Causal discovery in heavy-tailed models

Gnecco, Nicola; Meinshausen, Nicolai; Peters, Jonas; Engelke, Sebastian

Published in:
Annals of Statistics

DOI:
10.1214/20-AOS2021

Publication date:
2021

Document version
Publisher’s PDF, also known as Version of record

Document license:
Other

Citation for published version (APA):
CAUSAL DISCOVERY IN HEAVY-TAILED MODELS

BY NICOLA GNECCO¹,* , NICOLAI MEINSHAUSEN², JONAS PETERS³ AND SEBASTIAN ENGELKE¹,†

¹Research Center for Statistics, University of Geneva, *nicola.gnecco@unige.ch; †sebastian.engelke@unige.ch
²Department of Mathematics, ETH Zurich, meinshausen@stat.math.ethz.ch
³Department of Mathematical Sciences, University of Copenhagen, jonas.peters@math.ku.dk

Causal questions are omnipresent in many scientific problems. While much progress has been made in the analysis of causal relationships between random variables, these methods are not well suited if the causal mechanisms only manifest themselves in extremes. This work aims to connect the two fields of causal inference and extreme value theory. We define the causal tail coefficient that captures asymmetries in the extremal dependence of two random variables. In the population case, the causal tail coefficient is shown to reveal the causal structure if the distribution follows a linear structural causal model. This holds even in the presence of latent common causes that have the same tail index as the observed variables. Based on a consistent estimator of the causal tail coefficient, we propose a computationally highly efficient algorithm that estimates the causal structure. We prove that our method consistently recovers the causal order and we compare it to other well-established and nonextremal approaches in causal discovery on synthetic and real data. The code is available as an open-access R package.

1. Introduction and background. Reasoning about the causal structure underlying a data generating process is a key scientific question in many disciplines. In recent years, much progress has been made in the formalisation of causal language (Pearl (2009), Spirtes, Glymour and Scheines (2000), Imbens and Rubin (2015)). In several situations, causal relationships manifest themselves only in extreme events. As stated by Cox and Wermuth (1996), Section 8.7, large interventions (also named natural experiments) often carry information that is likely to be causal. In this light, one can view extreme observations as natural experiments that perturb a given system and facilitate causal analysis. On the one hand, existing causal methodology, focuses on moment related quantities of the distribution and is not tailored to estimate causal relationships from extreme events. On the other hand, the statistics of univariate extremes is relatively well understood (Resnick (2008)) and there is a large set of tools for the analysis of heavy-tailed distributions. This work attempts to bring the fields of causality and extremes closer.

Let us first consider a bivariate random vector \((X_1, X_2)\) and assume that we are interested in the causal relationship between the two random variables, \(X_1\) and \(X_2\). We consider a linear structural causal model (Pearl (2009), Section 1.4) over variables including \((X_1, X_2)\) (without feedback mechanisms). We can then distinguish between the six scenarios of causal configurations shown in Figure 1 that include \(X_1\), \(X_2\) and possibly a third unobserved random variable \(X_0\). This collection is complete in the following sense. Any structural causal model including \(X_1\) and \(X_2\) is interventionally equivalent (e.g., Peters, Janzing and Schölkopf (2017), Section 6.8) to one of the examples shown in Figure 1 when taking into account interventions on \(X_1\) or \(X_2\) only. The dashed edges can be interpreted as directed paths induced by a linear structural causal model (SCM); see Section 1.1 for a formal definition. Here and in

Received November 2019; revised September 2020.

MSC2020 subject classifications. Primary 62Hxx; secondary 62G08, 62G32.

Key words and phrases. Causality, extreme value theory, heavy-tailed distributions, nonparametric estimation.

1755
The six possible causal configurations between $X_1$ and $X_2$, and possibly a third unobserved variable $X_0$. The variable $X_0$ will be referred to as a hidden confounder. Formal definitions are included in Section 1.1. We will see in Section 2.2 that both configurations (a-1) and (a-2), for example, show the same tail coefficient behaviour. The enumeration letters (a)–(d) visualise the cases in Table 1.

In the sequel, we say that “$X_1$ is the cause of $X_2$” or “$X_1$ causes $X_2$” if there is a directed path from $X_1$ to $X_2$ in the SCM’s underlying directed acyclic graph (DAG). Assume that $X_1$ is the cause of $X_2$, and that both variables are heavy-tailed. Intuitively, if the causal relationship is monotonic, then an extremely large value of $X_1$ should cause an extreme value of $X_2$. The causal direction should, therefore, be strongly visible in the largest absolute values of the random variables. Here, “extreme” is to be seen in the respective scale of each variable, so it will make sense to consider the copula $\{F_1(X_1), F_2(X_2)\}$, where $F_j$ is the marginal distribution of $X_j, j = 1, 2$. To exploit this intuition, we define the causal tail coefficient between variables $X_1$ and $X_2$ as

$$\Gamma_{12} := \lim_{u \to 1} E[F_2(X_2) \mid F_1(X_1) > u] \in [0, 1],$$

if the limit exists. It reflects the causal relationship between $X_1$ and $X_2$ since, intuitively, if $X_1$ has a monotonically increasing causal influence on $X_2$, we expect $\Gamma_{12}$ to be close to one. Conversely, extremes of $X_2$ will not necessarily lead to extremes of $X_1$ and therefore, the coefficient $\Gamma_{21}$, where the roles of $X_1$ and $X_2$ in (1) are reversed, may be strictly smaller than one. This asymmetry will be made precise in Theorem 1 for linear structural causal models with heavy-tailed noise variables, and it forms the basis of our causal discovery algorithm.

A different perspective of our approach goes beyond the usual way of defining causality in the bulk of distribution. Namely, by looking at the signal in the tails, we might recover an extremal causal mechanism that is not necessarily present in the central part of the distribution. An example of this can be observed in financial markets. During calm periods, it is not clear whether any causal relationship exists among the financial variables. However, during turmoil, it is common to observe one specific stock or sector causing very negative (or positive) returns of other stocks (or sectors), displaying an extremal causal mechanism. In the finance literature, the concept of extremal causal mechanism is explained in terms of contagion, that is, the spread of shocks across different markets (see Forbes and Rigobon (2002)). For example, Rodriguez (2007) uses a copula model with time-varying parameters to explain such contagion phenomena within countries in Latin America and Asia. On the other hand, there are also applications where the causal mechanism is present in the bulk of the distribution but absent in the tails. Seneviratne et al. (2010) present this type of causal relationship between air temperature and the evapotranspiration of soil moisture. As the air temperature increases, the evapotranspiration process increases, too. This continues until the soil moisture resources are reduced to the point that a further increase in the temperature has no causal effect on the evapotranspiration.

Heavy-tailed distributions are an example of non-Gaussian models, which have received some attention in the causal literature. The LiNGAM algorithm (Shimizu et al. (2006)) exploits non-Gaussianity through independent component analysis (Comon (1994)) to estimate
an underlying causal structure. Misra and Kuruoglu (2016) consider stable noise variables in a Bayesian network and develop a structure learning algorithm based on BIC. The work of Gissibl, Klüppelberg and Lauritzen (2021) also studies causal questions related to extreme events. They consider max-linear models (Gissibl and Klüppelberg (2018)) where only the largest effect propagates to the descendants in a Bayesian network. The work by Naveau et al. (2018) falls within the domain of attribution science. Namely, by studying extreme climate events, they try to answer counterfactual questions such as “what the Earth’s climate might have been without anthropogenic interventions”. Mhalla, Chavez-Demoulin and Dupuis (2020) develop a method to estimate the causal relationships between bivariate extreme observations. It relies on the Kolmogorov complexity concept (see Kolmogorov (1968)) adapted to high conditional quantiles. Engelke and Ivanovs (2021) reviews recent work on causality and sparsity in extremes.

The rest of the paper is organised as follows. Sections 1.1 and 1.2 briefly review structural causal models and some important concepts from extreme value theory. Section 2 contains a causal model for heavy-tailed distributions with positive coefficients. We prove that, given the underlying distribution, the causal tail coefficient allows us to distinguish between the causal scenarios shown in Figure 1. In Section 3, we introduce an algorithm named extremal ancestral search (EASE) that can be applied to a matrix of causal tail coefficients and that retrieves the causal order of the true graph, in the population case. We prove that our algorithm estimates a causal order even in the case where the causal tail coefficients are estimated empirically from data, as the sample size tends to infinity. In Section 4, we first generalise the results of the previous sections to the case of a structural causal model with real-valued coefficients. To do that, we introduce a more general causal tail coefficient that is sensitive to both the upper and lower tail of the variables. Second, we discuss the robustness properties of EASE in the presence of hidden confounders. Third, we analyse the properties of the causal tail coefficient when the noise variables have different tail indices. Section 5 contains experiments on simulated data and real-world applications. All proofs of the main paper are given in the Supplementary Material (Gnecco et al. (2020)). Sections A to F are in the Supplementary Material.

1.1. Structural causal models. A linear structural causal model, or linear SCM, (Bollen (1989); Pearl (2009), Section 1.4) over variables \( X_1, \ldots, X_p \) is a collection of \( p \) assignments

\[
X_j := \sum_{k \in \text{pa}(j)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V,
\]

where \( \text{pa}(j) \subseteq V = \{1, \ldots, p\} \) and \( \beta_{jk} \in \mathbb{R} \setminus \{0\} \), together with a joint distribution over the noise variables \( \varepsilon_1, \ldots, \varepsilon_p \). Here, we assume that the noise variables are jointly independent and that the induced graph \( G = (V, E) \), obtained by adding directed edges from the parents \( \text{pa}(j) \) to \( j \), is a directed acyclic graph (DAG) with nodes \( V \) and (directed) edges \( E \subset V \times V \). To ease notation, we adopt the convention to sometimes identify a node with its corresponding random variable. To highlight the fact that \( \text{pa}(j) \) depends on a specific DAG \( G \), we write \( \text{pa}(j, G), j \in V \).

Structural causal models describe not only observational distributions but also interventional distributions. An intervention on \( X_j \), for example, is defined as replacing the corresponding assignment (2) while leaving the other equations as they were. In practice, a causal model can be falsified via randomised experiments (see, e.g., Peters, Janzing and Schölkopf (2017), Section 6.8).

We define a directed path between node \( j \) and \( k \) as a sequence of distinct vertices such that successive pairs of vertices belong to the edge set \( E \) of \( G \). If there is a directed path from \( j \) to \( k \), we say that \( j \) is an ancestor of \( k \) in \( G \). The set of ancestors of \( j \) is denoted by \( \text{An}(j, G) \),
and we write \( \text{an}(j, G) = \text{an}(j, G) \setminus \{j\} \) when we consider the ancestors of \( j \) except itself. A node \( j \) that has no ancestors, that is, \( \text{an}(j, G) = \emptyset \), is called a source node (or root node). Given two nodes \( j, k \in V \), we say that \( X_j \) \textit{causes} \( X_k \) if there is a directed path from \( j \) to \( k \) in \( G \). Furthermore, given nodes \( i, j, k \in V \), we say that \( X_i \) is a \textit{confounder} (or common cause) of \( X_j \) and \( X_k \) if there is a directed path from node \( i \) to node \( j \) and \( k \) that does not include \( k \) and \( j \), respectively. Whenever a confounder is unobserved, we say it is a \textit{hidden} confounder or hidden common cause. Finally, if \( \text{An}(j, G) \cap \text{An}(k, G) = \emptyset \), then we say that there is no \textit{causal link} between \( X_j \) and \( X_k \).

A graph \( G_1 = (V_1, E_1) \) is called a subgraph of \( G \) if \( V_1 \subseteq V \) and \( E_1 \subseteq (V_1 \times V_1) \cap E \). Recall that any subgraph of a DAG \( G \) is also a DAG. For details on graphical models, we refer to Lauritzen (1996).

1.2. \textit{Regularly varying functions and random variables}. A positive, measurable function \( f \) is said to be \textit{regularly varying} with index \( \alpha \in \mathbb{R} \), \( f \in \text{RV}_\alpha \), if it is defined on some neighbourhood of infinity \( [x_0, \infty) \), \( x_0 > 0 \), and if for all \( c > 0 \), \( \lim_{x \to \infty} f(cx)/f(x) = c^\alpha \). If \( \alpha = 0 \), \( f \) is said to be \textit{slowly varying}, \( f \in \text{RV}_0 \).

A random variable \( X \) is said to be \textit{regularly varying} with index \( \alpha \) if \( P(X > x) \sim \ell(x)x^{-\alpha}, x \to \infty \), for some \( \ell \in \text{RV}_0 \), where for any function \( f \) and \( g \), we write \( f \sim g \) if \( f(x)/g(x) \to 1 \) as \( x \to \infty \). If \( X \) is regularly varying with index \( \alpha \) then \( cX \) is also regularly varying with the same index, for any \( c > 0 \). For example, random variables with a Student’s-\( t \), Pareto, Cauchy, or Fréchet distribution are regularly varying.

A characteristic property of regularly varying random variables is the \textit{max-sum-equivalence}. The idea is that large sums of independent random variables tend to be driven by only one single large value. For this reason, the tail of the distribution of the maximum is equal to the tail of the distribution of the sum. For a rigorous formulation; see Section A in the Supplementary Material. We refer to Embrechts, Klüppelberg and Mikosch (1997), Section A3, for further details on regular variation and max-sum-equivalence.

2. \textit{The causal tail coefficient}. To measure the causal effects in the extremes, we define the following parameter.

\textbf{Definition 1.} Given two random variables \( X_1 \) and \( X_2 \), we define the \textit{causal tail coefficient}

\begin{equation}
\Gamma_{jk} = \lim_{u \to 1^-} E\left[ F_k(X_k) \mid F_j(X_j) > u \right],
\end{equation}

if the limit exists, for \( j, k = 1, 2 \) and \( j \neq k \).

The coefficient \( \Gamma_{jk} \) lies between zero and one and is invariant under any marginal strictly increasing transformation since it depends on the rescaled margins \( F_j(X_j) \), for \( j = 1, 2 \). Below, we lay down the setup.

2.1. \textit{Setup}. Consider a linear structural causal model (SCM) with an induced directed acyclic graph (DAG) \( G \),

\[ X_j := \sum_{k \in \text{pa}(j, G)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V, \]

where we assume that the coefficients \( \beta_{jk} \) are strictly positive, \( j, k \in V \). Let the independent noise variables \( \varepsilon_1, \ldots, \varepsilon_p \) be real-valued and regularly varying with comparable tails, that is,
there exists a tail-index \( \alpha > 0 \) and \( \ell \in \text{RV}_0 \) such that for all \( j \in V \), there exists \( c_j > 0 \) that satisfies
\[
P(\varepsilon_j > x) \sim c_j \ell(x)x^{-\alpha}, \quad x \to \infty.
\]
(4)

To simplify the notation, we rescale the variables \( X_j \) such that \( c_j = 1 \), \( j \in V \). Furthermore, denote by \( \beta_{j \rightarrow k} \) the sum of distinct weighted directed paths from node \( k \) to node \( j \), with \( \beta_{j \rightarrow j} := 1 \). Since \( G \) is acyclic, we can express recursively each variable \( X_j, j \in V \), as a weighted sum of the noise terms \( \varepsilon_1, \ldots, \varepsilon_k \) that belong to the ancestors of \( X_j \), that is,
\[
X_j = \sum_{h \in \text{An}(j, G)} \beta_{h \rightarrow j} \varepsilon_h.
\]
(5)

The noise terms in (5) are independent and regularly varying with comparable tails as in (4). Therefore, by using Lemma S1 of the Supplementary Material, we can write
\[
P(X_j > x) \sim \sum_{h \in \text{An}(j, G)} \beta_{h \rightarrow j}^\alpha \ell(x)x^{-\alpha}, \quad x \to \infty.
\]

Any probability distribution induced by an SCM is Markov with respect to the induced DAG \( G \), and thus we can read off statistical independencies from it by \( d \)-separation (Lauritzen et al. (1990); Pearl (2009), Section 1.2.3). Conversely, to infer dependencies directly from the graph, one needs to assume that the distribution is faithful to the DAG \( G \) (see Spirtes, Glymour and Scheines (2000), Section 2.3.3). Most causal methods based on restricted SCMs, such as LiNGAM (see Shimizu et al. (2006)), RESIT (Peters et al. (2014)), and Peters and Bühlmann (2014), do not assume faithfulness. Similarly, in this work we require the milder assumption that \( \beta_{j \rightarrow k} \) is nonzero if \( j \) is an ancestor of \( k \), that is, \( X_j \) causes \( X_k \). This is automatically satisfied if the SCM has positive coefficients. In the sequel, we refer to this model as a heavy-tailed linear SCM.

2.2. Causal structure and the causal tail coefficient. In the setting of Section 2.1, the causal tail coefficient always exists and carries information about the underlying causal structure. In particular, it can be expressed in closed form.

**Lemma 1.** Consider a heavy-tailed linear SCM over \( p \) variables. Then, for \( j, k \in V \) and \( j \neq k \),
\[
\Gamma_{jk} = \frac{1}{2} + \frac{1}{2} \sum_{h \in A_{jk}} \beta_{h \rightarrow j}^\alpha \sum_{h \in \text{An}(j, G)} \beta_{h \rightarrow j}^\alpha ,
\]
where \( A_{jk} = \text{An}(j, G) \cap \text{An}(k, G) \), and the sum over an empty index set equals zero.

For a proof, see the Supplementary Material B.1. Lemma 1 provides a closed form expression for the causal tail coefficient \( \Gamma_{jk} \), which can be written as a sum of two terms. The first term corresponds to the case when \( X_j \) and \( X_k \) are independent. The second term is nonnegative and depends on the coefficients of the SCM and the tail index \( \alpha > 0 \). By using matrix notation, it is possible to express the coefficient \( \Gamma_{jk} \) more compactly. Consider the matrix of coefficients \( B \) of the DAG \( G \), where \( B_{jk} = \beta_{jk}, j, k \in V \), and let \( I \) be the identity matrix. Furthermore, for any \( M \in \mathbb{R}^{p \times p} \), denote by \( M_\alpha \) the matrix where each entry of \( M \) is raised to the power \( \alpha \). By applying the Neumann series, we obtain \( H = (I - B)^{-1} \) where \( H_{jk} = \beta_{k \rightarrow j} \) for \( j, k \in V \). Therefore, for \( j, k \in V \) and \( j \neq k \), we can write the causal tail coefficient as
\[
\Gamma_{jk} = \frac{1}{2} + \frac{1}{2} \sum_{j \in C} e_j^T H_\alpha e_{A_{jk}},
\]
where \( e_j \in \mathbb{R}^p \) is the \( j \)th standard basis vector, and \( e_C = \sum_{j \in C} e_j \in \mathbb{R}^p \) for any set \( C \subseteq \{1, \ldots, p\} \).
**Example 1.** Consider the “diamond” graph $G = (V, E)$ in Figure 2, with $V = \{1, \ldots, 4\}$. In this graph, for instance, it is easy to see that $\Gamma_{14} = 1$. To compute $\Gamma_{41}$, we list the weighted directed paths from the ancestors of node 4 to node 4 itself, that is,

$$
\beta_{1\rightarrow 4} = \beta_{42}\beta_{21} + \beta_{43}\beta_{31}, \quad \beta_{2\rightarrow 4} = \beta_{42}, \quad \beta_{3\rightarrow 4} = \beta_{43}, \quad \beta_{4\rightarrow 4} = 1.
$$

Additionally, the set of common ancestors of node 1 and 4 is $A_{14} = \{1\}$. Putting everything together, by using Lemma 1, or formula (6), we obtain

$$
\Gamma_{41} = \frac{1}{2} + \frac{\beta_{1\rightarrow 4}^\alpha}{2\beta_{1\rightarrow 4}^\alpha + \beta_{2\rightarrow 4}^\alpha + \beta_{3\rightarrow 4}^\alpha + \beta_{4\rightarrow 4}^\alpha} < 1.
$$

In this example, we see that it is possible to infer the causal relationship between $X_1$ and $X_4$ because $\Gamma_{14} > \Gamma_{41}$.

Consider now a general, heavy-tailed linear SCM over $p$ variables including $X_1$ and $X_2$, and inducing graph $G$. The following theorem shows that the causal tail coefficient, which is computable from the bivariate distribution of $X_1$ and $X_2$, see equation (3), encodes the causal relationship between the two variables.

**Theorem 1.** Consider a heavy-tailed linear SCM over $p$ variables including $X_1$ and $X_2$, as defined in Section 2.1. Then, knowledge of $\Gamma_{12}$ and $\Gamma_{21}$ allows us to distinguish the following cases: (a) $X_1$ causes $X_2$, (b) $X_2$ causes $X_1$, (c) there is no causal link between $X_1$ and $X_2$ (i.e., $A_1(1, G) \cap A_2(2, G) = \emptyset$), (d) there is a node $j \notin \{1, 2\}$, such that $X_j$ is a common cause of $X_1$ and $X_2$ and neither $X_1$ causes $X_2$ nor $X_2$ causes $X_1$. The corresponding values for $\Gamma_{12}$ and $\Gamma_{21}$ are depicted in Table 1.

For a proof, see the Supplementary Material B.2. This result will also play a key role when estimating causal relationships from finitely many data. As a first remark, condition (a) and (b) might also include the presence of a common cause $X_j$. As a second remark, the empty entries in Table 1 cannot occur under the assumptions made in Section 2.1. For example, $\Gamma_{12} = \Gamma_{21} = 1$ can only happen if some variables have different tail indices. One possibility is when the cause has a heavier tail than the effect. Another scenario is when a common cause $X_j$, for some $j \neq 1, 2$, has heavier tails than the confounded variables $X_1$ and $X_2$. For further discussion on different tail indices, see Section 4.3.

<table>
<thead>
<tr>
<th>$\Gamma_{12}$</th>
<th>$\Gamma_{21}$</th>
<th>Implications for causality</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma_{12} = 1$</td>
<td>$\Gamma_{21} = 1$</td>
<td>$X_1$ causes $X_2$</td>
</tr>
<tr>
<td>$\Gamma_{12} \in (1/2, 1)$</td>
<td>$\Gamma_{21} \in (1/2, 1)$</td>
<td>(a) $X_1$ causes $X_2$</td>
</tr>
<tr>
<td>$\Gamma_{12} = 1/2$</td>
<td>$\Gamma_{21} = 1/2$</td>
<td>(b) $X_2$ causes $X_1$</td>
</tr>
<tr>
<td>$\Gamma_{12} = 1/2$</td>
<td>$\Gamma_{21} = 1/2$</td>
<td>(d) common cause</td>
</tr>
<tr>
<td>$\Gamma_{12} = 1/2$</td>
<td>$\Gamma_{21} = 1/2$</td>
<td>(c) no causal link</td>
</tr>
</tbody>
</table>
2.3. A nonparametric estimator. Consider a heavy-tailed linear SCM over \( p \) variables including \( X_1 \) and \( X_2 \), with distributions \( F_1 \) and \( F_2 \), as described in Section 2.1. In order to construct a nonparametric estimator of \( \Gamma_{12} \) and \( \Gamma_{21} \) based on independent observations \((X_{i1}, X_{i2}), i = 1, \ldots, n\), of \((X_1, X_2)\), we define the empirical distribution function of \( X_j \) as

\[
\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{ij} \leq x\}, \quad x \in \mathbb{R},
\]

for \( j = 1, 2 \). Denote by \( g^\leftarrow \) the left continuous generalised inverse

\[
g^\leftarrow(y) = \inf\{x \in \mathbb{R} : g(x) \geq y\}, \quad y \in \mathbb{R}.
\]

In addition, let the \((n - k)\)th order statistics be denoted by \( X_{(n-k),j} = \hat{F}_j^\leftarrow(1 - k/n) \), for all \( k = 0, \ldots, n - 1 \) and \( j = 1, 2 \), such that \( X_{(1),j} \leq \cdots \leq X_{(n),j} \). Replacing \( F_1 \) and \( F_2 \) in the definition of \( \Gamma_{12} \) in (3) by the empirical counterparts, and the threshold \( u \) by \( u_n = 1 - k/n \), for some integer \( 0 < k \leq n - 1 \), we define the estimator

\[
\hat{\Gamma}_{12} = \hat{\Gamma}_{12}^{(n)} = \frac{1}{k} \sum_{i=1}^{n} \hat{F}_2(X_{i2}) 1\{X_{i1} > X_{(n-k),1}\}.
\]

For this estimator to be consistent, a classical assumption in extreme value theory is that the number of upper order statistics \( k = k_n \) depends on the sample size \( n \) such that \( k_n \to \infty \) and \( k_n/n \to 0 \) as \( n \to \infty \). The first condition is needed to increase the effective sample size, whereas the second condition eliminates the approximation bias. The estimator \( \hat{\Gamma}_{21} = \hat{\Gamma}_{21}^{(n)} \) is defined in an analogous way as (8).

**Theorem 2.** Let \( X_{i1} \) and \( X_{i2}, i = 1, \ldots, n \), be independent copies of \( X_1 \) and \( X_2 \), respectively, where \( X_1 \) and \( X_2 \) are two of the \( p \) variables of a heavy-tailed linear SCM.

\( \text{(A1)} \) Assume that the density functions \( f_j = F_j' \), \( j = 1, 2 \), exist and satisfy the von Mises’ condition

\[
\lim_{x \to \infty} \frac{xf_j(x)}{1 - F_j(x)} = \frac{1}{\gamma} \quad \text{for some } \gamma > 0.
\]

\( \text{(A2)} \) Let \( k_n \in \mathbb{N} \) be an intermediate sequence with

\[ k_n \to \infty \quad \text{and} \quad k_n/n \to 0, \quad n \to \infty. \]

Then the estimators \( \hat{\Gamma}_{12} \) and \( \hat{\Gamma}_{21} \) are consistent, as \( n \to \infty \), that is,

\[
\hat{\Gamma}_{12} \overset{P}{\to} \Gamma_{12} \quad \text{and} \quad \hat{\Gamma}_{21} \overset{P}{\to} \Gamma_{21}.
\]

**Remark 1.** The von Mises’ condition in (A1) is a very mild assumption that is satisfied by most univariate regularly varying distributions of interest. In our case \( \gamma = 1/\alpha \), where \( \alpha \) is the common tail index of the noise variables.

For a proof of Theorem 2, see the Supplementary Material B.3. It uses several results from tail empirical process theory (e.g., de Haan and Ferreira (2006), Section 2.2). The main challenge comes from the fact that the variables \( X_1 \) and \( X_2 \) are tail dependent, and that the use of the empirical distribution function \( \hat{F}_2 \) in (8) introduces dependence between the terms corresponding to different observations \( i = 1, \ldots, n \). A related problem is studied in Cai et al. (2015), where they derive asymptotic properties of the empirical estimator of the expected shortfall when another dependent variable is extreme. However, in contrast to Cai et al. (2015), in the proof of Theorem 2, we work with a more explicit model and we consider the variables scaled to uniform margins, that is, \( \hat{F}_j(X_j) \) instead of \( X_j \), \( j = 1, 2 \).
3. Causal discovery using extremes. We would like to recover the causal information from a dataset of \( p \) variables under the model specification of Section 2.1. We develop an algorithm named extremal ancestral search (EASE) based on the causal tail coefficient defined in (3). We show that EASE can recover the causal order of the underlying graph in the population case (Section 3.1), and that it is consistent (Section 3.2).

3.1. Learning the causal order. Our goal is to recover the causal order of a heavy-tailed linear SCM over \( p \) variables (as defined in Section 2.1) by observing \( n \) i.i.d. copies of the random vector \( X \in \mathbb{R}^p \). Given a DAG \( G = (V, E) \), a permutation \( \pi : \{1, \ldots, p\} \to \{1, \ldots, p\} \) is said to be a causal order (or topological order) of \( G \) if \( \pi(i) < \pi(j) \) for all \( i \) and \( j \) such that \( i \in \text{an}(j, G) \). We denote by \( \Pi_G \) the set of all causal orders of \( G \). For a permutation \( \pi \) we sometimes use the notation \( \pi = (\pi(1), \ldots, \pi(p)) \).

A given causal order \( \pi \) does not specify a unique DAG. As an example, the causal order \( \pi = (1, 2) \) comprises two DAGs: one where there is a directed edge between node 1 and node 2, and one where the two nodes are unconnected. On the other hand, there can be several causal orders for a given DAG. For example, a fully-disconnected DAG satisfies any causal order. However, even if the causal order does not identify a unique DAG, it still conveys important information. In particular, each causal order defines a class of DAGs that agree with respect to the nonancestral relations. Therefore, once a causal order is available, one can estimate the complete DAG by using regularised regression methods. This idea has been exploited, for example, by Shimizu et al. (2011) and Bühlmann, Peters and Ernest (2014). In addition, Bühlmann, Peters and Ernest (2014) and Peters and Bühlmann (2015) argue that knowledge of a causal order is useful per se. In fact, given a correct causal order, one can construct a fully-connected DAG that describes interventional distribution across the variables.

For any heavy-tailed linear SCM and induced DAG \( G = (V, E) \), we define the matrix \( \Gamma \in \mathbb{R}^{p \times p} \) with entries \( \Gamma_{ij} \), the causal tail coefficients between all pairs of variables \( X_i \) and \( X_j \), \( i, j \in V \); see Definition 1. Theorem 1 tells us how the entries of \( \Gamma \) encode the causal relationships between the random variables of the SCM. To recover the causal order of the DAG \( G \), we propose the algorithm named extremal ancestral search (EASE).

**EASE algorithm**

**INPUT:** A matrix \( \Gamma \in \mathbb{R}^{p \times p} \) of causal tail coefficients related to a DAG \( G = (V, E) \) with \( V = \{1, \ldots, p\} \).

**RETURNS:** Permutation of the nodes \( \pi : V \to \{1, \ldots, p\} \).

0. Set \( V_1 = V \).
1. For \( s \in \{1, \ldots, p\} \)
   (a) Let \( M^{(s)}_i = \max_{j \in V_s \setminus \{i\}} \Gamma_{ji} \), for all \( i \in V_s \).
   (b) Let \( i_s \in \arg \min_{i \in V_s} M^{(s)}_i \).
   (c) Set \( \pi(i_s) = s \).
   (d) Set \( V_{s+1} = V_s \setminus \{i_s\} \).
2. RETURN the permutation \( \pi \).

**COMPLEXITY:** \( O(p^2) \).

The above is a greedy algorithm that identifies root nodes of the current subgraph at each step. In the first step, the algorithm finds a root node \( i_1 \in V \) as the one that minimises the score \( M^{(1)}_i = \max_{j \neq i} \Gamma_{ji} \), \( i \in V \). In fact, by Theorem 1, \( M^{(1)}_i < 1 \) if and only if \( i \) is a source node. Once the first node is selected, the algorithm searches for a second root node in the subgraph where \( i_1 \) is removed. The procedure continues until all nodes have been selected. In
Section C of the Supplementary Material, one example illustrates how EASE finds a causal order for a given DAG.

The next result states that, in the population case, the EASE algorithm yields a correct causal order of the underlying DAG.

**Proposition 1.** Consider a heavy-tailed linear SCM over \( p \) variables, as defined in Section 2.1, and let \( G = (V, E) \) be the induced DAG. If the input \( \Gamma \) is the matrix of causal tail coefficients associated with the SCM, then EASE returns a permutation \( \pi \) that is a causal order of \( G \).

For a proof, see Supplementary Material B.4.

**3.2. Sample properties for the EASE algorithm.** For finite samples, the EASE algorithm will take an estimate of the causal coefficient matrix \( \hat{\Gamma} \) as input. Based on the empirical nonparametric estimator \( \hat{\Gamma} \) and its asymptotic properties, we assess the performance of the algorithm. Let \( \hat{\Gamma} \in \mathbb{R}^{p \times p} \) denote the matrix where each entry \( \hat{\Gamma}_{ij} \) is defined as in (8) in Section 2.3, for \( i, j \in V \). We say that a procedure makes a mistake when it returns a permutation \( \pi \notin \Pi_G \). We derive an upper bound for the probability that EASE makes a mistake when the matrix \( \Gamma \) is estimated by \( \hat{\Gamma} \).

**Proposition 2.** Consider a heavy-tailed linear SCM over \( p \) variables \( X = (X_1, \ldots, X_p) \), as defined in Section 2.1, with induced DAG \( G \). Let \( \hat{\Gamma} \) be the estimated causal coefficient matrix related to \( G \). Let \( \hat{\pi} \) denote the permutation returned by EASE based on \( \hat{\Gamma} \). Then,

\[
P(\hat{\pi} \notin \Pi_G) \leq p^2 \max_{i, j \in V : i \neq j} P(|\hat{\Gamma}_{ij} - \Gamma_{ij}| > \frac{1 - \eta}{2},
\]

where \( \eta = \max_{u \notin \text{An}(v, G)} \Gamma_{uv} < 1 \).

For a proof, see Supplementary Material B.5. The bound for the probability of making a mistake in the estimated causal order is expressed in terms of the distance between the true \( \Gamma_{ij} \) and the estimated \( \hat{\Gamma}_{ij} \). This bound in combination with the consistency result of Theorem 2 yields the consistency of the EASE algorithm in the sample case.

**Corollary 1.** Let \( \hat{\pi} \) be the permutation computed by EASE under the assumptions of Proposition 2. Let \( k_n \in \mathbb{N} \) be an intermediate sequence with

\[
k_n \to \infty \quad \text{and} \quad k_n/n \to 0, \quad n \to \infty.
\]

If the von Mises’ condition (9) holds, then the EASE algorithm is consistent, that is,

\[
P(\hat{\pi} \notin \Pi_G) \to 0 \quad \text{as} \quad n \to \infty.
\]

The result above is for fixed dimension \( p \). To prove consistency in a regime where \( p \) scales with the sample size \( n \) we would need to establish concentration inequalities for \( \hat{\Gamma} \) or asymptotic normality in Theorem 2. Both would require stronger assumptions on the tails of the noise variables and a second-order analysis in line with the proof of Theorem 2.

**3.3. Computational complexity.** The EASE algorithm is based on pairwise quantities and is therefore computationally efficient. To estimate the matrix \( \hat{\Gamma} \) of causal tail coefficients, which is the input for EASE, first, we need to rank the \( n \) observations for each of the variables, with a computational complexity of \( O(pn \log n) \). Then we compute the coefficients \( \hat{\Gamma}_{ij} \) for each pair \( i, j \in V \), with a computational complexity of \( O(k_n p^2) \). The computational complexity of EASE grows with the square of the number of variables, that is, \( O(p^2) \). The overall computational complexity of estimating the matrix \( \hat{\Gamma} \) and running the EASE algorithm is therefore \( O(\max(pn \log n, k_n p^2)) \).
4. Extensions.

4.1. Real-valued coefficients. Until now, we have worked with a heavy-tailed linear SCM with positive coefficients (see Section 2.1 for a detailed explanation of the model). In the current section, we relax this assumption and let the coefficients of the SCM be real-valued, that is, \( \beta_{jk} \in \mathbb{R}, \ j, k \in V \). Additionally, we assume that \( \beta_{j \rightarrow k} \) is nonzero, if \( j \) is an ancestor of \( k \). Given that the coefficients are real-valued, we want to consider both the upper and the lower tails of the variables. We assume that the noise variables \( \epsilon_1, \ldots, \epsilon_p \), of the SCM have comparable upper and lower tails, respectively.

The result stated in Lemma 2 allows us to extend Theorem 1 to the more general setting where the heavy-tailed SCM has coefficients \( \beta_{jk} \in \mathbb{R}, \ j, k \in V \).

**THEOREM 3.** Consider a heavy-tailed linear SCM over \( p \) variables including \( X_1 \) and \( X_2 \), and assume that \( \beta_{jk} \in \mathbb{R}, \ j, k \in V \). In addition, assume that \( \beta_{j \rightarrow k} \neq 0 \) if \( j \) is an ancestor of \( k \). Then, knowledge of \( \Psi_{12} \) and \( \Psi_{21} \) allows us to distinguish the following cases: (a) \( X_1 \) causes \( X_2 \), (b) \( X_2 \) causes \( X_1 \), (c) there is no causal link between \( X_1 \) and \( X_2 \), (d) there is a node \( j \not\in \{1,2\} \), such that \( X_j \) is a common cause of \( X_1 \) and \( X_2 \) and neither \( X_1 \) causes \( X_2 \) nor \( X_2 \) causes \( X_1 \). The corresponding values for \( \Psi_{12} \) and \( \Psi_{21} \) are shown in Table 2.
**TABLE 2**

Summary of the possible values of $\Psi_{12}$ and $\Psi_{21}$ and the implications for causality

<table>
<thead>
<tr>
<th>$\Psi_{12}$</th>
<th>$\Psi_{21}$</th>
<th>Implications for causality</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Psi_{12} = 1$</td>
<td>$\Psi_{21} = 1$</td>
<td>(a) $X_1$ causes $X_2$</td>
</tr>
<tr>
<td>$\Psi_{12} \in (1/2, 1)$</td>
<td>$\Psi_{21} \in (1/2, 1)$</td>
<td>(b) $X_2$ causes $X_1$</td>
</tr>
<tr>
<td>$\Psi_{12} = 1/2$</td>
<td>(c) no causal link</td>
<td></td>
</tr>
</tbody>
</table>

The proof is identical to the proof of Theorem 1, replacing $\Gamma_{ij}$ with $\Psi_{ij}$ and by referring to Lemma 2 instead of Lemma 1. Moreover, as in Theorem 1, condition (a) and (b) can also include the presence of a common cause $X_j$. Theorem 3 implies that if we run the EASE algorithm based on the matrix $\Psi \in \mathbb{R}^{p \times p}$, containing the pairwise $\Psi_{ij}$, $i, j \in V$, then we retrieve a causal order of the underlying DAG. This is the analogue to Proposition 1 for heavy-tailed linear SCM with real-valued coefficients.

We define an empirical estimator $\hat{\Psi}_{ij}$ of $\Psi_{ij}$ in a similar fashion as the estimator $\hat{\Gamma}_{ij}$ in (8). The proof of Lemma 2 shows that the coefficient $\Psi_{ij}$ can be decomposed in the same way as $\Gamma_{ij}$ in the proof of Lemma 1. Therefore, following the lines of the proof of Theorem 2 with some minor modifications, we obtain the consistency $\hat{\Psi}_{ij} \xrightarrow{P} \Psi_{ij}$ as $n \to \infty$ for any intermediate sequence $k_n \to \infty$ and $k_n/n \to 0$, and under the assumption of the von Mises’ condition for both the upper and the lower tail of $X_j$, $j \in V$.

We can then estimate a permutation $\hat{\pi}$ by the EASE algorithm based on the matrix $\hat{\Psi} \in \mathbb{R}^{p \times p}$ that contains the estimators $\hat{\Psi}_{ij}$, $i, j \in V$, as entries. For this permutation, we obtain the same bound for the probability that EASE makes a mistake as shown in Proposition 2 by replacing $\hat{\Gamma}_{ij}$ and $\Gamma_{ij}$ by $\hat{\Psi}_{ij}$ and $\Psi_{ij}$, respectively. This together with the consistency of $\hat{\Psi}_{ij}$ yields the following result.

**COROLLARY 2.** Assume the general setup of the heavy-tailed linear SCM with real-valued coefficients of this section. Let $\hat{\pi}$ be the permutation computed by EASE based on the matrix $\hat{\Psi} \in \mathbb{R}^{p \times p}$ that contains the estimators $\hat{\Psi}_{ij}$, $i, j \in V$, as entries. For this permutation, we obtain the consistency $\hat{\Psi}_{ij} \xrightarrow{P} \Psi_{ij}$ as $n \to \infty$ for any intermediate sequence $k_n \to \infty$ and $k_n/n \to 0$, and under the assumption of the von Mises’ condition for both the upper and the lower tail of $X_j$, $j \in V$.

Then, the EASE algorithm is consistent, that is,

$$P(\hat{\pi} \not\in \Pi_G) \to 0 \quad as \ n \to \infty.$$  

**4.2. Presence of hidden confounders.** A frequent assumption in causality is that one can observe all the relevant variables. However, in many real-world situations, it is hard, if not impossible, to do so. When some of the hidden variables are confounders (i.e., common causes), the causal inference process might be compromised. Therefore, an attractive property of a causal inference algorithm involves its robustness to hidden confounders. In this section, we show that EASE is capable of dealing with hidden common causes and, under certain assumptions, it recovers the causal order of the observed graph both in the population and in the asymptotic case.

Consider a heavy-tailed linear SCM with real-valued coefficients, as defined in Section 4.1, consisting of both observed and hidden variables. This SCM induces a DAG $G = (V, E)$, with $V = V_O \cup V_H$, $V_O \cap V_H = \emptyset$, where $V_O$ ($V_H$) denotes the set of nodes corresponding to the observed (hidden) variables. Our goal is to recover a causal order for the
subset of the observed variables $X_j$, $j \in V_O$. In particular, we say that the EASE algorithm recovers a causal order $\pi$ over the observed variables if

$$\pi(i) < \pi(j) \implies j \notin \text{an}(i, G) \quad \text{for all } i, j \in V_O.$$  

(13)

In fact, the results of the previous sections hold even in the presence of hidden confounders. Regarding the population properties, Theorems 1 and 3 still apply: they state that the causal tail coefficients $\Gamma$ and $\Psi$ reflect the causal relationships between pairs of variables without taking into account other variables, for example, by conditioning. In addition, the result of Proposition 1, and the corresponding extension in Section 4.1, are also valid. The proof of Proposition 1 depends only on the assumption that the input matrix contains the pairwise causal effects between the variables. Therefore, if we use matrix $\Gamma$ (or $\Psi$) as input for the EASE algorithm, we recover a causal order $\pi$ that satisfies (13).

Regarding the asymptotic properties, $\hat{\Gamma}$ and $\hat{\Psi}$ are consistent even in the presence of hidden common causes: in the proof of Theorem 2, the other variables do not appear. In addition, we can still find an upper bound for the probability that the EASE algorithm makes a mistake. To do so, one needs to adjust the proof of Proposition 2 by replacing the full DAG $G$ with the subgraph $G_O = (V_O, E_O)$ containing only the observed variables, where $E_O = E \cap (V_O \times V_O)$. Combining the two previous arguments, it follows that Corollaries 1 and 2 hold even in the presence of hidden confounders.

The ability to deal with hidden confounders is a property that, in general, is not shared by all methods in causality. For example, the PC algorithm (Section 5.4.2 Spirtes, Glymour and Scheines (2000)) might retrieve a Markov equivalence class that contains DAGs with a wrong causal order if some of the variables are not included in the analysis. Similarly, the standard version of the LiNGAM algorithm (Shimizu et al. (2006)) might produce a wrong DAG in the presence of hidden common causes. Hoyer et al. (2008), Entner and Hoyer (2010), and Tashiro et al. (2014) proposed extensions of LiNGAM that deal with hidden variables. While Entner and Hoyer (2010), and Tashiro et al. (2014) show good performance in practice, all three methods suffer from some drawbacks. For example, the LiNGAM version of Hoyer et al. (2008) requires a priori the number of hidden variables in the SCM (or needs to estimate it from data). The main limitation of Entner and Hoyer (2010) is that it recovers causal information only for subsets of variables that are not affected by hidden confounders. For some nonancestral graphs, the method by Tashiro et al. (2014) does not identify all ancestral relationships (Wang and Drton (2020)). In addition, both the work of Hoyer et al. (2008) and Tashiro et al. (2014) are computationally intensive, with the latter showing a computational time that grows exponentially with the sample size and the number of observed variables.

Among the constraint-based methods, Spirtes, Glymour and Scheines (2000), Section 6.7, proposed the FCI method, which is an extension to the PC algorithm that deals with arbitrarily many hidden confounders and produces a partial ancestral graph (see Zhang (2008)). Due to the high number of independence tests, the FCI algorithm can be slow when the number of variables is large. For this reason, Claassen, Mooij and Heskes (2013) proposed the FCI+ algorithm, a faster version of FCI that is consistent in sparse high-dimensional settings with arbitrarily many hidden variables. In general, FCI type algorithms produce an equivalence class of graphs. They are not guaranteed to recover the causal order of the variables.

Compared to the methods mentioned above, our algorithm has the advantage of being computationally fast, and being able to produce a causal order without assumptions on the number of hidden variables and the sparsity of the true underlying DAG.

4.3. Noise variables with different tails. We have so far considered the case where the noise variables of a given SCM share the same tail coefficient $\alpha > 0$ and the same slowly varying function $\ell$. Consider now a heavy-tailed SCM over $p$ variables, as defined in Section 2.1,
with the difference that the noise variables have possibly different tail indices \( \alpha_1, \ldots, \alpha_p > 0 \) and slowly varying functions \( \ell_1, \ldots, \ell_p \in \text{RV}_0 \), that is, for \( j = 1, \ldots, p \),

\[
P(\varepsilon_j > x) \sim \ell_j(x)x^{-\alpha_j}, \quad x \to \infty.
\]

We say that \( \varepsilon_j \) has heavier (upper) tail than \( \varepsilon_k \) if either \( 0 < \alpha_j < \alpha_k \), or \( \alpha_j = \alpha_k \) and \( \ell_j(x)/\ell_k(x) \to \infty \) as \( x \to \infty \). Denote by \( G = (V, E) \) the DAG induced by the SCM. With similar arguments to the proof of Lemma 1, the causal tail coefficient for \( j, k \in V \) can then be expressed as

\[
\Gamma_{jk} = \frac{1}{2} + \frac{1}{2} \lim_{x \to \infty} \frac{\sum_{h \in A_{jk}} \beta_{\alpha_j}^{\alpha_h} P(\varepsilon_h > x)}{\sum_{h \in \text{An}(j, G)} \beta_{\alpha_j}^{\alpha_h} P(\varepsilon_h > x)},
\]

where \( A_{jk} = \text{An}(j, G) \cap \text{An}(k, G) \), and the sum over an empty index set equals zero. From (14), we can study the different constellations of \( X_j \) and \( X_k \) and the corresponding values of \( \Gamma_{jk} \). We obtain the following three statements.

1. If \( X_j \) and \( X_k \) are independent, the causal tail coefficient satisfies, as before, \( \Gamma_{jk} = 1/2 \).
2. If \( X_j \) is an ancestor of \( X_k \) then, as before, \( \Gamma_{jk} = 1 \).
3. In all other scenarios, it holds that \( \Gamma_{jk} < 1 \) as long as the noise variables \( \varepsilon_h, h \in A_{jk} \), of the common ancestors of \( X_j \) and \( X_k \) have tails that are lighter than (or as light as) the one of \( \varepsilon_j \). On the other hand, if there is some common ancestor of \( X_j \) and \( X_k \) for which the noise variable’s tail is heavier than the one of \( \varepsilon_j \), then \( \Gamma_{jk} = 1 \).

These statements help to understand in which cases the values of \( \Gamma_{jk} \) indicate a correct causal relation.

**Example 2.** Suppose that \( X_j \) is an ancestor of \( X_k \), and there is possibly a common ancestor \( X_0 \) (which can also be a hidden confounder). Since \( \Gamma_{jk} = 1 \), we will never mistakenly detect the existence of a causal effect from \( X_k \) to \( X_j \). If either \( X_j \) or \( X_0 \) has a heavier tail than \( X_k \), then \( \Gamma_{kj} = 1 \) and we cannot detect the causal effect from \( X_j \) to \( X_k \).

**Example 3.** Suppose neither \( X_j \) causes \( X_k \) nor \( X_k \) causes \( X_j \), and \( X_0 \) is a common ancestor of \( X_j \) and \( X_k \). If \( X_0 \) has a tail that is lighter than (or as light as) the one of \( X_k \), but heavier than the one of \( X_j \), then \( \Gamma_{jk} = 1 > \Gamma_{kj} \). Therefore, the causal tail coefficient indicates a wrong causal effect from \( X_j \) to \( X_k \).

Whenever there exists a causal effect between two variables, we can, at worst, fail to detect it (that is, the causal tail coefficient does not indicate a causal effect in the wrong direction). When there is no causal connection between two variables, the causal tail coefficient might indicate a wrong causal effect. However, this does not affect the correctness of the EASE algorithm, on the population level. Indeed, if \( \Gamma_{jk} = 1 > \Gamma_{kj} \), for \( j, k \in V \), there are two possibilities. If \( X_j \) is an ancestor of \( X_k \), then the algorithm correctly chooses \( j \) before \( k \). If \( X_j \) and \( X_k \) share a common ancestor, but none of them is causing the other (see Example 3), then any permutation of \( j \) and \( k \) yields a valid causal order. Example 2 shows that EASE could make mistakes when \( X_j \) is an ancestor of \( X_k \) and \( \Gamma_{jk} = \Gamma_{kj} = 1 \), since the causal tail coefficient does not indicate any causal effect. In this case, one could remove one variable at a time to obtain a subset \( \tilde{V} \subset V \) that satisfies \( \Gamma_{hm} < 1 \) or \( \Gamma_{mh} < 1 \) for each \( h, m \in \tilde{V} \). By applying the EASE algorithm to the subset of the remaining variables, one would recover a correct causal order on such subset.

For simplicity, we only considered the causal tail coefficient \( \Gamma \), but similar conclusions hold for \( \Psi \).
5. Numerical results.

5.1. Simulation study. We assess the performance of EASE in estimating a causal order of a graph induced by a heavy-tailed SCM. We simulate the SCMs with real-valued coefficients and different numbers of variables $p$ and samples $n$. The noise variables have Student’s $t$ distributions with different degrees of freedom $\alpha$ and we consider four different settings, including unobserved confounders and model misspecification; see Section D of the Supplementary Material for details. Since the coefficients in the SCM are real-valued, we use the causal tail coefficient $\Psi$ defined in Section 4.1 for our EASE algorithm. Our code is available as an R package at https://github.com/nicolagnecco/causalXtreme. Scripts generating all our figures and results can be found at the same url.

5.1.1. Competing methods and evaluation metric. We compare our algorithm to three well-established methods in causality, the Rank PC algorithm (Harris and Drton (2013)), ICA-LiNGAM (Shimizu et al. (2006)), and Pairwise LiNGAM (Hyvärinen and Smith (2013)).

The classic PC algorithm (Spirtes, Glymour and Scheines (2000), Section 5.4.2) belongs to the class of constraint-based methods for causal discovery. It estimates the Markov equivalence class of a DAG, encoded as a completed partially directed acyclic graph (CPDAG). The PC algorithm retrieves a CPDAG by performing conditional independence tests on the variables. The Rank PC algorithm, proposed by Harris and Drton (2013), is an extension of the PC algorithm and uses the rank-based Spearman correlation to perform the independence tests. This modification ensures that the method is more robust to non-Gaussian data.

The algorithms that fit our problem best are ICA-LiNGAM and Pairwise LiNGAM. ICA-LiNGAM, proposed by Shimizu et al. (2006), leverages the results of independent component analysis (ICA) (Comon (1994)) to estimate the DAG of a linear SCM under the only assumption that the noise is non-Gaussian. Pairwise LiNGAM, proposed by Hyvärinen and Smith (2013), is a likelihood-ratio-based method to identify the exogenous variables within the DirectLiNGAM framework. DirectLiNGAM, introduced by Shimizu et al. (2011), is an algorithm based on two iterative steps, namely, finding an exogenous variable (i.e., a node in the DAG with no parents), and regressing this variable out of all the others. In this simulation study, we let ICA-LiNGAM and Pairwise LiNGAM return only a causal order (and not a complete DAG structure), in order to make a fair comparison with EASE.

The algorithms return different types of causal information. On the one hand, EASE, ICA-LiNGAM, and Pairwise LiNGAM estimate a causal order. On the other hand, the Rank PC algorithm computes a CPDAG that represents a Markov equivalence class of DAGs. Therefore, when it comes to evaluating the performance of the algorithms, it becomes crucial to use a measure that is meaningful for all of them. We choose the structural intervention distance (SID) proposed by Peters and Bühlmann (2015). The SID takes as input either a pair of DAGs or a DAG and a CPDAG and returns the number of falsely inferred interventional distributions (Peters and Bühlmann (2015), Definition 3). We standardise the SID to lie between zero and one. For each method, we compute the distance between the simulated DAG, that is, the ground truth, and the estimated DAG or CPDAG. An estimated causal order $\hat{\pi}$ corresponds to a fully connected DAG $G = (V, E)$, where $(i, j) \in E$ if $\hat{\pi}(i) < \hat{\pi}(j)$. As a caveat, we slightly adapt the SID in the case of hidden confounders (see Setting 2 of our simulations), since it is not designed to work in such a situation.

5.1.2. Results. In this simulation experiment, we use the implementation of the Rank PC, and ICA-LiNGAM algorithm developed by Kalisch et al. (2012). We implemented Pairwise LiNGAM in C++ and included it in our software package.
FIG. 3. The figure refers to Section 5.1.2. It shows the structural intervention distance (SID) of the EASE algorithm for different fractional exponents $\nu \in [0.2, 0.7]$ of $k_n = [n^\nu]$ and different tail indices $\alpha \in \{1.5, 2.5, 3.5\}$. Each point represents the SID measure averaged over 10 random samples for different sample sizes $n \in \{500, 1000, 10,000\}$ and dimensions $p \in \{4, 7, 10, 15, 20, 30, 50\}$ in a linear SCM. In the experiments of Section 5.1, we set $k_n = [n^{0.4}]$.

Regarding the hyperparameter settings, for the Rank PC algorithm, we perform a conditional independence test based on Spearman’s correlation coefficient, as proposed by Harris and Drton (2013), and we set the level of the independence tests to 0.0005.

Concerning the choice of the number of exceedances $k_n$ in the EASE algorithm, we perform a small preliminary simulation. Figure 3 shows the SID of EASE for $k_n = [n^\nu]$ and different fractional exponents $\nu > 0$. The best fractional exponent in Figure 3 seems to depend on the tail heaviness of the noise variables, and in particular it appears to be smaller for larger values of the degree of freedom $\alpha$ of the Student’s $t$ distribution. Our estimators $\hat{\Gamma}_{ij}$ and $\hat{\Psi}_{ij}$ are similar in construction to Hill’s estimator (Hill (1975)). For the latter, the optimal number $k^*_n$ of exceedances depends on the tail index and an index related to a second-order condition; see Section 3.2 in de Haan and Ferreira (2006) for details. For the Student’s $t$ distribution with $\alpha$ degrees of freedom it can be shown that $k^*_n \sim C_\alpha n^{1/(\alpha+1)}$, where $C_\alpha > 0$ is a constant. This intuitive explanation coincides well with the optimal fractional exponents in Figure 3. In the sequel, we choose $k_n = [n^{0.4}]$ because it lies within the best range for the fractional exponent. This result also agrees with the assumptions of Theorem 2, where $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$.

Regarding the simulation settings, we let $n$ denote the number of observations, $p$ the number of variables, and $\alpha > 0$ the tail index of the simulated distribution. For each combination of $n \in \{500, 1000, 10,000\}$, $p \in \{4, 7, 10, 15, 20, 30, 50\}$ and $\alpha \in \{1.5, 2.5, 3.5\}$ we simulate 50 random SCMs under four different settings. The simulated data is independent of the data used to choose the best fractional exponent of $k_n$ (see Figure 3). The first setting corresponds to linear SCMs with real-valued coefficients described in Section 4.1. In the second setting, we introduce hidden confounders. The third setting corresponds to nonlinear SCMs. In the fourth setting, we first generate linear SCMs and then transform each variable to uniform margins. Further details on the generation of the SCMs are in Section D of the Supplementary Material. For each simulation and setting we evaluate the performance of EASE, ICA-LiNGAM, Pairwise LiNGAM, and Rank PC algorithm with the SID. As a baseline, in each simulation, we also compute the SID of a randomly generated DAG, where we randomly choose the causal order, the sparsity and the edges of the graph.

Figure 4 displays the results of the simulations when the tail index $\alpha = 1.5$. We can observe that EASE is quite robust across the four different settings. We explain this finding as follows. In the presence of hidden confounders (Setting 2), EASE can retrieve a correct causal
The figure refers to Section 5.1.2. It shows the SID averaged over 50 simulations, for each method, setting, sample size $n$ and dimension $p$, when the tail index is $\alpha = 1.5$. Each row of the figure corresponds to one setting. In order, the settings are: (1) Linear SCM; (2) Linear SCM with hidden confounders; (3) Nonlinear SCM; (4) Linear SCM where each variable is transformed to a uniform margin.

Comparing to the other methods, we observe that EASE performs better than Rank PC across all settings, and better than ICA-LiNGAM in Settings 2 and 4. Pairwise LiNGAM is overall the best performing method, except in Setting 4. Also, Pairwise LiNGAM is less affected by misspecifications in the bulk of the data distribution (Setting 3), compared to ICA-LiNGAM. One reason is that Pairwise LiNGAM relies on ordinary least square regression that is sensitive to high-leverage points. In this particular setting, ICA-LiNGAM and Pairwise LiNGAM gain in robustness if we discard the data in the bulk of the distribution; see Table 1 in the Supplementary Material. Furthermore, both ICA-LiNGAM and Pairwise LiNGAM are the algorithms with the best convergence for high dimension $p$, as $n$ increases. This result is not surprising because EASE uses only the $k_n < n$ upper order statistics to recover the causal structure. In addition, we notice that Pairwise LiNGAM outperforms ICA-
LiNGAM, in agreement to the findings of Hyvärinen and Smith (2013). Regarding the Rank PC algorithm, we can see that it is quite stable under different settings, but it performs only marginally better than the random method. Moreover, in Figure 7 of the Supplementary Material, we observe that the performance of the Rank PC algorithm is almost constant for significance levels of the independence test between $5 \cdot 10^{-4}$ and 0.5. The results do not change when we consider the standard PC algorithm, which is based on partial correlation as a conditional independence test. The results for tail indices $\alpha = 2.5, 3.5$ are in Figures 3 and 4 in the Supplementary Material. Increasing values of $\alpha$ correspond to lighter tails, and we observe that it becomes more challenging for EASE to recover the correct causal order.

In addition to the competitive performance in the simulations, a further advantage of EASE is its computational efficiency. The algorithm performs computations only on the tails of the dataset and relies on simple nonparametric estimators of the causal tail coefficient; see Section 3.3. Figure 6 in the Supplementary Material shows that EASE can be up to two orders of magnitude faster than the other methods.

5.2. Financial application. In general, one cannot easily reason about causality in financial markets. Several factors influence financial returns, and most of them are unobserved. In addition, the effect of these factors varies in time. However, under particular circumstances, it is possible to conjecture the existence of a specific causal relationship, with a reasonable degree of confidence. For example, in the Swiss financial market, one can argue that very large (both positive and negative) changes to the Euro Swiss franc exchange rate (EURCHF) induce changes in the Swiss Market Index (SMI), the main stock index in Switzerland. This is due to multiple reasons such as the multinational nature and the high export dependency of the Swiss economy. Consider, for instance, the decision of the Swiss National Bank (SNB) to discontinue the minimum exchange rate between Swiss franc and Euro, on January 15, 2015. This event can be deemed as a large intervention with a plausible causal interpretation (in the spirit of Cox and Wermuth (1996), Section 8.7). Following the SNB decision, the EURCHF plummeted more than 30 standard deviations, and all the stocks included in the SMI dropped in value on the same day.

For this reason, we consider the returns of the Euro Swiss franc exchange rate (EURCHF) and the three largest Swiss stocks in terms of market capitalisation, namely, Nestlé (NESN), Novartis (NOVN) and Roche (ROG). We choose to analyse three individual stocks instead of the SMI for three reasons. First, we deem it more appropriate to test our assumptions on more than two variables. Moreover, the three stocks make up 50\% of the SMI composition, and thus they are representative of the index itself. Furthermore, all three companies are multinational corporations with a homogeneous exposure to foreign markets and a negligible fraction of revenues coming from the Swiss market (see Nestlé (2019), Novartis (2019), Roche (2019)). The last point suggests that the effect (if any) of EURCHF on these stocks does not depend on the idiosyncrasies of each firm.

The dataset consists of daily returns spanning from January 2005 to September 2019 and includes $n = 3832$ observations. The goal is to assess whether EASE can retrieve a correct causal order for the set of four variables. As ground truth, we conjecture that large changes in EURCHF (both positive and negative) will cause large changes in the stock returns, but not vice versa. Figure 5 shows the causal structure corresponding to our hypothesis.

Before running the EASE algorithm, we assess the tail behaviour of each variable by estimating the shape parameter $\xi$ of a generalised Pareto distribution (see Embrechts, Klüppelberg and Mikosch (1997), Section 3.4) on the threshold data. Recall that $\xi$ is the reciprocal of the tail index $\alpha = 1/\xi$, if $\xi > 0$. For each variable and each tail (upper and lower) we estimate the $\xi$ parameter using 200 observations, corresponding to the 95\%-quantile, approximately. The estimated parameters and their standard errors for the upper tails are 0.31 (0.08)
for EURCHF, 0.25 (0.08) for NESN, 0.16 (0.07) for NOVN, and 0.25 (0.09) for ROG. Regarding the lower tail, the estimated parameters and their standard errors are 0.27 (0.08) for EURCHF, 0.12 (0.08) for NESN, 0.17 (0.08) for NOVN, and 0.23 (0.10) for ROG. By adding and subtracting two standard errors from each point estimate of $\xi$, we observe that the lower shape parameter of Nestlé and Novartis is not significantly different from zero. Moreover, the shape parameters for the Euro Swiss franc exchange rate and for Roche are significantly different from zero in both tails. In addition, since the confidence intervals of all estimates are overlapping, the assumption of a common shape parameter seems reasonable. It seems however that the returns of Nestlé and Novartis have slightly lighter tails compared to the other two variables.

With the goal of recovering a causal order with EASE, first we estimate the $\hat{\Psi}$ matrix from the full dataset, by setting the number of exceedances to $k = 10$ (this corresponds to $\lfloor n^{0.3} \rfloor$, approximately). We run the EASE algorithm on the matrix $\hat{\Psi}$, and we obtain the causal order $\hat{\pi}^{-1} = (EURCHF, NOVN, ROG, NESN)$; this agrees with the proposed ground truth of Figure 5. As a comparison, also ICA-LiNGAM and Pairwise LiNGAM recover a causal order that agrees with our hypothesis.

Since our results are based on the $k = 10$ upper order statistic, we assess the variability of the estimates $\hat{\Psi}$ for different values of $k = \lfloor n^\nu \rfloor$, with $\nu \in [0.2, 0.7]$. Figure 6 shows the estimated coefficients $\hat{\Psi}$ for the pairs (EURCHF, NESN), (EURCHF, NOVN), and (EURCHF, ROGN), with the corresponding 90% bootstrap confidence intervals. In the three plots, the black (blue) line corresponds to the estimated coefficient $\hat{\Psi}_{EURCHF,i} (\hat{\Psi}_{i,EURCHF})$, with $i = NESN, NOVN, ROG$. We can interpret the difference between the black and blue lines as a causal signal, since $\Psi_{ij} - \Psi_{ji} > 0$ if variable $i$ causes variable $j$, for $i, j \in V$ (see Section 4.1). For the pairs (EURCHF, NESN) and (EURCHF, NOVN) the blue and the black lines overlap for all values of the upper order statistic $k$, and therefore any possible causal effect is not identified by the estimated coefficient $\hat{\Psi}$. This result agrees with Example 2 of Section 4.3 which shows that the causal tail coefficients do not identify a causal signal when the ancestor has a heavier tail than its descendant—as is the case for the pairs (EURCHF, NESN) and (EURCHF, NOVN). In contrast, if we consider the pair (EURCHF, ROGN) we notice that the difference $\hat{\Psi}_{EURCHF,ROGN} - \hat{\Psi}_{ROGN,EURCHF}$ is positive for all fractional exponents $\nu \leq 0.4$. This can be explained by the fact that EURCHF and ROGN have comparable tail indices, and therefore it is easier for the coefficient $\hat{\Psi}$ to detect a possible causal effect between the variables. In Section F of the Supplementary Material, we show the dynamic evolution of the $\hat{\Psi}$ coefficient across time.

Given the highly complex nature of financial markets, we do not take the conclusion of this experiment as a definite answer but rather consider it as an indication for a possible causal relationship in this data.

5.3. River data. We apply the EASE algorithm to the average daily discharges of the rivers located in the upper Danube basin. This dataset has been studied in Asadi, Davison and Engelke (2015), Engelke and Hitz (2020) and Mhalla, Chavez-Demoulin and Dupuis (2020), and it is made available by the Bavarian Environmental Agency (http://www.gkd.bayern.de).
We consider average daily discharges for 12 stations along the basin, representing the different tributaries and different sections of the Danube, and 11 of them are a subset of the 31 stations selected by Asadi, Davison and Engelke (2015). We exclude some of the 31 stations that are spatially very close since those are highly dependent and almost indistinguishable. For convenience, we name the stations with the same numbers used in Asadi, Davison and Engelke (2015). While Asadi, Davison and Engelke (2015) decluster the data prior to their analysis in order to obtain independent samples, we use all observations despite the possible temporal dependence. In fact, for extreme value copulas, Zou, Volgushev and Böcher (2021) show that the use of a larger but possibly dependent dataset can decrease the asymptotic estimation error. Moreover, Fawcett and Walsh (2007) argue that considering all exceedances over a high threshold reduces the bias of the maximum likelihood estimators compared to a declustered analysis. To account for the time dependence of the exceedances, they adjust the standard errors using methods presented by Smith (1990). In this experiment, we compute the standard errors according to the adjustment proposed by Fawcett and Walsh (2007).

The dataset spans from 1960 to 2009, where we consider only the summer months, that is, June, July and August. The rationale is that most of the extreme observations occur in summer due to heavy rainfall. The final dataset contains \( n = 4600 \) observations. The rivers
have an average volume that ranges between $20 \text{ m}^3/\text{s}$ (for the upstream rivers) and $1400 \text{ m}^3/\text{s}$ (for the downstream rivers). A map of the basin can be seen in Figure 5 of the Supplementary Material. In order to implement our method, we first assess whether the equal tail index assumption is satisfied. To do so, we consider a regional model similar to the one presented by Asadi, Davison and Engelke (2015). We split the stations into four separate regions. Region 1 contains three stations in the southwest of the upper Danube basin and the catchment areas are located at mid-altitude; region 2 includes three stations in the Inn–Salzach basin whose tributaries are located in high-altitude alpine regions; region 3 contains four stations along the main Danube with large average water volume; region 4 comprises two stations in the north of the Danube. For each region, we fit a Poisson point process likelihood (Coles (2001), Chapter 7) by considering exceedances over the 90% quantile and by constraining the shape parameter $\xi$ to be equal across the stations within the same region. To address the presence of temporal dependence in the exceedances, we adjust the standard errors as shown by Fawcett and Walshaw (2007) and, based on these, we compute approximate confidence intervals. For each region, the estimated shape parameter and the corresponding confidence intervals are $0.167$ ($0.062, 0.273$), $0.145$ ($0.047, 0.242$), $0.133$ ($0.027, 0.239$) and $0.229$ ($0.099, 0.358$), respectively. The fact that the confidence intervals overlap suggests that our assumption of equal tail index across the variables is satisfied. Moreover, all confidence intervals do not include the zero value, and therefore the data can be deemed to be heavy-tailed.

We perform two separate analyses to identify causal structures both in space and time. Regarding the spatial structure, the goal is to recover the causal order of the network flow of the 12 stations on the rivers. This is a nontrivial task for two reasons. First, the water discharges at the stations can be confounded by rainfall that spreads out across the region of interest. Second, the large difference in water volume between the stations can further mask the causal structure of the water flow. The true DAG of the spatial disposition of the stations is shown in Figure 7. We run the EASE algorithm based on the $\Gamma$ coefficient, considering the upper tails, and setting the number of exceedances to $k = 29 = \lceil n^{0.4} \rceil$. The estimated causal order $\hat{\pi}^{-1} = (23, 32, 26, 28, 19, 21, 11, 9, 7, 14, 13, 1)$ is correct, and the corresponding fully connected DAG has an SID equal to 0. We also run ICA-LiNGAM and Pairwise LiNGAM on the same dataset, and we obtain an SID of 0 and 0.053, respectively. Clearly, in this example, the causal structure in the bulk of the distribution is the same as in the extremes. To assess the variability of our results, we compute the average SID of EASE over 50 bootstrap samples, for different values of the threshold parameter $k = \lceil n^{\nu} \rceil$, $\nu \in [0.2, 0.7]$—see Figure 8. From this figure, we observe that the fractional exponent $\nu \approx 0.4$ yields a good performance both in terms of SID and variability. The value for the optimal fractional exponent agrees with the empirical findings of Section 5.1.

Concerning the time-series analysis, we consider each station individually and try to recover the direction of time from the lagged data. Consider an autoregressive (AR) process of order $p \geq 1$,

$$ X_t = \sum_{j=1}^{p} \beta_j X_{t-j} + \varepsilon_j, \quad t \geq 0, $$
where the $\varepsilon_j$ are regularly varying with comparable tails. For a detailed discussion of such time series models we refer to Basrak and Segers (2009) and Embrechts, Klüppelberg and Mikosch (1997), Chapter 7. Peters et al. (2009) prove that an AR($p$) process is time reversible, that is, can be represented by an AR($p$) process in the reversed time direction, if and only if the noise is Gaussian. This means that for heavy-tailed random variables, one can in principle detect the direction of time from the data. For each station, we construct a dataset $D$ where the rows correspond to different days and the columns to different lags. We denote by $X_0, X_1, \ldots, X_6$ the columns containing the current and lagged values of the station discharge. We then run EASE on the dataset $D$ and recover a causal order $\hat{\pi}$ over the seven variables $X_0, \ldots, X_6$. We say that the direction of time is correctly inferred if the estimated causal order places the lags in the correct position, that is, $\hat{\pi}(j) < \hat{\pi}(i)$ if $j < i$, for $i, j = 0, \ldots, 6$. EASE successfully recovers the direction of time for 11 out of the 12 stations. As a comparison, ICA-LiNGAM, and Pairwise LiNGAM find the correct order for 9 and 11 stations, respectively.

So far, we have not considered a multivariate time series analysis of the dataset. On the one hand, in this particular application of the river data, the effects among the stations are almost instantaneous—due to the closeness of the water catchment areas, the fact that we have daily values and the river speed, which is about ten kilometres per hour. On the other hand, time information usually helps in estimating causal relations. For this reason, we apply multivariate Granger causality (Granger (1969)) to the river data, considering one day lag. For each pair of stations $(i, j)$, we say that $i$ Granger-causes $j$ if the corresponding $p$-value is significant at a 0.05 level, after the Bonferroni correction. We sort the significant $p$-values in ascending order, and we sequentially add the directed edge $(i, j)$ if nodes $i$ and $j$ are not connected. We continue until the skeleton of the resulting graph is connected, or all the $p$-values have been selected. The resulting directed tree achieves an SID of 0.083. Alternatively, if we sequentially add directed edges $(i, j)$ that do not create cycles (until all $p$-values have been selected) we obtain a DAG with an SID of 0.196.

6. Discussion and future work. In several real-world phenomena, the causal mechanisms between the variables appear more clearly during extreme events. Moreover, there are situations where the causal relationship in the bulk of the distribution differs from the structure in the tails. We have introduced an algorithm, named extremal ancestral search (EASE), that is shown to consistently recover the causal order of a DAG from extreme observations.

Fig. 8. The figure refers to Section 5.3. It shows the robustness of the structural intervention distance (SID) for varying fractional exponents $\nu \in [0.2, 0.7]$ of $k_\nu = \lfloor n^\nu \rfloor$. Each point represents an average over 50 SID evaluations for the EASE algorithm, after bootstrapping the original dataset. The shaded interval corresponds to the 90% bootstrap confidence interval.
only. EASE has the advantage of relying on the pairwise causal tail coefficient, and therefore it is computationally efficient. In addition, our algorithm can deal with the presence of hidden confounders and performs well for small sample sizes and high dimensions. The EASE algorithm is robust to model misspecifications, such as nonlinear relationships in the bulk of the distribution, and strictly monotone increasing transformations applied marginally to each variable.

This work sheds light on a connection between causality and extremes, and thereby opens new directions of research. In particular, it might be interesting to study the properties of the causal tail coefficient under more general conditions. This includes high-dimensional settings where the dimension grows with the sample size, more general SCMs where the functional relations between the variables can be nonlinear, and settings where the noise variables have lighter tails. For example, in future research, one could combine EASE with regression techniques to obtain the complete DAG structure, compute the residuals and then test them for the assumption of common tail indices.

Another possible extension may consider multivariate time series data, where the temporal order of cause and effect could help to estimate causal relationships among variables; see, for example, Granger (1969). Future work might study how to combine our approach with the Granger causality framework. For instance, one could first perform a Granger causality analysis, and then, apply EASE to the residuals of the vector autoregression model. For a careful study in this direction, it would also be necessary to investigate the statistical properties of the residuals.

Acknowledgments. We thank Cesare Miglioli, Stanislav Volgushev and Linbo Wang for helpful discussions. We are grateful to the editorial team and two anonymous referees for constructive comments that improved the paper.

Funding. JP was supported by research grants from VILLUM FONDEN and the Carlsberg Foundation, and SE was supported by the Swiss National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplement to “Causal discovery in heavy-tailed models” (DOI: 10.1214/20-AOS2021SUPP; .pdf). The supplementary material (Gnecco et al. (2020)) consists of six sections. Section A summarises important facts about regularly varying random variables. Section B contains the proofs of the results of the paper. Section C illustrates how the EASE algorithm retrieves a causal order of a DAG. Section D describes the settings used in the simulation study. Section E contains additional figures and tables. Section F presents further results for Section 5.2.

REFERENCES


