Received: 11 December 2020 | Accepted: 23 August 2022 | Published on: 18 November 2022

DOI: 10.1111/rssb.12556

ORIGINAL ARTICLE

Series B Statistical Methodology

Structure learning for extremal tree models

Sebastian Engelke[1] | **Stanislav Volgushev**[2]

[1]Research Center for Statistics, University of Geneva, Geneva, Switzerland

[2]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Correspondence
Sebastian Engelke, Research Center for Statistics, University of Geneva, Boulevard du Pont d'Arve 40, 1205 Geneva, Switzerland.
Email: sebastian.engelke@unige.ch

Funding information
Natural Sciences and Engineering Research Council of Canada; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 161297

Abstract

Extremal graphical models are sparse statistical models for multivariate extreme events. The underlying graph encodes conditional independencies and enables a visual interpretation of the complex extremal dependence structure. For the important case of tree models, we develop a data-driven methodology for learning the graphical structure. We show that sample versions of the extremal correlation and a new summary statistic, which we call the extremal variogram, can be used as weights for a minimum spanning tree to consistently recover the true underlying tree. Remarkably, this implies that extremal tree models can be learned in a completely non-parametric fashion by using simple summary statistics and without the need to assume discrete distributions, existence of densities or parametric models for bivariate distributions.

KEYWORDS

domain of attraction, extreme value theory, graphical models, minimum spanning tree, multivariate Pareto distribution

1 | INTRODUCTION

Extreme value theory provides essential statistical tools to quantify the risk of rare events such as floods, heatwaves or financial crises (e.g. Engelke et al., 2019; Katz et al., 2002; Poon et al., 2004). The univariate case is well understood and the generalised extreme value and Pareto distributions describe the distributional tail with only few parameters. In dimension $d \geq 2$, the dependence between large values in the different components of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ can become very complex. Estimating this dependence in higher dimensions is particularly challenging because the number of extreme observations k_n is by definition much smaller than the number n of all samples in a data set. Constructing sparse models for the multivariate dependence between

marginal extremes is therefore crucial for obtaining tractable and reliable methods in multivariate extremes; see Engelke and Ivanovs (2021) for a review of recent advances.

One line of research aims at exploiting conditional independence structures (Dawid, 1979) and corresponding graphical models. In the setting of max-stable distributions, which arise as limits of component-wise block maxima of independent copies of $\mathbf{X}$, Gissibl and Klüppelberg (2018) and Klüppelberg and Lauritzen (2019) study max-linear models on directed acyclic graphs. The distributions considered in there do not have densities, and a general result by Papastathopoulos and Strokorb (2016) shows that there exist no non-trivial density factorisation of max-stable distributions on graphical structures.

A different perspective on multivariate extremes is given by threshold exceedances and the resulting class of multivariate Pareto distributions. Such distributions are the only possible limits that can arise from the conditional distribution of $\mathbf{X}$ given that it exceeds a high threshold (Rootzén et al., 2018; Rootzén & Tajvidi, 2006). For a d-dimensional random vector $\mathbf{Y}$ that follows a multivariate Pareto distribution, Engelke and Hitz (2020) introduce suitable notions of conditional independence and extremal graphical models with respect to a graph G. They further show that these notions are natural as they imply the factorisation of the density of $\mathbf{Y}$ through a Hammersley–Clifford type theorem. Extremal graphical models are also related to limits of regularly varying Markov trees studied in Segers (2020) and Asenova et al. (2021).

In most of the above work, the graphical structure G is assumed to be known a priori. It is either based on expert knowledge in the domain of application or it might be identified with an existing graph, as for instance a river network for discharge measurements. However, often no or insufficient domain knowledge on a prior candidate for a graphical structure is available, and a data-driven approach should be followed in order to detect conditional independence relations and to estimate a sensible graph structure. In this work we discuss structural learning for extreme observations.

An important sub-class of general graphs for which structure learning for extremes turns out to be possible in great generality is given by trees. A tree $T = (V, E)$ with nodes V and edge set E is a connected undirected graph without cycles. Most structure learning approaches for trees are based on the notion of the minimum spanning tree. For a set of symmetric weights $\rho_{ij} > 0$ associated with any pair of nodes $i, j \in V$, $i \neq j$, the latter is defined as the tree structure

$$T_{\text{mst}} = \underset{T=(V,E)}{\arg\min} \sum_{(i,j)\in E} \rho_{ij}, \tag{1}$$

that minimises the sum of distances on that tree. Given the set of weights, there exist greedy algorithms that constructively solve this minimisation problem (Kruskal, 1956; Prim, 1957).

The crucial ingredient for this algorithm are the weights ρ_{ij} between the nodes, and for statistical inference it is generally desirable to choose them in such a way that T_{mst} recovers the true underlying tree structure that represents the conditional independence relations. A common approach in graphical modelling is to use the Chow–Liu tree (Chow & Liu, 1968), which is the conditional independence structure that maximises the likelihood for a given parametric model (cf, Cowell et al., 2006, chapter 11). This method uses the negative mutual information as edge weights ρ_{ij} in (1), and in general this requires formulating parametric models for the bivariate marginal distributions. In the Gaussian case the weights then simplify to $\rho_{ij} = \log\left(1 - r_{ij}^2\right)/2$, where r_{ij} are the correlation coefficients (cf, Drton & Maathuis, 2017).

For a multivariate Pareto distribution $\mathbf{Y}$ that is an extremal graphical model on a tree T, Engelke and Hitz (2020) proposed to use the negative maximised bivariate log-likelihoods as edge

weights. This approach has two disadvantages. First, in higher dimensions d it may become prohibitively costly to compute d^2 censored likelihood optimisations, and second, a set of parametric bivariate models has to be chosen in advance.

In this paper we study structure learning for extremal tree models in much larger generality. We show that a function of the extremal correlation χ_{ij}, a widely used coefficient to summarise the strength of extremal dependence between marginals $i, j \in V$ (e.g., Coles et al., 1999), can be used as weights ρ_{ij} in (1) to retrieve the underlying tree structure T as the minimum spanning tree under mild non-parametric assumptions. We further introduce a new summary coefficient for extremal dependence, the extremal variogram Γ_{ij}, which turns out to take a similar role in multivariate extremes as covariances in Gaussian models. More precisely, the extremal variogram of $\mathbf{Y}$ is shown to be an additive tree metric on the tree T and, as a consequence, it can be used as well as weights ρ_{ij} of the minimum spanning tree to recover the true tree structure. Surprisingly, these results are stronger than for non-extremal tree structures, since we do not require any further parametric assumptions or the existence of densities. This phenomenon originates from the homogeneity of multivariate Pareto distributions and particularly nice stochastic representations of extremal tree models.

In practice, we usually observe n samples of $\mathbf{X}$ in the domain of attraction of $\mathbf{Y}$, that is, the conditional distribution of $\mathbf{X}$ given $\mathbf{X}$ exceeds a high threshold converges to the distribution of $\mathbf{Y}$ after proper scaling; see Section 2.1 for a formal definition. We then rely on estimators of the quantities χ_{ij} and Γ_{ij} to plug into (1). To take into account that $\mathbf{X}$ is only in the domain of attraction of $\mathbf{Y}$, typical estimators in extreme value theory use only the most extreme observations. We use an existing estimator $\hat{\chi}_{ij}$ of extremal correlation and a new empirical estimator of the extremal variogram to show that the extremal tree structure can be estimated consistently in a non-parametric way, even when the dimension increases with the sample size.

The remaining paper is organised as follows. In Section 2 we revisit the notion of extremal graphical models and extend existing representations to the case where densities may not exist. The extremal variogram is introduced in Section 3 and its properties are discussed in detail. In Section 4 we prove the main results on the consistent recovery of extremal tree structures based on extremal correlations and extremal variograms, both on the population level and using empirical estimates. The simulation studies in Section 5 illustrate the finite sample behaviour of our structure estimators and show that extremal variogram based methods typically outperform methods working with the extremal correlation. We apply the new tools in Section 6 to a financial data set of foreign exchange rates. The Appendix and the Supplementary Material S1 contain the proofs and additional illustrations. The methods of this paper are implemented in the R package `graphicalExtremes` (Engelke et al., 2019).

2 | EXTREMAL GRAPHICAL MODELS

2.1 | Multivariate Pareto distributions

Let $\mathbf{X} = (X_i)_{i \in V}$ be a random vector with eventually continuous marginal distribution functions F_i. Extreme value theory studies marginal and joint tail properties of $\mathbf{X}$. Univariate extreme value theory focuses on the behaviour of marginal components X_i, see for example Embrechts et al. (1997) and Coles et al. (1999). Multivariate extreme value theory is concerned with the dependence structure among different components of extreme observations from $\mathbf{X}$; see Resnick (2008, chapter 5);

de Haan and Ferreira (2006); Beirlant et al. (2004) or Engelke and Ivanovs (2021) for an introduction.

One way to describe such tail properties is based on threshold exceedances; here only observations that land above a high threshold are considered. Multivariate Pareto distributions arise as the limits of such high threshold exceedances and are thus natural models for extreme events (Rootzén & Tajvidi, 2006). To formally define threshold exceedances in dimension $d > 1$, we need to specify the notion of a high threshold in a multivariate setting. Throughout the paper, we consider multivariate exceedances of the random vector $\mathbf{X}$ as those realisations where at least one component of $\mathbf{X}$ exceeds a high marginal quantile. In order to guarantee the existence of the limit of the exceedance distribution, a regularity condition called multivariate regular variation (Resnick, 2008, Chapter 5) is needed. Intuitively, this assumption ensures that the dependence between different components of this conditional distribution stabilises if the marginal quantile is sufficiently large. More formally, this means that there exists a random vector $\mathbf{Y}$ supported on $\mathcal{L} = \{\mathbf{x} \geq \mathbf{0} : \|\mathbf{x}\|_\infty > 1\}$ such that for all continuity points $\mathbf{x} \in \mathcal{L}$ of the distribution function of $\mathbf{Y}$ we have

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{x}) = \lim_{q \to 0} \mathbb{P}(F(\mathbf{X}) \leq 1 - q/\mathbf{x} | F(\mathbf{X}) \not\leq 1 - q), \tag{2}$$

where we define $F(\mathbf{x}) = (F_1(x_1), \dots, F_d(x_d))$. Note that the condition $\{F(\mathbf{X}) \not\leq 1 - q\}$ states that at least one component of $\mathbf{X}$ exceeds its marginal $1 - q$ quantile, explaining the terminology of threshold exceedances. Distributions of random vectors $\mathbf{Y}$ that can arise in the above limit are called multivariate Pareto distributions. We say that the random vector $\mathbf{X}$ is in the max-domain of attraction of the multivariate Pareto distribution $\mathbf{Y} = (Y_i)_{i \in V}$.

The class of multivariate Pareto distributions is very general and contains many different parametric sub-families. Nevertheless, since the random vector $\mathbf{Y}$ arises as a limiting distribution, it has an important structural property called homogeneity:

$$\mathbb{P}(\mathbf{Y} \in tA) = t^{-1}\mathbb{P}(\mathbf{Y} \in A), \qquad t \geq 1, \tag{3}$$

where for any Borel subset $A \subset \mathcal{L}$ we define $tA = \{t\mathbf{x} : \mathbf{x} \in A\}$. This explains the name multivariate Pareto distribution since it implies that for any $i \in V$ we have $\mathbb{P}(Y_i \leq x | Y_i > 1) = 1 - 1/x$ for $x \geq 1$, that is, $Y_i | Y_i > 1$ follows a standard Pareto distribution. Moreover, since $F(\mathbf{X})$ has identically distributed margins, it follows from (2) that $\mathbb{P}(Y_1 > 1) = \cdots = \mathbb{P}(Y_d > 1)$. Conversely, if the latter holds and $\mathbf{Y}$ is homogeneous as in (3), then $\mathbf{Y}$ is a multivariate Pareto distribution; for a proof of this equivalence and the relationship to limits appearing in Segers (2020) see Section S.5 of Supplementary Material S1.

2.2 | Extremal Markov structures

Since the support $\mathcal{L}$ of multivariate Pareto distributions is not a product space, the definition of conditional independence is non-standard and relies on auxiliary random vectors derived from $\mathbf{Y}$. For any $m \in V$, we consider the random vector $\mathbf{Y}^m$ defined as $\mathbf{Y}$ conditioned on the event that $\{Y_m > 1\}$, which has support on the space $\mathcal{L}^m = \{\mathbf{x} \in \mathcal{L} : x_m > 1\}$. For general random vectors $\mathbf{X} \in \mathbb{R}^d$ and ordered sets $A \subset \{1, \dots, d\}$, let $\mathbf{X}_A$ denote the sub-vector of $\mathbf{X}$ with indices in A. The notation $\mathbf{Y}^m_A$ will be used to denote the subvector of $\mathbf{Y}^m$ with indices in A.

With this notation we can state a definition of conditional independence for multivariate Pareto distributions that is more general than the one in Engelke and Hitz (2020), since we do not assume existence of densities.

Definition 1. For disjoint subsets $A, B, C \subset V = \{1, \dots, d\}$, we say that $\mathbf{Y}_A$ is conditionally independent of $\mathbf{Y}_C$ given $\mathbf{Y}_B$ if

$$\forall m \in \{1, \dots, d\} : \quad \mathbf{Y}_A^m \perp\!\!\!\perp \mathbf{Y}_C^m \mid \mathbf{Y}_B^m. \tag{4}$$

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

The subscript e in $\perp_e$ indicates that this conditional independence notion is defined for extreme observations, which are described by the multivariate Pareto distribution $\mathbf{Y}$ according to (2).

We view the index set V as a set of nodes of a graph $G = (V, E)$, with connections given by 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. For disjoint subsets $A, B, C \subset V$, B is said to separate A and C in G if every path from A to C contains as least one node in B. For an illustration of these definitions see Section S.1 in Supplementary Material S1.

The notion of an extremal graphical model is then naturally defined as a multivariate Pareto distribution that satisfies the global Markov property on the graph G with respect to the conditional independence relation $\perp_e$, that is, for any disjoint subsets $A, B, C \subset V$ such that B separates A from C in G,

$$\mathbf{Y}_A \perp_e \mathbf{Y}_C \mid \mathbf{Y}_B. \tag{5}$$

In line with the definition in the graphical models literature (Lauritzen, 1996, chapter 3), the definition allows for additional conditional independence relations that are not encoded by graph separation. This means that there are typically several graphs G that are consistent with the distribution of $\mathbf{Y}$; for instance, any multivariate Pareto distribution is an extremal graphical model on the fully connected graph.

In the case of a decomposable graph G and if $\mathbf{Y}$ possesses a positive and continuous density $f_{\mathbf{Y}}$, Engelke and Hitz (2020) show that this density factorises into lower-dimensional densities, and that the graph G is necessarily connected. If $\mathbf{Y}$ does not have a density, then the extremal graph can be disconnected and the connected components are mutually independent of each other (Engelke & Hitz, 2020, see Kirstin Strokorb's discussion contribution). Note that we require the global Markov property in the definition of extremal graphical models as opposed to the pairwise Markov property used in Engelke and Hitz (2020). Both properties are equivalent in the case of positive, continuous densities, but in general, the former implies the latter but not the other way around (see Lauritzen, 1996, chapter 3).

2.3 | Extremal tree models

An important example of a sparse graph structure is a tree. A tree $T = (V, E)$ is a connected undirected graph without cycles and thus $|E| = |V| - 1$. Equivalently, a tree is a graph with a unique path between any two nodes. If $\mathbf{Y}$ is an extremal graphical model satisfying the global Markov

property (5) with respect to a tree T, we obtain a simple stochastic representation of $\mathbf{Y}^m$. This stochastic representation will be the crucial building block for the results on tree learning given in the next section.

To this end we need to introduce the concept of extremal functions. Define the extremal function relative to coordinate m as the d-dimensional, non-negative random vector $\mathbf{W}^m$ with $W_m^m = 1$ almost surely that satisfies the stochastic representation

$$\mathbf{Y}^m \overset{(d)}{=} P\mathbf{W}^m, \tag{6}$$

where P is a standard Pareto random variable, $\mathbb{P}(P \leq x) = 1 - 1/x, x \geq 1$, which is independent of $\mathbf{W}^m$, and $\overset{(d)}{=}$ stands for equality in distribution. Such a representation is possible by homogeneity (3) of $\mathbf{Y}$, which is inherited by $\mathbf{Y}^m$. Indeed, given homogeneity of $\mathbf{Y}^m$ we see that Y_m^m follows a standard Pareto distribution. Moreover, writing $\mathbf{W}^m := \mathbf{Y}^m / Y_m^m$, homogeneity of $\mathbf{Y}^m$ and a simple calculation implies that $\mathbf{W}^m$ and Y_m^m are independent, resulting in the representation (6).

The representation (6) is an alternative way of describing the distribution of $\mathbf{Y}$, and indeed, the set of the d extremal functions $\mathbf{W}^1, \ldots, \mathbf{W}^d$ uniquely defines the multivariate Pareto distribution. We refer to Dombry et al. (2013, 2016) for additional technical background on extremal functions.

Example 1. In the case $d = 2$, due to homogeneity, the bivariate Pareto distribution $\mathbf{Y} = (Y_1, Y_2)$ can essentially be characterised by a univariate distribution. Indeed, for any non-negative random variable W_2^1 with $\mathbb{E}W_2^1 \leq 1$, the random vector $\mathbf{W}^1 = \left(1, W_2^1\right)$ is the extremal function relative to the first coordinate of a unique bivariate Pareto distribution $\mathbf{Y}$. The extremal function relative to the second coordinate $\mathbf{W}^2 = \left(W_1^2, 1\right)$ is obtained through a change of measure

$$\mathbb{P}\left(W_1^2 \leq z, W_1^2 > 0\right) = \mathbb{E}\left(\mathbb{1}\left\{1/W_2^1 \leq z\right\} W_2^1\right), \quad z > 0, \tag{7}$$

which implies that $\mathbb{E}\left(W_2^1\right) = 1 - \mathbb{P}\left(W_1^2 = 0\right) \leq 1$.

An elementary proof of (7) can be found in Section S.7 of Supplementary Material S1.

We now proceed to a stochastic representation for $\mathbf{Y}^m$ that involves only the univariate random variables W_i^j. Define a new, directed tree $T^m = (V, E^m)$ rooted at an arbitrary but fixed node $m \in V$. The edge set E^m consists of all edges $e \in E$ of the tree T pointing away from node m. For the resulting directed tree we define a set $\{W_e : e \in E^m\}$ of independent random variables, where for $e = (i, j)$, the distribution of $W_e = W_j^i$ is jth coordinate of the extremal function of $\mathbf{Y}$ relative to coordinate i.

The following result generalises proposition 2 in Engelke and Hitz (2020) to extremal tree models with arbitrary edge distributions.

Proposition 1. *Let $\mathbf{Y}$ be a multivariate Pareto distribution that is an extremal graphical model on the tree $T = (V, E)$. Let P be a standard Pareto distribution, independent of $\{W_e : e \in E^m\}$. Then we have the joint stochastic representation for $\mathbf{Y}^m$ on $\mathcal{L}^m$*

$$Y_i^m \overset{(d)}{=} \begin{cases} P, & \text{for } i = m, \\ P \times \prod_{e \in \mathrm{ph}(mi; T^m)} W_e, & \text{for } i \in V \backslash \{m\}, \end{cases} \tag{8}$$

where $\mathrm{ph}(mi; T^m)$ *denotes the set of edges on the unique path from node m to node i on the tree T^m; see Figure 1 for an example with $m = 2$.*

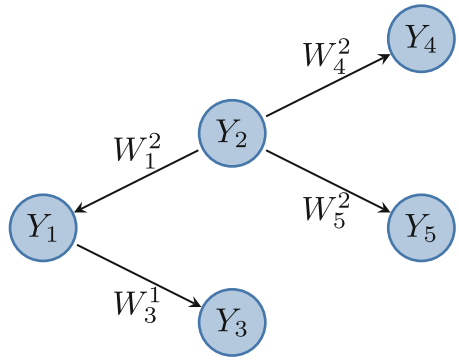

FIGURE 1 A tree T^2 rooted at node $m = 2$ with the extremal functions on the edges [Colour figure can be viewed at wileyonlinelibrary.com]

Conversely, for any set of independent random variables $\{W_i^j, W_j^i; \{i,j\} \in E\}$, *where* W_i^j *and* W_j^i *satisfy the duality (7), the construction (8) defines a consistent family of extremal functions* $\mathbf{W}^1, \ldots, \mathbf{W}^d$ *that correspond to a unique d-dimensional Pareto distribution* $\mathbf{Y}$ *that is an extremal graphical model on* T.

The above result formally establishes the link of the conditional independence in Definition 1 to the limiting tail trees in Segers (2020); see also Proposition 1 in Supplementary Material S1 for details on this link. In this sense, the first part of Proposition 1 can be deduced from theorem 1 in Segers (2020).

Note that $\mathbf{Y}$ defined as in Proposition 1 can also be an extremal graphical model on a disconnected graph G; see the paragraph after (5). The representation (8) then remains true and some of the W_e are almost surely equal to zero.

Remark 1. It is remarkable that for an extremal tree model $\mathbf{Y}$, the distribution of its extremal functions, and therefore also of the multivariate Pareto distribution itself, is characterised by the set of univariate random variables $\{W_i^j, W_j^i; \{i,j\} \in E\}$. This indicates that the probabilistic structure is simpler than in the non-extremal case, where in general both univariate and bivariate distributions are needed to describe a tree graphical model.

3 | THE EXTREMAL VARIOGRAM

Covariance matrices play a central role in structure learning for Gaussian graphical models due to their connection to conditional independence properties. In multivariate extreme value theory, several summary statistics have been developed to measure the strength of dependence between the extremes of different variables. The most popular one is the extremal correlation, which for $i,j \in V$ is defined as

$$\chi_{ij} := \lim_{q \to 0} \chi_{ij}(q) := \lim_{q \to 0} \mathbb{P}\left\{F_i(X_i) > 1 - q | F_j(X_j) > 1 - q\right\}, \tag{9}$$

whenever the limit exists. It ranges between 0 and 1 where the boundary cases are asymptotic independence and complete extremal dependence, respectively (cf, Coles et al., 1999; Schlather & Tawn, 2003). In particular, if $\mathbf{X}$ is in the max-domain of attraction of the multivariate Pareto distribution $\mathbf{Y}$, then the extremal correlation always exists and $\chi_{ij} = \mathbb{P}(Y_i > 1 | Y_j > 1)$. There are many other coefficients for extremal dependence in the literature, including the madogram (Cooley et al., 2006) and a coefficient defined on the spectral measure introduced in Larsson and

Resnick (2012) and used for dimension reduction in Cooley and Thibaud (2019) and Fomichov and Ivanovs (2022).

While designed as summaries for extremal dependence, none of these coefficients has an obvious relation to conditional independence for multivariate Pareto distributions or density factorisation in extremal graphical models of Engelke and Hitz (2020). In this section we define a new coefficient that will turn out to take a similar role in multivariate extremes as covariances in non-extremal models.

3.1 | Limiting extremal variogram

The variogram is a well-known object in geostatistics that measures the degree of spatial dependence of a random field (cf, Chilès & Delfiner, 2012; Wackernagel, 2013). It is similar to a covariance function, but instead of positive definiteness, a variogram is conditionally negative definite; for details, see for instance Engelke and Hitz (2020, Appendix B). For Brown–Resnick processes, the seminal work of Kabluchko et al. (2009) has shown that negative definite functions play a crucial role in spatial extreme value theory. We define a variogram for general multivariate Pareto distributions.

Definition 2. For a multivariate Pareto distribution $\mathbf{Y}$ we define the extremal variogram rooted at node $m \in V$ as the matrix $\Gamma^{(m)}$ with entries

$$\Gamma^{(m)}_{ij} = \mathrm{Var}\left\{\log Y^m_i - \log Y^m_j\right\}, \quad i,j \in V, \tag{10}$$

whenever the right-hand side exists and is finite.

We can interpret the $\Gamma^{(m)}_{ij}$ as a distance between the variables Y^m_i and Y^m_j that is large if their extremal dependence is weak and *vice versa*.

Proposition 2. *Let* $\mathbf{Y}$ *be a multivariate Pareto distribution.*

(i) For $m \in V$*, we can express the extremal variogram in terms of the extremal function relative to coordinate* m*,*

$$\Gamma^{(m)}_{ij} = \mathrm{Var}\left\{\log W^m_i - \log W^m_j\right\}, \quad i,j \in V.$$

(ii) For $m \in V$*, the matrix* $\Gamma^{(m)}$ *is a variogram matrix, that is, it is conditionally negative definite.*

(iii) Let $\mathbf{Y}_n$ *be a sequence of multivariate Pareto distributions with extremal coefficients* $\chi_{n,im}$ *between the* i*th and* m*th coordinate of* $\mathbf{Y}_n$ *satisfying* $\chi_{n,im} \to 0$ *as* $n \to \infty$ *for some* $i, m \in V$*. Then the corresponding extremal variograms satisfy* $\Gamma^{(m)}_{n,im} \to \infty$ *as* $n \to \infty$*.*

Part (iii) in the above proposition underlines the interpretation of the extremal variogram. When the variables become asymptotically independent, then the extremal variogram grows and eventually diverges to $+\infty$. Note that the converse statement is not true in general, since there are cases where $\Gamma^{(m)}_{im} = \infty$ but $\chi_{im} > 0$. We proceed with several examples where the extremal variogram can be computed explicitly. Figure 2 shows the extremal variogram values for these models as a function of the corresponding extremal correlation.

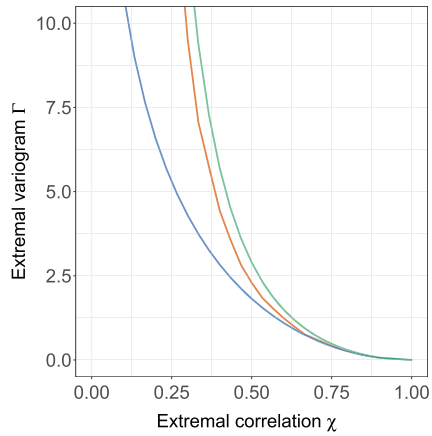

FIGURE 2 Values of the extremal variogram $\Gamma_{12}^{(1)}$ as a function of the extremal correlation χ_{12} for the bivariate Hüsler–Reiss (blue), symmetric Dirichlet (orange) and logistic (green) models. Note that in all three cases we have that $W_1^2 \overset{(d)}{=} W_2^1$ and therefore $\Gamma_{12}^{(1)} = \Gamma_{12}^{(2)}$. [Colour figure can be viewed at wileyonlinelibrary.com]

Example 2. The extremal logistic distribution with parameter $\theta \in (0, 1)$ can be defined through its extremal functions (see Dombry et al., 2016)

$$\mathbf{W}^m = (U_1/U_m, \ldots, U_d/U_m),$$

where $U_1, \ldots, U_d$ are independent and U_i, $i \neq m$ follow Fréchet$(1/\theta, G(1-\theta)^{-1})$ distributions, and $(G(1-\theta)U_m)^{-1/\theta}$ follows a Gamma$(1-\theta, 1)$ distribution; here $G(x)$ is the Gamma function evaluated at $x \geq 0$. It turns out that for the logistic model we have

$$\Gamma_{ij}^{(m)} = \begin{cases} \pi^2\theta^2/3, & \text{if } i,j \neq m, \\ \theta^2\{\psi^{(1)}(1-\theta) + \pi^2/6\}, & \text{if } i = m, j \neq m, \end{cases}$$

where $\psi^{(1)}$ is the trigamma function defined as the second derivative of the logarithm of the gamma function.

The corresponding extremal correlations have the form $\chi_{ij} = 2 - 2^{\theta}$, $i,j \in V$.

The proof of this representation of the extremal variogram in the logistic model can be found in Section S.8 in Supplementary Material S1.

Example 3. The extremal Dirichlet distributions with parameters $\alpha_1, \ldots, \alpha_d$ (cf, Coles & Tawn, 1991) has extremal functions

$$\mathbf{W}^m = (U_1/U_m, \ldots, U_d/U_m),$$

where $U_1, \ldots, U_d$ are independent and U_i, $i \neq m$ follow Gamma$(\alpha_i, 1/\alpha_i)$ distributions, and U_m follows a Gamma$(\alpha_m + 1, 1/\alpha_m)$ distribution. By straight-forward calculations,

$$\Gamma_{ij}^{(m)} = \begin{cases} \psi^{(1)}(\alpha_i) + \psi^{(1)}(\alpha_j), & \text{if } i,j \neq m, \\ \psi^{(1)}(\alpha_m + 1) + \psi^{(1)}(\alpha_j), & \text{if } i = m, j \neq m, \end{cases}$$

with $\psi^{(1)}$ denoting the trigamma function as in Example 2.

The corresponding extremal correlations do have have a closed form but can be calculated numerically.

For the class of Hüsler–Reiss distributions the extremal variogram turns out to be very natural.

Example 4. The Hüsler–Reiss distribution is parameterised by a variogram matrix $\Gamma \in \mathbb{R}^{d\times d}$; see Engelke and Hitz (2020) for details. For any d-variate centred normal random vector $\mathbf{U}$ with variogram matrix Γ, the extremal function relative to coordinate $m \in V$ has representation

$$\mathbf{W}^m = \exp\{\mathbf{U} - U_m - \Gamma_{\cdot m}/2\}, \tag{11}$$

see Dombry et al. (2016, prop. 4). The extremal variogram $\Gamma^{(m)}$ for any $m \in V$ is then equal to the variogram matrix Γ from the definition of the Hüsler–Reiss distributions, and, in particular, it is independent of the root node,

$$\Gamma_{ij} = \Gamma_{ij}^{(1)} = \cdots = \Gamma_{ij}^{(d)}, \quad i,j \in V.$$

The corresponding extremal correlations have the form $\chi_{ij} = 2 - 2\Phi\left(\sqrt{\Gamma_{ij}}/2\right)$, where Φ is the standard normal distribution function.

3.2 | Pre-asymptotic extremal variogram

Similar to the extremal correlation in (9) we can define the extremal variogram as the limit of pre-asymptotic versions.

Definition 3. For a multivariate distribution $\mathbf{X}$ with continuous marginal distributions we define the pre-asymptotic extremal variogram at level $q \in (0,1)$ rooted at node $m \in V$ as the matrix $\Gamma^{(m)}(q)$ with entries

$$\Gamma_{ij}^{(m)}(q) = \mathrm{Var}\left[\log\{1 - F_i(X_i)\} - \log\{1 - F_j(X_j)\} | F_m(X_m) > 1 - q\right], \quad i,j \in V,$$

whenever right-hand side exists and is finite.

Note that for q close to zero the conditional distribution of the terms $-\log\{1 - F_i(X_i)\}$ given $F_m(X_m) > 1 - q$ is approximately that of $\log Y_i^m$, $i \in V$.

Next we provide conditions which ensure the convergence $\Gamma_{ij}^{(m)}(q) \to \Gamma_{ij}^{(m)}$ as $q \to 0$. We introduce the following notation: for a vector $\mathbf{x} \in \mathbb{R}^d$ and $I \subset \{1, \ldots, d\}$, let $\mathbf{x}_I$ denote a vector in $\mathbb{R}^{|I|}$ with entries $x_j, j \in I$. For a distribution function F of a d-dimensional random vector $\mathbf{X}$ define F_I as the distribution function of the corresponding random vector $\mathbf{X}_I$ and let $\mathbf{Y}_{(I)}$ denote the limit obtained in relation (2) when $F, \mathbf{X}, \mathbf{x}$ are replaced by $F_I, \mathbf{X}_I, \mathbf{x}_I$. Note that $\mathbf{Y}_{(I)}$ is not the same as $\mathbf{Y}_I$, the sub-vector of $\mathbf{Y}$ with entries in I, because the latter is not supported on $\mathcal{L}_I = \{\mathbf{x} \geq \mathbf{0} : \|\mathbf{x}\|_\infty > 1\} \subset \mathbb{R}^{|I|}$. The distribution of $\mathbf{Y}_{(I)}$ can be obtained from that of $\mathbf{Y}_I$ by conditioning.

(B) There exist constants $\xi > 0, K_B < \infty$ such that for any $I \subset V$ with $|I| \in \{2,3\}$ and all $q \in (0,1)$

$$\sup_{\mathbf{x}_I \in [1,\infty]^{|I|}} \left| \mathbb{P}\left(F_I(\mathbf{X}_I) \leq 1 - q/\mathbf{x}_I | F_I(\mathbf{X}_I) \not\leq 1 - q\right) - \mathbb{P}\left(\mathbf{Y}_{(I)} \leq \mathbf{x}_I\right) \right| \leq K_B q^\xi. \tag{12}$$

(T) There exists a $\gamma > 0$ such that for any $i, m \in V$ the extremal function satisfies

$$\mathbb{E}\left(W_i^m\right)^{-\gamma} \leq K_W < \infty. \tag{13}$$

Assumption (B) is a strengthening of (2) for bivariate and trivariate distributions as it imposes that convergence to the limit should take place uniformly and at a certain rate. It is closely related to typical second order conditions on the stable tail dependence function that are fairly standard in the literature; see for instance Einmahl et al. (2012) and Fougères et al. (2015) among many others. Additional details on this matter are given in the Section S.12.1 in Supplementary Material S1. Condition (T) is a mild assumption on the extremal functions W_i^m, which holds for all examples considered in the previous section. This condition prevents the distribution of W_i^m from putting too much mass close to zero.

Proposition 3. *Under conditions (B), (T) we have for any $m, i, j \in V$*

$$\Gamma_{ij}^{(m)}(q) \to \Gamma_{ij}^{(m)}, \qquad \text{as } q \to 0.$$

We note that condition (T) already implies that $\Gamma_{ij}^{(m)} \in [0, \infty)$ for any i, j, so the convergence above is always to a finite limit.

4 | STRUCTURE LEARNING FOR EXTREMAL TREE MODELS

4.1 | Extremal tree models

Extremal graphical models where the underlying graph is a tree were considered as a sparse statistical model in Engelke and Hitz (2020). As explained in the introduction, their approach of using a censored maximum-likelihood tree becomes prohibitively costly in higher dimension d and requires parametric assumptions on the bivariate distributions of the tree.

Ideally, one would like to have summary statistics, similar to the correlation coefficients r_{ij} in the Gaussian case, that can be estimated empirically and that guarantee to recover the true underlying tree structure when used as edge weights. The extremal variogram defined in Section 3 turns out to be a so-called tree metric, and as such a natural quantity to infer the conditional independence structure in extremal tree models. We underline that the extremal variogram $\Gamma^{(m)}$ is defined for arbitrary multivariate Pareto distributions and in the case of the Hüsler–Reiss distribution it coincides with the parameter matrix.

Proposition 4. *Let $\mathbf{Y}$ be an extremal graphical model with respect to the tree $T = (V, E)$ and suppose that the extremal variogram matrix $\Gamma^{(m)}$ exists for all $m \in V$. Then we have that*

$$\Gamma_{ij}^{(m)} = \sum_{(s,t) \in \mathrm{ph}(ij;T)} \Gamma_{st}^{(m)}. \tag{14}$$

In other words, for any $m \in V$, the extremal variogram matrix $\Gamma^{(m)}$ defines an additive tree metric.

Corollary 1. *Let* $\mathbf{Y}$ *be an extremal graphical model with respect to the tree* $T = (V, E)$*. Suppose that the extremal variogram matrix* $\Gamma^{(m)}$ *exists and is finite for all* $m \in V$ *and that* $\mathbb{P}(Y_i \neq Y_j) > 0$ *for all* $i, j \in V$, $i \neq j$ *(or equivalently,* $\Gamma^{(m)}_{ij} > 0$*). For any* $m \in V$*, the minimum spanning tree with* $\rho_{ij} = \Gamma^{(m)}_{ij}$ *is unique and satisfies*

$$T_{\text{mst}} = T.$$

For extremal tree models, Corollary 1 shows that independently of any distributional assumption, the extremal variogram contains the conditional independence structure of the tree T. This result is quite surprising, since it is stronger than what is known in the classical, non-extremal theory of trees. Indeed, as discussed in the introduction, for Gaussian graphical models, a analogous result holds for a minimum spanning tree with weights $\rho_{ij} = \log\left(1 - r_{ij}^2\right)/2$ for r_{ij} denoting the correlation between the ith and jth component of the Gaussian random vector under consideration. The assumption of Gaussianity is crucial and the result no longer holds outside this specific parametric class.

Beyond the world of Gaussian graphical models, there exists some literature on the non-parametric estimation of graphical models on tree structures, see Chow and Liu (1968) for an early contribution and Drton and Maathuis (2017, section 3.1) for an overview. However, one either needs to assume discrete distributions (Chow & Liu, 1968) or the existence of densities (Lafferty et al., 2012; Liu et al., 2011), and non-parametric density estimation is required in the latter case. To the best of our knowledge, multivariate Pareto distributions are the first example for a non-parametric sub-class of multivariate distributions where tree dependence structures can be learned using simple moment-based summary statistics without additional parametric assumptions. It is also remarkable that there is no need to assume the existence of densities and that the distributions we consider can simultaneously have continuous and discrete components.

The reason why such a strong result can hold can be explained by the homogeneity of the multivariate Pareto distribution $\mathbf{Y}$. For trees, all cliques contain two nodes and therefore the density $f_{\mathbf{Y}}$ factorises into bivariate Pareto densities. Because of the homogeneity, such a bivariate density can be decomposed into independent radial and angular parts; see Example 1. Bivariate Pareto distributions only differ in terms of the angular distribution, whose support is the subset of a one-dimensional sphere with all coordinates positive. Consequently, an extremal tree model in d dimensions can essentially be reduced to $d - 1$ univariate angular distributions; see also Proposition 1. This provides an intuitive explanation why the result in Corollary 1 can hold.

We can go further and show that a linear combination of the matrices $\Gamma^{(m)}$, $m \in V$, which are possibly different from each other, still induces the true tree as the minimum spanning tree.

Corollary 2. *Under the same assumptions as in Corollary 1, the minimum spanning tree with distances*

$$\rho_{ij} = \sum_{m=1}^{d} w_m \Gamma^{(m)}_{ij},$$

given by a linear combination of the extremal variograms rooted at different nodes with coefficients $w_m \geq 0$, $m \in V$, $\max_{m \in V} w_m > 0$*, is unique and satisfies* $T_{\text{mst}} = T$.

The extremal correlation coefficients χ_{ij} do not form a tree metric, that is, they are not additive according to the tree structure as the extremal variogram in (A4). It is therefore a non-trivial question whether these coefficients can also be used as weights in a minimum spanning tree to infer the underlying conditional independence structure. Interestingly, the next result gives a partially affirmative answer.

Proposition 5. *Let* $\mathbf{Y}$ *be an extremal graphical model on the tree* $T = (V, E)$*. Then the extremal correlation coefficients satisfy for any* $h, l \in V$ *with* $h \neq l$ *that*

$$\chi_{hl} \leq \chi_{ij} \quad \forall (i,j) \in \mathrm{ph}(hl; T). \tag{15}$$

Under the additional assumption that this inequality is strict as soon as $(i,j) \neq (h,l)$*, the minimum spanning tree corresponding to distances* $\rho_{ij} = -\log(\chi_{ij})$ *is unique and satisfies*

$$T_{\mathrm{mst}} = T.$$

The assumption that $\chi_{hl} < \chi_{ij}$ for any $(i,j) \in \mathrm{ph}(hl; T)$ with $(i,j) \neq (h,l)$ is not satisfied for all tree models. Indeed, a counterexample (Segers, J. [Personal communication, 7th July 2022]) with index set $V = \{1,2,3\}$ and edges $E = \{(1,2),(2,3)\}$ is the following. Let the extremal function $W_2^1 \sim \mathrm{Unif}([1/2, 3/2])$ and let W_3^2 have a discrete distribution $\mathbb{P}\big(W_3^2 = 1/4\big) = 4/5$ and $\mathbb{P}\big(W_3^2 = 4\big) = 1/5$; both are valid extremal functions as in Example 1. In this case $\chi_{13} = \chi_{23} = 2/5$ and the set of minimum spanning trees is not unique. We can then only guarantee that the true underlying tree T lies in the set of all possible minimum spanning trees; this follows from a close inspection of the proof of Proposition 5.

There are simple conditions to ensure that inequality (15) is strict for all $(i,j) \neq (h,l)$. For instance, a sufficient condition for this to hold is that all extremal functions W_j^i for $(i,j) \in E$ have support equal to the whole space $[0, \infty)$; see Lemma 1 in Section S.10 in Supplementary Material S1. This covers many relevant examples such as the Hüsler–Reiss, the extremal logistic and the extremal Dirichlet distributions in Examples 2, 3 and 4, respectively. A weaker condition for strict inequality was recently obtained by Hu et al. (2022).

Remark 2. Both the extremal variogram $\Gamma_{ij}^{(m)}$ and the extremal correlation χ_{ij} contain information on conditional independence structure for extremal tree models. The extremal correlation is defined for any model but needs additional assumptions to correctly recover the tree. The extremal variogram does not exist if $\mathbf{Y}$ has mass on lower-dimensional sub-faces of $\mathcal{L}$ but is guaranteed to recover the underlying tree whenever all extremal variograms exist. When their sample versions are used (see Section 4.2), the probability of correctly identifying the underlying tree may differ even when both approaches work on population level; see Section 5.

4.2 | Estimation

Throughout this section assume that we observe independent copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the d-dimensional random vector $\mathbf{X}$, which is in the max-domain of attraction of a multivariate Pareto distribution $\mathbf{Y}$, an extremal graphical model on the tree T according to (5). Our aim is to estimate T from the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Motivated by Proposition 5 and Corollaries 1, 2 we propose to achieve this through a two-step procedure. We first construct estimators for the quantities χ_{ij} and $\Gamma_{ij}^{(m)}$, and then compute the minimal spanning trees corresponding to those estimators.

The empirical estimator for χ_{ij} is defined as

$$\hat{\chi}_{ij} := \frac{n}{k} \sum_{t=1}^{n} \mathbb{1}\big\{\tilde{F}_i(X_{ti}) > 1 - k/n, \tilde{F}_j(X_{tj}) > 1 - k/n\big\},$$

where $k = k_n$ is an intermediate sequence and $\tilde{F}_i$ denotes the empirical distribution function of $X_{1i}, \ldots, X_{ni}$. Standard arguments imply that under (2) and provided that $k \to \infty$ and $k/n \to q \in [0,1]$ as $n \to \infty$, we have for any $i, j \in V$

$$\hat{\chi}_{ij} = \chi_{ij}(q) + o_{\mathbb{P}}(1), \qquad \text{as } n \to \infty, \tag{16}$$

where $\chi_{ij}(q)$ is defined in (9) and $\chi_{ij}(0) := \chi_{ij}$. In particular, if $q = 0$ then $\hat{\chi}_{ij}$ is a consistent estimator of χ_{ij}.

The extremal variogram matrix $\Gamma^{(m)}$ for the sample $\mathbf{X}_t$, $t = 1, \ldots, n$, is estimated by

$$\hat{\Gamma}^{(m)}_{ij} := \widehat{\mathrm{Var}}\left(\log(1 - \tilde{F}_i(X_{ti})) - \log(1 - \tilde{F}_j(X_{tj})) \,:\, \tilde{F}_m(X_{tm}) \geq 1 - k/n\right),$$

where $\widehat{\mathrm{Var}}$ denotes the sample variance. Under the assumption $k/n \to q \in [0,1]$ as $n \to \infty$ and mild conditions on the underlying data generation, this estimator can be shown to be consistent for the pre-asymptotic version $\Gamma^{(m)}_{ij}(q)$ as introduced in Definition 3.

Theorem 1. *Let assumptions (B), (T) hold and assume that $k \geq n^{\theta}$ for some $\theta > 0$ and that $k/n \to q \in [0,1]$ as $n \to \infty$. Then we have for any $m, i, j \in V$*

$$\hat{\Gamma}^{(m)}_{ij} = \Gamma^{(m)}_{ij}(q) + o_{\mathbb{P}}(1), \qquad \text{as } n \to \infty,$$

where $\Gamma^{(m)}_{ij}(0) := \Gamma^{(m)}_{ij}$.

The proof of this result turns out to be surprisingly technical, details are given in the Section S.12.4 in Supplementary Material S1. The main challenge arises from the fact that in the definition of $\Gamma^{(m)}_{ij}$ only the observations in component m are extreme while observations in other components may also be non-extreme. This is different from the setting that is typically considered in asymptotically dependent extreme value theory.

Remark 3. By choosing $q = 0$, the above theorem implies consistency of the empirical extremal variogram $\hat{\Gamma}^{(m)}$. This result is of independent interest, since it is the first proof of consistency of the moment estimators

$$\hat{\Sigma}^{(m)}_{ij} = \frac{1}{2}\left\{\hat{\Gamma}^{(m)}_{im} + \hat{\Gamma}^{(m)}_{jm} - \hat{\Gamma}^{(m)}_{ij}\right\}, \quad i, j \neq m,$$

which were introduced in Engelke et al. (2015) as estimators for the parameters of the Hüsler–Reiss distribution.

Remark 4. The assumption that the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent was only made to keep the presentation simple. The consistency result in Theorem 1 continues to hold under a high-level assumption that allows for temporal dependence and is spelled out in detail at the beginning of Section S.12.4 in Supplementary Material S1.

Now we have all results that are needed for consistent estimation of the underlying tree structure. Given a general distance ρ with estimator $\hat{\rho}$ on pairs $(i, j) \in V \times V$, we consider plug-in procedures of the form

$$\hat{T}_{\rho} := \arg\min_{T=(V,E)} \sum_{(i,j)\in E} \hat{\rho}_{ij}, \tag{17}$$

with three cases of particular interest given by

$$\hat{\rho}_{ij} = -\log(\hat{\chi}_{ij}), \qquad \hat{\rho}_{ij} = \hat{\Gamma}^{(m)}_{ij}, \qquad \hat{\rho}_{ij} = \sum_{m=1}^{d} w_m \hat{\Gamma}^{(m)}_{ij},$$

resulting in the estimators $\hat{T}_\chi, \hat{T}^{(m)}_\Gamma, \hat{T}^w_\Gamma$, respectively. The special case of $w_1 = \cdots = w_d = 1/d$ is denoted by $\hat{T}_\Gamma$. We solve the minimum spanning tree problem (17) by Prim's algorithm, which is guaranteed to find a global optimiser of problem (17) that is unique if the distances $\hat{\rho}_{ij}$ are distinct for all pairs (Prim, 1957).

Theorem 2. *Assume that* $\mathbf{Y}$ *is an extremal graphical model on the tree* T. *Assume (2) holds and that the inequality in (15) is strict whenever* $(i,j) \neq (h,l)$. *If* $k \to \infty$ *as* $n \to \infty$ *then there exists* $q^* > 0$ *such that under the additional assumption* $k/n \to q \in [0, q^*]$ *as* $n \to \infty$,

$$\mathbb{P}(\hat{T}_\chi = T) \to 1, \qquad \text{as } n \to \infty.$$

If instead of (2) and strict inequality in (15) assumptions (B), (T) hold, $\mathbb{P}(Y_i \neq Y_j) > 0$ *for all* $i,j \in V$, $i \neq j$ *(or equivalently,* $\Gamma^{(m)}_{ij} > 0$*), and if* $k \geq n^\theta$ *for some* $\theta > 0$ *then for any* $m \in V$ *there exists* $q^*_m > 0$ *such that for* $k/n \to q \in [0, q^*_m]$ *as* $n \to \infty$, *we have*

$$\mathbb{P}\big(\hat{T}^{(m)}_\Gamma = T\big) \to 1, \qquad \text{as } n \to \infty.$$

The same is true for $\hat{T}^w_\Gamma$ *provided the weights* w_m *satisfy* $w_m \geq 0, \max_m w_m > 0$.

Remark 5. As pointed out by a referee, it would be of interest to find weights that maximise (asymptotically) the probability of correct tree structure recovery by $\hat{T}^w_\Gamma$. This would require precise information on the joint asymptotic distribution of $\hat{\Gamma}^{(m)}$ for different m, which is currently an open question.

Remark 6. At first glance it might seem surprising that the tree structure can be estimated consistently even when k_n/n does not converge to zero. The latter would be a classical minimal assumption in extreme value theory and would be required for consistent estimation of χ_{ij} or $\Gamma^{(m)}_{ij}$. We explain the intuition behind this result for the extremal correlation, the arguments for the extremal variogram are exactly the same. Assume that the inequality in (15) is strict whenever $(i,j) \neq (h,l)$, making the minimal spanning tree with respect to $-\log \chi_{ij}$ unique. The key insight is that even biased estimators of $\chi_{ij}(q)$ can lead to the correct minimal spanning tree since all we need is

$$\sum_{(i,j)\in E'} -\log \chi_{ij}(q) > \sum_{(i,j)\in E} -\log \chi_{ij}(q),$$

for all trees $T' = (V, E') \neq T$, where T denotes the true underlying tree. Multivariate regular variation (2) implies that $\chi_{ij}(q) \to \chi_{ij}$ as $q \to 0$ for all i,j, so there exists $q_0 > 0$ such that the above inequality is satisfied for all $q < q_0$. Since in addition $\hat{\chi}_{ij} = \chi_{ij}(k/n) + o_{\mathbb{P}}(1)$ as $n \to \infty$ under the assumption $k \to \infty$, consistency follows.

Theorem 2 shows that the proposed procedures are able to consistently recover the tree structure under rather weak assumptions on the sequence $k = k_n$. It is natural to wonder which choices

of k correspond to higher probabilities of recovering the tree structure consistently. Here we provide some indicative discussion of this issue for minimal spanning trees based on χ_{ij} without going into technical details. Standard results from empirical process theory show that under mild assumptions and for $k/n \to q \in [0,1]$ as $n \to \infty$, all $\sqrt{k}(\hat{\chi}_{ij} - \chi_{ij}(k/n))$ converge jointly to a multivariate normal distribution with covariance matrix Σ_q. The latter matrices satisfy $\Sigma_q \to \Sigma_0$ as $q \to 0$. Combined with the delta method this implies that for any tree $T' = (V, E') \neq T$

$$\sum_{(i,j)\in E'} \hat{\rho}_{ij} - \sum_{(i,j)\in E} \hat{\rho}_{ij} = \Delta_{k,n} + \frac{1}{\sqrt{k}} Z_{k,n} := \sum_{(i,j)\in E'} \rho_{ij}(k/n) - \sum_{(i,j)\in E} \rho_{ij}(k/n) + \frac{1}{\sqrt{k}} Z_{k,n},$$

where $\rho_{ij}(k/n) := -\log \chi_{ij}(k/n)$ and $\hat{\rho}_{ij} := -\log \hat{\chi}_{ij}$, and $Z_{k,n}$ is a weighted linear combination of differences $\sqrt{k}(\hat{\chi}_{ij} - \chi_{ij}(k/n))$ and thus approximately centred normal with variance σ_q^2. The probability that the sum over estimated distances on T' is shorter than the sum over true tree T is given by $\mathbb{P}(-Z_{k,n} > \sqrt{k}\Delta_{k,n})$. Under the assumptions for asymptotic normality of $\hat{\chi}_{ij}$, $\Delta_{k,n}$ converges to $\Delta(q) := \sum_{(i,j)\in E'} \rho_{ij}(q) - \sum_{(i,j)\in E} \rho_{ij}(q)$. Combining all of the above approximations we find $\mathbb{P}(-Z_{k,n} > \sqrt{k}\Delta_{k,n}) \approx \mathbb{P}(\sigma_q \mathcal{N}(0,1) > \sqrt{n}\sqrt{q}\Delta(q))$. Since $\sigma_q \to \sigma_0 > 0$ and $\Delta(q) \to \Delta(0) > 0$ as $q \to 0$, it is easy to see that there exists $q_0 > 0$ such that $\sqrt{q}\Delta(q)/\sigma_q < \sqrt{q_0}\Delta(q_0)/\sigma_{q_0}$ for all $q < q_0$, and thus the probability of selecting T' instead of the true tree T starts to increase as the limit of k/n decreases after q_0. This suggests that an optimal value for k in terms of maximising the probability of estimating the true tree would satisfy $k/n \to \tilde{q}$ as $n \to \infty$ for some $\tilde{q} > 0$. Turning the above arguments into a formal proof would require many technicalities which are beyond the scope of the present paper, but the intuition obtained here is also confirmed in the simulations in Section 5.

4.3 | Estimation in growing dimensions

The consistency results in the previous section were derived for data of fixed dimension for sample size tending to infinity. Here we provide an extension of those results by adding non-asymptotic bounds on the probability of consistently estimating the true tree. Throughout this section, the underlying tree can change with the sample size n.

We start with discussing results for $\hat{T}_\chi$. This requires the following additional notation. Assume that $\mathbf{Y}$ is an extremal graphical model on the tree $T = (V, E)$ and define the corresponding extremal correlation $\chi_{ij}^Y := \mathbb{P}(Y_i > 1 | Y_j > 1)$. Let

$$\mu_\chi^Y := \min_{(h,l)\notin E} \min_{(i,j)\in \mathrm{ph}(hl;T)} \left(\chi_{ij}^Y - \chi_{hl}^Y \right).$$

To gain some intuition on the reason for this definition, recall that in (15) in Proposition 5 we show that

$$\chi_{hl}^Y \leq \chi_{ij}^Y \quad \forall (i,j) \in \mathrm{ph}(hl;T).$$

To ensure that the minimal spanning tree corresponding to $-\log \chi_{ij}^Y$ is unique, we need to rule out equality in the above statement whenever $(i,j) \neq (h,l)$, which follows from $\mu_\chi^Y > 0$. Thus the quantity μ_χ^Y can be interpreted as a lower bound on the increase of the sum of distances on the

edges if we move from the true tree T to $T' \neq T$. This is formalised in the proof of Theorem 3. We are now ready to state the first main result.

Theorem 3. *Assume that* $\mathbf{Y}$ *is an extremal graphical model on the tree* T *and that* $\mathbf{X}_1, \dots, \mathbf{X}_n$ *are independent copies of* $\mathbf{X}$*, a random vector with continuous marginal distributions. Let* $\chi_{ij}(k/n)$ *denote the pre-asymptotic extremal coefficients corresponding to* $\mathbf{X}$ *in the sense of (9) and define*

$$\delta_{k/n} := \max_{i \neq j} \left| \chi_{ij}(k/n) - \chi_{ij}^Y \right|.$$

Then there exists a universal constant $K > 0$ *such that*

$$\mathbb{P}\left(\hat{T}_\chi \neq T\right) \leq 5d^2 \exp\left(-\frac{3k}{10} \left\{ \left(\frac{\mu_\chi^Y - 2\delta_{k/n}}{2K} \right)_+^2 \wedge 1 \right\} \right). \tag{18}$$

Note that above we did not assume that $\mathbf{X}$ is in the max-domain of attraction of $\mathbf{Y}$. A link between $\mathbf{X}$ and $\mathbf{Y}$ is implicitly provided through $\delta_{k/n}$ which measures the distance between $\chi_{ij}(k/n)$ computed from $\mathbf{X}$ and the extremal coefficients χ_{ij}^Y which correspond to $\mathbf{Y}$.

Some comments on the implications of the above result are in order. On a high level, larger dimensions d, smaller values of μ_χ^Y, and larger bias $\delta_{k/n}$ lead to a larger bound. The effects of dimension d and bias $\delta_{k/n}$ are intuitive: larger dimensions or more bias make the tree recovery problem more difficult. The effect of μ_χ^Y is also expected because smaller values of μ_χ^Y imply that, on population level, there exist trees that are closer to the true tree and estimation becomes more difficult.

For a more quantitative discussion assume that (B) in Section 3.2 holds with constants K_B, ξ independent of n, d. In this case $\delta_{k/n} \leq K_R(k/n)^\xi$ for a possibly different constant K_R which is still independent of n, d, ξ; see (S.18) in Section S.13.3 in Supplementary Material S1. Note that the exponent can be bounded by $-3n(k/n)\{[(\mu_\chi^Y - 2K_B(k/n)^\xi)_+^2/(4K^2)] \wedge 1\}$. Straightforward but tedious computations optimising this rate over k show that the largest achievable rate for this exponent is of order $n(\mu_\chi^Y)^{2+1/\xi}$ if we let $k = cn(\mu_\chi^Y)^{1/\xi}$ for a suitable constant $c \in (0, \infty)$ which depends on K, K_B, ξ only. With this choice of k consistent tree structure recovery is possible if $\log d = o\left(n\left(\mu_\chi^Y\right)^{2+1/\xi}\right)$ as $n \to \infty$. If μ_χ^Y stays bounded away from zero this simplifies to $\log d = o(n)$ as $n \to \infty$, which allows the dimension to grow exponentially in n. In contrast, if the dimension d is fixed but we consider observations from a triangular array with the same tree but changing value of μ_χ^Y, we require $n(\mu_\chi^Y)^{2+1/\xi} \to \infty$ as $n \to \infty$, provided that k is chosen as described above. This condition becomes more stringent if ξ is smaller, which is intuitive since it corresponds to slower decaying bias.

We now discuss tree structure recovery with $\hat{T}_\Gamma^{(m)}$ and $\hat{T}_\Gamma^w$. A key result here are concentration bounds on $\hat{\Gamma}_{ij}^m$. Such bounds are established in Engelke et al. (2021) and reproduced in the proof of Theorem 4 given in the Section S.13.3 in Supplementary Material S1. To state those bounds we need an additional assumption.

(D) For all $I \subset V$ with $|I| = 2$ the random variables $\mathbf{Y}_{(I)}$ have densities f_I. There exists an $\varepsilon > 0$ such that for all $\beta \in [-\varepsilon, 1-\varepsilon]$ there is a constant $K(\beta)$ such that

$$f_I(x, y) \leq K(\beta) \frac{1}{y^{1+\beta} x^{2-\beta}} \quad x, y \in (1, \infty)^2.$$

This is equivalent to assumption 2 in Engelke et al. (2021); see the discussion around (S.61) in Section S.13.3 in Supplementary Material S1. Engelke et al. (2021) show that it holds for Hüsler–Reiss distributions, for instance. This condition is implied by the simpler but stronger condition $f_I(x, 2-x) \leq K_r(x(2-x))^{1+\varepsilon}$ for some $\varepsilon > 0$ and all $x \in (0,2)$; this follows from elementary calculations involving the homogeneity of f_I which is derived in (S.60) in Section S.13.3 in Supplementary Material S1.

Theorem 4. *Assume that* $\mathbf{Y}$ *is an extremal graphical model with respect to the tree* T *and that* $\mathbf{X}_1, \dots, \mathbf{X}_n$ *are independent samples of* $\mathbf{X}$*, a random vector with continuous marginal distributions. Assume that (B), (T), (D) hold and that* $k \geq n^\theta$ *for some* $\theta > 0$*. Then there exist constants* $c, C, M > 0$ *depending only on the constants from (B), (T), (D) and* θ *such that for all* $k \geq 1$ *and* $b_{k/n} := (k/n)^\kappa (\log(n/k))^2$ *where* $\kappa := \gamma\xi/(1+\gamma+\xi)$

$$\mathbb{P}\left(\hat{T}_\Gamma^{(m)} \neq T\right) \leq Md^3 \exp\left(-ck\left\{\left(\frac{\min_{(i,j)\in E}\Gamma_{ij}^{(m)}}{2C} - b_{k/n}\right)_+^2 \wedge \frac{1}{(\log n)^8}\right\}\right). \tag{19}$$

For $w_m \geq 0$ *with* $\sum_{m=1}^d w_m = 1$ *the same bound holds for* $\mathbb{P}\left(\hat{T}_\Gamma^{w} \neq T\right)$ *with* $\min_{(i,j)\in E}\Gamma_{ij}^{(m)}$ *replaced by* $\min_{(i,j)\in E}\sum_{m=1}^d w_m \Gamma_{ij}^{(m)}$.

We note that Assumption (D) can be dropped at the cost of introducing an additional $1/(\log n)^4$ factor; details are provided in Section S.13.3 in Supplementary Material S1. Similarly to Theorem 3 we do not explicitly assume that $\mathbf{X}$ is in the domain of attraction of $\mathbf{Y}$. Assumption (B) provides the link between $\mathbf{X}$ and $\mathbf{Y}$ in terms of their bivariate and trivariate distributions.

We briefly comment on the result in Theorem 4. Observe that the general structure of the bound is similar to the corresponding result for tree structure recovery based on χ. The fact that $\wedge 1$ in (18) is replaced by $\wedge(\log n)^{-8}$ in (19) is due to technical details in the derivation of tail bounds for $\hat{\Gamma}_{ij}^{(m)}$, which has a more complex structure than the simple estimator $\hat{\chi}_{ij}$. Similarly to μ_χ^Y in (18), $\min_{(i,j)\in E}\Gamma_{ij}^{(m)}$ can be interpreted as measuring the minimal separation between the length of shortest and second-shortest minimal spanning tree. The quantity $b_{k/n}$ appearing in Theorem 4 stems from bounds on bias terms in estimating $\Gamma_{ij}^{(m)}$ and plays a similar role as $\delta_{k/n}$ for χ_{ij}. Comments on the fastest possible growth of the dimension d and minimal separation conditions that still allow for consistent tree structure recovery follow along the same lines as in the discussion following Theorem 3 and are omitted for the sake of brevity.

5 | SIMULATIONS

The minimum spanning trees based on the empirical versions of the extremal variogram and extremal correlation both recover asymptotically the underlying extremal tree structure. In this section we study the finite sample behaviour of the different tree estimators on simulated data. The results and figures of Sections 5 and 6 can be reproduced with the code at https://github.com/sebastian-engelke/extremal_tree_learning.

Let $T = (V, E)$ be a random tree structure that is generated by sampling uniformly $d-1$ edges and adding these to the empty graph under the constraint to avoid circles. Throughout the whole section, we simulate n samples $\mathbf{X}_1, \dots, \mathbf{X}_n$ from a random vector $\mathbf{X}$ in the domain of attraction of a multivariate Pareto distribution $\mathbf{Y}$ that is an extremal graphical model on the tree T in dimension $d = |V|$. As random vector $\mathbf{X}$ we take the corresponding max-stable distribution

(e.g., de Haan, 1984), which is indeed in the domain of attraction of $\mathbf{Y}$ in the sense of (2). In order to perturb the samples, a common way is to add lighter-tailed noise (e.g., Einmahl et al., 2016). More precisely,

$$\mathbf{X}_i = \mathbf{Z}_i + \varepsilon_i, \quad \varepsilon_i \perp\!\!\!\perp \mathbf{Z}_i, \qquad i = 1, \dots, n, \tag{20}$$

where $\mathbf{Z}_i$ is a max-stable random vector with standard Fréchet margins associated to $\mathbf{Y}$, and ε_i is a lighter-tailed noise vector which is independent of $\mathbf{Z}_i$. We consider two scenarios for the noise distribution, where in both cases the marginal distribution is transformed to a Fréchet distribution with $\mathbb{P}(\varepsilon_{ij} \leq x) = \exp(-1/x^2)$, $x \geq 0, j \in V$.

(N1) The noise vector ε_i has independent entries.
(N2) The noise vector ε_i in (20) is generated from an extremal tree model on a fixed tree T_{noise} that is generally different from the true tree T.

Since the marginals of the noise vector have lighter tails, the limit of $\mathbf{X}_i$ in (2) is not altered by ε_i. The main difference between the two noise mechanisms lies in the type of bias they introduce for large k, and we observe that this has an interesting impact on the recovery of the tree structure underlying $\mathbf{Y}$.

We consider two different parametric classes of distributions for $\mathbf{Y}$.

(M1) The Hüsler–Reiss tree model is a multivariate Pareto distribution that factorises on $T = (V, E)$, where each bivariate distribution (Y_i, Y_j) for $(i, j) \in E$ is Hüsler–Reiss with parameter $\Gamma_{ij} > 0$; see Example 4. The joint distribution is then also Hüsler–Reiss with parameter matrix Γ induced by the tree structure through (14). The coefficients on the edges are generated as

$$\Gamma_{ij} \sim \text{Unif}([0.2, 1]), \quad (i, j) \in E.$$

(M2) For the second model we let each bivariate distribution be given by the family of asymmetric Dirichlet distributions; see Example 3. We generate the two parameters of the bivariate Dirichlet models independently as

$$\alpha_1, \alpha_2 \sim \text{Unif}([1, 10]), \quad (i, j) \in E.$$

Note that the resulting d-dimensional Pareto distribution is not in the family of Dirichlet distributions.

We compare four different estimators for the weights on the minimum spanning tree $\hat{T}_\rho = (V, \hat{E}_\rho)$ in (17):

(i) $\hat{\rho}_{ij} = -\log \hat{\chi}_{ij}$, where $\hat{\chi}_{ij}$ is the empirical extremal correlation;
(ii) $\hat{\rho}_{ij} = \hat{\Gamma}_{ij}^{(m)}$, the extremal variogram estimator for one fixed $m \in V$;
(iii) $\hat{\rho}_{ij} = \hat{\Gamma}_{ij}$, the combined extremal variogram estimator;
(iv) $\hat{\rho}_{ij}$ are the censored negative log-likelihoods of the bivariate Hüsler–Reiss model (Y_i, Y_j), evaluated at the optimiser.

The estimators (i)–(iii) were introduced in Section 4.2 and their consistency has been derived. The estimator in (iv) is the one used in Engelke and Hitz (2020) to learn the structure of

Hüsler–Reiss tree models. Note that for this estimator, no theoretical justification is available. As performance measures we choose the average proportion of wrongly estimated edges

$$\mathbb{E}_{T=(V,E)}\mathbb{E}\left(1-\frac{|\hat{E}_\rho \cap E|}{d-1}\right), \tag{21}$$

and the probability of not recovering the correct tree structure

$$\mathbb{E}_{T=(V,E)}\mathbb{P}(\hat{T}_\rho \neq T), \tag{22}$$

where the outer expectations signify that the tree T is randomly generated in each repetition. Each experiment is repeated 300 times in order to estimate these errors empirically. We report only the results on the structure recovery rate error (22) and provide the corresponding results on the wrong edge rate (21) in the Section S.2 in Supplementary Material S1.

We first investigate the choice of the intermediate sequence $k = k_n$ of the number of exceedances used for estimation. We simulate from the Hüsler–Reiss tree model (M1) in dimension $d = 20$ and consider the minimum spanning trees $\hat{T}_\Gamma$ and $\hat{T}_{\mathrm{CL}}$ based on the combined extremal variogram and the censored likelihoods, respectively. Figure 3 shows the structure

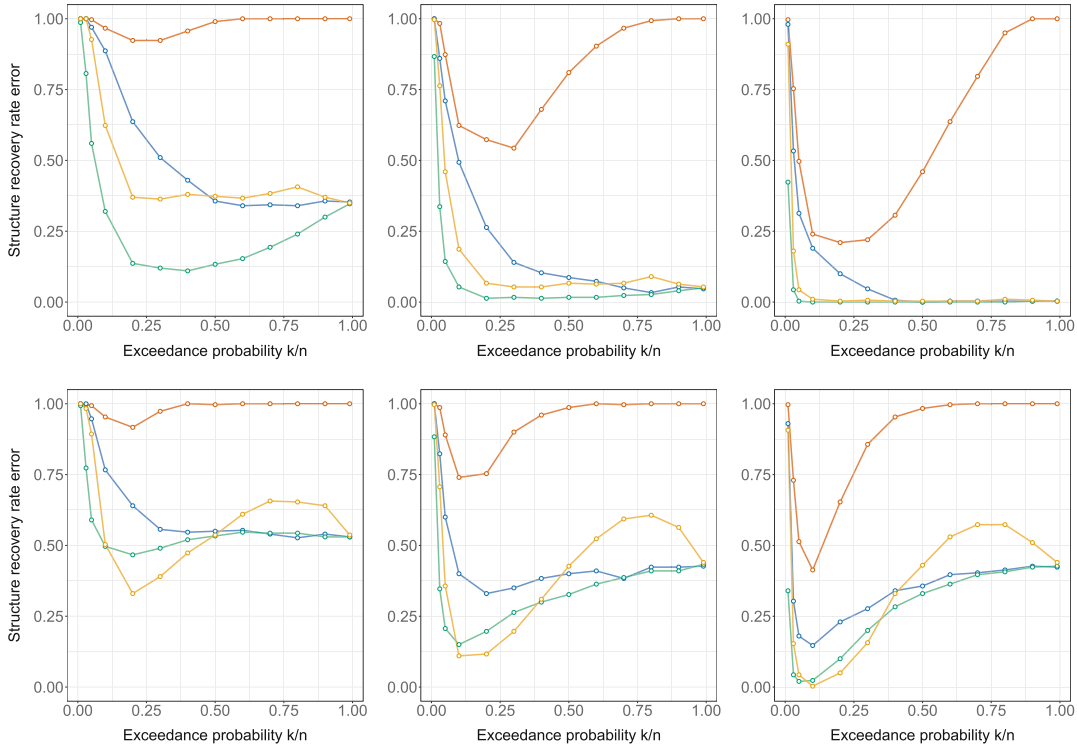

FIGURE 3 Structure recovery rate error of trees from the Hüsler–Reiss model (M1) and independent noise (N1) (top) and tree noise (N2) (bottom) in dimension $d = 20$ estimated based on empirical correlation (orange), extremal variogram with fixed $m \in V$ (blue), combined empirical variogram (green) and censored maximum likelihood (yellow) as a function of the exceedance probability k/n; sample sizes $n = 500$ (left column), $n = 1000$ (center column) and $n = 2000$ (right column) [Colour figure can be viewed at wileyonlinelibrary.com]

recovery rate error as a function of the exceedance probability k/n for different samples sizes n. Interestingly, the two noise patterns lead to qualitatively different results: while consistent recovery of the limiting tree seems possible even when $k = n$ for noise model (N1), noise model (N2) with a dependence structure also introduces a bias in the corresponding minimal spanning tree and the true tree cannot be recovered when the limit of k/n is too large. It is interesting to observe that the optimal exceedance probability k/n seems to converge to a positive value q^*, especially for noise (N1). This is consistent with the intuition given at the end of Section 4.2 in the paragraph after Remark 6. This is in contrast to classical asymptotic theory for consistent estimation in extremes where $k = o(n)$ is required to remove the approximation bias and therefore $q^* = 0$.

Next we compare the performance of the different structure learning methods for varying sample size n. Since the value of q^* which is required for consistent estimation is unknown in practice we choose $k = \lfloor n^{0.8} \rfloor$, which satisfies all assumptions of our theory. The results for dimension $d = 20$ are shown in the top row of Figure 4 for the Hüsler–Reiss model (M1) and in the bottom row for the asymmetric Dirichlet model (M2). We observe that the two methods based on the extremal variogram perform consistently better that the extremal correlation-based method. Intuitively this can be explained by the fact that the extremal variogram is a tree metric for conditional independence of multivariate Pareto distributions. The additivity on the tree results in a bigger loss in the minimum spanning tree algorithm when choosing a wrong edge, and therefore it is easier to identify the true structure. The extremal correlation only satisfies a weaker relation (15) on the tree, which might be a reason for the higher error rate. Additionally, the empirical variogram uses information from the entire multivariate Pareto distribution, while the extremal correlation evaluates its distribution at a single point only. A comparison with the censored maximum likelihood estimator (iv) yields several insights. First, this approach seems to lead to consistent estimation of the tree structure even in model (M2) where $\mathbf{Y}$ is not a Hüsler–Reiss distribution and the likelihood is thus misspecified. This might be explained by the fact that the strength of dependence is still sufficiently well estimated and the minimum spanning tree does only require correct ordering of the edge weights, which is much weaker than consistency of the estimated weights. Second, the different types of noise distributions in (N1) and (N2) lead to opposing orderings of the best method: whereas $\hat{T}_{\mathrm{CL}}$ has a slight advantage for noise (N2), $\hat{T}_{\Gamma}$ performs substantially better under (N1). Notably, this is even the case in model (M1) where the likelihood is well-specified. A possible explanation is that the likelihood is not exactly specified due to the added noise in the model and the use of ranks during estimation. This implies that classical results about asymptotic optimality of maximum likelihood methods do not apply here. Moreover, the added noise has different effects on the biases of the estimators, which changes the order of performance depending on the noise distribution.

For a given tree, the task of estimating the correct structure can largely differ according to the strength of dependence of the multivariate Pareto distribution. We therefore conduct a simulation study where we fix $n = 500$ and $k = \lfloor n^{0.8} \rfloor$ and illustrate the performance of the structure estimation methods for a varying strength of tail dependence. For the Hüsler–Reiss model, we randomly generate a tree $T = (V, E)$ in dimension $d = 20$ and for $(i, j) \in E$ we fix all $\Gamma_{ij} = \lambda$ to some constant $\lambda > 0$. Equivalently, that means that all neighbouring nodes have extremal correlation $\chi_{ij} = 2 - 2\Phi\left(\sqrt{\lambda}/2\right)$. The left panel of Figure 5 shows the results for varying strength of extremal dependence between neighbours measured by the extremal correlation under noise model (N1). Unsurprisingly, the performance of all methods deteriorates at the boundaries, which correspond to the non-identifiable cases of independence and complete dependence. In general, it seems that the empirical variogram-based estimators perform better under stronger dependence, which is

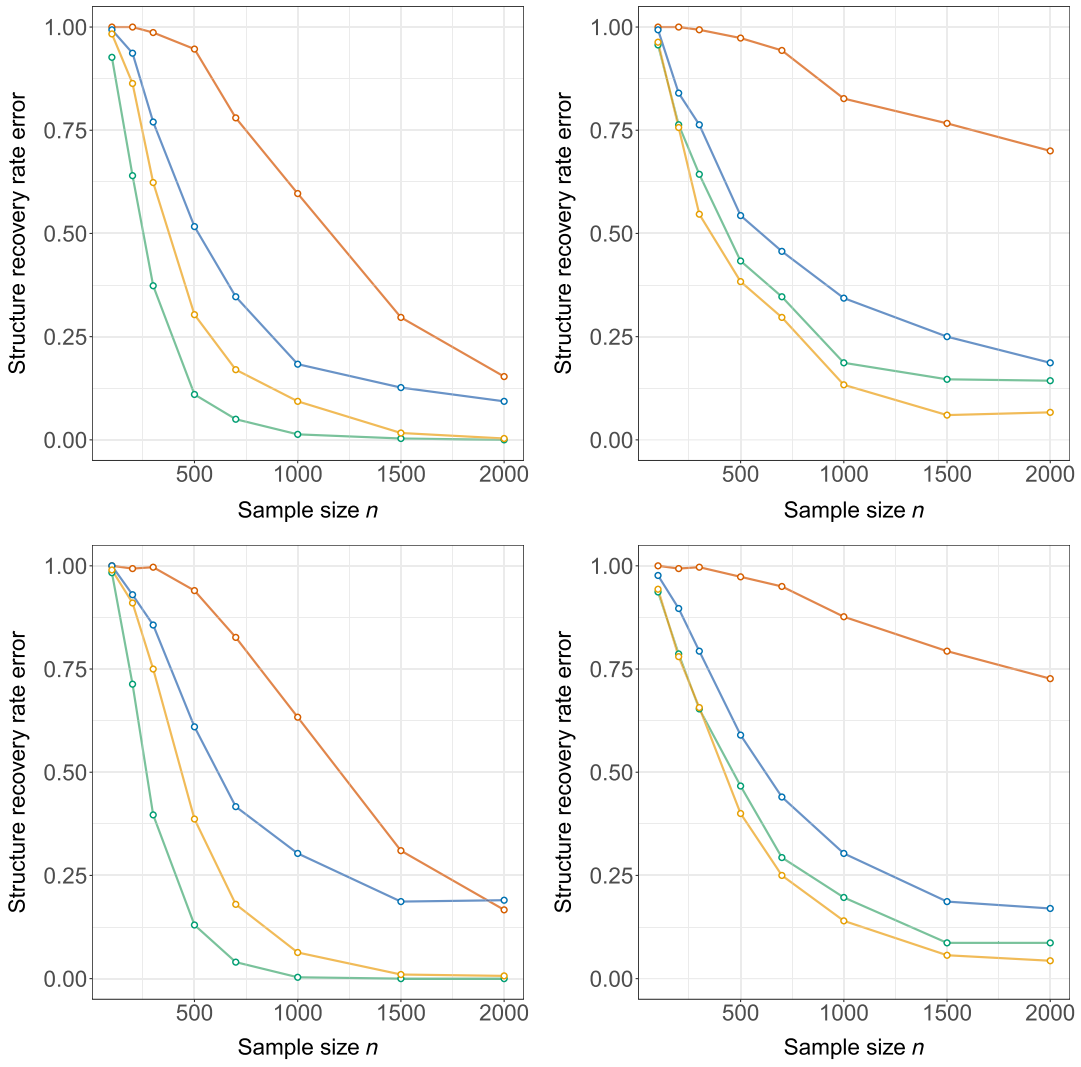

FIGURE 4 Structure recovery rate error of trees from the Hüsler–Reiss model (M1) (top) and Dirichlet model (M2) (bottom) in dimension $d = 20$ estimated based on empirical correlation (orange), extremal variogram with fixed $m \in V$ (blue), combined empirical variogram (green) and censored maximum likelihood (yellow); independent noise (N1) (left) and tree noise (N2) (right) [Colour figure can be viewed at wileyonlinelibrary.com]

probably due to the higher bias of the empirical extremal variogram under weak dependence. The same asymmetry can be observed for the censored maximum likelihood method, while the performance of the extremal correlation seems to be symmetric around $\chi = 1/2$. Comparing the performance of different methods, we observe that under noise (N1) the combined extremal variogram performs best uniformly in the values of χ, and the advantage over all other methods can be substantial. The same analysis with noise (N2) is shown in the right panel of Figure 5. In line with the results in Figure 4, the performance of $\hat{T}_{\mathrm{CL}}$ and $\hat{T}_{\Gamma}$ is fairly similar, with a slight advantage for $\hat{T}_{\mathrm{CL}}$ at values of χ around 0.5 and the converse for χ closer to 0.2 and 0.8.

For the final set of comparisons we study the performance of the methods for a growing dimension $d \in \{10, 20, 30, 50, 100, 200, 300\}$, where we fixed the sample size $n = 1000$ and number of

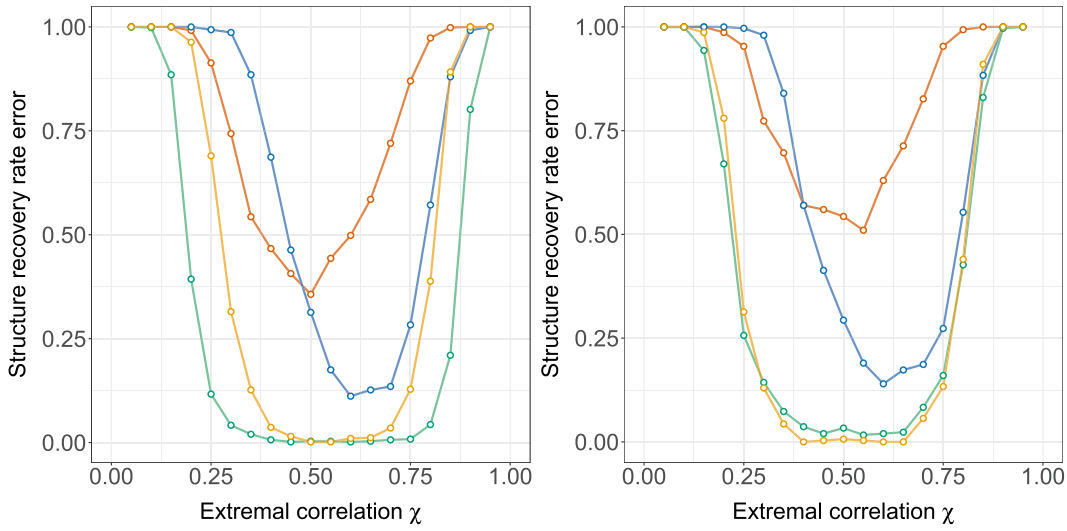

FIGURE 5 Structure recovery rate error for the Hüsler–Reiss model (M1) with noise model (N1) (left) and (N2) (right) in dimension $d = 20$ as a function of the extremal dependence between neighbours measured by the extremal correlation χ; the different methods are based on empirical correlation (orange), extremal variogram with fixed $m \in V$ (blue), combined empirical variogram (green) and censored maximum likelihood (yellow) [Colour figure can be viewed at wileyonlinelibrary.com]

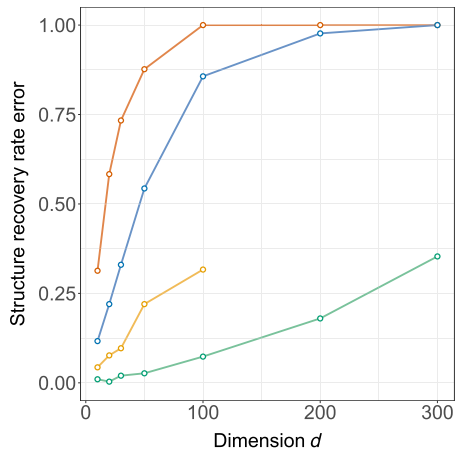

FIGURE 6 Structure recovery rate error for the Hüsler–Reiss model (M1) with noise model (N1) (left) as a function of the number of nodes $|V| = d$ of the tree; the different methods are based on empirical correlation (orange), extremal variogram with fixed $m \in V$ (blue), combined empirical variogram (green) and censored maximum likelihood (yellow) [Colour figure can be viewed at wileyonlinelibrary.com]

exceedances $k = \lfloor n^{0.8} \rfloor = 251$. Figure 6 shows the structure recovery rate errors for the different methods. As expected, the errors increase for larger dimensions, but much slower for the combined extremal variogram than for the other methods. Theoretical pre-asymptotic error bounds for the error rates can be found in Section 4.3. We remark that for we were not able to run simulations in more than $d = 100$ dimensions for the censored maximum likelihood estimator because of the prohibitive computational cost. For the same reason we have only included simulations in one model and one noise setting.

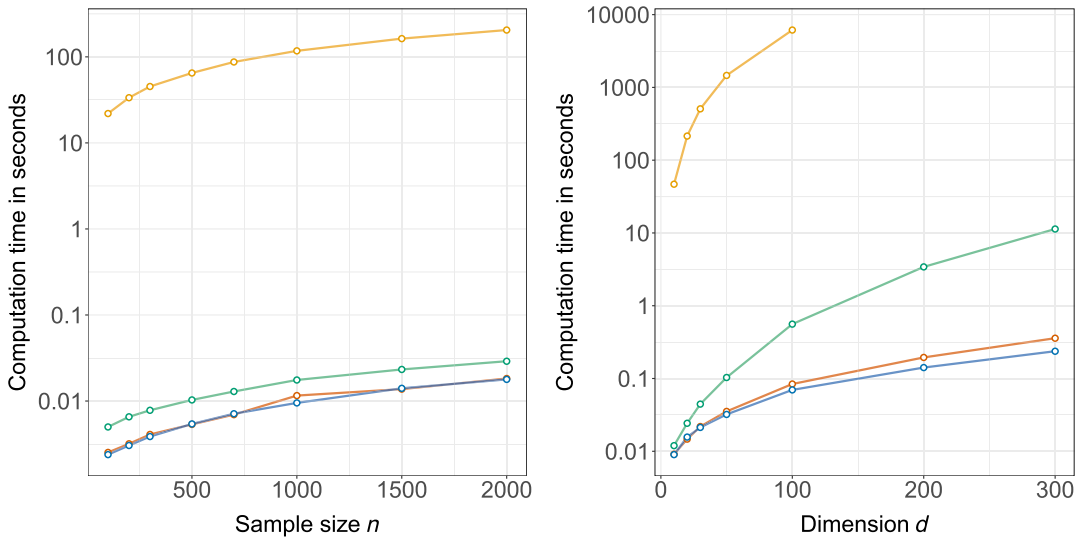

FIGURE 7 Average computation times in seconds in the simulation studies in Section 5 of the four algorithms based on empirical correlation (orange), extremal variogram with fixed $m \in V$ (blue), combined empirical variogram (green) and censored maximum likelihood (yellow). Left: for fixed dimension $d = 20$; right: for fixed sample size $n = 1000$ [Colour figure can be viewed at wileyonlinelibrary.com]

We close this section with some comments on computation times for the four estimators. The extremal correlation and variogram-based trees rely on empirical estimators and are very efficient to compute. The censored likelihood estimator however requires numerical optimisation for every weight ρ_{ij}, $i,j \in V$. Especially in higher dimensions this becomes prohibitively costly. Figure 7 shows the average computation times for the four estimators in the simulations in Figures 4 and 6. It can be seen that the censored likelihood method is several orders of magnitude slower than the empirical methods. As seen in the right-hand panel of Figure 7, this quickly becomes prohibitive if the dimension grows.

6 | APPLICATION

We illustrate the proposed methodology on foreign exchange rates of $d = 26$ currencies expressed in terms of the British Pound sterling; see Table A1 in Appendix A.1 for the three-letter abbreviations of the respective countries. The data are available from the website of the Bank of England[1]. They consist of daily observations of spot foreign exchange rates in the period from 1 October 2005 to 30 September 2020, resulting in $n = 3790$ observations.

In order to obtain time series without temporal dependence, we pre-process the data set. We first compute the daily log-returns R_{ij}, $i = 1, \ldots, n, j = 1, \ldots, d$, from the original time series. To remove the serial dependence, we then filter the univariate series by ARMA-GARCH processes; see Hilal et al. (2014) for a similar approach, and Bollerslev et al. (1992) and Engle (1982) for background on financial time series modelling. The AIC suggest that an ARMA(0, 2)-GARCH(1, 1) model is the most appropriate for most of the univariate series. We derive the absolute values of the standardised filtered returns as

[1]https://www.bankofengland.co.uk/

$$X_{ij} = \left| \frac{R_{ij} - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}} \right|,$$

where $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}$ are the estimated mean and SD of the ARMA-GARCH model. The absolute value means that we are interested in extremes in both directions.

The data $\mathbf{X}_i = (X_{i1}, \dots, X_{id})$ are approximately independent and identically distributed for $i = 1, \dots, n$, and we will model their tail dependence using an extremal tree model. We first check whether the assumption of asymptotic dependence is satisfied by inspecting the behaviour of the function $q \mapsto \hat{\chi}_{ij}(q)$ for values $q = k/n$ close to 1. For most of the pairs this function seems to converge to a positive value and thus there is fairly strongly dependence in the tail between the filtered log-returns; see Figure S3 in the Section S.4 in Supplementary Material S1 for some examples.

Before estimating for this data set the extremal tree structure, we discuss the choice of the number of exceedances k, or equivalently the probability of exceedance $q = k/n$. This is an important practical issue and a long-standing problem in extreme value theory. In essence, it is a bias-variance trade-off as illustrated in the simulations in Figure 3.

For tree structure estimation, we propose to leverage the specific structure of the tree learning problem. From (14) it follows that for an extremal graphical model on a tree T, the corresponding population $\Gamma = d^{-1}\sum_{m=1}^{d}\Gamma^{(m)}$ forms a tree metric on that tree. In the sequel, for generic Γ and tree T, denote the extremal variogram matrix completed on the tree T by

$$\Gamma_{ij}^{T} = \sum_{(s,t)\in \mathrm{ph}(ij;T)} \Gamma_{st}, \quad i,j \in V.$$

We propose to select k so as to minimise the deviation of the empirical values of $\hat{\Gamma}$ from forming a tree metric on the estimated tree $\hat{T}$. More precisely, define

$$\hat{\Delta}(k/n) = \sum_{i,j\in V} \left\{ g\big(\hat{\Gamma}_{ij}^{\hat{T}}(k/n)\big) - g\big(\hat{\Gamma}_{ij}(k/n)\big) \right\}^2, \tag{23}$$

where as function $g : R^+ \to [0,1]$, we choose the transformation from Γ to χ in Hüsler–Reiss models, that is, $g(x) = 2 - 2\Phi(\sqrt{x}/2)$; see Example 4. Here we indicate that these estimates depend on the exceedance probability $q = k/n$; note that also the estimated tree $\hat{T}$ depends on k. Additional motivation for the form of $\hat{\Delta}(k/n)$ and a literature review of classical approaches is given in Section S.3 in Supplementary Material S1.

Motivated by the simulations in the previous section we estimate the extremal tree structure $\hat{T}_\Gamma = (V, \hat{E}_\Gamma)$ non-parameterically using the combined empirical extremal variogram $\hat{\Gamma}$. The left panel of Figure 8 shows the error $\hat{\Delta}(k/n)$ as a function of $q = k/n$ for the exchange rate data set. It can be seen that, indeed, the error seems to stabilise for values of $1 - q$ above 0.97. We therefore choose $q = 0.03$ in this application, which corresponds to $k = 114$. The corresponding minimum spanning tree is shown in Figure 9; we note that the tree is very stable across different values of q close to 0.

The structure of the tree allows for a nice interpretation of extremal dependence. Extreme observations in the exchange rates with the Euro are strongly connected with extremes of other European currencies in Northern and Eastern countries. The graph suggests that extremes of exchange rates of these currencies are conditionally independent of exchange rates of other countries, given the value of Euro exchange rate. The Malaysian ringgit, the Chinese yuan, the Hong Kong dollar and the Taiwan dollar are strongly pegged to the US dollar and their closeness

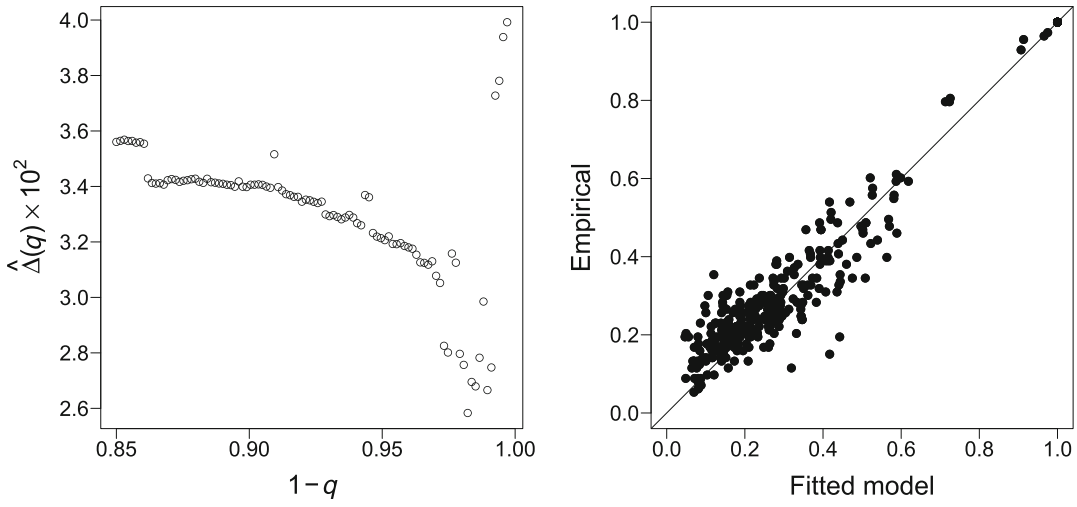

FIGURE 8 Left: The squared error $\hat{\Delta}(q)$ defined in (23) between empirical extremal correlations and those implied by the tree metric structure for different values of the exceedance probability $q = k/n$. Right: Extremal correlations for the spot foreign exchange rate data implied by the fitted Hüsler–Reiss tree model against the empirical counterparts

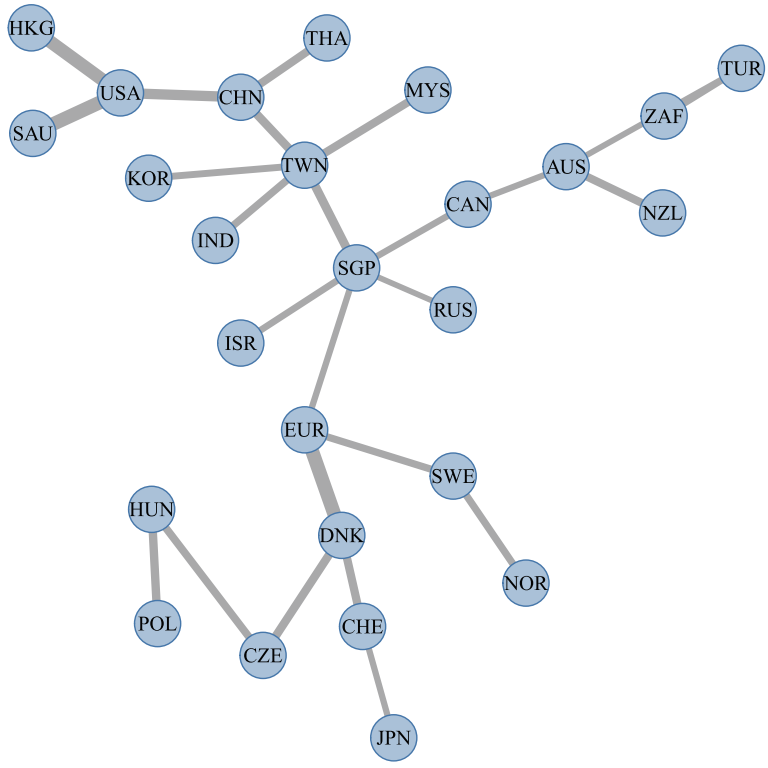

FIGURE 9 Minimum spanning tree $\hat{T}_\Gamma$ of extremal dependence for the spot foreign exchange rate data based on the combined extremal variogram. The width of each edge $(i,j) \in \hat{E}_\Gamma$ is proportional to the extremal correlation $2 - 2\Phi\left(\sqrt{\hat{\Gamma}_{ij}}/2\right)$, and therefore wider edges indicate stronger extremal dependence. [Colour figure can be viewed at wileyonlinelibrary.com]

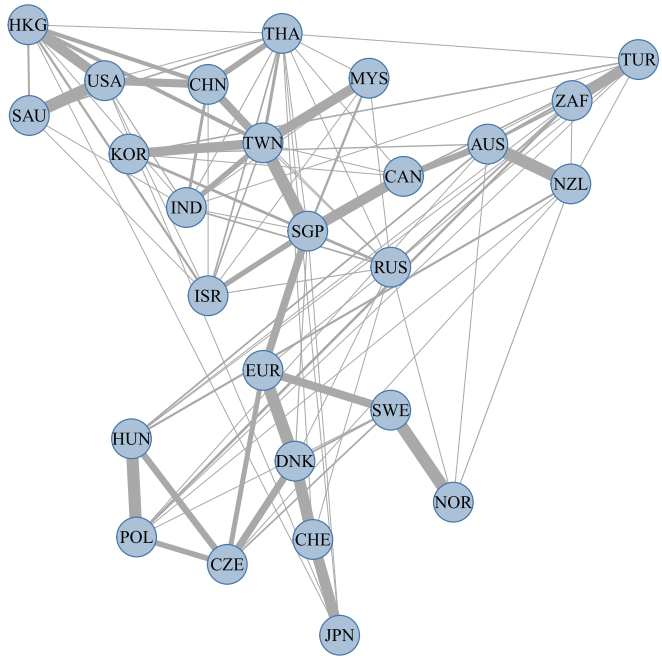

FIGURE 10 Graph where the width of each edge is proportional to the number of times it has been selected in an extremal tree in the bootstrap procedure. [Colour figure can be viewed at wileyonlinelibrary.com]

in the tree is therefore not surprising. Another branch of the tree contains several currencies of the Commonwealth. Finally, the connection between Japan and Switzerland is plausible because both currencies can be considered safe-haven currencies, which are both popular investments in times of crises.

In order to address the stability of the tree structure we bootstrap our data $B = 100$ times and fit each time the tree structure. For generating each bootstrap sample, we draw with replacement n data from the sample of filtered observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. This is a heuristic approach to assess the overall stability of our empirical conclusions to small perturbations in the data and does not have a formal theoretical justification at this point. Figure 10 shows that graph where the width of each edge is proportional to the number of times it has been selected in an extremal tree. Overall, the tree seems to be fairly stable since there is only a small number of dominant edges. Moreover, we can identify clear clusters that are connected in most of the trees, such as the European currencies. On the other hand some currencies such as the Russian ruble that do not have a dominant connection to any of these clusters. In future research, it could therefore be interesting to study structure estimation for forests, which allow to have unconnected graphs whose connected components are trees (Liu et al., 2011).

So far we have not assumed any specific model for the extremal dependence on the edges since we are able to estimate the tree structure fully non-parametrically with the methods from this paper. If we were only interested in interpretation of the extremal graphical structure we could stop our analysis here. If we require a model for rare event simulation or risk assessment, in a second step we can choose arbitrary bivariate Pareto models for each edge. For simplicity, we choose here for all edges the Hüsler–Reiss model (see Example 4) resulting in a Hüsler–Reiss tree. For this model, the bivariate parameter estimates $\hat{\Gamma}_{ij}$ can be chosen directly as the empirical extremal variogram estimates for all $\{i, j\} \in \hat{E}_\Gamma$. Alternatively, we could estimate them by censored

maximum likelihood. In both cases, the remaining entries of the Hüsler–Reiss parameter matrix can be obtained from the additivity of the extremal variogram on the tree in (14). We denote the corresponding parameter matrix completed on the tree $\hat{T}_\Gamma$ by $\hat{\Gamma}^{\hat{T}_\Gamma}$. Recall the relation between Γ and χ for Hüsler–Reiss models from Example 4. The right panel of Figure 8 shows the extremal correlations implied by the fitted Hüsler–Reiss tree model, that is, $\hat{\chi}_{ij}^{\hat{T}_\Gamma} = 2 - 2\Phi\left(\sqrt{\hat{\Gamma}_{ij}^{\hat{T}_\Gamma}}/2\right)$, against the empirical counterparts $\hat{\chi}_{ij}$, $i,j \in V$. Even though the tree structure is a very sparse graph with only $d-1$ edges, the extremal dependence between all variables is well-explained.

ACKNOWLEDGEMENTS

The authors would like to thank Jiaying Gu for pointing us to Hall's marriage theorem, and Johan Segers for pointing us to a mistake in an earlier version of Proposition 5. We further thank Nicola Gnecco, Adrien S. Hitz, Michaël Lalancette and Chen Zhou for helpful comments. We also thank the Associate Editor and two anonymous Referees for comments which helped us to improve the presentation and for encouraging us to consider results in growing dimensions. Sebastian Engelke was supported by an Eccellenza grant of the Swiss National Science Foundation and Stanislav Volgushev was partially supported by a discovery grant from NSERC of Canada.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in extremal_tree_learning at https://github.com/sebastian-engelke/.

ORCID

Sebastian Engelke https://orcid.org/0000-0001-6356-918X

REFERENCES

Asenova, S., Mazo, G. & Segers, J. (2021) Inference on extremal dependence in the domain of attraction of a structured Hüsler–Reiss distribution motivated by a Markov tree with latent variables. *Extremes*, 24, 461–500.

Beirlant, J., Goegebeur, Y., Teugels, J. & Segers, J. (2004) *Statistics of extremes. Wiley series in probability and statistics*. Chichester: John Wiley & Sons, Ltd.

Bollerslev, T., Chou, R.Y. & Kroner, K.F. (1992) Arch modeling in finance: a review of the theory and empirical evidence. *Journal of Econometrics*, 52(1), 5–59.

Chilès, J.-P. & Delfiner, P. (2012) *Geostatistics: modeling spatial uncertainty*. *Wiley Series in Probability and Statistics*, 2nd edition. Hoboken, NJ: John Wiley & Sons, Inc.

Chow, C. & Liu, C. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14, 462–467.

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

Coles, S.G. & Tawn, J.A. (1991) Modelling extreme multivariate events. *Journal of the Royal Statistical Society Series B. Methodological*, 53(2), 377–392.

Cooley, D., Naveau, P. & Poncet, P. (2006) Variograms for spatial max-stable random fields. In: Bertail, P., Soulier, P. & Doukhan, P. (Eds.) *Dependence in probability and statistics*. *Lecture Notes in Statistics*, Vol. 187. New York: Springer, pp. 373–390.

Cooley, D. & Thibaud, E. (2019) Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3), 587–604.

Cowell, R.G., Dawid, P., Lauritzen, S.L. & Spiegelhalter, D.J. (2006) *Probabilistic networks and expert systems: exact computational methods for Bayesian networks*. New York: Springer.

Dawid, A.P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 1–31.

de Haan, L. (1984) A spectral representation for max-stable processes. *The Annals of Probability*, 12, 1194–1204.

de Haan, L. & Ferreira, A. (2006) *Extreme value theory*. New York: Springer.

Dombry, C., Engelke, S. & Oesting, M. (2016) Exact simulation of max-stable processes. *Biometrika*, 103, 303–317.
Dombry, C., Eyi-Minko, F. & Ribatet, M. (2013) Conditional simulation of max-stable processes. *Biometrika*, 100(1), 111–124.
Drton, M.t. & Maathuis, M.H. (2017) Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1), 365–393.
Einmahl, J.H., Krajina, A., Segers, J. et al. (2012) An m-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3), 1764–1793.
Einmahl, J.H.J., Kiriliouk, A., Krajina, A. & Segers, J. (2016) An M–estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 275–298.
Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997) *Modelling extremal events: for insurance and finance*. London: Springer.
Engelke, S., de Fondeville, R. & Oesting, M. (2019) Extremal behaviour of aggregated data with an application to downscaling. *Biometrika*, 106, 127–144.
Engelke, S. & Hitz, A. (2020) Graphical models for extremes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 871–932.
Engelke, S., Hitz, S.A. & Gnecco, N. (2019) *graphical extremes: statistical methodology for graphical extreme value models. R package version 0.1.0*. Available from: https://CRAN.R-project.org/package=graphicalExtremes
Engelke, S. & Ivanovs, J. (2021) Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8, 241–270.
Engelke, S., Lalancette, M. & Volgushev, S. (2021) Learning extremal graphical structures in high dimensions. *arXiv preprint arXiv:2111.00840*.
Engelke, S., Malinowski, A., Kabluchko, Z. & Schlather, M. (2015) Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society Series B. Methodological*, 77(1), 239–265.
Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.
Fomichov, V. & Ivanovs, J. (2022) Spherical clustering in detection of groups of concomitant extremes. *Biometrika*, asac020. https://doi.org/10.1093/biomet/asac020.
Fougères, A.-L., De Haan, L., Mercadier, C. et al. (2015) Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2), 903–934.
Gissibl, N. & Klüppelberg, C. (2018) Max-linear models on directed acyclic graphs. *Bernoulli*, 24, 2693–2720.
Hall, P. (1935) On representatives of subsets. *The Journal of the London Mathematical Society*, 10, 26–30.
Hilal, S., Poon, S.-H. & Tawn, J. (2014) Portfolio risk assessment using multivariate extreme value methods. *Extremes*, 17, 531–556.
Hu, S., Peng, Z. & Segers, J. (2022) *Modelling multivariate extreme value distributions via Markov trees*. Available from: https://arxiv.org/abs/2208.02627
Kabluchko, Z., Schlather, M. & de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. *The Annals of Probability*, 37, 2042–2065.
Katz, R.W., Parlange, M.B. & Naveau, P. (2002) Statistics of extremes in hydrology. *Advances in Water Resources*, 25, 1287–1304.
Klüppelberg, C. & Lauritzen, S. (2019) *Bayesian networks for max-linear models*. Cham: Springer International Publishing, pp. 79–97.
Kruskal, J.B., Jr. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7, 48–50.
Lafferty, J., Liu, H. & Wasserman, L. (2012) Sparse nonparametric graphical models. *Statistical Science*, 27, 519–537.
Larsson, M. & Resnick, S.I. (2012) Extremal dependence measure and extremogram: the regularly varying case. *Extremes*, 15, 231–256.
Lauritzen, S.L. (1996) *Graphical models*. Oxford: Oxford University Press.
Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J. & Wasserman, L. (2011) Forest density estimation. *The Journal of Machine Learning Research*, 12, 907–951.
Papastathopoulos, I. & Strokorb, K. (2016) Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108, 9–15.
Poon, S.-H., Rockinger, M. & Tawn, J. (2004) Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *The Review of Financial Studies*, 17, 581–610.

Prim, R.C. (1957) Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36, 1389–1401.
Resnick, S.I. (2008) *Extreme values, regular variation and point processes*. New York: Springer.
Rootzén, H., Segers, J. & Wadsworth, J.L. (2018) Multivariate peaks over thresholds models. *Extremes*, 21(1), 115–145.
Rootzén, H. & Tajvidi, N. (2006) Multivariate generalized Pareto distributions. *Bernoulli*, 12, 917–930.
Schlather, M. & Tawn, J.A. (2003) A dependence measure for multivariate and spatial extreme values: properties and inference. *Biometrika*, 90, 139–156.
Segers, J. (2020) One-versus multi-component regular variation and extremes of Markov trees. *Advances in Applied Probability*, 52(3), 855–878.
Wackernagel, H. (2013) *Multivariate geostatistics: an introduction with applications*. New York: Springer.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Engelke, S. & Volgushev, S. (2022) Structure learning for extremal tree models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5), 2055–2087. Available from: https://doi.org/10.1111/rssb.12556

APPENDIX

A.1 Country codes used in the application in Section 6

Table A1 shows the three-letter country codes of the exchange rates into British Pound sterling.

TABLE A1 Three-letter country codes

Code	Foreign exchange rate (into GBP)	Code	Foreign exchange rate (into GBP)
AUS	Australian Dollar	NOR	Norwegian Krone
CAN	Canadian Dollar	POL	Polish Zloty
CHN	Chinese Yuan	RUS	Russian Ruble
CZE	Czech Koruna	SAU	Saudi Riyal
DNK	Danish Krone	SGP	Singapore Dollar
EUR	Euro	ZAF	South African Rand
a HKG	Hong Kong Dollar	KOR	South Korean Won
HUN	Hungarian	SWE	Swedish Krona
IND	Indian Rupee	CHE	Swiss Franc
ISR	Israeli Shekel	TWN	Taiwan Dollar
JPN	Japanese Yen	THA	Thai Baht
MYS	Malaysian ringgit	TUR	Turkish Lira
NZL	New Zealand Dollar	USA	US Dollar

A.2 Proof of Proposition 4

In order to show that the extremal variogram $\Gamma^{(m)}$ defines a tree metric on T, we recall the stochastic representation of $\mathbf{Y}^m$ in Proposition 1. We compute

$$\begin{aligned}\Gamma_{ij}^{(m)} &= \mathrm{Var}\left\{\sum_{e\in \mathrm{ph}(mi;T^m)} \log W_e - \sum_{\tilde{e}\in \mathrm{ph}(mj;T^m)} \log W_{\tilde{e}}\right\} \\ &= \sum_{e\in \mathrm{ph}(mi;T^m)\Delta \mathrm{ph}(mj;T^m)} \mathrm{Var}\left\{\log W_e\right\} \\ &= \sum_{(s,t)\in \mathrm{ph}(ij;T)} \Gamma_{st}^{(m)},\end{aligned}$$

where for two sets A and B, $A\Delta B$ denotes the symmetric difference. The second to last equality follows from the independence of the $\{W_e : e \in E\}$. Moreover, for the last equation we note that for two neighbouring nodes $(s,t) \in E^m$ in the directed tree T^m, by applying the same argument as above, we have $\Gamma_{st}^{(s)} = \mathrm{Var}\left\{\log W_t^s\right\} = \Gamma_{st}^{(m)}$. □

A.3 Proof of Corollary 1

We have to show that for any tree $T' = (V, E')$ that differs from T in at least one edge, it holds

$$\sum_{(i,j)\in E'} \Gamma_{ij}^{(m)} - \sum_{(i,j)\in E} \Gamma_{ij}^{(m)} > 0. \tag{A1}$$

The terms for $(i,j) \in E \cap E'$ cancel directly between the two sums. For $(i,j) \in E \setminus E'$, the graph $(V, E \setminus \{(i,j)\})$ is disconnected with connected components, say, $V_1, V_2 \subset V$. Since T' is connected, there must be a $h \in V_1$ and $l \in V_2$ such that $(h,l) \in E'$. Since the path $\mathrm{ph}(hl;T)$ must contain the edge (i,j) and

$$\Gamma_{hl}^{(m)} = \sum_{e\in \mathrm{ph}(hl;T)} \Gamma_e^{(m)}, \tag{A2}$$

this means that the first sum in (A1) contains $\Gamma_{ij}^{(m)}$ as part of $\Gamma_{hl}^{(m)}$, which cancels the corresponding term in the second sum.

There are the same number of edges in $E \setminus E'$ as in $E' \setminus E$ and every $\Gamma_{hl}^{(m)}$ for $(h,l) \in E \setminus E'$ is the sum of several terms in the decomposition A2. Therefore, the difference on the left-hand side of (A1) is indeed strictly positive as long as none of the distances vanishes. □

A.4 Proof of Corollary 2

For the true edge set E we observe

$$\begin{aligned}\sum_{(i,j)\in E} \rho_{ij} &= \sum_{(i,j)\in E} \sum_{m=1}^{d} w_m \Gamma_{ij}^{(m)} < \sum_{m=1}^{d} w_m \min_{T\neq T'=(V,E')} \sum_{(i,j)\in E'} \Gamma_{ij}^{(m)} \\ &\leq \min_{T\neq T'=(V,E')} \sum_{(i,j)\in E'} \sum_{m=1}^{d} w_m \Gamma_{ij}^{(m)} = \min_{T\neq T'=(V,E')} \sum_{(i,j)\in E'} \rho_{ij},\end{aligned}$$

where the first inequality follows from the uniqueness of the minimum spanning tree with weights $\Gamma_{ij}^{(m)}$, $m \in V$. It follows that $T = (V, E)$ must be the minimum spanning tree corresponding to the weights $\rho_{ij} = \sum_{m=1}^{d} w_m \Gamma_{ij}^{(m)}$. □

A.5 Proof of Proposition 5

We begin by proving (15). To this end, note that we can write the extremal correlation χ_{hl} in the extremal tree model $\mathbf{Y}$ as

$$\chi_{hl} = \mathbb{P}(Y_l > 1 | Y_h > 1) = \mathbb{P}\big(Y_l^h > 1\big).$$

From (8) we have that

$$Y_l^h = P \prod_{e \in \mathrm{ph}(hl; T^m)} W_e,$$

and therefore, by independence between P and $\{W_e : e \in \mathrm{ph}(hl; T^m)\}$ and since P follows a standard Pareto distribution,

$$\chi_{hl} = \int_1^\infty u^{-2} \mathbb{P}\left(u > \prod_{e \in \mathrm{ph}(hl; T^m)} 1/W_e\right) \mathrm{d}u = \mathbb{E}\left[\min\left(\prod_{e \in \mathrm{ph}(hl; T^m)} W_e, 1\right)\right],$$

by changing the order of integration. Observe that for any two positive, independent random variables A and B with $\mathbb{E}A, \mathbb{E}B \leq 1$, we have from Jensen's inequality by concavity of $x \mapsto \min(x, 1)$

$$\mathbb{E}\left[\min\left(AB, 1\right)\right] = \mathbb{E}\left\{\mathbb{E}\left[\min\left(AB, 1\right) | A\right]\right\} \leq \mathbb{E}\left\{\min\left[A\mathbb{E}(B|A), 1\right]\right\} = \mathbb{E}\left[\min\left(A, 1\right)\right]. \tag{A3}$$

Recall that we have $\mathbb{E}W_e \leq 1$ for all $e \in E$. Since $(i, j) \in \mathrm{ph}(hl; T^m)$ we can apply the above successively to obtain

$$\chi_{hl} = \mathbb{E}\left[\min\left(\prod_{e \in \mathrm{ph}(hl; T^m)} W_e, 1\right)\right] \leq \mathbb{E}\left[\min\left(W_{(i,j)}, 1\right)\right] = \chi_{ij}.$$

Thus (15) follows.

We show that the minimal spanning tree is unique provided that the inequality in (15) is strict. We have to show that for any tree $T' = (V, E')$ that differs from T in at least one edge, it holds

$$\sum_{(i,j) \in E'} \rho_{ij} - \sum_{(i,j) \in E} \rho_{ij} > 0, \tag{A4}$$

where we let $\rho_{ij} = -\log(\chi_{ij}) > 0$.

We will now compare the summands in the two sums in (A4) in a pairwise fashion. To this end, we will construct a bijective mapping $\tau : E \to E'$ such that for any $(i, j) \in E$, the corresponding edge $(h, l) = \tau\{(i, j)\} \in E'$ satisfies $(i, j) \in \mathrm{ph}(hl; T)$.

Consider the undirected graph $G = (E + E', \mathcal{E})$ where $(i, j) \in E$ is connected to $(h, l) \in E'$ if and only if $(i, j) \in \mathrm{ph}(hl; T)$. In this formulation, our goal is to find an E-saturating matching, that is, a matching such that every element of E is assigned one element in E'. A graphical illustration of this idea is provided in Figure A1.

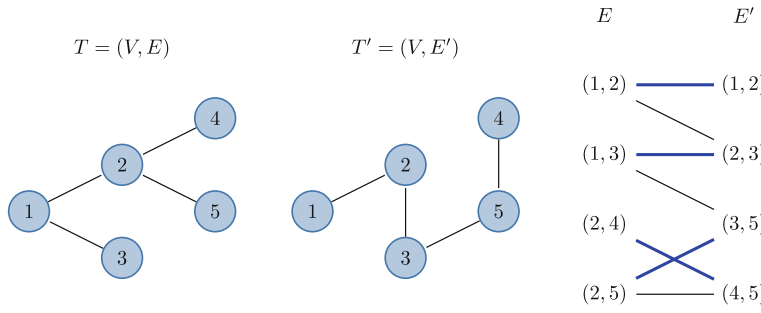

FIGURE A1 Left and center: two trees T and T'. Right: bipartite graph between elements in E and E'. A link from $(i,j) \in E$ to $(h,l) \in E'$ means that $(i,j) \in \mathrm{ph}(hl; T)$. The blue links indicate one possible matching $\tau : E \to E'$ in this case. [Colour figure can be viewed at wileyonlinelibrary.com]

By Hall's marriage theorem (Hall, 1935), such a matching exists provided that for any subset $C \subset E$, the corresponding neighbourhood $n(C) \subset E'$ of elements in E' that are connected to at least one of the elements in C satisfies

$$|C| \leq |n(C)|. \tag{A5}$$

Let $e_1, \ldots, e_p$ be the edges in C, where $p = |C|$. Removing these edges from the tree $T = (V, E)$ results in a graph $(V, E \setminus C)$ with $p + 1$ connected components, which we denote by $V_1, \ldots, V_{p+1}$.

Starting with component V_1, we know from the connectedness of the tree T' that there must be an edge in E' between at least one of the elements of V_1, say h_1, to $l_1 \in V_{k_1}$ for some $k_1 \neq 1$. Since h_1 and l_1 are in different connected components in $(V, E \setminus C)$, the path $\mathrm{ph}(h_1 l_1; T)$ must contain one of the edges in C, and therefore $e'_1 = (h_1, l_1) \in n(C)$.

Similarly, there must exist an edge $e'_2 = (h_2, l_2)$ between an element $h_2 \in V_1 \cup V_{k_1}$ and some $l_2 \in V_{k_2}$, $k_2 \notin \{1, k_1\}$. This edge is necessarily different from e'_1 as it has a node in V_{k_2}, and the path $\mathrm{ph}(h_2 l_2; T)$ must contain one of the edges in C because h_2, l_2 are in different connected components of $(V, E \setminus C)$. Thus $e'_2 \in n(C)$.

Continuing this argument inductively we obtain p different edges in $n(C)$ and therefore the condition (A5) holds.

In order to show inequality (A4) we rewrite the left-hand side as

$$\sum_{(i,j) \in E} \left(\rho_{\tau\{(i,j)\}} - \rho_{ij} \right). \tag{A6}$$

By construction of τ, for $(h,l) = \tau\{(i,j)\}$, the path $\mathrm{ph}(hl; T)$ must contain the edge (i,j) and thus by (15)

$$\rho_{hl} \geq \max_{e \in \mathrm{ph}(hl;T)} \rho_e \geq \rho_{ij}. \tag{A7}$$

This means that all summands in (A6) are non-negative. Recall that we assume in the second part of Proposition 5 that the inequalities (15) are strict for $(i,j) \neq (h,l)$. Since there is at least one $(h,l) \in E' \setminus E$, the first inequality in (A7) is strict for this edge and therefore, the difference on the left-hand side of (A4) is indeed strictly positive. Thus the proof is complete. □