume conditional independence for all other pairs of variables (i.e. the edge set contains only first-order neighbours). Such a graph is, however, *not* decomposable, and it is currently unclear to us whether or how the results can be extended to this case. However, to retain computational tractability, a promising approach might be to consider finite mixtures of tree-based multivariate Pareto distributions similarly to Yu *et al.* (2017) and Vettori *et al.* (2020).

A way to impose some spatial structure for data on a regular lattice modelled with tree-based multivariate Pareto distributions is to assume the *same* dependence parameter γ for each clique. In this context, the efficiency of the proposed cliquewise inference approach is no longer obvious. To assess this, we replicated the simulation study in Section 5.5, now specifying the same value $\gamma = \Gamma_{12} = \Gamma_{13} = \Gamma_{24} = \Gamma_{25}$ for each clique. For illustration, we chose $\gamma = 1$. We then estimated γ with the average $\hat{\Gamma}_{ij} = (\hat{\Gamma}_{12} + \hat{\Gamma}_{13} + \hat{\Gamma}_{24} + \hat{\Gamma}_{25})/4$, obtained by maximizing either the joint or the cliquewise censored likelihoods (treating the Γ_{ij}s as distinct free parameters), and by fitting the 'full' model enforcing the same value γ on all cliques $\{i, j\}$. The average of cliquewise estimates, $\hat{\Gamma}_{ij}^{\text{clique}}$, is comparable with using a pairwise likelihood with pairs chosen as the cliques themselves. We also explored a larger 16-dimensional tree structure defined on a 4×4 spatial lattice. The results, reported in Fig. 22, suggest that the estimator $\hat{\Gamma}_{ij}^{\text{clique}}$ performs almost as well as $\hat{\gamma}^{\text{full}}$ (based on the full model), though the former has a larger loss of efficiency compared with the latter in higher dimensions (about 9% when $d = 5$ *versus* 33% when $d = 16$, in terms of root-mean-squared error). Nevertheless, the approach proposed is much faster and still quite accurate in this context.

Jevgenijs Ivanovs (*Aarhus University*)
I would like to point out that the methodology that is developed by Engelke and Hitz has applications beyond extreme value theory. Essentially, the authors define conditional independence and then also graphical models for a *homogeneous Radon measure* Λ on the positive orthant exploding at the origin. Homogeneity is the key property enabling us to extend the usual notion of conditional

independence in a certain subspace of product form to the whole space. Importantly, such Radon measures are central not only to extremes but also to *stable Lévy motions* and related objects including stable random measures, processes and distributions (Samorodnitsky and Taqqu, 1994). Certain similarities between max- and sum stable theories have been discussed by, for example, Resnick (2007) and Kabluchko (2009).

Stable Lévy motions appear in various limit theorems supplementing Brownian motion in the cases when, for example, heavy tails are present. Importantly, a d-dimensional stable motion $(S_t)_{t\geqslant 0}$ is characterized (up to a drift parameter) by a $-\alpha$-homogeneous Radon measure Π on $\mathbb{R}^d\backslash\{0\}$ with $\alpha\in(0,2)$; see Sato (2013), theorem 14.3. The ideas of Engelke and Hitz apply with some minor changes in this setting as well, thus leading to graphical models for stable motions and stable distributions. As in extremes, Π restricted to some set bounded away from the origin has a clear interpretation: it gives the distribution of jumps of S restricted to this set times their rate. Thus conditional independence is defined not for the value of the stable motion S at fixed times, but rather for the jumps (on every scale) assembling into the motion S, which relates to Lévy–Itô decomposition of a Lévy process (Sato (2013), theorem 19.2). There are various further interpretations of Π concerning jumps at small times as well as extremal behaviour of the respective stable variable S_1 which links back to the present paper.

Claudia Klüppelberg (*Technical University of Munich*)

Engelke and Hitz are to be congratulated for developing an innovative approach to graphical extreme value models. This comment focuses on differences between the framework of the authors and an alternative approach (also motivated by extreme value theory) in Gissibl and Klüppelberg (2018). I shall use the same notation as the authors.

In Gissibl and Klüppelberg (2018) a (causal) max-linear Bayesian network is proposed:

$$X_i = \bigvee_{j\in \mathrm{an}(i)\cup\{i\}} b_{ji} Z_j, \qquad i\in V, \tag{58}$$

with ancestors $\mathrm{an}(i)$ in a directed acyclic graph and $\mathbf{Z}$ has independent positive and atom-free components.

Both the authors' approach and that of Gissibl and Klüppelberg (2018) lead to graphical models: the authors work with densities on undirected graphs, and the model in Gissibl and Klüppelberg (2018) has discrete dependence structure on a directed acyclic graph.

As a first highlight, the authors prove for a decomposable graph that the density of $\mathbf{Y}$ factorizes according to the graph and this is equivalent to an appropriate Markov property. In Klüppelberg and Lauritzen (2019) it was observed that d-separation typically will not identify all valid conditional independence relations, and in Améndola *et al.* (2020) we give a complete description of all conditional independence relations via a new separation criterion. We also describe how extreme events spread through the network.

Estimation of both models is different. The authors propose extreme value densities yielding conditional independence properties, the Hüsler–Reiss distribution being most promising. As only low dimensional densities are computationally feasible, the authors restrict themselves to decomposable graphs with singleton separator sets, so-called block graphs, starting with bivariate densities and trees.

The max-linear Bayesian network model (58) has no densities and the model parameters are the max-linear coefficients b_{ji} regardless of the independent and identically distributed components of $\mathbf{Z}$. These parameters can be estimated by a minimum ratio estimator, which has nice properties (Gissibl *et al.*, 2019). Alternative approaches assume regular variation of $\mathbf{Z}$; see Gissibl *et al.* (2018) and Klüppelberg and Krali (2019).

The final task is learning the graph. The authors focus on a model factorizing on a block graph so that they can apply classical greedy algorithms. In Buck and Klüppelberg (2020) an algorithm is suggested to estimate a topological order of the nodes first and afterwards the b_{ji}, applying a hard threshold to identify the non-edges, but ensuring a max-linear Bayesian network.

Both new graphical models have their pros and cons, the graphical model of the authors being closer to classical Gaussian models, and the Bayesian network of Gissibl and Klüppelberg (2018) having discrete dependence structure. Both approaches are starting points of graphical models for extremes, and it will be interesting to follow their future development.

Jorge Mateu (*University Jaume I, Castellón*) **and Matthias Eckardt** (*Humboldt-Universität zu Berlin*)

Engelke and Hitz are to be congratulated on a valuable contribution and thought-provoking paper on graphical models for extremes which bridges the prominent inverse variance lemma to the non-Gaussian field case embedded in the field of extreme value theory. In particular, a new undirected graphical model in

the spirit of undirected Gaussian graphical models termed the *extreme graphical model* is proposed, which enables factorization into the marginal of the exponent measure density similarly to the Hammersley–Clifford theorem. Assuming a d-dimensional Hüssler–Reiss Pareto specification, the authors show that conditional independence between distinct components can be determined directly from the inverse of the covariance matrix of the exponent measure density. This is certainly a very timely and promising topic in the era of data science and big data analysis. Graphical models treat conditional independence in a more natural and easy going way, and, under the presence of large amounts of data and variables, conditioning is the right way to go.

Our discussion focuses on the linkage between undirected graphical models and data defined on spatial and spatiotemporal domains which enable the identification of conditional independence statements among sets of (potentially different types) of components. Under such a framework, some components might be of point process nature or recorded over a countable set of interconnected areal entities. In particular, having hydrographic or traffic data under study, the domain itself obeys a particular network structure. For such data, one is interested in the structural exploration and the detection or extraction of the characteristics and features within and between distinct components conditionally on all alternative components. Although a large body of literature on the analysis of spatial point patterns exists, applications of graphical models apart from the Markov random field and Besag's conditional auto-regression (Besag, 1974, 1977) specifications still remain very limited, and clearly are methodologically and computationally challenging. One exception defined through the partial spectral characteristics are Eckardt and Mateu (2017, 2018, 2019a, b) and Eckardt *et al.* (2020). We want to point the authors to the new research avenues of extreme data analysis using max-stable processes on network support with interesting applications in traffic mortality data and crime events occurring on roads and city streets. In this line, directed graphs are more naturally happening, and extensions of the above references to partial spectral tools for directed graphs are very welcome.

Nicolas Meyer and Olivier Wintenberger (*Sorbonne Université, Paris*)

Independence and conditional independence require distributions supported on product spaces. Engelke and Hitz define a notion of conditional independence for a multivariate Pareto distribution $\mathbf{Y} = \lim_{u\to\infty} \mathbf{X}/u \mid \|\mathbf{X}\|_\infty > u$ when

$$\mathbb{P}(\mathbf{Y} \in \tilde{\mathcal{E}}) = 1, \tag{59}$$

where $\tilde{\mathcal{E}} = (0,\infty)^d = \{\mathbf{x} \in [0,\infty)^d, \min_{1\leqslant k\leqslant d} x_k > 0\}$. By conditioning the limit vector $\mathbf{Y}$ on the event that $\{Y_k > 1\}$ the authors actually work on the product space $\mathcal{L}^k = \{\mathbf{x} \in \mathcal{E}, x_k > 1\}$ with $\mathcal{E} = [0,\infty)^d \backslash \{\mathbf{0}\}$.

The definition of $\tilde{\mathcal{E}}$ encourages the use of a characterization of regularly varying random vectors in terms of the minimum of their marginals. A random vector $\mathbf{X} \in \mathcal{E}$ is regularly varying on the space $\tilde{\mathcal{E}}$ if and only if

$$\min_{1\leqslant k\leqslant d} X_k \text{ is regularly varying and } (u^{-1}\mathbf{X} \mid \min_{1\leqslant k\leqslant d} X_k > u) \xrightarrow{\mathrm{d}} \mathbf{Y}', \qquad u \to \infty, \tag{60}$$

where 'd' denotes convergence in distribution and $\mathbf{Y}'$ takes values in the product space

$$\{\mathbf{x} \in \mathcal{E}, \min_{1\leqslant k\leqslant d} x_k > 1\} = (1,\infty)^d = \bigcap_{1\leqslant k\leqslant d} \mathcal{L}^k;$$

see Segers *et al.* (2017), proposition 3.1. Restricting the regular variation condition of $\mathbf{X}$ to the space $\tilde{\mathcal{E}}$ is necessary to capture its asymptotic behaviour through the events $\{\min_{1\leqslant k\leqslant d} X_k > u\}$ for $u > 0$.

In the framework of Engelke and Hitz (i.e. regular variation on $\mathcal{E}$ and assumption (59)) both conditions in expression (60) hold and the limit vector $\mathbf{Y}'$ corresponds to the vector $\mathbf{Y}$ conditioned on the event that $\{\min_{1\leqslant k\leqslant d} Y_k > 1\}$; see Fig. 23. Conversely, the two assumptions in expression (60) are more general since they include for instance the case where the marginals of $\mathbf{X}$ are independent (and then $\mathbf{Y}$ exists and its distribution concentrates on the axes).

An alternative notion of conditional independence for a multivariate Pareto distribution can therefore be defined through conditional independence of $\mathbf{Y}'_A$ and $\mathbf{Y}'_C$ given $\mathbf{Y}'_B$. Such a property holds if the density $f_{\mathbf{Y}'}$ of $\mathbf{Y}'$ factorizes as

$$f_{\mathbf{Y}'}(\mathbf{y}') f_{\mathbf{Y}',B}(\mathbf{y}'_B) = f_{\mathbf{Y}',A\cup B}(\mathbf{y}'_{A\cup B}) f_{\mathbf{Y}',B\cup C}(\mathbf{y}'_{B\cup C}), \qquad \mathbf{y}' \in (1,\infty)^d.$$

Regarding the original vector $\mathbf{X}$, the study of $\mathbf{Y}'$ instead of $\mathbf{Y}$ provides two advantages. First, the vector $\mathbf{Y}'$ models the extremal behaviour of $\mathbf{X}$ when all its marginals are simultaneously large. It therefore

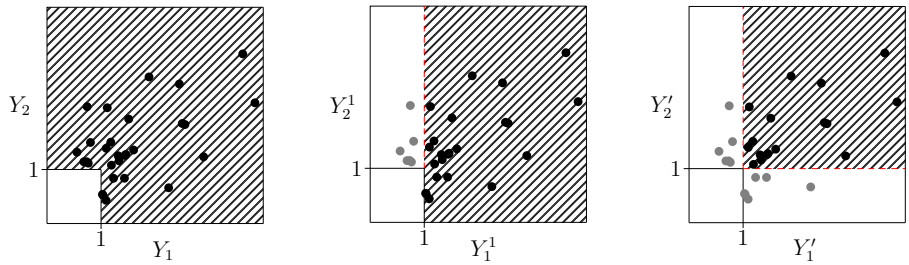


Fig. 23. The shaded areas correspond to (a) the support of $\mathbf{Y}$, (b) the support of $\mathbf{Y}^1$ and (c) the support of $\mathbf{Y}'$

provides accurate models for extremal dependent data. If the strong condition that all marginals are extreme together is not satisfied, then condition (60), and thus condition (59), is very unlikely. Second, conditional independence of $\mathbf{Y}'$ can be interpreted in terms of $\mathbf{X}$. Indeed if $\mathbf{X}_A$ is conditionally independent of $\mathbf{X}_C$ given $\mathbf{X}_B$ then $\mathbf{Y}'_A$ is conditionally independent of $\mathbf{Y}'_C$ given $\mathbf{Y}'_B$. In the case $B = \emptyset$ we obtain independence of $\mathbf{X}_A$ and $\mathbf{X}_C$ and therefore of $\mathbf{Y}'_A$ and $\mathbf{Y}'_C$, whereas $\mathbf{Y} \in \mathcal{E} \backslash \tilde{\mathcal{E}}$ (as soon as $\mathbf{Y}$ exists). There, condition (60) is satisfied but not condition (59).

Linda Mhalla (*HEC Montréal*)

I congratulate Engelke and Hitz on an interesting and inspiring contribution to multivariate extreme value theory. I would like to issue a challenge on graphical models for tail distributions. It seems that an important and natural next step would be an extension of the developed methodology to the setting where the random vector $\mathbf{X}^{\mathrm{P}} = (X_1^{\mathrm{P}}, \ldots, X_d^{\mathrm{P}})$ with Pareto margins is asymptotically independent, i.e. the setting where the exponent measure Λ places mass on lower dimensional faces of the cone $\mathcal{E} = [0, \infty)^d \backslash \{\mathbf{0}\}$ and where the factorization of the multivariate Pareto density according to a connected graph is no longer possible. Under such a setting, Wadsworth and Tawn (2013) proposed to look at the joint tail probability decay rate, under different fixed marginal growth rates. Exploiting the notion of hidden regular variation, they represented the joint tail probability of a positive quadrant dependent $\mathbf{X}^{\mathrm{P}}$ as

$$\Pr(X_1^{\mathrm{P}} > t^{\beta_1}, \ldots, X_d^{\mathrm{P}} > t^{\beta_d}) \sim \mathcal{L}(t; \boldsymbol{\beta}) t^{-\kappa(\beta_1, \ldots, \beta_d)},$$

where $\mathcal{L}$ is a slowly varying function in t and κ links the different marginal growth rates $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}_+^d \backslash \{\mathbf{0}\}$ to the decay rate in the tail dependence. The homogeneous function κ describes the dependence at subasymptotic levels and plays thus the role of the exponent measure Λ in the asymptotic dependence regime. For instance, in the special case of an inverted extreme value distribution, these quantities are related through $\kappa(\beta_1, \ldots, \beta_d) = \Lambda(1/\beta_1, \ldots, 1/\beta_d)$. Given this link, would it be possible to define an alternative notion of conditional subasymptotic independence for extremes? Instead of considering the coarse vague convergence of the measures $t\{1 - \Pr(\mathbf{X}^{\mathrm{P}} \leqslant t\mathbf{z})\}$, for $\mathbf{z} \in \mathcal{E}$, one could examine the limiting measure (when $t \to \infty$) of the survival probability

$$\Pr\left\{ \left(\frac{X_1^{\mathrm{P}}}{t^{\omega_1}}, \ldots, \frac{X_d^{\mathrm{P}}}{t^{\omega_d}} \right) \in \cdot \,\middle|\, \left(\frac{X_1^{\mathrm{P}}}{t^{\omega_1}}, \ldots, \frac{X_d^{\mathrm{P}}}{t^{\omega_d}} \right) \in [1, \infty)^d \right\},$$

which, in the bivariate setting, enjoys the homogeneity property desired. Whereas the theory that is presented by the authors relies on the uniquely determined limiting exponent measure, results in terms of κ need to be formulated with care as they would depend on the different marginal normalizations. Finally, if a measure of the correct sparsity structure of Λ could be derived from an edge weighting procedure, the graph resulting from κ would complement the graph derived by the authors.

Nancy Reid (*University of Toronto*)

My comment concerns inference for θ, addressed in Section 5.2 and in simulations in Section 5.4. It seems natural to make a link to composite likelihood, as the product of the terms over $C \in \mathcal{C}$ is of the form of a composite likelihood, and inferential properties of composite likelihood estimators and tests are available from general theory (e.g. Varin (2008)). The composite likelihood function here, $\mathrm{CL}(\theta) \propto$

$\Pi_C\{\lambda_C(\mathbf{y}_C;\theta_C)/\Lambda(1;\theta_C)\}$, is somewhat unique. First, the parameters in each component are completely separate from the other components, as $\theta=\{\theta_C, C\in\mathcal{C}\}$, so there is no 'overlap', and in principle this should make composite likelihood inference quite effective. Second, it seems that the composite likelihood function may not satisfy the first Bartlett identity, $E[\partial_\theta \log\{\mathrm{CL}(\theta;\mathbf{Y})\}]=E[\Sigma_C \partial_{\theta_C}\log\{f_C(\mathbf{y}_C;\theta_C)\}]=0$, because this expectation is taken under the full joint distribution for $\mathbf{Y}$. This is consistent with the observations made by the authors that Z_θ varies slowly with θ, and that without incorporating censoring the point estimators are biased. The rate of change of Z_θ can be related directly to the Bartlett identities with elementary calculations:

$$\partial_\theta \log(Z_\theta)=E(\Sigma_C[\partial_{\theta_C}\log\{f_C(\mathbf{y}_C;\theta_C)\}]),$$
$$\partial^2_{\theta\theta^{\mathrm{T}}}\log(Z_\theta)=\mathrm{var}(\Sigma_C[\partial_{\theta_C}\log\{f_C(\mathbf{y}_C;\theta_C)\}])+E(\Sigma_C[\partial^2_{\theta_C\theta_C^{\mathrm{T}}}\log\{f_C(\mathbf{y}_C;\theta_C)\}]).$$

This might be useful in suggesting an estimator for the variance of the point estimator of θ, although it may be that inference for θ is less important than model selection in many practical contexts. Pseudo (composite) likelihood inference has been used in work on spatial extremes in different contexts: Schlather and Tawn (2003) used it for estimation of the extremal coefficient and Davison *et al.* (2012), section 6.2, and references therein described its use in fitting models for spatial extremes.

Christian Y. Robert (*École National de la Statistique et de l'Administration Economique, Paris, and Université Claude Bernard Lyon 1*)

I congratulate Sebastian Engelke and Adrien Hitz for a stimulating paper in which they nave made a substantial contribution to the problem of modelling high dimensional extremal dependence structures with graphical models. I shall discuss two aspects about the radial decomposition of a multivariate extreme value distribution and the choice of the norm for usual graphical models to be considered.

Let '$\|\cdot\|$' be a norm on $\mathbb{R}^d$, $\mathcal{N}=\{\mathbf{x}\in\mathcal{E}:|\mathbf{x}|=1\}$ and S a finite measure on $\mathcal{N}$ such that $\int_{\mathcal{N}} x_i S(\mathrm{d}\mathbf{x})=1, 1\leqslant i\leqslant d$. The following statements are equivalent:

(a) $\mathbf{X}$ is in the max-domain of attraction of $\mathbf{Z}$;
(b) $u\mathbb{P}(\|\mathbf{X}\|/u, \mathbf{X}/\|\mathbf{X}\|\in\cdot)\to^{\mathrm{v}} r^{-2}\mathrm{d}r\times S$ on $(0,\infty]\times\mathcal{N}$ as $u\to\infty$, where '$\to^{\mathrm{v}}$' denotes the vague convergence on $\mathcal{E}$.

The latter provides a radial decomposition of the distribution of the large values of $\mathbf{X}$ and disentangles the norm component and the angular component. As a consequence, we can define a random variable $\mathbf{S}$ on $\mathcal{N}$ in the following way:

$$\mathbb{P}(\mathbf{S}\in\cdot)=\lim_{u\to\infty}\mathbb{P}(\mathbf{X}/\|\mathbf{X}\|\in\cdot|\|\mathbf{X}\|>u)=S(\cdot)/S(\mathcal{N})$$

which is independent on the limit distribution of the norm. In the paper, $\mathbf{X}$ is normalized by u (rather than $\|\mathbf{X}\|$) and the infinite norm $\|\cdot\|_\infty$ is favoured to obtain the multivariate Pareto distribution. This choice does not allow separation of the radial components. This has the consequence that all components of $\mathbf{Y}$ are dependent and leads the authors to propose an adapted definition of conditional independence for such distributions. Of course working with $\mathbf{S}$ brings issues because of its value space $\mathcal{N}$.

L_1-norm

With $\|\mathbf{x}\|=\Sigma_{i=1}^n|x_i|$, the realizations of $\mathbf{S}$ are compositional data. The Aitchison geometry and well-characterized isomorphisms that transform the Aitchison simplex to $\mathbb{R}^{d-1}$ can be used to consider graphical models on $\mathbb{R}^{d-1}$. I wonder what the meaning of conditional independence for distributions on such a space would be and whether there are links with the extremal conditional independence $\perp_{\mathrm{e}}$ introduced by the authors.

The geometric mean pseudonorm and the Hüsler–Reiss distribution

Let $\mathbf{W}$ be a centred normal random vector with invertible variance matrix Σ and $\mathbf{U}=\exp\{\mathbf{W}-\mathrm{diag}(\Sigma)/2\}$. Assume that $\mathbf{X}=P\mathbf{U}$, where P is a random variable with unit Pareto distribution and independent of $\mathbf{U}$. $\mathbf{X}$ is in the max-domain of attraction of the Hüsler-Reiss distribution. With $\|\mathbf{x}\|=(\Pi_{i=1}^d|x_i|)^{1/d}$ (which is only a seminorm but can be used to define a polar co-ordinate transformation on $(0,\infty)^d$), $\log(\mathbf{S})$ has a multivariate normal distribution characterized by

$$\log(\mathbf{S})\stackrel{\mathrm{d}}{=}(\mathbf{I}_d-d^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\left(\mathbf{W}-\frac{1}{d}\Sigma\mathbf{1}\right)-\frac{1}{2}\mathrm{diag}\left\{\Sigma-\frac{1}{d}\mathrm{tr}(\Sigma)\mathbf{I}_d\right\}$$

whose variance matrix is of rank $d-1$. In this particular case, graphical Gaussian models could be of

interest also (e.g. after eliminating in a suitable way one component of **S**). This remark can be connected to the discussion about Fig. 9(b) in the paper.

Johan Segers (*Université catholique de Louvain, Louvain-la-Neuve*)

Starting from the observation in Papastathopoulos and Strokorb (2016) that, for max-stable distributions, conditional independence only leads to trivial structures, Engelke and Hitz propose in their definition I a novel concept of conditional independence for the associated multivariate Pareto distributions. The idea paves the way for modelling extremal dependence between many variables linked through a physical or conceptual network.

A convenient parametric model is the d-variate Hüsler–Reiss Pareto distribution (named after Hüsler and Reiss (1989)) with variogram matrix $\Gamma \in \mathbb{R}^{d\times d}$ as in Section 4.3. For a vector $\mathbf{Y}$ with such a distribution, proposition 3 in the paper characterizes conditional independence $Y_i \perp_e Y_j | \mathbf{Y}_{V\setminus\{i,j\}}$ for $i, j \in V := \{1, \ldots, d\}$ in terms of $(d-1)\times(d-1)$ dimensional precision matrices $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ for $k \in V$, where $\Sigma^{(k)}$ is the covariance matrix indexed by $V\setminus\{k\}$ with elements $\Sigma^{(k)}_{ij} = 0.5(\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij})$. For $i, j \in V\setminus\{k\}$, the criterion is simply that $\Theta^{(k)}_{i,j} = 0$ but, if $i = k \neq j$, the criterion is that $\Sigma_{l\neq k}\Theta^{(k)}_{lk} = 0$, and similarly if $j = k \neq i$.

A natural question is whether there is a single $d \times d$ covariance matrix Γ with variogram matrix Γ such that conditional independence within $\mathbf{Y}$ is equivalent to the presence of a zero entry in the precision matrix $\Theta = \Sigma^{-1}$:

$$\forall i, j \in V, i \neq j: \quad Y_i \perp_e Y_j | \mathbf{Y}_{V\setminus\{i,j\}} \Leftrightarrow \Theta_{i,j} = 0. \tag{61}$$

It turns out that there are many such covariance matrices Σ. Fix $k \in V$ and let $\mathbf{Z}^{(k)}$ be a $(d-1)$-variate centred normal random vector with covariance matrix $\Sigma^{(k)}$; the elements of $\mathbf{Z}^{(k)}$ are indexed by $V\setminus\{k\}$. Further, let ε be an independent centred normal random variable with variance $\sigma^2 > 0$ and define the d-variate normal vector $\mathbf{X}$ by $X_i = Z^{(k)}_i + \varepsilon$ if $i \in V\setminus\{k\}$ and $X_k = \varepsilon$. The covariance matrix Σ of $\mathbf{X}$ has elements $\Sigma_{i,j} = \Sigma^{(k)}_{ij} + \sigma^2$ for $i, j \in V$, with $\Sigma^{(k)}_{ij} = 0$ as soon as $i = k$ or $j = k$. Then the precision matrix $\Theta = \Sigma^{-1}$ satisfies condition (61) above.

To see why, suppose for convenience that $k = d$. Writing $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{d-1}$, we have

$$\Sigma = \begin{pmatrix} \Sigma^{(d)} + \sigma^2 \mathbf{1}\mathbf{1}^{\mathrm{T}} & \sigma^2 \mathbf{1} \\ \sigma^2 \mathbf{1}^{\mathrm{T}} & \sigma^2 \end{pmatrix}.$$

Its precision matrix is

$$\Theta = \Sigma^{-1} = \begin{pmatrix} \Theta^{(d)} & -\Theta^{(d)}\mathbf{1} \\ -\mathbf{1}^{\mathrm{T}}\Theta^{(d)} & \mathbf{1}^{\mathrm{T}}\Theta^{(d)}\mathbf{1} + \sigma^{-2} \end{pmatrix}.$$

In particular, for distinct indices $i, j \in V$, we have

$$\Theta_{ij} = \begin{cases} \Theta^{(d)}_{ij} & \text{if } i \neq d \text{ and } j \neq d, \\ -\sum_{l=1}^{d-1} \Theta^{(d)}_{lj} & \text{if } i = d \text{ and } j \neq d, \\ -\sum_{l=1}^{d-1} \Theta^{(d)}_{il} & \text{if } i \neq d \text{ and } j = d. \end{cases}$$

The claimed equivalence relationship (61) then follows from proposition 3 in the paper.

Phyllis Wan (*Erasmus University Rotterdam*)

I congratulate the authors on a fine endeavour of bringing together two exciting areas. This opens a door for extreme value analysis to the vast literature of the graphical modelling approach.

In the paper, the defined framework and the theoretical results are based on the assumption that the tail spectral density of the random vector studied exists and is continuous. This type of assumption is not necessary in the canonical graphical model analysis and in some way it can be restrictive. A main target for graphical models is to analyse in high dimensional spaces, where data naturally become sparse and concentrated. This could result in either asymptotic independence (which has been shown to be prevalent in real data) or point mass concentration of the spectral measure. This raises the following questions.

(a) How can this assumption be verified in data?

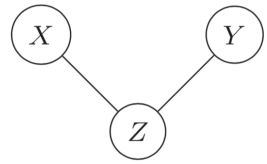


Fig. 24. Simple graphical model

(b) How robust is the model estimation process to the violation of this assumption? What would happen in boundary cases where asymptotic dependence between dimensions is weak?
(c) In the case of asymptotic independence, are there any possible ways, e.g. augmenting or transforming the data, such that the theory and methodology of this paper can still be applied?

For the last question, the following trick may serve as an inspiration. A simple directed graphical model is illustrated in Fig. 24. In addition to the graphical presentation, we further assume that variables X and Y are independent (not just conditionally). If we transform the variables as

$$X' := X \cdot Z$$
$$Y' := Y \cdot Z,$$

then X' and Y' remain independent conditionally on Z, but no longer marginally. If instead X and Y are asymptotically independent, is there a transformation such that the asymptotic independence can be wiped out whereas the conditional (extremal) independence remains? For example, from the results in Huser *et al.* (2019), we may construct

$$X' := W(X) \cdot R(Z),$$
$$Y' := W(Y) \cdot R(Z),$$

where $W(\cdot)$ and $Z(\cdot)$ are transformations of the marginals to a Gaussian and a heavy-tailed distribution respectively, such that the asymptotic dependence between X' and Y' can be monitored by parameters.

Yuexia Zhang and Linbo Wang (*University of Toronto*)

We congratulate Engelke and Hitz on a thought-provoking paper on graphical models for extremes. A key contribution of the paper is the introduction of a novel definition of conditional independence for a multivariate Pareto distribution. Here, we outline a proposal for independence and conditional independence of general random variables whose support is a general set Ω in $\mathbb{R}^d$. Our proposal includes the authors' definition of conditional independence, and the analogous definition of independence as special cases. By making our proposal independent of the context of extreme value theory, we highlight the importance of the authors' contribution beyond this particular context.

Definition 4. Suppose that $\mathbf{Y} = (Y_A, Y_B, Y_C)$ is a random vector with support Ω in $\mathbb{R}^d$, where A, B and C are disjoint sets whose union is $\{1, \ldots, d\}$. Let $U \times V$ denote the Cartesian product of U and V.

(a) We say that Y_A is *conditionally outer independent* of Y_C given Y_B if there is a random vector $\mathbf{W} = (W_A, W_B, W_C)$ with support $L_A \times L_B \times L_C$ in $\mathbb{R}^d$ such that
   (i) $\Omega \subset L_A \times L_B \times L_C$;
   (ii) $(\mathbf{W} | \mathbf{W} \in \Omega) =^{\mathrm{d}} \mathbf{Y}$ and
   (iii) $W_A \perp\!\!\!\perp W_C | W_B$.
   In this case, we write $Y_A \perp\!\!\!\perp_{\mathrm{o}} Y_C | Y_B$.

   If $B = \emptyset$, we say that Y_A is outer independent of Y_C, denoted as $Y_A \perp\!\!\!\perp_{\mathrm{o}} Y_C$.

(b) We say that Y_A is *conditionally inner independent* of Y_C given Y_B if for any $S_A \times S_B \times S_C \subset \Omega$ such that S_k is a measurable subset of $\mathbb{R}^{\dim(Y_k)}$, $k \in \{A, B, C\}$ and $P(\mathbf{Y} \in S_A \times S_B \times S_C) > 0$, we have $Y_A \perp\!\!\!\perp Y_C | (Y_B, \mathbf{Y} \in S_A \times S_B \times S_C)$. In this case, we write $Y_A \perp\!\!\!\perp_{\mathrm{i}} Y_C | Y_B$.

   If $B = \emptyset$, we say that Y_A is *inner independent* of Y_C, which is denoted as $Y_A \perp\!\!\!\perp_{\mathrm{i}} Y_C$.

Proposition 5. Suppose that $\Omega = [0, \infty)^d \backslash [0, 1]^d$ as in the authors' case. Then

$$Y_A \perp\!\!\!\perp_{\mathrm{o}} Y_C | Y_B \quad \Leftrightarrow \quad Y_A \perp\!\!\!\perp_{\mathrm{i}} Y_C | Y_B \quad \Leftrightarrow \quad Y_A \perp\!\!\!\perp_{\mathrm{e}} Y_C | Y_B,$$

where '$\perp\!\!\!\perp_e$' denotes the notion of conditional independence introduced by Engelke and Hitz.

In particular, if $B = \emptyset$, then

$$Y_A \perp\!\!\!\perp_o Y_C \quad \Leftrightarrow \quad Y_A \perp\!\!\!\perp_i Y_C \quad \Leftrightarrow \quad Y_A \perp\!\!\!\perp_e Y_C.$$

Remark 4. We do not place any distributional assumptions on $\mathbf{Y}$ in proposition 1.

Remark 5. Engelke and Hitz showed that if $\mathbf{Y}$ is multivariate Pareto and admits a positive and continuous density, then $Y_A \not\perp\!\!\!\perp_e Y_C$. This does not rule out the possibility of $Y_A \perp\!\!\!\perp_e Y_C$ for general Pareto distributions. For example, consider two independent standard Pareto distributions X_1 and X_2. Following the paper's equation (6), all the probability mass of $\mathbf{Y}$ lies on $(1, \infty) \times \{0\}$ and $\{0\} \times (1, \infty)$ so that it does not admit a density with respect to Lebesgue measure. Nevertheless $Y_1 \perp\!\!\!\perp_e Y_2$.

The proofs of these propositions are presented in Zhang and Wang (2020).

Chen Zhou (*Erasmus University Rotterdam*)

The paper by Sebastian Engelke and Adrien Hitz considers conditioning on one specific dimension exceeding a high threshold and then showing that conditional independence in extremes does not depend on the conditioning dimension chosen from the conditioning set (proposition 1). This beautiful mathematical result paves the way for introducing graphical models for extremes, in which conditional independence is a key building block.

The first highlight of the paper is to establish a Hammersley–Clifford-type theorem in the context of extremes (theorem 1). The second highlight of the paper is that the graphical structure can be recovered from data without prior knowledge based on combining the minimum spanning tree and a greedy forward selection.

My comment is regarding the two assumptions in the statistical inference (Section 5). Firstly, all separators are single nodes. Secondly, the size of cliques is at maximum 3. Although making such assumptions is a sensible step towards statistical inference, breaking those assumptions will increase the computational burden. Firstly, if the separators may contain multiple nodes (say m nodes), all cliques including such separators have at least $m + 1$ dimension. Applying the maximum likelihood method to such cliques will be cumbersome. Secondly, when $m = 1$, there is no need to estimate the conditional density for the set of separators, whereas, for $m > 1$, such an estimation is needed. Last, but not least, if the size of cliques may exceed 3, then the complexity in the greedy forward selection procedure rises rapidly. The current successful result in Fig. 5 may be attributed to the fact that both the true graphical structure and the allowed graphical structure in the greedy forward selection are restricted to having cliques with a size less than or equal to 3.

I would like to raise a particular application case for which breaking the two assumptions is necessary. Graphical models have been extensively studied in the banking literature regarding various networks across banks: interbank exposures, interbank loans, payment network etc. Applying the paper for exploring the graphical structure of banks in extreme events will provide further insight for policy makers to identify the so-called *systemically important banks*.

One common finding in the existing literature regarding banking networks is the *core–periphery* structure:

(a) there are a few *core* banks fully connected with each other;
(b) all other *periphery* banks are connected to one or a few core banks only;
(c) there is no connection between periphery banks.

The dependence between extreme events across banks may inherit such a core–periphery structure.

If one periphery bank connects to m core banks, then there is a separator with m nodes. Further, since all core banks form a clique, its size can be relatively large. To detect such a graphical structure by using the minimum spanning tree and greedy forward selection will be computationally difficult. To conclude, there is still room for developing new exploratory techniques to reveal the graphical structure among extreme events.

The **authors** replied later, in writing, as follows.

We thank all the discussants for many thought-provoking comments and for pointing out exciting future research directions.

The definition of extremal graphical models

Whereas we assume throughout the paper that the multivariate Pareto distribution $\mathbf{Y}$ admits a positive and continuous density, it has been noted by Strokorb, and Zhang and Wang that this is not necessary for the definition of extremal graphical models. We agree that this assumption can be dropped and is required only in results on the factorization of densities, as for instance in proposition 1 and theorem 1. We propose here a more general version of our extremal conditional independence notion $\perp_e$.

Definition 5. Suppose that $\mathbf{Y}$ is multivariate Pareto and let $A, B, C \subset V = \{1, \ldots, d\}$. We say that $\mathbf{Y}_A$ is conditionally independent of $\mathbf{Y}_C$ given $\mathbf{Y}_B$, denoted $\mathbf{Y}_A \perp_e \mathbf{Y}_C | \mathbf{Y}_B$, if

$$\forall\, k \in \{1, \ldots, d\}: \quad \mathbf{Y}_A^k \perp\!\!\!\perp \mathbf{Y}_C^k | \mathbf{Y}_B^k.$$

Similarly, we now define an extremal graphical model for the undirected graph $\mathcal{G} = (V, E)$ to be a multivariate Pareto distribution $\mathbf{Y}$ that satisfies the pairwise Markov property according to $\mathcal{G}$ with respect to the new definition of $\perp_e$ above. It is important that in the new definition we allow that $B = \emptyset$, which corresponds to a new notion of independence in extremes. This is motivated by the comment of Strokorb who then further proves that this new independence notion is in fact equivalent to the classical notion of asymptotic independence. This makes our remark 1 rigorous since it proves that unconnected components of graph $\mathcal{G}$ correspond to asymptotically independent components of $\mathbf{Y}$. Strokorb therefore shows that our extremal graphical models naturally extend to the case where $\mathbf{Y}$ takes mass on subfaces of the space $\mathcal{E}$ and where some components may be asymptotically independent.

Lauritzen raises the point that our conditional independence notion $\perp_e$ forms a semigraphoid on subsets of V. This is indeed correct with the new definition 5 above that allows for independence by conditioning on $B = \emptyset$. It is not obvious whether his axiom 5 that would make $\perp_e$ a graphoid holds, since assuming a positive and continuous density excludes independence in $\mathbf{Y}$: see the discussion after proposition 1.

Inference for more general graph structures

Belzile and Dupuis, Darné and Davison, and Zhou point out that block graph structures may be too restrictive for certain applications. We agree that the restriction to block graphs limits the full potential of graphical models for statistical inference. In on-going work we follow several directions to develop estimation and structure learning methods for more general decomposable and non-decomposable graphs.

Our Hammersley–Clifford theorem holds for any connected, decomposable graph $\mathcal{G}$ and, since the definition of extremal graphical models is closely related to classical conditional independence, inferential techniques from graphical models can be adapted. For instance, for a Hüsler–Reiss distribution with parameter matrix Γ that factorizes according to a decomposable graph $\mathcal{G}$, a possible estimation procedure is as follows. Let $\{C_1, \ldots, C_m\}$ and $\{D_2, \ldots, D_m\}$ be the sequences of cliques and separators of $\mathcal{G}$ with the running intersection property. We first obtain an estimate $\hat{\Gamma}_{C_1}$ for all parameters that are related to clique C_1 by maximizing the censored clique likelihood (37). This already determines the subsets of parameters $\hat{\Gamma}_{D_2}$ that are related to clique C_2 since $D_2 = C_1 \cap C_2$. We then obtain an estimate of $\hat{\Gamma}_{C_2}$ under the constraint that the parameters $\hat{\Gamma}_{D_2}$ are fixed to the values from the first fitting step. We sequentially go through all cliques to obtain a $(d \times d)$-dimensional matrix $\hat{\Gamma}$ that is partially specified on the cliques of $\mathcal{G}$. To complete $\hat{\Gamma}$ to a valid variogram matrix factorizing on $\mathcal{G}$, the solution to a matrix completion problem for variograms as in proposition 4, but for decomposable graphs, will be essential; see Dempster (1972) and Grone *et al.* (1984) for completion problems for covariance matrices in the Gaussian case.

For model selection within decomposable graphs, Chavez-Demoulin proposes to use likelihood ratio statistics to compare nested graphs and to perform a greedy procedure through local computations. For the same model class, Evans outlines how junction trees may be used for efficient computations of conditional risk probabilities. Reid suggests that a better understanding of the link between the clique likelihood (37) and composite likelihood methods may improve statistical inference for extremal graphical models.

For general, possibly non-decomposable graphs, the class of Hüsler–Reiss distributions seems to be most promising with regard to our results in Section 4.6. Supplementing our proposition 3, Segers shows the existence of a $(d \times d)$-dimensional covariance matrix whose inverse contains the extremal graphical structure of a Hüsler–Reiss distribution as zero patterns. For estimation and structure learning of general Hüsler–Reiss graphical models we are currently investigating an extremal graphical lasso with a suitable l_1-regularization of the precision matrices $\Theta^{(k)}$. For spatial applications where the graph is often a regular lattice, Huser and Cisneros suggest using finite mixtures of trees as approximation of the non-decomposable graph structure.

Klüppelberg describes the line of research related to directed acyclic graphs for max-linear models, which gives a different perspective on graphical methods for extremes. Lauritzen suggests that directed extremal graphical models in our framework may lead to new density factorizations.

All of these comments point out new research directions that will help to tighten the links between the fields of graphical models and extremes, and to establish new areas such as high dimensional statistics and causality for extremes.

Asymptotically independent graphical models

Several comments address the point that our assumption of a multivariate. Pareto distribution with positive density implies that all components are asymptotically dependent, which excludes the weaker form of extremal dependence called asymptotic independence. We acknowledge this restriction of our current theory. Strokorb's extension of our results shows that asymptotic independence naturally arises as unconnected components of the extremal graph $\mathcal{G}$. A more refined modelling of this asymptotic independence using graphical structures requires further extensions of our theory.

Wadsworth, Papastathopoulos and Casey propose the use of the conditional extreme value model of Heffernan and Tawn (2004) in combination with graphical methods as a possibility to allow for a large range of dependence scenarios. Wadsworth fits such a model to the Danube data set imposing a graph structure by applying a classical Gaussian graphical lasso to the limiting vector in this approach. Although the model hints at asymptotic independence between some stations, it is noted that the graph structure is fairly similar to our results in Fig. 6(b). This is certainly an appealing approach that deserves further investigation; however, one difficulty is that conditioning on different components results in different graphical structures; see also Papastathopoulos's comment on a mixture model to resolve some of the issues. As already pointed out by Wadsworth, it remains an open question whether a coherent unifying notion of graphical models can be established this way.

Meyer and Wintenberger suggest defining exceedances of the vector $\mathbf{X}$ by conditioning on the event $\{\min_{1\leqslant k\leqslant d} X_k > u\}$ instead of the maximum. The limiting distribution may then be non-trivial even for asymptotic independence under the assumption of hidden regular variation (Resnick, 2002), and it can potentially be used for graphical modelling. Mhalla proposes a similar conditioning in connection with a graph structure for a dependence function defined in Wadsworth and Tawn (2013). Conditioning on the minimum of the variables being large has the advantage of capturing residual dependence structures in asymptotically independent regimes. However, one limitation of such approaches is that only part of the extremal behaviour is modelled, whereas events where not all components are simultaneously larger are not captured. Wan suggests another procedure, namely to make our theory applicable to asymptotic independence by a suitable transformation of the data in a random-scale construction.

Asymptotic independence is an important feature of extremes that arises in many data sets, and these suggestions are promising first steps to link this dependence regime with the notions of sparsity and graphical models.

Further conditional independence definitions

Several contributors have pointed out that our definition of conditional dependence $\perp_e$ may be useful in a broader context than only for extremes. Ivanovs rightly remarks that our results essentially apply to any homogeneous Radon measure Λ. Such objects also appear in the theory of stable distributions and Lévy motions and it will be exciting to see in future research whether graph structures can be defined similarly in this area.

Zhang and Wang propose to investigate a theory of independence and conditional independence on general subsets $\Omega \subset \mathbb{R}^d$ that may not be product spaces. Their notion includes our definition of $\perp_e$ as a special case and extends our work beyond extreme value theory. This directly relates to the comments by Robert, and Belzile and Dupuis who propose defining exceedances according to a risk functional that is different from the maximum (such as the arithmetic or geometric mean), which would result in a different space Ω on which a conditional independence notion is required.

Applications

The data application on peak river flow was motivated by the fact that in this case we have an underlying tree structure given by flow connections that we can compare with the estimated graphs. Cox points out that there might be fruitful links to empirical or physical laws in the hydrological literature. The methodology of our paper is intended to be widely applicable, including situations without domain knowledge on the graph structure. Several discussants suggest applications to other fields such as finance, traffic mortality or crime events (Mateu and Eckardt).

Darné and Davison apply our structure learning algorithm to negative daily returns of a network of 22 banks. They observe that the estimated extremal graph nicely separates European from US banks, and they also infer a different structure before and after the banking crisis of 2008. Zhou proposes a similar application to help policy makers to identify systemically important banks. He points out that

from economic theory it is expected to see few core banks that are fully connected, and periphery banks that are connected to one or a few core banks only.

References in the discussion

Améndola, C., Klüppelberg, C., Lauritzen, S. and Tran, N. (2020) Conditional independence in max-linear Bayesian networks. *Preprint arXiv:2002.09233*.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Besag, J. (1975) Statistical analysis of non-lattice data. *Statistician*, **24**, 179–195.

Besag, J. (1977) Errors-in-variables estimation for Gaussian lattice schemes. *J. R. Statist. Soc.* B, **39**, 73–78.

Bhansali, R. J. (1990) On a relationship between the inverse of a stationary covariance matrix and the linear interpolator. *J. Appl. Probab.*, **27**, 156–170.

Bhansali, R. J. (1996) Asymptotically efficient autoregressive model selection for multistep prediction. *Ann. Inst. Statist. Math.*, **48**, 577–602.

Bhansali, R. J. and Downham, D. Y. (1977) Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, **64**, 547–551.

Buck, J. and Klüppelberg, C. (2020) Recursive max-linear models with propagating noise. To be published.

Castruccio, S., Huser, R. and Genton, M. G. (2016) High-order composite likelihood inference for max-stable distributions and processes. *J. Computnl Graph. Statist.*, **25**, 1212–1229.

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modeling of spatial extremes. *Statist. Sci.*, **27**, 161–186.

Dempster, A. P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Drton, S. and Maathuis, M. H. (2017) Structure learning in graphical modeling. *A. Rev. Statist. Appl.*, **4**, 365–393.

Eckardt, M., Gonzalez-Monsalve, J. and Mateu, J. (2020) Graphical modelling and partial characteristics for multitype and multivariate marked spatio-temporal point processes. To be published.

Eckardt, M. and Mateu, J. (2017) Analysing highly complex and highly structured point patterns in space. *Spatl Statist.*, **22**, 296–305.

Eckardt, M. and Mateu, J. (2018) Point patterns occurring on complex structures in space and space-time: an alternative network approach. *J. Computnl Graph. Statist.*, **27**, 312–322.

Eckardt, M. and Mateu, J. (2019a) Partial characteristics for marked spatial point processes. *Environmetrics*, **30**, 1–13.

Eckardt, M. and Mateu, J. (2019b) Analysing multivariate spatial point processes with continuous marks: a graphical modelling approach. *Int. Statist. Rev.*, **87**, 44–67.

Engelke, S., Malinowski, A., Kabluchko, Z. and Schlather, M. (2015) Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *J. R. Statist. Soc.* B, **77**, 239–265.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D. and Borsboom, D. (2012) qgraph: network visualizations of relationships in psychometric data. *J. Statist. Softwr.*, **48**, no. 4, 1–18.

de Fondeville, R. and Davison, A. C. (2018) High-dimensional peaks-over-threshold inference. *Biometrika*, **105**, 575–592.

Frigessi, A., Haug, O. and Rue, H. (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, **5**, 219–235.

Gissibl, N. and Klüppelberg, C. (2018) Max-linear models on directed acyclic graphs. *Bernoulli*, **24**, no. 4A, 2693–2720.

Gissibl, N., Klüppelberg, C. and Lauritzen, S. L. (2019) Identifiability and estimation of recursive max-linear models. *Scand. J. Statist.*, to be published, doi 10.1111/sjos.12446.

Gissibl, N., Klüppelberg, C. and Otto, M. (2018) Tail dependence of recursive max-linear models with regularly varying noise variables. *Econmetr. Statist.*, **6**, 149–167.

Grone, R., Johnson, C. R., de Sá, E. M. and Wolkowicz, H. (1984) Positive definite completions of partial Hermitian matrices. *Lin. Alg. Appl.*, **58**, 109–124.

Heffernan, J. E. and Resnick, S. I. (2007) Limit laws for random vectors with an extreme component. *Ann. Appl. Probab.*, **17**, 537–571.

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *J. R. Statist. Soc.* B, **66**, 497–546.

Horton, R. E. (1945) Erosional development of streams and their drainage basins: hydrological approach to quasi-morphology. *Am. Soc. Hydrol.*, **56**, 275–370.

Huser, R., Dombry, C., Ribatet, M. and Genton, M. G. (2019) Full likelihood inference for max-stable data. *Stat*, **8**, no. 1, article e218.

Huser, R., Opitz, T. and Thibaud, E. (2017) Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatl Statist.* A, **21**, 166–186.

Hüsler, J. and Reiss, R.-D. (1989) Maxima of normal random vectors: between independence and complete dependence. *Statist. Probab. Lett.*, **7**, 283–286.

Kabluchko, Z. (2009) Spectral representations of sum- and max-stable processes. *Extremes*, **12**, article 401.

Keef, C., Papastathopoulos, I. and Tawn, J. A. (2013) Estimation of the conditional distribution of a multivariate variable given that one of its components is large: additional constraints for the Heffernan and Tawn model. *J. Multiv. Anal.*, **115**, 396–404.
Kiriliouk, A., Rootzén, H., Segers, J. J. and Wadsworth, J. L. (2019) Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, **61**, 123–135.
Klüppelberg, C. and Krali, M. (2019) Estimating an extreme Bayesian network via scalings. *Preprint arXiv: 1912.03968*.
Klüppelberg, C. and Lauritzen, S. (2019) Bayesian networks for max-linear models. In *Network Science—an Aerial View from Different Perspectives* (eds F. Biagini, G. Kauermann and T. Meyer-Brandis). Berlin: Springer.
Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Oxford University Press.
Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H.-G. (1990) Independence properties of directed Markov fields. *Networks*, **20**, 491–505.
Liu, Y. and Tawn, J. (2014) Self-consistent estimation of conditional multivariate extreme value distributions. *J. Multiv. Anal.*, **127**, 19–35.
Moore, R. J., Jones, D. A., Cox, D. R. and Isham, V. S. (2000) Design of the HYREX raingauge network. *Hydrol. Earth Syst. Sci.*, **4**, 523–530.
Papastathopoulos, I. and Strokorb, K. (2016) Conditional independence among max-stable laws. *Statist. Probab. Lett.*, **108**, 9–15.
Papastathopoulos, I. and Tawn, J. A. (2019) Extreme events of higher-order Markov chains: hidden tail chains and extremal Yule-Walker equations. *Preprint arXiv:1903.04059*.
Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd edn. New York: Wiley.
Resnick, S. (2002) Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, **5**, 303–336.
Resnick, S. I. (2007) *Heavy-tail Phenomena*. New York: Springer.
Rodriguez-Iturbe, I. and Rinaldo, I. (1997) *Fractal River Basins*. Cambridge: Cambridge University Press.
Rootzén, H., Segers, J. J. and Wadsworth, J. L. (2018a) Multivariate generalized Pareto distributions: parametrizations, representations, and properties. *J. Multiv. Anal.*, **165**, 117–131.
Rootzén, H., Segers, J. J. and Wadsworth, J. L. (2018b) Multivariate peaks over thresholds models. *Extremes*, **21**, 115–145.
Rootzén, H. and Tajvidi, N. (2006) Multivariate generalized Pareto distributions. *Bernoulli*, **12**, 917–930.
Samorodnitsky, G. and Taqqu, M. S. (1994) *Stable non-Gaussian Random Processes*. New York: Chapman and Hall.
Sato, K.-I. (2013) *Lévy Processes and Infinitely Divisible Distributions*. Cambridge: Cambridge University Press.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Segers, J., Zhao, Y. and Meinguet, T. (2017) Polar decomposition of regularly varying time series in star-shaped metric spaces. *Extremes*, **20**, 539–566.
Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.
Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating the parameters of a linear process. *Ann. Statist.*, **8**, 147–164.
Shibata, R. (1981) An optimal autoregressive spectral estimate. *Ann. Statist.*, **9**, 300–306.
Strokorb, K. (2020) Extremal independence old and new. *Technical Report*. Cardiff University, Cardiff. (Available from `https://arXiv.org/pdf/2002.07808.pdf`.)
Varin, C. (2008) On composite marginal likelihoods. *Adv. Statist. Anal.*, **92**, 1–28.
Vettori, S., Huser, R., Segers, J. and Genton, M. G. (2020) Bayesian model averaging over tree-based dependence structures for multivariate extremes. *J. Computnl Graph. Statist.*, **29**, 174–190.
Wadsworth, J. L. and Tawn, J. A. (2013) A new representation for multivariate tail probabilities. *Bernoulli*, **19**, 2689–2714.
Wadsworth, J. L. and Tawn, J. (2019) Higher-dimensional spatial extremes via single-site conditioning. *Preprint arXiv:1912.06560*. Lancaster University, Lancaster.
Yu, H., Uy, W. and Dauwels, J. (2017) Modeling spatial extremes via ensemble-of-trees of pairwise copulas. *IEEE Trans. Signal Process.*, **65**, 571–586.
Zhang, Y. and Wang, L. *Preprint*. University of Toronto, Toronto. (Available from `https://arxiv.org.abs/2006.01036`.)