Statistica Sinica Preprint No: SS-2024-0199	
Title	Efficient Learning of DAG Structures in Heavy-tailed Data
Manuscript ID	SS-2024-0199
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0199
Complete List of Authors	Wei Zhou, Xueqian Kang, Wei Zhong and Junhui Wang
Corresponding Authors	Xueqian Kang
E-mails	kangxueqian@stu.xmu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

Efficient Learning of DAG Structures in Heavy-tailed Data

Wei Zhou, Xueqian Kang, Wei Zhong and Junhui Wang

Southwestern University of Finance and Economics, Xiamen University and Chinese University of Hong Kong

Abstract:

Directed acyclic graph (DAG) models are widely used to discover causal relationships among random variables. However, most existing DAG learning algorithms are not directly applicable to heavy-tailed data which are commonly observed in finance and other fields. In this article, we propose a two-step efficient algorithm based on topological layers, referred as TopHeat, to learn linear DAGs with heavy-tailed error distributions which include Pareto, Fréchet, log-normal, Cauchy distributions, and so on. First, we reconstruct the topological layers hierarchically in a top-down fashion based on the new reconstruction criteria for heavy-tailed DAGs without assuming the popularly-employed faithfulness condition. Second, we recover the directed edges via the modified conditional independence testing for heavy-tailed distributions. We theoretically demonstrate the consistency of the exact DAG structures. Monte Carlo simulations validate the outstanding finite-sample performance of the proposed algorithm compared with competing methods. In the real data analysis, we analyze the exchange rates among 17 countries and uncover the source of financial contagion and the pathways, which indicates that the financial risk contagion effect became increasingly stable among European countries as the euro was introduced.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

Key words and phrases: Causality, exact DAG structures, heavy-tailed data, topological layers, conditional independence testing.

1. Introduction

Directed acyclic graphs (DAG) provide a powerful tool to describe the causal relationships via the directional parent-child arrows among random variables, which has received growing attention in many application domains (Pearl, 2000). Despite the success, existing DAG learning methods are developed largely relying on the assumption of Gaussian, sub-Gaussian, and moment bounded distributions. Yet, heavy-tailed data frequently appear in finance and insurance due to the occurrence of rare events (Resnick, 2007; Peng and Qi, 2017), which brings great challenges to existing methods for learning DAG structures.

Recently, only a few studies have explored DAG learning in heavy-tailed financial data. For instance, the popular PC algorithm (Spirtes et al., 2000) produces a partial DAG when learning the causal structures of the credit risk among financial institutions (Yang and Zhou, 2013), and of the implied volatilities of U.S. Treasury bonds, global stock indices, and commodities (Yang and Zhou, 2017). In the seminal work, Gnecco et al. (2021) proposed the extremal ancestral search (EASE) algorithm to recover causal orderings among returns of the Euro Swiss franc exchange rate and three largest Swiss stocks. However, the aforementioned methods fail to recover complete DAG structures, which is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

of great importance to understand the systemic risk. Specifically, the financial contagion pathway can be represented with the directed relationships in a DAG, which shows the propagation of financial shocks or disturbance from one currency to another in the financial exchange market.

In literature, structure learning methods of DAGs are mainly of two types, including constraint-based algorithms (Spirtes et al., 2000) and score-based methods (Chickering, 2003). Recently, identifying a unique DAG from the joint distribution by imposing a structural causal model (SCM, Peters et al., 2017) has been extensively studied (Peters and Bühlmann, 2014). Heavy-tailed distributions are special examples of non-Gaussian DAG models, early attempts of which include Shimizu et al. (2006, 2011); Hyvärinen and Smith (2013), and high-dimensional non-Gaussian DAGs are also considered in Wang and Drton (2020) with the moment quantities and Zhao et al. (2022) with the precision matrix, respectively. However, these aforementioned methods designed for learning non-Gaussian DAGs often lead to underestimation of extremal events in heavy-tailed distributed data, such as the financial risk (Klüppelberg and Krali, 2021) and flooding in river network (Asadi et al., 2015).

In this paper, we propose a two-step learning algorithm for heavy-tailed DAGs based on topological layers. Explicitly, a DAG can be reformulated via the layer structure with the number of layers defined by the longest length of a

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

directed path from a root node to a leaf node, and the parents of each node must lie in its upper layers. In particular, we first show that the topological layers can be fully reconstructed in a top-down fashion based on a modified expected shortfall measure. Second, the directed edges can be determined by applying the refined conditional independence testing (CIT) procedure for heavy-tailed distributions hierarchically. The proposed method, denoted as TopHeat, is computationally efficient and its asymptotic properties are provided in terms of exact DAG recovery. The superior performance is supported by simulation studies and real-life examples, where we study the exchange rates data of 17 countries and discover financial contagion paths arising from the foreign exchange market.

The main contribution of this paper is the proposed efficient learning algorithm for a heavy-tailed DAG with a diverging number of nodes and its statistical guarantees of the recovery consistency for the underlying DAG structures. Specifically, we first show that the topological layers of a heavy-tailed DAG can be sequentially reconstructed in Lemma 1. Secondly, we establish the asymptotic normality of the reconstruction measure in Theorem 2 with the help of extreme value theory and tail empirical process, which is particularly attractive in line of the research in actuarial science and risk management. More importantly, we connect the heavy-tailed DAG learning with the CIT measure (Azadkia and Chatterjee, 2021), by extending it from sub-exponential distributions to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

accommodate heavy-tailed distributions, and derive the tail bound for the sample CIT measure in Proposition 1. Overall, we establish the statistical guarantees in terms of exact DAG recovery, which is among first attempts in heavy-tailed DAG learning literature. We need to emphasize the differences between the proposed method and some existing works, which fail to give solutions to obtain the complete DAG structures (Gnecco et al., 2021), or cannot adapt to a general heavy-tailed distribution family (Zhao et al., 2022). Further, TopHeat is computationally efficient among related methods when dealing with shallow graphs for large node size. More details for the comparison with recent DAG learning methods are provided in Section 1.1.

The rest of this paper is organized as follows. In Section 2, we introduce the heavy-tailed DAG models. In Section 3, we propose an efficient learning algorithm for heavy-tailed DAGs by developing the criteria to reconstruct the topological layers. In Section 4, we investigate the consistency of recovering the underlying DAG structures under regularity conditions. In Section 5, we conduct numerical studies to compare the proposed algorithm with competing methods. Section 6 applies the proposed method to analyze the foreign exchange rates data. Section 7 contains a brief discussion. The proof of theoretical results and additional experimental results are presented in the supplementary material.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

1.1 Related methods

In this subsection, we mainly compare the proposed TopHeat method with some existing competitors in terms of theoretical results and computational complexity. Specifically, Gnecco et al. (2021) proposed EASE to learn the causal ordering of a heavy-tailed DAG sequentially. Most recently, Zhao et al. (2022) developed a non-Gaussian DAG learning method, named by TL, by utilizing the topological layers in a bottom-up fashion.

Theoretically, EASE only investigates the reconstruction criteria of the causal ordering and its consistency when the number of nodes is fixed, without providing solutions for complete DAG structures. However, the proposed TopHeat overcomes these obstacles, with consistency DAG structure recovery that allows the number of nodes and layers to diverge with the sample size. Moreover, the established consistency result of TL largely depends on the maximum cardinality of Markov blankets (Peters et al., 2017) and only accommodates some special distributions, which cannot adapt to a general heavy-tailed distribution family we consider in this paper, not to mention that the violation of the conditions for the precision matrix to produce false layers and edges. Interestingly, TopHeat imposes no additional assumptions on the graph structures.

The computational complexity of TopHeat is much smaller than EASE in a shallow graph, and they become the same when the number of layers is equiv-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

alent to the number of nodes in a chain graph. Further, TopHeat is computationally much more efficient than TL for small sample size or small ancestors. Detailed complexity comparisons analytically and numerically are provided in Sections 3.2 and 5.2, respectively.

Further, we give some guidance on when to choose these algorithms in real-world application. In practice, practitioners should first conduct descriptive statistical analysis and give a basic idea about the data distribution. If the histogram shows a light tail, then various DAG learning methods can be used. However, if the histogram shows polynomial decaying tail, which is heavier than Gaussian, and the probabilty of tail is relatively small, TL is recommended to learn a DAG. Furthermore, if the probabilty of tail is not small, then we may choose TopHeat and EASE to estimate the causal graphs. Clearly, methods designed for learning heavy-tailed DAGs can reveal more information in real-world financial data compared with general non-Gaussian DAGs learning algorithms. Therefore, we strongly recommend using TopHeat to obtain a complete DAG for heavy-tailed data, out of its efficiency and superior performance over EASE.

2. Heavy-tailed DAG

A DAG model is widely used to encode the joint distribution of $\mathbf{X} = (X_1, ..., X_p)^\top$. Precisely, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a DAG, where $\mathcal{V} = \{1, ..., p\}$ represents a set of nodes each corresponding to one X_j, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes a set of directed

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

edges without directed cycles. A directed edge from node j to node m is denoted as $j \to m$, and then node j is a parent of node m. We denote node m's parents in a DAG $\mathcal{G}$ as pa_m, and let $\mathbf{X}_{\text{pa}_m} := \{X_j : j \in \text{pa}_m \subset \mathcal{V}\}$. In general, for any subset $\mathcal{S} \subset \mathcal{V}$, we denote $\mathbf{X}_{\mathcal{S}} := \{X_j : j \in \mathcal{S} \subset \mathcal{V}\}$. We also define a directed path from node m_1 to node m_q in $\mathcal{G}$ as a sequence of distinct nodes $m_1, \ldots, m_q$ such that $m_j \in \text{pa}_{m_{j+1}}$ for $j = 1, \ldots, q-1$. If there is a directed path from node j to node m, we say j is an ancestor of m in $\mathcal{G}$. We denote the set of ancestors of node m as an_m, and $\text{An}_m = \text{an}_m \cup \{m\}$. Assume that the joint distribution $P(\mathbf{X})$ satisfies the Markov property with respect to $\mathcal{G}$, and thus it allows for the factorization, $P(\mathbf{X}) = \prod_{j \in \mathcal{V}} P(X_j | \mathbf{X}_{\text{pa}_j})$, where $P(X_j | \mathbf{X}_{\text{pa}_j})$ denotes the conditional distribution of X_j given its parents $\mathbf{X}_{\text{pa}_j}$. We also assume causal minimality (Peters et al., 2017) holds.

Next, we consider a linear structural causal model (SCM)

$$X_m = \sum_{j \in \text{pa}_m} \beta_{mj} X_j + \varepsilon_m, \; m = 1, \ldots, p, \tag{2.1}$$

where β_{mj} is assumed to be strictly positive. Assume that $\varepsilon_1, \ldots, \varepsilon_p$ are independently sampled from a (right) heavy-tailed distribution with regularly varying tails with the tail index θ, given in Definitions S1–S2 in Section S1 of the supplementary material. That is, there exists $c_m > 0$ and for each $m \in \mathcal{V}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

$$P(\varepsilon_m > x) \sim c_m h(x) x^{-\theta}, \text{ as } x \to \infty, \tag{2.2}$$

for some $h \in \text{RV}_0$, where RV_0 is a slowly regulary varying function defined in Definitions S2. Here, for any functions f and g, we denote $f \sim g$ if $\lim_{x\to\infty} f(x)/g(x) \to 1$. In the sequel, we denote the model in (2.1) and (2.2) as the heavy-tailed SCM. Heavy-tailed distributions have been frequently employed in analyzing real-world financial data, including Pareto, Fréchet, log-Gamma, Student's-t, Cauchy distribution, and many others. It is worthy noting that some other discrete distributions are also included in the heavy-tailed distribution. For example, the insurer's net loss (the total number of claims less premiums) is quantified as a discrete real-valued random variable within time periods and assumed with a regularly varying tail in literature (Li and Tang, 2015). The literature has also documented substantial heavy-tailed distributed datasets, such as stock market returns, exchange rates, and interest rates, which have infinite fourth moments and are collected to capture complex relationships for financial forecasting, risk management, and portfolio optimization (Lee, 1992; Chen and Schienle, 2022).

The linear SCM model in (2.1) can be rewritten as a matrix form $\mathbf{X} = \mathbf{B}\mathbf{X} + \boldsymbol{\varepsilon}$ with $\mathbf{B} = (\beta_{mj}) \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$, where β_{mj} is considered as the direct causal effect of X_j on X_m. This implies that $\mathbf{X} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\varepsilon} := \boldsymbol{\Pi}\boldsymbol{\varepsilon}$, where $\boldsymbol{\Pi} = (\pi_{mj}) \in \mathbb{R}^{p\times p}$ with π_{mj} as the total effect of X_j on X_m and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

$\pi_{jj} = 1$ for all j. To read off the dependence from the graph, faithfulness is commonly assumed (Peters et al., 2017) and it is required that $\pi_{mj} \neq 0$. Note that if all β_{mj}'s in (2.1) are positive, then $\pi_{mj} > 0$ for any $m \neq j$, and thus the faithfulness assumption is automatically satisfied. In this paper, we first investigate the positive-valued coefficients case, and then extend it to the real-valued scenario with $\beta_{mj} \in \mathbb{R}$ in Section S8 of the supplementary material.

3. Two-step DAG learning algorithm

In this section, we first introduce the concept and reconstruction criteria of topological layers for a heavy-tailed DAG, and then the proposed two-step efficient learning algorithm to recover the exact DAG structures.

3.1 Reconstruction of topological layers

The definition of topological layers of a DAG is explicitly given in Section S1 of the supplementary material. In literature, the topological layers have been widely employed for learning DAG structures (Gao et al., 2020; Zhao et al., 2022; Zhou et al., 2022). Examples of the topological layers structure of a DAG are displayed in Figure 1. However, the distribution classes considered in Gao et al. (2020); Zhou et al. (2022) fail to satisfy regularly varying conditions and the precision matrix in Zhao et al. (2022) for non-Gaussian DAGs is not identifiable for heavy-tailed random variables (Zhao and Liu, 2014). Next, we provide

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

the provable reconstruction result of topological layers for a heavy-tailed DAG.

We first introduce some notations. Let $F_j(X_j)$ denote the marginal cumulative distribution function (cdf) of X_j, and define the causal tail coefficient matrix $\mathbf{\Gamma} = (\Gamma_{jm})_{j,m=1}^{p} \in \mathbb{R}^{p\times p}$ with

$$\Gamma_{jm} = \lim_{u\to 1^-} E\{F_m(X_m)|F_j(X_j) > u\}. \tag{3.1}$$

Note that $\Gamma_{jm} \in [0,1]$ by definition, and it can be used to capture the causal relationship between nodes j and m. Intuitively, Γ_{jm} tends to 1, if j has a directed path to m, which means that extremes of X_j are more likely to lead to those of X_m. However, if there is no directed path from j to m and no causal relationships exist between j and m, then we expect that Γ_{jm} is strictly much smaller than 1. It is worth pointing out that the measure Γ_{jm} is developed from the expected shortfall where the tail-dependent variables are replaced with their marginal cdfs, and proposed to describe the ancestor-descendant relationship between two variables in Gnecco et al. (2021) in terms of the causal ordering.

Lemma 1. *We consider the heavy-tailed linear SCM model in* (2.1)–(2.2). *Given* $\mathcal{A}_0, \ldots, \mathcal{A}_{t-1}$, *we let* $\mathcal{C}_0 = \mathcal{V}$ *and* $\mathcal{C}_t = \mathcal{V}\setminus\bigcup_{d=0}^{t-1}\mathcal{A}_d$, *and then it holds true that* $\mathcal{A}_t = \{m \in \mathcal{C}_t : \max_{j\in\mathcal{C}_t}\Gamma_{jm} < 1\}$.

Lemma 1 provides a constructive proof of reconstructing the topological

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

layers of a heavy-tailed DAG via mathematical induction. Particularly, we first identify $\mathcal{A}_0 = \{m \in \mathcal{V} : \max_{j\in\mathcal{V}} \Gamma_{jm} < 1\}$, since $\mathrm{an}_m = \emptyset$ if $m \in \mathcal{A}_0$ and thus $\Gamma_{jm} < 1$ for all $j \in \mathcal{V}$ and $j \neq m$. Otherwise, $\Gamma_{jm} = 1$ holds for any $j \in \mathrm{An}_m$ if $m \notin \mathcal{A}_0$. Then we apply the similar treatment to $\mathcal{C}_1 = \mathcal{V}\backslash\mathcal{A}_0$ to identify $\mathcal{A}_1$, and proceed to identify other layers sequentially until all the nodes are assigned. Interestingly, Lemma 1 ensures that the layers can be reconstructed in a top-down fashion, whereas Theorem 1 of Gnecco et al. (2021) shows the causal ordering can be recovered by searching each root node greedily in the current subgraph with $\mathbf{\Gamma}$. It is important to remark that Lemma 1 holds true without assuming the popularly-employed faithfulness condition in literature (Spirtes et al., 2000).

Generally, suppose that $\mathcal{A}_0, \ldots, \mathcal{A}_t$ are identified. For node $m \in \mathcal{A}_t$, we have $\mathrm{pa}_m \subset \mathcal{S}_t = \cup_{d=0}^{t-1}\mathcal{A}_d$ and $\mathrm{de}_m \cap \mathcal{S}_t = \emptyset$. Further, the causal minimality holds if and only if $X_m \not\perp\!\!\!\perp X_j|\mathbf{X}_{\mathrm{pa}_m\backslash\{j\}}$ for $j \in \mathrm{pa}_m$ (Peters et al., 2017), yielding that $X_m \not\perp\!\!\!\perp X_j|\mathbf{X}_{\mathcal{S}_t\backslash\{j\}}$. Thus, pa_m is the set of nodes with conditional dependence. It is interesting to notice that in Section 4, the minimal signal strength for the measure to test the conditional dependence is required in Assumption 5 to establish the asymptotic consistency under the finite sample setting.

Theorem 1. *Suppose that all the assumptions in Lemma 1 are satisfied and the causal minimality holds. Then, the heavy-tailed DAG $\mathcal{G}$ is uniquely identifiable.*

Theorem 1 establishes the identifiability of the heavy-tailed DAG under the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

linear SCM model (2.1) and (2.2), regardless of continuous or discrete distributions. The proof of Theorem 1 directly follows from Lemma 1 that all the topological layers can be exactly recovered by comparing $\mathbf{\Gamma}$, and from the causal minimality assumption that the underlying directed edges can be exactly reconstructed by testing the conditional dependence if the true layers are given. Therefore, the details are omitted here. Note that we are the first to establish identifiability results for the heavy-tailed DAG, but only the causal ordering is identified in Gnecco et al. (2021). To the best of our knowledge, since no off-the-shelf regularized regression techniques can be applied to determine parent-child relationships for heavy-tailed data in our setting (2.1), we refine a conditional independence testing (CIT) measure to recover the exact DAG structures.

3.2 TopHeat

We now develop a two-step efficient algorithm to learn a heavy-tailed DAG. The first step is to recover the topological layers in a top-down fashion, motivated by Lemma 1, and then the directed edges can be reconstructed by applying a CIT method among layers for the heavy-tailed data in a parallel fashion.

Given a random sample $\mathbf{X}^n = (\mathbf{X}_i^n)_{i=1}^n$ with $\mathbf{X}_i^n = (X_{i,1}^n, ..., X_{i,p}^n)^T$, we first estimate the causal tail dependence as

$$\widehat{\Gamma}_{jm} = \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_m(X_{i,m}^n) \mathbf{1}\{X_{i,j}^n > X_{(n-k),j}^n\}, \quad j \neq m, \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

where $\widehat{F}_m(X^n_{i,m}) = n^{-1}\sum_{i=1}^n \mathbf{1}\{X^n_{i,m} < x\}$ with $x > 0$, $\mathbf{1}\{\cdot\}$ is an indicator function, and $X_{(n-k),j}$ is the $(n-k)$-th order statistic of X_j, satisfying $X^n_{(1),j} \leq X^n_{(2),j} \leq X^n_{(n-k),j} \leq X^n_{(n),j}$ with the integer $0 < k \leq n-1$.

With $\widehat{\Gamma}_{jm}$, it is assured by Lemma 1 that $\widehat{\mathcal{A}}_0$ can be estimated as $\widehat{\mathcal{A}}_0 = \{m \in \mathcal{C}_0 : \min_{j\in\mathcal{C}_0} |\widehat{\Gamma}_{jm} - 1| > \epsilon_0\}$, where ϵ_0 is a small positive constant. Note that we expect $\widehat{\Gamma}_{jm}$ is strictly smaller than 1 with a tolerance ϵ_0 for all $j \in \mathcal{C}_0$, if m is estimated as a root node such that $m \in \widehat{\mathcal{A}}_0$. Therefore, all the root nodes located in $\widehat{\mathcal{A}}_0$ should keep the distance from 1 at least ϵ_0. Suppose that the topological layers $\widehat{\mathcal{A}}_0, \ldots, \widehat{\mathcal{A}}_{t-1}$ have been estimated and $\widehat{\mathcal{C}}_t = \mathcal{V}\backslash\widehat{\mathcal{S}}_t$ with $\widehat{\mathcal{S}}_t = \cup_{d=0}^{t-1}\widehat{\mathcal{A}}_d$, we next estimate the topological layer $\widehat{\mathcal{A}}_t$ in a similar manner. That is, it follows from Lemma 1 that $\widehat{\mathcal{A}}_t = \{m \in \widehat{\mathcal{C}}_t : \min_{j\in\widehat{\mathcal{C}}_t} |\widehat{\Gamma}_{jm} - 1| > \epsilon_t\}$, where ϵ_t is a small positive constant. We repeat these procedures until $\widehat{\mathcal{C}}_t = \emptyset$.

After $\widehat{\mathcal{A}}_t$'s are reconstructed, the task of DAGs learning boils down to estimation of the skeletons (Shojaie and Michailidis, 2010), as directed edges can only point from upper layers to lower layers and no edges are allowed within the same layer. One direct way is to apply the regression-based methods. However, to the best of our knowledge, no suitable regression methods can be used here, since Huber loss based methods require the finite moment condition for the error (Fan et al., 2017; Sun et al., 2020), which is difficult to satisfy for heavy-tailed distributions we consider in this paper. Therefore, we turn to perform a CIT

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

procedure to estimate the parents of each node $m \in \mathcal{A}_t$ from $\mathcal{S}_t$,

$$H_{m,j,0} : X_m \perp\!\!\!\perp X_j | \mathbf{X}_{\mathcal{S}_t \setminus \{j\}} \quad \text{v.s.} \quad H_{m,j,1} : X_m \not\perp\!\!\!\perp X_j | \mathbf{X}_{\mathcal{S}_t \setminus \{j\}}, \tag{3.3}$$

for each $j \in \mathcal{S}_t$. There is a vast and rapidly growing literature on CIT. Existing methods fall roughly into four main categories. The metric-based tests (Wang et al., 2015) may suffer from the curse of dimensionality as kernel smoothers are involved, and thus the kernel-based tests (Zhang et al., 2011) also have inflated type-I errors. Instead, the conditional randomization-based tests (Candes et al., 2018) require that the conditional distribution $X_m | \mathbf{X}_{\mathcal{S}_t \setminus \{j\}}$ is known as a prior. If unknown, the type-I error rates largely depend on the approximation of the conditional distribution. Regression-based tests (Shah and Peters, 2020) may not have sufficient power to detect the alternative hypothesis. Interestingly, Azadkia and Chatterjee (2021) proposed a novel and rather different conditional dependence measure,

$$Q_{m,j,t} = \frac{\int E[Var\{P(X_m \geq t | \mathbf{X}_{\mathcal{S}_t}) | \mathbf{X}_{\mathcal{S}_t \setminus \{j\}}\}] dF_m(t)}{\int E\{Var(I_{\{X_m \geq t\}} | \mathbf{X}_{\mathcal{S}_t})\} dF_m(t)}, \tag{3.4}$$

where $I(\cdot)$ is an indicator function. Out of its simplicity, computational efficiency, and asymptotically properties, we employ this measure to perform a CIT in (3.3). Note that the null and alternative hypotheses in (3.3) correspond to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

$Q_{m,j,t} = 0$ and $Q_{m,j,t} \neq 0$, respectively (Azadkia and Chatterjee, 2021).

The sample CIT measure is denoted as $\widehat{Q}_{m,j,t}$ and its asymptotic normality is also established in Theorem 3.1 of Shi et al. (2024) with $1/\sqrt{n}$ convergence rate and asymptotic variance σ^2. For testing whether $\widehat{Q}_{m,j,t}$ is zero or not, we apply the t-type test statistic $\sqrt{n}\widehat{Q}_{m,j,t}/\widehat{\sigma}^2$, where the details of the estimator $\widehat{\sigma}^2$ is given in the Section S6 of the supplementary material. The equivalent null hypothesis $H_{m,j,0} : Q_{m,j,t} = 0$ is rejected against the two-sided alternative $H_{m,j,1} : Q_{m,j,t} \neq 0$ if $\sqrt{n}\widehat{Q}_{m,j,t}/\widehat{\sigma}^2 > \Phi^{-1}(1 - \alpha/2)$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution and α is the significance level.

It is important to remark that the CIT measure can be adopted for heavy-tailed distributions, which is shown in Section S6. Also, the CIT measure can not adapt to the unconditional independence testing. To circumvent this difficulty, a random error ε is generated and included as the conditional variable, so that an unconditional independence testing is transformed into a CIT. Particularly, when $|\widehat{\mathcal{S}}_t| = 1$, we generate a heavy-tailed distributed error ϵ to perform a CIT.

The proposed two-step learning algorithm for Topological layers based Heavy-Tailed DAGs is summarized in Algorithm 1, denoted as the TopHeat algorithm.

In terms of the computational complexity, the input $\widehat{\mathbf{\Gamma}}$ in Algorithm 1 involves the ranks of observations and the calculations of pairwise causal tail coefficients, which have the complexity of $O(pn \log n)$ and $O(kp^2)$, respec-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

Algorithm 1: The TopHeat algorithm

Input: $\mathbf{X}^n \in \mathbb{R}^{n\times p}$, $\widehat{\mathbf{\Gamma}} \in \mathbb{R}^{p\times p}$, $t = 0$, $\widehat{\mathcal{C}} = \{1, 2, \ldots, p\}$, and $\widehat{\mathcal{S}}_0 = \emptyset$;
Output: $\{\widehat{\mathcal{A}}_t\}_{t=0}^{\widehat{T}-1}$ and $\widehat{\mathcal{E}} = \cup_{m\in\{\mathcal{A}_1,\ldots,\mathcal{A}_{\widehat{T}-1}\}} \cup_{j\in\widehat{\text{pa}}_m} (j, m)$.

```
while Ĉ ≠ ∅ do
    Estimate Â_t = {m ∈ Ĉ : min_{j∈Ĉ} |Γ̂_jm − 1| > ε_t};
    Update Ĉ ← Ĉ\Â_t, Ŝ_{t+1} ← Ŝ_t ∪ Â_t, and t ← t + 1;
end
Denote T̂ = t;
for t = 1, 2, ..., T̂ − 1 do
    for m ∈ Â_t and j ∈ Ŝ_t do
        if √n Q̂_{m,j,t}/σ̂² > Φ^{-1}(1 − α/2) then
            Denote j ∈ p̂a_m and (j, m) ⊆ Ê;
        end
    end
end
```

tively (Gnecco et al., 2021). The computational complexity of Step 1 in the TopHeat algorithm to reconstruct the topological layers is of order $O(\sum_{t=0}^{T-1}(p - |\mathcal{S}_t|)) = O(p)$. In the second step of TopHeat, the complexity is of order $O(\sum_{t=1}^{T-1}(n^2|\mathcal{S}_t| + n\log n)|\mathcal{S}_t||\mathcal{A}_t|) = O(n^2\sum_{t=1}^{T-1}|\mathcal{S}_t|^2|\mathcal{A}_t| + dn\log n)$, where the complexity of the Euclidean distance is of order $O(n^2|\mathcal{S}_t|)$ and the rank of observations is of order $O(n\log n)$ for the k-nearest neighbor in each CIT, and we denote $d = \sum_{t=1}^{T-1}|\mathcal{A}_t||\mathcal{S}_t|$. Therefore, the overall computational complexity of estimating $\widehat{\mathbf{\Gamma}}$ and running the TopHeat algorithm is $O(\max(kp^2, n^2\sum_{t=1}^{T-1}|\mathcal{S}_t|^2|\mathcal{A}_t|))$. For a chain graph with $T = p$, the complexity becomes $O(n^2p^3)$. When $T = 2$ in a shallow hub graph, the complexity of TopHeat is $O(\max(kp^2, n^2p))$. In con-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

trast, EASE needs $O(p^2)$ to compute the order and $O(\sum_{t=1}^{T-1}(n^2|\mathcal{S}_t|+n\log n)p^2)$ in the CIT procedure, and thus the total complexity of EASE is $O(n^2p^3)$ in practice. It is clear that EASE is more expensive than TopHeat, if T is much smaller than p, and their computational cost becomes the same when $T = p$. It is also interesting to note that the computational complexity of TL is $O(\sum_{t=0}^{T-1}(R|\mathcal{C}_t|^3 + n\log n|\mathcal{C}_t|^2))$ where R denotes the number of coordinate descent cycles until convergence. When the DAG is relatively sparse, R is considered as a constant and the complexity is $O(p^4+np^3\log n)$ in the worst case with $T = p$. For $T = 2$, the complexity of TL becomes $O(p^3 + np^2\log n)$. Since the complexity of TopHeat depends on $k, |\mathcal{S}_t|, |\mathcal{A}_t|$ and the complexity of TL relies on $|\mathcal{C}_t|$, they cannot be directly compared and their relationship also depends on n and p. However, it is expected that if each node in a DAG has less ancestors, TopHeat is more efficient than TL. Runtime comparisons of these algorithms are provided in Section 5.2.

Note that the numerical performance of TopHeat largely depends on the choice of the hyperparameters, including k, α, and ϵ_t. More details are provided in Section S2 of the supplementary material.

4. Theoretical Guarantees

In this section, the asymptotic theory of the proposed method is investigated. We first give notations below. We define the true and estimated topological layers

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

as $\mathcal{L} = \{\mathcal{A}_0, ..., \mathcal{A}_{T-1}\}$ and $\widehat{\mathcal{L}} = \{\widehat{\mathcal{A}}_0, ..., \widehat{\mathcal{A}}_{\widehat{T}-1}\}$, respectively, and $\widehat{\mathcal{G}} = (\mathcal{V}, \widehat{\mathcal{E}})$ as the estimated DAG. The right-hand upper tail dependence between two random variables X_j and X_m is introduced, denoted as $\lim_{t\to\infty} tP(1 - F_j(X_j) \leq x/t, 1 - F_m(X_m) \leq y/t) = R(x, y)$ for $(x, y) \in [0, \infty]^2 \backslash \{(\infty, \infty)\}$ and $j, m \in \{1, \ldots, p\}$. We define a Gaussian process W_R on $[0, \infty]^2 \backslash \{\infty, \infty\}$ with mean 0 and covariance structure $\mathrm{E}\{W_R(x_1, y_1)W_R(x_2, y_2)\} = R(x_1 \wedge x_2, y_1 \wedge y_2)$, and thus W_R is a Wiener process. We denote the tail function by $U_j = (1/(1-F_j))^{\leftarrow}$, where the left-continuous inverse of a non-decreasing function f is defined as $f^{\leftarrow}(x) = \inf\{y \in \mathbb{R} : f(y) \geq x\}$. Note that the heavy-tailed assumption in (2.2) indicates that $\lim_{t\to\infty}\{1 - F_j(tx)\}/\{1 - F_j(t)\} = x^{-\theta}$, equivalent to $\lim_{t\to\infty} U_j(tx)/U_j(x) = x^{1/\theta}$. The probability density function of X_j is written as f_j. For two positive random sequences a_n and b_n, we denote $a_n = \Omega(b_n)$ if $a_n \geq cb_n$ for sufficiently large n. Let k be an intermediate sequence of integers such that $k/n \to 0$ holds as $k, n \to \infty$. The following technical conditions are required to establish the layer recovery consistency of the proposed algorithm.

Assumption 1. *There exist* $\tau_1, \tau_2 < 0$ *and* $\tau_3 < -1$ *such that as* $t \to \infty$,

$$\sup_{0<x<\infty, 1/2\leq y\leq 2} |tP\{1 - F_j(X_j) \leq x/t, 1 - F_m(X_m) \leq y/t\}/R(x, y) - 1| = O\left(t^{\tau_1}\right), \tag{4.1}$$

$$\sup_{0<x<\infty} |g_t(x) - \theta x^{1+1/\theta}| = O\left(t^{\tau_2}\right), \tag{4.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

$$|E\{F_m(X_m)|F_j(X_j) > 1 - 1/t\} - \Gamma_{jm}| = O\left(t^{(\tau_3-1)/2}\right), \tag{4.3}$$

with $g_t(x_j) = tU_j(t)f_j\big(U_j(t)x_j^{-1/\theta}\big)$ *for* $x_j > 0$.

Assumption 2. *There exist* $\rho < 0$ *and a function* A_1 *such that as* $t \to \infty$, $A_1(tx)/A_1(t) \to x^\rho$ *for all* $x > 0$ *and* $\sup_{x>1}\left|x^{-1/\theta}\frac{U_j(tx)}{U_j(x)} - 1\right| = O(A_1(t))$.

Assumption 3. *As* $n \to \infty$, $k = O(n^\gamma)$ *for some* γ *satisfying* $0 < \gamma < \min\left\{\frac{2\tau_1}{2\tau_1-1}, \frac{2\tau_2}{2\tau_2-1}, \frac{2\rho}{2\rho+\theta(\rho-1)}\right\}$.

To derive the convergence rate by controlling the estimation bias in Theorem 2, Assumption 1 provides some technical conditions. Specifically, (4.1) is the second-order strengthening of the upper tail dependence $\lim_{t\to\infty} tP\{1 - F_j(X_j) \le x/t, 1 - F_m(X_m) \le y/t\} = R(x, y)$, similar to condition (7.2.8) in De Haan and Ferreira (2006); (4.2) is the second-order strengthening of the density convergence result $ds_n(x)/dx \to 1$, equivalent to $\lim_{n\to\infty} g_{\frac{n}{k}}(x) \to \theta x^{1+1/\theta}$, since $\lim_{t\to\infty} U_m(tx)/U_m(t) = x^{1/\theta}$ implies $s_n(x) := (n/k)[1-F_m\{U_m(n/k)x^{-1/\theta}\}] \to x$ as $n \to \infty$ for $x > 0$; (4.3) also imposes the second-order strengthening of $\Gamma_{jm} = \lim_{u\to 1^-} E\{F_m(X_m)|F_j(X_j) > u\}$ in (3.1). Assumption 2 is a second-order condition for U_j and implied by Theorem B.2.2 in De Haan and Ferreira (2006). Assumption 3 imposes conditions on the upper bound of γ, which is a typical constraint in the extreme value theory literature to guarantee that the first $k+1$ largest observations for estimation are actually in the tail (Cai et al., 2015).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

Theorem 2. *Assume that Assumptions 1–3 hold and* $\theta > 2$. *We have* $\frac{n}{\sqrt{k}}\big(\widehat{\Gamma}_{jm} - \Gamma_{jm}\big) \xrightarrow{\mathrm{d}} \Theta$, *where* $\Theta = (1/\theta - 1)W_R(\infty, 1)\int_0^\infty R(s,1)ds - \int_0^\infty W_R(s,1)ds^{-1/\theta}$.

With the help of tail empirical process and extreme value theory, Theorem 2 establishes the asymptotic normality for $\widehat{\Gamma}_{jm}$, extending the convergence result for fixed p in Gnecco et al. (2021), which helps to derive the consistency of the topological layers estimation in Theorem 3. The assumption $\theta > 2$ in Theorem 2 is the tail rate condition for the error term and also commonly used in finance (Daouia et al., 2018). The main challenge of Theorem 2 comes from the tail dependence between random variables X_j and X_m, which is introduced by (3.2).

Theorem 3. (Layer recovery consistency) *Suppose that Assumptions 1–3 hold and* $\theta > 2$, *and for all* ϵ_t, *we take* $\epsilon_t = \frac{\eta_{min}}{2}$ *with* $\eta_{min} \leq 1 - \min_{j \in \mathcal{C}_t, m \in \mathcal{A}_t, j \notin an_m} \Gamma_{jm}$. *Then, there holds that for some constant* $C_0 > 0$,

$$P(\widehat{\mathcal{L}} = \mathcal{L}) \geq 1 - C_0 T p^2 \sqrt{k}/n = 1 - C_0 T p^2 / n^{1-\gamma/2}. \tag{4.4}$$

Theorem 3 shows that the topological layers of a heavy-tailed DAG can be exactly reconstructed with high probability. Note that the layer consistency result depends on the threshold ϵ_t and its upper bound could improve the precision of layer recovery, as many spurious nodes may be included in the current layer if ϵ_t is too large. However, small ϵ_t increases the computational cost in TopHeat. It is worth pointing out that the consistency result also holds if we take

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

$\epsilon_t \in (c_1 \frac{Tp^2\sqrt{k}}{n}, \frac{\eta_{\min}}{2}]$ for some positive constant c_1. Therefore, we take a stability procedure to choose ϵ_t adaptively for each layer $\mathcal{A}_t$ and verify that Assumptions 1 and 2 in Sun et al. (2013) are satisfied with their λ replaced with $\lambda = 1/\epsilon_t$; More details are provided in Section S2 of the supplementary material.

Next, we assume mild conditions about the CIT method to reconstruct directed edges and derive the graph consistency result.

Assumption 4. *There are nonnegative real numbers C_1 and C_2 such that for any $t \in \mathbb{R}$, and $\mathbf{x}_{\mathcal{S}_t}, \mathbf{x}'_{\mathcal{S}_t} \in \mathbb{R}^{|\mathcal{S}_t|}$,*

$$\begin{aligned}
&\left|P(X_m \geq t|\mathbf{X}_{\mathcal{S}_t} = \mathbf{x}_{\mathcal{S}_t}) - P(X_m \geq t|\mathbf{X}_{\mathcal{S}_t} = \mathbf{x}'_{\mathcal{S}_t})\right| \\
\leq & C_1\big(1 + \|[\mathbf{x}_{\mathcal{S}_t}]_j\|^{C_2} + \left\|[\mathbf{x}'_{\mathcal{S}_t}]_j\right\|^{C_2} + \|[\mathbf{x}_{\mathcal{S}_t}]_{-j}\|^{C_2} + \left\|[\mathbf{x}'_{\mathcal{S}_t}]_{-j}\right\|^{C_2}\big) \\
& \times \left(\left\|[\mathbf{x}_{\mathcal{S}_t}]_j - [\mathbf{x}'_{\mathcal{S}_t}]_j\right\| + \left\|[\mathbf{x}_{\mathcal{S}_t}]_{-j} - [\mathbf{x}'_{\mathcal{S}_t}]_{-j}\right\|\right),
\end{aligned}$$

where $[\mathbf{x}_{\mathcal{S}_t}]_j$ is denoted as the element of $\mathbf{x}_{\mathcal{S}_t}$ corresponding to node j and $[\mathbf{x}_{\mathcal{S}_t}]_{-j} = \mathbf{x}_{\mathcal{S}_t \setminus \{[\mathbf{x}_{\mathcal{S}_t}]_j\}}$.

Assumption 5. *For any $t = 1, \ldots, T-1$, there holds that $\inf_{m \in \mathcal{A}_t, j \in \mathcal{S}_t}\{|Q_{m,j,t}| : Q_{m,j,t} \neq 0\} \geq \phi_n$ where $\phi_n = O(n^{-c})$ for some $\frac{1-\xi}{2} < c < \frac{1}{2}$ with $\xi \in (0, \min\{\frac{1}{2}, \frac{2}{|\mathcal{S}_t|}\}]$.*

Assumption 4 is exactly condition (A1) in Azadkia and Chatterjee (2021) and a locally Lipschitz condition of the conditional distribution of X_m given its

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

upper layers with a polynomial rate. Assumption 5 gives a stronger condition for the lower bound for the non-zero CIT measure in the finite sample setting. Similar conditions as Assumption 5 are imposed in Kalisch and Bühlmann (2007).

Proposition 1. *Suppose that the heavy-tailed distribution assumption* (2.2) *and Assumption 4 hold. Then, for any* $\eta > 0$ *and* $t = 1, \ldots, T-1$*, there exist positive constants* $C_3, C_4 > 0$ *such that* $\sup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} P\big(|\widehat{Q}_{m,j,t} - Q_{m,j,t}| > \eta\big) \leq C_3 \exp(-C_4 n \eta^2)$.

Proposition 1 establishes the tail bound for the sample CIT measure for heavy-tailed distributions, which replaces the sub-exponential decaying rate condition in Azadkia and Chatterjee (2021).

Theorem 4. (Graph recovery consistency) *Suppose that Assumptions 1–5 are satisfied and* $\theta > 2$*. If* $n = \Omega\left(p^{4/(2-\gamma)}\right)$*, we have*

$$P(\widehat{\mathcal{G}} = \mathcal{G}) \to 1, \text{ as } n \to \infty.$$

Theorem 4 guarantees that TopHeat consistently recovers the exact DAG structure while allowing p to diverge with n at a certain rate, which is in sharp contrast to the literature only recovering causal orderings for a heavy-tailed DAG for fixed p (Gnecco et al., 2021). To investigate the relationship between n and the tail index θ, TopHeat has the sample complexity $n = \Omega\big(p^{\frac{4(2\rho+\theta(\rho-1))}{2\rho+2\theta(\rho-1)}}\big)$ where $\rho < 0$ and $\frac{4(2\rho+\theta(\rho-1))}{2\rho+2\theta(\rho-1)} > 1$ holds. However, Corollary 4 in Zhao et al. (2022)

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

shows that TL has the sample complexity $n = \Omega\big(p^{\max\{\frac{4}{m-\tau+4}, \frac{2m}{(2m\phi-1)(m-\tau+4)}\}}\big)$ with $m + 4 > \tau > 4$ and $\phi > \frac{1}{2m}$ for $4m$-th bounded moment distributions. It is worthy to note that their sample complexities depend on the relationship of θ and m. If θ becomes smaller and thus m may be also smaller, the order of p in TopHeat can be much smaller than that in TL. If m becomes larger and thus θ may be also larger, the order of p in TL can be much smaller than that in TopHeat. Also, Theorem 4 requires no graph structure restrictions by controlling the number of parents or the Markov blankets, which are needed in non-Gaussian DAG literature (Wang and Drton, 2020; Zhao et al., 2022). It is worthy to note that our work is the first to provide a solid theoretical guarantee for learning a heavy-tailed DAG in terms of the exact DAG recovery.

5. Simulation Studies

In this section, we demonstrate the performance of TopHeat and compare it against some common baselines: EASE (Gnecco et al., 2021), TL (Zhao et al., 2022), the ICA-LiNGAM algorithm (Shimizu et al., 2006), the Direct LiNGAM algorithm (Shimizu et al., 2011), the high-dimensional LiNGAM (HD-LiNGAM; Wang and Drton, 2020) and the Rank PC algorithm (Harris and Drton, 2013). Particularly, EASE and Rank PC are implemented in the R package "causalXtreme" (Gnecco et al., 2021). Since EASE only returns a causal order, we follow

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

the CIT procedure in TopHeat to output a complete DAG for fair comparison. Further, Rank PC returns a completed partially DAG, which is transformed into a DAG by applying the function pdag2dag in the R package "pcalg" (Kalisch et al., 2012). Note that TL, Direct-LiNGAM, and ICA-LiNGAM are implemented in the R packages "TransGraph", "rlingam", and "highDLingam", respectively.

To evaluate the performance of these methods, we adopt commonly-used measures, including the normalized Hamming distance (HM), Recall, Precision, and F1-score to evaluate the accuracy of estimating a DAG. Note that HM measures the number of adding, removing, and reversing directed edges to make the estimated DAG into the true one. Therefore, a smaller HM value indicates better accuracy in graph estimation. The remaining three measures assess the accuracy of estimated directed edges, with higher values indicating better performance.

5.1 Simulated examples

In the following numerical studies, we consider three generating schemes for graphs, including a hub graph in Example 1 and two random graphs in Examples 2–3, generated from the Barabási-Albert (BA) model (Barabási and Albert, 1999) and the Erdös-Rényi (ER) model Erdös and Rényi (1960), respectively.

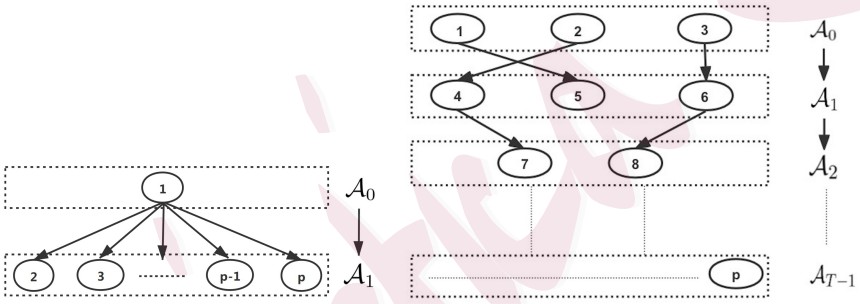
Example 1. A hub graph with $T = 2$ is considered with $\mathcal{A}_0 = \{1\}$ and $\mathcal{A}_1 = \{2, 3, \ldots, p\}$, and node 1 directs to all other nodes, shown in Figure 1(a). Note

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

that these highly connected hub nodes are of great interest in social networks.

Example 2. The BA graph is a scale-free network model where nodes are preferentially attached to existing ones with higher degrees. We generate a BA graph where one directed edge is added for each node, and the corresponding DAG is illustrated in Figure 1(b).

Example 3. The ER graph is a random graph model where edges are connected with probability p_c. We consider a sparse ER graph with $p_c = 1/(p-1)$, and thus the mean of the number of neighbors is 1 for each node.



(a) The hub graph (b) The random graph

Figure 1: The illustration for the topological layers of the DAG structure in Examples 1–2.

For each example, we generate p independent errors from a Cauchy distribution with the location parameter of 3 and the scale parameter of 3, and from a Student-t distribution with 1 degree of freedom, respectively. We consider a linear SCM and a simple nonlinear model, replacing X_i with the empirical cdf

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

$\widehat{F}_i(X_i)$ in (2.1) by the transformation in Gnecco et al. (2021). The coefficients for each directed edge are sampled from a uniform distribution $U[0.3, 0.7]$.

5.2 Results

To select an optimal value of k for estimating $\mathbf{\Gamma}$, we choose $k = \lfloor n^{\gamma} \rfloor$ with different choices of $\gamma \in \{0.2, 0.25, \cdots, 0.7\}$ and evaluate the performance of TopHeat in Figure S1. In practice, we take $k = \lfloor n^{0.5} \rfloor$ since it is located within the best range of γ under different settings. More discussion are provided in Section S11 of the supplementary material.

To choose the tuning parameter ϵ_t from the grids $\{10^{-2+0.05s}, s = 0, 1, \ldots, 35\}$, we apply the stability selection method in Section S2. The performance is given in Figure S2, which indicates that we set $(a, \mathrm{B}) = (10^{-1}, 5)$ for $p \in \{5, 20\}$, and $(a, \mathrm{B}) = (10^{-1.5}, 25)$ for $p = 50$ in subsequent experiments.

During the CIT procedure of TopHeat, the conditional variable is considered to follow a standard normal distribution when $|\widehat{\mathcal{S}}_t| = 1$. This is verified by preliminary experiments depicted in Figure S3, which suggests that the estimation accuracy of TopHeat is not significantly affected by different choices of the distribution. In order to control the graph-wise false discoveries, the significance level α, should be smaller and tend towards zero as p and n approach infinity. Therefore, we set $(\alpha, p) \in \{(10^{-2}, 5), (10^{-5}, 20), (10^{-10}, 50)\}$ here. More details are referred to Section S11 the supplementary material.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

In the sequel, we conduct the experiments for 50 repetitions under the settings with $(n, p) \in \{(500, 5), (2000, 20), (5000, 50)\}$. The averaged performance metrics of all methods, along with their standard errors, are reported. Here, Table 1 displays the results of the simulations for a hub graph in Example 1 with the Student-t distribution. Additional results are provided in Tables S3–S7 in Section S11 of the supplementary material.

It is evident that TopHeat demonstrates superior numerical performance and outperforms other competitors across almost all metrics with hub graphs in Example 1 and BA graphs in Example 2, except that the Recall of TopHeat is a little lower than LiNGAM-based methods for small graphs in the linear case. However, Directed-LiNGAM, ICA-LiNGAM, and HD-LiNGAM achieve much lower Precision and F1-score, since they estimate many false edges in dense graphs. In Example 3 for ER graphs, TopHeat exhibits comparable performance with EASE with smaller (n, p) and yields better performance than other methods as n and p increase, which is also supported by the theoretical consistency result in Section 4. Note that TL achieves much lower Recall and F1-score compared with TopHeat, even though higher Precision for hub and BA graphs in a linear SCM, and completely fails in the nonlinear setting, since its violation of the required data distribution conditions for the precision matrix to produce false layers and edges. In conclusion, TopHeat keeps its superiority across var-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

Table 1: The averaged performance metrics of various methods, as well as their standard errors in parentheses, are presented for a hub graph in Example 1 with the Student-t distribution.

Model	(n, p)	Methods	HM (%)	Recall	Precision	F1-score
linear	$(500, 5)$	TopHeat	3.30(0.65)	0.88(0.02)	0.95(0.02)	0.91(0.02)
		EASE	12.00(0.77)	0.54(0.02)	0.83(0.03)	0.63(0.02)
		TL	6.20(1.14)	0.69(0.06)	0.78(0.06)	0.73(0.06)
		Directed-LiNGAM	3.30(0.66)	0.98(0.01)	0.89(0.02)	0.93(0.01)
		ICA-LiNGAM	3.80(0.66)	0.98(0.01)	0.87(0.02)	0.92(0.01)
		HD-LiNGAM	30.00(0.00)	1.00(0.00)	0.40(0.00)	0.57(0.00)
		Rank PC	31.70(0.55)	0.28(0.01)	0.25(0.01)	0.26(0.01)
	$(2000, 20)$	TopHeat	1.30(0.21)	0.75(0.04)	0.97(0.01)	0.82(0.03)
		EASE	4.52(0.03)	0.10(0.00)	0.96(0.02)	0.18(0.01)
		TL	1.85(0.31)	0.64(0.06)	0.70(0.07)	0.67(0.06)
		Directed-LiNGAM	7.72(0.43)	1.00(0.00)	0.42(0.01)	0.58(0.01)
		ICA-LiNGAM	8.54(0.45)	0.99(0.00)	0.39(0.01)	0.55(0.01)
		HD-LiNGAM	45.00(0.00)	1.00(0.00)	0.10(0.00)	0.18(0.00)
		Rank PC	7.41(0.04)	0.05(0.00)	0.09(0.00)	0.07(0.00)
	$(5000, 50)$	TopHeat	0.68(0.15)	0.91(0.03)	0.83(0.04)	0.86(0.03)
		EASE	2.07(0.01)	0.06(0.00)	0.41(0.02)	0.10(0.00)
		TL	0.42(0.09)	0.80(0.05)	0.85(0.05)	0.82(0.05)
		Directed-LiNGAM	8.66(0.39)	1.00(0.00)	0.20(0.01)	0.33(0.01)
		ICA-LiNGAM	9.92(0.36)	1.00(0.00)	0.18(0.01)	0.30(0.01)
		HD-LiNGAM	48.00(0.00)	1.00(0.00)	0.04(0.00)	0.08(0.00)
nonlinear	$(500, 5)$	TopHeat	4.30(0.77)	0.82(0.03)	0.95(0.02)	0.88(0.02)
		EASE	12.00(0.65)	0.55(0.03)	0.82(0.03)	0.64(0.02)
		TL	20.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	54.30(2.19)	0.13(0.03)	0.09(0.02)	0.11(0.02)
		ICA-LiNGAM	20.20(3.07)	0.64(0.05)	0.60(0.06)	0.62(0.05)
		HD-LiNGAM	25.00(0.00)	0.25(0.00)	0.33(0.00)	0.29(0.00)
		Rank PC	31.70(0.55)	0.28(0.01)	0.25(0.01)	0.26(0.01)
	$(2000, 20)$	TopHeat	1.93(0.41)	0.85(0.03)	0.83(0.04)	0.83(0.03)
		EASE	4.80(0.06)	0.15(0.01)	0.62(0.03)	0.23(0.01)
		TL	5.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	23.44(0.81)	0.38(0.02)	0.09(0.01)	0.15(0.01)
		ICA-LiNGAM	33.56(1.17)	0.17(0.03)	0.04(0.01)	0.06(0.01)
		HD-LiNGAM	9.21(0.00)	0.05(0.00)	0.06(0.00)	0.05(0.00)
		Rank PC	9.69(0.15)	0.07(0.01)	0.07(0.01)	0.07(0.01)
	$(5000, 50)$	TopHeat	0.66(0.15)	0.91(0.03)	0.83(0.04)	0.86(0.03)
		EASE	2.04(0.01)	0.06(0.00)	0.45(0.02)	0.11(0.00)
		TL	2.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	21.69(0.24)	0.51(0.01)	0.05(0.00)	0.09(0.00)
		ICA-LiNGAM	24.88(0.82)	0.32(0.05)	0.03(0.01)	0.06(0.01)
		HD-LiNGAM	3.88(0.00)	0.02(0.00)	0.02(0.00)	0.02(0.00)

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

ious sample and node sizes, graph types, and (non)linear model settings, which makes it a good choice for learning the DAG structure in heavy-tailed data.

To investigate the exact recovery rates of TopHeat, we consider the same data generating scheme as Example 1. Here, we set $\mathrm{B} = 25$ and $\gamma = \frac{1}{2}$, replicate the experiments for 100 times, and fix $p = 10$ with $n/p^{8/3} \in \{1, 2, 3, 5, 10, 15, 20, 25, 30, 35\}$. Figure S5 displays the exact recovery rates, which grow with the ratio $n/p^{4/(2-\gamma)}$ and converge to 1. This validates the theoretical consistency result for the proposed TopHeat algorithm in Theorem 4.

In terms of the computational comparison, we only compare the average running time of the proposed TopHeat algorithm with EASE and TL by repeating 50 times, while considering the same data generating mechanisms as Examples 1–3. Figure S6 reports that TopHeat is much more efficient than EASE in terms of computational cost, where all the tuning parameters chosen procedures are included. Overall, TopHeat costs half the time that EASE takes. Besides, the average running time of TopHeat is less than TL when n is relatively small, which is suggested by the complexity analysis in Section 3.2. Further, TopHeat is also more efficient than TL as $|\mathcal{S}_t|$ may tend to be smaller, when p become smaller or the number of ancestors decreases from random graphs to hub graphs, which also echoes the computational complexity analysis.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

6. Financial data analysis

In this section, we apply TopHeat to analyze the financial contagion among 17 currencies, and investigate the effect of the euro's introduction from 1999 in the financial market, since the euro has been the second most widely held international reserve currency after the U.S. dollar. This helps to find a currency as a good option for risk diversification and thus reduce the systemic risk. DAGs can reveal the financial contagion effect encoded with the causal relationships among currencies, with directed edges from one currency to another.

The exchange rates data for the empirical analysis are available in supplementary material of Chen and Schienle (2022). The bilateral exchange rate $X_{i,j}^n$ is recorded as the exchange rate of country j against 1 U. S. dollar in the end of i-th each quarter. We consider the period from the first quarter of 1973 to that of 2008 with $n = 141$ quarters in total, and $p = 17$ OECD (Organization for Economic Co-operation and Development) countries, including Australia, Canada, Denmark, Great Britain, Japan, Korea, Norway, Sweden, Switzerland, Austria, Belgium, France, Germany, Spain, Italy, Finland, and the Netherlands. Since the U.S. brought the Bretton Woods system to an end in 1971, then IMF members were free to choose any forms of exchange arrangements they wish, and the financial crisis broke out in 2008, the exchange rate during 1973 to 2007 relatively floated and the causal relations actually existed.

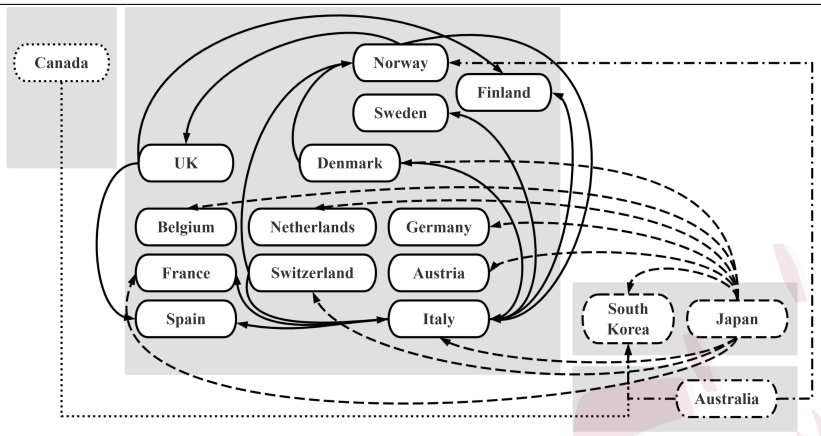
Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

It is commonly assumed the distribution of the exchange rates follows from the log-normal distribution in empirical studies, which satisfies the heavy-tailed distributions assumption in (2.2). Figure S7 displays the histograms of exchange rates for 17 currencies, illustrate the tendency to log-normal distributions.

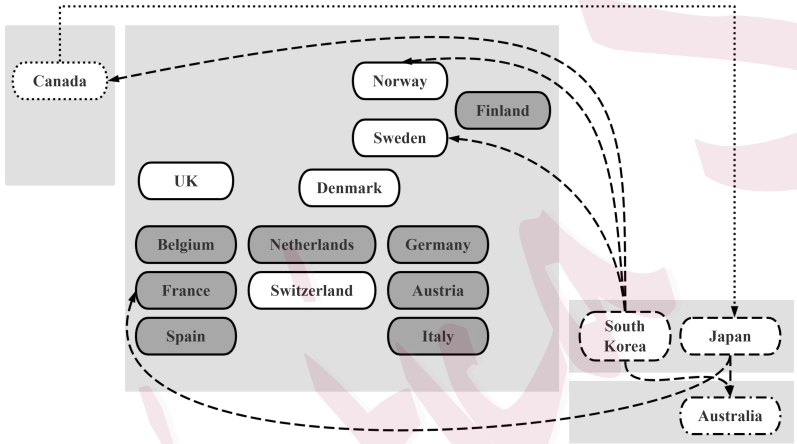
This dataset is firstly processed by adopting a three-quarter moving average of the recorded quarterly exchange rates to remove the seasonality of the original data. We take the tail index $k = 0.28$, consider the standard normal distribution as the conditional distribution suggested by Section 5 when $|\widehat{\mathcal{S}}_t| = 1$, and then apply the TopHeat algorithm to estimate the DAG structures of the financial contagion from the foreign exchange rates.

Figure 2 displays the estimated DAGs among currencies, which consist of 24 and 7 directed edges in pre-euro system and post-euro system, respectively. Clearly, there are more estimated directed edges among European countries in the pre-euro DAG than that in the post-euro one. This finding is supported by the fact that close trade exchanges and economic connections lead to frequent fluctuation among currencies in the pre-euro era, which indicates the financial risk spreads rapidly. However, the strong relevance between these individual currencies declines since countries began to use the Euro, which reflects that risk propagation becomes more stable in the regional level. Furthermore, in the pre-euro era, the hub nodes of the contagion network are Japan, UK, and some European

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199



(a) The estimated DAG before 1999



(b) The estimated DAG after 1999

Figure 2: The estimated spillover networks for the financial contagion from the exchange rates of 17 OECD countries by our proposed TopHeat method. The top in (a) and bottom in (b) display all the estimated directed edges for the pre-euro system and post-euro era, respectively. Countries from the same continent are drawed with lines of the same type and in the same shaded box, and the nodes with a dark gray background are the countries in (b) that firstly became members of the euro area in 1999.

countries, since Japan's highly development was fueled by its robust manufac-

turing sector and export-oriented economy, UK's strong connections with Spain

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

and Finland stemmed from historical ties and trade relationships within the European Economic Community, and the connections between Scandinavian countries like Norway, Sweden, and Denmark reflects the strong economic and trade links within the Nordic region. The Republic of Korea's accession to the World Trade Organization (WTO) in 1995 also influenced the financial market, leading to the growing importance of East Asian economies in global trade and finance during the late 20th and early 21st centuries. It is worth to noting that the use of the euro removed the causal relations from the intra-European countries as the launch of the euro in 1999, and we conjecture that dependence may exist, which fails to be captured by DAGs and deserves to be studied in future work.

7. Conclusion

In this paper, we propose an efficient learning method to learn the DAG structures in heavy-tailed data. The proposed TopHeat method utilizes a concept of topological layers to facilitate learning in a two-step algorithm where we first reconstruct the topological layers hierarchically in a top-down fashion and then recover the directed edges via modified CIT for heavy-tailed distributions. The asymptotic consistency of TopHeat is established to recover the underlying exact DAG structures under mild conditions when the number of nodes diverges. The simulation studies and real data analysis support the advantages of TopHeat against the existing learning algorithms in literature. It is interesting to point out

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

that one of the possible future work is to develop the regression-based methods to learn the skeletons under the heavy-tailed SCM setting.

Supplementary Material

The online Supplementary Material contains all the technical details and additional results.

Acknowledgments

The authors thank the editor, associate editor, and reviewers for their constructive comments, which led to significant improvement in this work. The authors are supported by National Key R&D Program of China (Grant No. 2022YFA1003800), National Natural Science Foundation of China (Grant Nos. 12471265, 72495122, 12231011, 12501381, 72473114, and 71988101), HK RGC Grants GRF (11311022, 14306523, and 14303424), CUHK Startup Grant 4937091, and Sichuan Science and Technology Program (2024NSFSC1393). Zhong also thanks the supports of Fujian Key Lab of Statistics, Fujian Key lab of Digital Finance.

References

Asadi, P., A. C. Davison, and S. Engelke (2015). Extremes on river networks. *The Annals of Applied Statistics* 9(4), 2023–2050.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

REFERENCES

Azadkia, M. and S. Chatterjee (2021). A simple measure of conditional dependence. *The Annals of Statistics 49*(6), 3070–3102.

Barabási, A. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5349), 509–512.

Cai, J., J. H. J. Einmahl, L. De Haan, and C. Zhou (2015). Estimation of the marginal expected shortfall: the mean when a related variable is extreme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(2), 417–442.

Candes, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(3), 551–577.

Chen, S. and M. Schienle (2022). Large spillover networks of nonstationary systems. *Journal of Business and Economic Statistics 42*(2), 422–436.

Chickering, D. W. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research 3*(3), 507–554.

Daouia, A., S. Girard, and G. Stupfler (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(2), 263–292.

De Haan, L. and A. Ferreira (2006). *Extreme Value Theory: An Introduction*. New York: Springer.

Erdös, P. and A. Rényi (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5*, 17–60.

Fan, J., Q. Li, and Y. Wang (2017). Estimation of high dimensional mean regression in the absence of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

symmetry and light tail assumptions. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(1), 247–265.

Gao, M., Y. Ding, and B. Aragam (2020). A polynomial-time algorithm for learning nonparametric causal graphs. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 11599–11611.

Gnecco, N., N. Meinshausen, J. Peters, and S. Engelke (2021). Causal discovery in heavy-tailed models. *The Annals of Statistics 49*(3), 1755–1778.

Harris, N. and M. Drton (2013). PC algorithm for nonparanormal graphical models. *The Journal of Machine Learning Research 14*(11), 3365–3383.

Hyvärinen, A. and S. M. Smith (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *The Journal of Machine Learning Research 14*(1), 111–152.

Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research 8*(22), 613–636.

Kalisch, M., M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software 47*(11), 1–26.

Klüppelberg, C. and M. Krali (2021). Estimating an extreme Bayesian network via scalings. *Journal of Multivariate Analysis 181*(C), 104672.

Lee, B. S. (1992). Causal relations among stock returns, interest rates, real activity, and inflation. *The Journal of Finance 47*(4), 1591–1603.

Li, J. and Q. Tang (2015). Interplay of insurance and financial risks in a discrete-time model with strongly

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

regular variation. *Bernoulli 21*(3), 1800–1823.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Peng, L. and Y. Qi (2017). *Inference for Heavy-Tailed Data: Applications in Insurance and Finance*. Cambridge, Massachusetts: Academic Press.

Peters, J. and P. Bühlmann (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika 101*(1), 219–228.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, Massachusetts: The MIT Press.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Berlin, Germany: Springer Science & Business Media.

Shah, R. D. and J. Peters (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics 48*(3), 1514–1538.

Shi, H., M. Drton, and F. Han (2024). On Azadkia–Chatterjee's conditional dependence coefficient. *Bernoulli 30*(2), 851–877.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen (2006). A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research 7*(72), 2003–2030.

Shimizu, S., T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

model. *The Journal of Machine Learning Research 12*(33), 1225–1248.

Shojaie, A. and G. Michailidis (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika 97*(3), 519–538.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press.

Sun, Q., W. Zhou, and J. Fan (2020). Adaptive huber regression. *Journal of the American Statistical Association 115*(529), 254–265.

Sun, W., J. Wang, and Y. Fang (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research 14*(107), 3419–3440.

Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association 110*(512), 1726–1734.

Wang, Y. S. and M. Drton (2020). High-dimensional causal discovery under non-Gaussianity. *Biometrika 107*(1), 41–59.

Yang, J. and Y. Zhou (2013). Credit risk spillovers among financial institutions around the global credit crisis: Firm-level evidence. *Management Science 59*(10), 2343–2359.

Yang, Z. and Y. Zhou (2017). Quantitative easing and volatility spillovers across countries and asset classes. *Management Science 63*(2), 333–354.

Zhang, K., J. Peters, D. Janzing, and B. Schölkopf (2011). Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0199

2011), pp. 804–813. AUAI Press.

Zhao, R., X. He, and J. Wang (2022). Learning linear non-Gaussian directed acyclic graph with diverging number of nodes. *The Journal of Machine Learning Research 23*(269), 1–34.

Zhao, T. and H. Liu (2014). Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE transactions on Information Theory 60*(12), 7874–7887.

Zhou, W., X. He, W. Zhong, and J. Wang (2022). Efficient learning of quadratic variance function directed acyclic graphs via topological layers. *Journal of Computational and Graphical Statistics 31*(4), 1269–1279.

Joint Laboratory of Data Science and Business Intelligence, School of Statistics and Data Science, Southwestern University of Finance and Economics, China

E-mail: zhouwei23@swufe.edu.cn

Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, China

E-mail: kangxueqian@stu.xmu.edu.cn

MOE Key Lab of Econometrics, WISE and Department of Statistics and Data Science, School of Economics, Xiamen University, China

E-mail: wzhong@xmu.edu.cn

Department of Statistics, Chinese University of Hong Kong, Hong Kong

E-mail: junhuiwang@cuhk.edu.hk