Efficient Learning of DAG Structures in Heavy-tailed Data

Wei Zhou

Joint Laboratory of Data Science and Business Intelligence,

School of Statistics and Data Science,

Southwestern University of Finance and Economics

Xueqian Kang

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

Wei Zhong

MOE Key Lab of Econometrics, WISE and Department of Statistics and Data Science,

School of Economics, Xiamen University

Junhui Wang

Department of Statistics, Chinese University of Hong Kong

Supplementary Material

The supplementary file contains all the technical details and additional simulation results.

S1 Definitions

We first give some definitions about heavy-tailed functions from Peng and Qi (2017), which is used for defining the heavy-tailed SCM.

Definition S1. *A distribution function $F(x)$ is said to have a (right) heavy tail with tail index $\delta > 0$ if it satisfies that*

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\delta} \quad \textit{for all} \quad x > 0.$$

Definition S2. *A measurable function $a(x)$ defined over $(0, x_0)$ for some $x_0 > 0$ is said to be regularly varying or a regular variation at zero with an exponent $\upsilon \in \mathbb{R}$, denoted by $a(x) \in RV_\upsilon$ if*

$$\lim_{t\to 0} \frac{a(tx)}{a(t)} = x^{\upsilon} \quad \textit{for all} \quad x > 0.$$

In this section, we also introduce the definition of topological layers of a DAG. A node m is a root node if $\mathrm{pa}_m = \emptyset$. We denote the length of the longest directed path of node m to one root node as L_m. It is clear that the root node has zero length and $0 \le L_m \le p-1$. Let the topological layers $\mathcal{A}_t = \{m : L_m = t\}$ be the collection of nodes with the same length t to the root node. Without loss of generality, a DAG can be reorganized as the layer structure with a total of T topological layers. Then, $\mathcal{A}_0 := \mathcal{S}_1 \subseteq$

$\mathcal{A}_0 \cup \mathcal{A}_1 := \mathcal{S}_2 \subseteq \ldots \subseteq \mathcal{A}_0 \cup \ldots \cup \mathcal{A}_{t-1} := \mathcal{S}_t \subseteq \mathcal{A}_0 \cup \ldots \cup \mathcal{A}_{T-1} = \mathcal{V}$, where $\mathcal{S}_t$ includes all ancestors of $m \in \mathcal{A}_t$, but excludes its descendants. Thus, the layer structure assures the acyclicity in $\mathcal{G}$. We remark that the topological layer structures of $\mathcal{G}$ are unique, which is in contrast to the topological ordering (Peters et al., 2017) only requiring that the parent node arrives before its children.

S2 Hyperparameters

Note that the numerical performance of the proposed TopHeat algorithm largely depends on the choice of the hyperparameters, including the index of the order statistic k, the significant level α in the CIT procedure, and the threshold for the layer ϵ_t. Here, we perform initial simulations to determine k, which is related with the tail index θ of the underlying distribution, and select an optimal value for α; More details are provided in Section 5.2. To adaptively select the optimal values for ϵ_t, we employ the stability selection criterion in Sun et al. (2013).

The key idea is to measure the reconstruction stability by randomly splitting the training sample into two parts and comparing the disagreement between the two estimated active sets. Specifically, given a value ϵ_t, we randomly split the training sample D into two parts D^1 and D^2. Then the

proposed method is applied to D^1 and D^2 and we obtain two estimated active sets $\widehat{E}^1_{\epsilon_t}$ and $\widehat{E}^2_{\epsilon_t}$, respectively. The disagreement between $\widehat{E}^1_{\epsilon_t}$ and $\widehat{E}^2_{\epsilon_t}$ is measured by Cohen's kappa coefficient

$$\kappa(\widehat{E}^1_{\epsilon_t}, \widehat{E}^2_{\epsilon_t}) = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where $Pr(a) = \frac{n_{11}+n_{22}}{p_t}$ and $Pr(e) = \frac{(n_{11}+n_{12})(n_{11}+n_{21})}{p_t^2} + \frac{(n_{12}+n_{22})(n_{21}+n_{22})}{p_t^2}$ with $n_{11} = |\widehat{E}^1_{\epsilon_t} \cap \widehat{E}^2_{\epsilon_t}|, n_{12} = |\widehat{E}^1_{\epsilon_t} \cap \widehat{E}^{2,C}_{\epsilon_t}|, n_{21} = |\widehat{E}^{1,C}_{\epsilon_t} \cap \widehat{E}^2_{\epsilon_t}|, n_{22} = |\widehat{E}^{1,C}_{\epsilon_t} \cap \widehat{E}^{2,C}_{\epsilon_t}|$, $p_t = p - |\mathcal{S}_t|$. Here, $|\cdot|$ denotes the set cardinality and $\widehat{E}^{j,C}_{\epsilon_t}$ is the complementary set of $\widehat{E}^j_{\epsilon_t}$ for $j = 1, 2$. The procedure is repeated for B times and the estimated reconstruction stability is measured as

$$\hat{s}(\Psi_{\epsilon_t}) = \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} \kappa(\widehat{E}^1_{b,\epsilon_t}, \widehat{E}^2_{b,\epsilon_t}),$$

where $\widehat{E}^1_{b,\epsilon_t}$ and $\widehat{E}^2_{b,\epsilon_t}$ are the estimated active sets in the b-th splitting. Finally, we set $\widehat{\epsilon_t} = \min\left\{\epsilon_t : \frac{\hat{s}(\Psi_{\epsilon_t})}{\max_{\epsilon_t} \hat{s}(\Psi_{\epsilon_t})} \geq 1 - a\right\}$, where $a \in (0, 1)$ is some given percentage. As recommended by Sun et al. (2013), the threshold parameter a approaches zero as n grows to ensure asymptotic selection consistency, and more discussion are presented in Section 4. Empirical studies in Section 5.2 demonstrate satisfactory performance while we choose a and B varying with n.

S3 Proof of Theorem 1

Before giving the proof of Theorem 1, we first restate the results about Γ_{jm} in Gnecco et al. (2021) below.

Lemma 1. *(Theorem 1 in Gnecco et al. (2021)) Under the heavy-tailed linear SCM model in* (2.1)*, for any two variables X_j and X_m with $j \neq m$, there holds that*

$$\Gamma_{jm} = \frac{1}{2} + \frac{\sum_{l \in An_j \cap An_m} \pi_{jl}^{\theta}}{2 \sum_{l \in An_j} \pi_{jl}^{\theta}}. \tag{S3.1}$$

Further, Table S1 gives the corresponding values of Γ_{jm} when: (a) $j \in An_m$; (b) $m \in An_j$; (c) $An_j \cap An_m = \emptyset$; or (d) $An_j \cap An_m \neq \emptyset$ neither $j \notin An_m$ nor $m \notin An_j$.

Table S1: Values of Γ_{jm} and Γ_{mj} under different scenarios, where $-$ indicates impossible scenarios for the heavy-tailed linear SCM model (2.1).

	$\Gamma_{mj} = 1$	$\Gamma_{mj} \in (1/2, 1)$	$\Gamma_{mj} = 1/2$
$\Gamma_{jm} = 1$	$-$	(a)	$-$
$\Gamma_{jm} \in (1/2, 1)$	(b)	(d)	$-$
$\Gamma_{jm} = 1/2$	$-$	$-$	(c)

It is clear that Table S1 provides the criteria to identify the ancestor and descendant relationship between any two nodes, which helps to reconstruct the topological layers of a heavy-tailed DAG in a top-down fashion.

Proof of Theorem 1: Assume that the number of topological layers is in total T. We now prove the reconstruction of topological layers by induction.

For the layer $\mathcal{A}_0$, if $m \in \mathcal{A}_0$, and thus $\text{pa}_m = \emptyset$, there exists no node $j \in \mathcal{V}$ such that $\Gamma_{jm} = 1$, and thus $\max_{j\in\mathcal{V}} \Gamma_{jm} < 1$ holds. Next, we show that if $\max_{j\in\mathcal{V}} \Gamma_{jm} < 1$ with $m \in \mathcal{V}$, then $m \in \mathcal{A}_0$. If not, without loss of generality, we assume that $m \notin \mathcal{A}_0$ and thus $m \in \mathcal{A}_t$ with $t \geq 1$, and then there exists a directed path from a root node $j \in \mathcal{A}_0$ to node m such that $j \in \text{An}_m$. It follows from Lemma 1 that $\Gamma_{mj} \in (1/2, 1)$ and $\Gamma_{jm} = 1$, which implies that $\max_{j\in\mathcal{V}} \Gamma_{jm} = 1$. Therefore, $\mathcal{A}_0$ is identified.

For the layer $\mathcal{A}_t$, assume that the first t layers have been correctly identified, denoted as $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{t-1}$. Now, we show that the statement also holds for $\mathcal{A}_t$.

If node $m \in \mathcal{A}_t$, then $\text{pa}_m \subseteq \mathcal{S}_t$ where $\mathcal{S}_t = \cup_{d=0}^{t-1}\mathcal{A}_d$. This implies that $\max_{j\in\mathcal{S}_t} \Gamma_{jm} = 1$ and thus $\max_{j\in\mathcal{C}_t} \Gamma_{jm} \in [\frac{1}{2}, 1)$ holds as $\text{An}_m \cap \mathcal{C}_t = \{m\}$ with $\Gamma_{mm} = \frac{1}{2}$ and $\text{an}_m \cap \mathcal{C}_t = \emptyset$. Thus, $\max_{j\in\mathcal{C}_t} \Gamma_{jm} < 1$ holds. On the other hand, we show that if $\max_{j\in\mathcal{C}_t} \Gamma_{jm} < 1$ for $m \in \mathcal{C}_t$, then $m \in \mathcal{A}_t$. If not, for $m \notin \mathcal{A}_t$, we suppose $m \in \mathcal{C}_t \backslash \mathcal{A}_t$, there must exist a directed path from one node j to node m such that $j \in \text{an}_m$ and thus $\Gamma_{jm} = 1$. This leads to $\max_{j\in\mathcal{C}_t} \Gamma_{jm} = 1$ holds by Lemma 1. Now, $\mathcal{A}_t$ is identified. By induction,

the proof is completed. ■

S4 Proof of Theorem 2

We first give necessary lemmas as follows. The following Lemma 2 shows the limit result of an approximation for the quantity $E(F_m(X_m)|F_j(X_j) > u)$.

Lemma 2. *For $\theta > 0$, there holds that*

$$\lim_{u\to 1^-} \frac{1}{1-u} E(F_m(X_m)|F_j(X_j) > u) = \int_0^\infty R(1 - F_m(s), 1) dF_m(s).$$

Proof of Lemma 2: By applying the definition of conditional expectation and the change of variables, we have

$$\begin{aligned}
&E(F_m(X_m)|F_j(X_j) > u) \\
=&\int_0^\infty P(X_m > s|F_j(X_j) > u) dF_m(s) \\
=&\frac{1}{1-u}\int_0^\infty P(X_m > s, F_j(X_j) > u) dF_m(s) \\
=&\frac{1}{1-u}\int_0^\infty P(F_m(X_m) > s(1-u)+u, F_j(X_j) > u)(1-u)ds \\
=&\int_0^\infty P(F_m(X_m) > F_m(s)(1-u)+u, F_j(X_j) > u) dF_m(s). \qquad \text{(S4.2)}
\end{aligned}$$

Therefore, the definition of the upper tail dependence function implies that

$$\lim_{u\to 1^-} \frac{1}{1-u} P(F_m(X_m) > F_m(s)(1-u)+u, F_j(X_j) > u)$$

$$= \lim_{u\to 1^-} \frac{1}{1-u} P(1 - F_m(X_m) \le (1 - F_m(s))(1-u), 1 - F_j(X_j) \le 1 - u)$$

$$= R(1 - F_m(s), 1). \tag{S4.3}$$

Next, we show that the integral in (S4.2) and the limit procedure $u \to 1$ can be exchanged by applying dominated convergence theorem. Note that $\frac{1}{1-u} P(F_m(X_m) \le F_m(s)(1-u) + u, F_j(X_j) > u) \le \frac{1}{1-u} P(F_j(X_j) > u) = 1$ with $u < 1-\varepsilon$ for a small $\varepsilon > 0$, and 1 is integrable on $[0, 1]$. This completes the proof. ■

To emphasize the dependence on the first $k+1$ larger order statistics of the estimator $\widehat{\Gamma}_{jm}$ in (3.4), we rewrite it as

$$\widehat{\Gamma}_{jm}^{1-k/n} = \widehat{\Gamma}_{jm} = \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_m(X_{i,m}^n) \mathbf{1} \left\{ X_{i,j}^n > U_j \left(\frac{n}{k e_n} \right) \right\},$$

since $e_n = (n/k)(1 - F_j(X_{(n-k),j}^n)) \xrightarrow{\mathrm{P}} 1$ as $n \to \infty$. Furthermore, we denote $\widetilde{\Gamma}_{jm}^{1-k\cdot/n}$ as a random function over the interval $[1/2, 2]$,

$$\widetilde{\Gamma}_{jm}^{1-ky/n} = \frac{1}{ky} \sum_{i=1}^{n} \widehat{F}_m(X_{i,m}^n) \mathbf{1} \left\{ X_{i,j}^n > U_j \left(\frac{n}{ky} \right) \right\}. \tag{S4.4}$$

The main idea to prove Theorem 2 is to apply $y = e_n$ and combine the asymptotic behavior of e_n, and then we prove the asymptotic normality of $\widehat{\Gamma}_{jm}^{1-k/n}$.

We write $R_n(x_j, x_m) = (n/k)P(1 - F_j(X_j) \le kx_j/n, 1 - F_m(X_m) \le kx_m/n)$ and its pseudo estimator is given as $T_n(x_j, x_m) = \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\{1 - F_j(X_{i,j}^n) \le kx_j/n, 1 - F_m(X_{i,m}^n) \le kx_m/n\}$. For $x > 0$, we denote $s_n(x) = \frac{n}{k}\left(1 - F_m\left(U_m\left(\frac{n}{k}\right)x^{-1/\theta}\right)\right)$ and its derivative is defined by $g_{\frac{n}{k}}(x) = \frac{n}{k}U_m\left(\frac{n}{k}\right) f_m\left(U_m\left(\frac{n}{k}\right)x^{-1/\theta}\right)$. Note that the regularly varying function property guarantees that $\lim_{t\to\infty} U_m(tx)/U_m(t) = x^{1/\theta}$, which implies that $s_n(x) \to x$ as $n \to \infty$. Density convergence, given as $ds_n(x)/dx \to 1$, leads to $g_{\frac{n}{k}}(x) \to \theta x^{1+1/\theta}$.

Next, we restate the following Lemmas 3–5 in Cai et al. (2015), where Lemma 3 gives the asymptotic behaviour of the pseudo estimator $T_n(\cdot,\cdot)$, Lemma 4 shows the W_R-process is bounded with proper functions, and Lemma 5 clarifies the role of $s_n(x)$ replaced by x in the limit for the integrals. The following three lemmas will be used and their proofs are omitted here.

Lemma 3. *(Lemma 1 in Cai et al. (2015)) Assume that this limit* $\lim_{t\to\infty} tP(1-F_j(X_j) \le x/t, 1-F_m(X_m) \le y/t) = R(x,y)$ *exists for* $(x,y) \in [0,\infty]^2\backslash\{(\infty,\infty)\}$ *and* $j, m \in \{1,\dots,p\}$. *For any* $\eta \in [0, \frac{1}{2})$ *and positive* ℓ, *with probability one, there holds that*

$$\sup_{x,y\in(0,\ell]} \left| \frac{\sqrt{k}(T_n(x,y) - R_n(x,y)) - W_R(x,y)}{x^\eta} \right| \to 0,$$
$$\sup_{x\in(0,\ell]} \left| \frac{\sqrt{k}(T_n(x,\infty) - x) - W_R(x,\infty)}{x^\eta} \right| \to 0,$$

$$\sup_{y\in(0,\ell]}\Big|\frac{\sqrt{k}(T_n(\infty,y)-y)-W_R(x,\infty)}{y^\eta}\Big|\to 0.$$

Lemma 4. *(Lemma 2 in Cai et al. (2015)) For any $\ell>0$ and $\eta\in[0,\frac{1}{2})$, with probability one,*

$$\sup_{0<x\le\ell,0<y<\infty}\frac{|W_R(x,y)|}{x^\eta}<\infty,\quad \sup_{0<x<\infty,0<y<\ell}\frac{|W_R(x,y)|}{y^\eta}<\infty.$$

Lemma 5. *(Lemma 3 in Cai et al. (2015)) Assume that this limit $\lim_{t\to\infty}tP(1-F_j(X_j)\le x/t,1-F_m(X_m)\le y/t)=R(x,y)$ exists for $(x,y)\in[0,\infty]^2\backslash\{(\infty,\infty)\}$ and $j,m\in\{1,\dots,p\}$. Denote ℓ as a bounded and continuous function on $[0,M_0)\times[a,b]$ with $0<M_0\le\infty$ and $0\le a<b<\infty$. Moreover, suppose that there exist $\eta_1>1/\theta$ and $K>0$ such that*

$$\sup_{0<x\le M_0,a\le y\le b}\frac{|\ell(x,y)|}{x^{\eta_1}}\le K,$$

If $M_0<\infty$, we further require that $0<M<M_0$. Then

$$\lim_{n\to\infty}\sup_{a\le y\le b}\Big|\int_0^s\ell(s_n(x),y)-\ell(x,y)dx^{-1/\theta}\Big|=0.$$

Furthermore, suppose that $|\ell(x_1,y)-\ell(x_2,y)|\le|x_1-x_2|$ holds for all $0\le$

$x_1, x_2 < M_0$ and $a \le y \le b$. Under Assumptions 2 and 3, we have that

$$\lim_{n\to\infty} \sup_{a\le y\le b} \sqrt{k}\left| \int_0^s \ell(s_n(x), y) - \ell(x, y) dx^{-1/\theta} \right| = 0.$$

Now, we first give the asymptotic behaviour of $\widetilde{\Gamma}_{jm}^{1-ky/n}$ in the following Lemma 6.

Lemma 6. *Suppose that Assumption 2 holds and for $\theta > 1$, we have*

$$\sup_{1/2\le y\le 2} \left| \frac{n}{\sqrt{k}} \left(\widetilde{\Gamma}_{jm}^{1-ky/n} - E\left(F_m(X_m) | F_j(X_j) > 1 - \frac{ky}{n} \right) \right) + \frac{1}{y} \int_0^\infty W_R(x, y) dx^{-1/\theta} \right| \xrightarrow{\mathrm{P}} 0.$$

Proof of Lemma 6: Following (S4.2) and (S4.3) in Lemma 2, we use the change of variables,

$$\begin{aligned}
&\frac{1}{ky/n} E(F_m(X_m) | F_j(X_j) > 1 - ky/n) \\
=&\frac{1}{ky/n} \int_0^\infty P\left(F_m(X_m) > F_m(s)\frac{ky}{n} + 1 - \frac{ky}{n}, F_j(X_j) > 1 - \frac{ky}{n} \right) dF_m(s) \\
=&\int_0^\infty \frac{1}{ky/n} P\left(1 - F_m(X_m) \le \frac{ky}{n}(1 - F_m(s)), 1 - F_j(X_j) \le \frac{ky}{n} \right) dF_m(s) \\
=&-\int_0^1 \frac{1}{(ky/n)^2} P\left(1 - F_m(X_m) \le t, 1 - F_j(X_j) \le \frac{ky}{n} \right) dt.
\end{aligned}$$

The above equality multiplies the term ky^2/n in both sides and we obtain

that

$$\begin{aligned}
&yE(F_m(X_m)|F_j(X_j) > 1 - ky/n) \\
=&-\frac{n}{k}\int_0^1 P\left(1 - F_m(X_m) \le \frac{k}{n} \times \frac{nt}{k}, 1 - F_j(X_j) \le \frac{ky}{n}\right) dt \\
=&-\int_0^1 R_n\left(\frac{nt}{k}, y\right) dt \\
=&\int_0^\infty R_n\left(\frac{n}{k}(1 - F_m(s)), y\right) dF_m(s) \\
=&U_m\left(\frac{n}{k}\right)\int_0^\infty R_n(s_n(x), y) f_m\left(U_m\left(\frac{n}{k}\right) x^{-1/\theta}\right) dx^{-1/\theta}, \qquad \text{(S4.5)}
\end{aligned}$$

since $s_n(x) = (n/k)\left(1 - F_m\left(U_m\left(\frac{n}{k}\right) x^{-1/\theta}\right)\right)$ and the last step in (S4.5) follows from the fact that

$$\begin{aligned}
&\int_0^\infty R_n(s_n(x), y) f_m\left(U_m\left(\frac{n}{k}\right) x^{-1/\theta}\right) dx^{-1/\theta} \\
=&\int_0^\infty R_n\left(\frac{n}{k}\left(1 - F_m\left(U_m\left(\frac{n}{k}\right) x^{-1/\theta}\right)\right), y\right) f_m\left(U_m\left(\frac{n}{k}\right) x^{-1/\theta}\right) dx^{-1/\theta} \\
=&\frac{1}{U_m\left(\frac{n}{k}\right)}\int_0^\infty R_n\left(\frac{n}{k}(1 - F_m(s)), y\right) f_m(s) ds.
\end{aligned}$$

Similar to (S4.5), we also have

$$y\widetilde{\Gamma}_{jm}^{1-ky/n} = U_m\left(\frac{n}{k}\right)\int_0^\infty T_n(s_n(x), y) f_m\left(U_m\left(\frac{n}{k}\right) x^{-1/\theta}\right) dx^{-1/\theta}.$$

Then, for any $L > 0$, there holds that

$$
\begin{aligned}
&\sup_{1/2\le y\le 2}\left|\frac{n}{\sqrt{k}}\left(y\widetilde{\Gamma}_{jm}^{1-ky/n}-yE(F_m(X_m)|F_j(X_j)>1-\frac{ky}{n})\right)+\int_0^\infty W_R(x,y)dx^{-1/\theta}\right| \\
=&\sup_{1/2\le y\le 2}\left|\int_0^\infty \sqrt{k}\big(T_n(s_n(x),y)-R_n(s_n(x),y)\big)g_{\frac{n}{k}}(x)dx^{-1/\theta}+\int_0^\infty W_R(x,y)dx^{-1/\theta}\right| \\
\le&\sup_{1/2\le y\le 2}\left|\int_L^\infty W_R(x,y)dx^{-1/\theta}\right| \\
&+\sup_{1/2\le y\le 2}\left|\int_L^\infty \sqrt{k}\big(T_n(s_n(x),y)-R_n(s_n(x),y)\big)g_{\frac{n}{k}}(x)dx^{-1/\theta}\right| \\
&+\sup_{1/2\le y\le 2}\left|\int_0^L \sqrt{k}\big(T_n(s_n(x),y)-R_n(s_n(x),y)\big)g_{\frac{n}{k}}(x)-W_R(x,y)dx^{-1/\theta}\right| \\
:=&K_1(L)+K_{2,n}(L)+K_{3,n}(L). \qquad \text{(S4.6)}
\end{aligned}
$$

It suffices to show that for any $\varepsilon > 0$, there exists $L_0 = L_0(\varepsilon)$, $n_0 = n_0(L_0)$ such that

$$P(K_1(L_0) > \varepsilon) < \varepsilon. \qquad \text{(S4.7)}$$

and for any $n > n_0$ there holds that

$$P(K_{2,n}(L_0) > \varepsilon) < \varepsilon, \quad P(K_{3,n}(L_0) > \varepsilon) < \varepsilon. \qquad \text{(S4.8)}$$

Note that the term $K_1(L)$ is the same as the term $I_1(T)$ in the proof of Proposition 2 in Cai et al. (2015). We apply Lemma 4 with $\eta = 0$ and there

exists $L_1 = L_1(\varepsilon)$ such that $P\left(\sup_{0<x<\infty,0\le y\le 2}|W_R(x,y)| > L_1^{1/\theta}\varepsilon\right) < \varepsilon$. Thus,

$$P(K_1(L_0) > \varepsilon) < P\Big[\sup_{x>L_1,0\le y\le 2}|W_R(x,y)| > L_1^{1/\theta}\varepsilon\Big] < \varepsilon, \tag{S4.9}$$

holds for any $L_0 > L_1$.

Next, we deal with the term $K_{2,n}(L)$. Denote $\tilde{P}_n$ and $\tilde{P}$ by the probability measure defined by $(1-F_j(X_j), 1-F_m(X_m))$ and the empirical probability measure by $(1-F_j(X_{i,j}^n), 1-F_m(X_{i,m}^n))_{1\le i\le n}$, respectively. Therefore, we have

$$\begin{aligned}
&P(K_{2,n}(L) > \varepsilon)\\
=&P\left(\sup_{1/2\le y\le 2}\left|\int_L^\infty \sqrt{k}\big(T_n(s_n(x)g_{\frac{n}{k}}(x), yg_n(x)) - R_n(s_n(x)g_{\frac{n}{k}}(x), yg_{\frac{n}{k}}(x))\big)dx^{-1/\theta}\right| > \varepsilon\right)\\
\le&P\left(\sup_{x>L,1/2\le y\le 2}\left|\sqrt{k}\big(T_n(s_n(x)g_{\frac{n}{k}}(x), yg_{\frac{n}{k}}(x)) - R_n(s_n(x)g_{\frac{n}{k}}(x), yg_{\frac{n}{k}}(x))\big)\right| > L^{1/\theta}\varepsilon\right)\\
\le&P\left(\sup_{x>L,1/2\le y\le 2}\left|\sqrt{n}\left(\tilde{P}_n - \tilde{P}\right)\left\{\left(0, \frac{ks_n(x)g_{\frac{n}{k}}(x)}{n}\right)\times\left(0, \frac{kyg_{\frac{n}{k}}(x)}{n}\right)\right\}\right| > \varepsilon L^{1/\theta}\sqrt{k/n}\right)\\
:=&\widetilde{K}_2.
\end{aligned}$$

Let $Q = [0,1]\times[0,1]$, and thus $\left(0, \frac{ks_n(x)g_{\frac{n}{k}}(x)}{n}\right)\times\left(0, \frac{kyg_{\frac{n}{k}}(x)}{n}\right) \subset Q$ holds with arbitrary high probability with $P(Q) = 1$ in $\widetilde{K}_2$. By applying Inequality 2.5 in Einmahl (1986), there exist a constant $C_5 > 0$ and a function ψ

satisfying $\lim_{t\to 0}\psi(t)=1$ such that

$$\begin{aligned}\widetilde{K}_2 \leq & C_5 \exp\left(-\frac{(\varepsilon L^{1/\theta}\sqrt{k/n})^2}{2P(Q)}\psi\left(\frac{\varepsilon L^{1/\theta}\sqrt{k/n}}{\sqrt{n}P(Q)}\right)\right)\\ = & C_5 \exp\left(-\frac{\varepsilon^2 L^{2/\theta}k}{2n}\psi\left(\frac{\varepsilon L^{1/\theta}\sqrt{k}}{n}\right)\right).\end{aligned}$$

We choose $L_2=L_2(\varepsilon)$ and $n_1=n_1(L_2)$ such that for $n>n_1$, $\psi\left(\frac{\varepsilon L_2^{1/\theta}\sqrt{k}}{n}\right)>\frac{1}{2}$, and thus for $L>L_2$ and $n>n_1(L)$, $\widetilde{K}_2\leq C_5\exp\left(-\frac{\varepsilon^2 L^{2/\theta}k}{4n}\right)\leq\varepsilon$ holds. Therefore, for $L_0>L_2$ and $n_0>n_1$, we have

$$P(K_{2,n}(L_0)>\varepsilon)<\varepsilon. \tag{S4.10}$$

Lastly, we consider $K_{3,n}(L)$. Note that

$$\begin{aligned}&P(K_{3,n}(L)>\varepsilon)\\ =&P\left(\sup_{1/2\leq y\leq 2}\left|\int_0^L\sqrt{k}\big[T_n(s_n(x)g_{\frac{n}{k}}(x),yg_{\frac{n}{k}}(x))-R_n(s_n(x)g_{\frac{n}{k}}(x),yg_{\frac{n}{k}}(x))\big]-\right.\right.\\ &\qquad\left.\left.W_R(x,y)dx^{-1/\theta}\right|>\varepsilon\right)\\ \leq&P\left(\sup_{1/2\leq y\leq 2}\left|\int_0^L\sqrt{k}\big[T_n(s_n(x)g_n(x),yg_n(x))-R_n(s_n(x)g_n(x),yg_n(x))\big]\right.\right.\\ &\qquad\left.\left.-W_R(s_n(x)g_{\frac{n}{k}}(x),yg_{\frac{n}{k}}(x))dx^{-1/\theta}\right|>\frac{\varepsilon}{2}\right)\\ &+P\left(\sup_{1/2\leq y\leq 2}\left|\int_0^L W_R(s_n(x)g_{\frac{n}{k}}(x),yg_{\frac{n}{k}}(x))-W_R(x,y)dx^{-1/\theta}\right|>\frac{\varepsilon}{2}\right)\end{aligned}$$

$:=\widetilde{K}_{31}+\widetilde{K}_{32}$.

Next, we first deal with the term $\widetilde{K}_{31}$. For any $L>0$, there exists $n_2=n_2(L)$ such that for all $n>n_2$, $s_n(x)\le L$ and $g_{\frac{n}{k}}(x)<\theta L^{1+1/\theta}$, which implies that $s_n(x)g_{\frac{n}{k}}(x)\le\theta L^{2+1/\theta}$ and $g_{\frac{n}{k}}(x)y\le 2\theta L^{1+1/\theta}$ for $y\in(1/2,2)$. Thus, we denote $\widetilde{L}=\max\{\theta L^{2+1/\theta},2\theta L^{1+1/\theta}\}$ and for $n>n_2$ and any $\theta>\max\left\{\frac{1}{2\eta_0}-\frac{1}{2},\frac{1}{\eta_0}-1\right\}=\frac{1}{\eta_0}-1$ with $\eta_0\in(\frac{1}{2\theta+1},\frac{1}{2})$,

$$\begin{aligned}\widetilde{K}_{31}\le & P\left(\sup_{0<s,t\le\widetilde{L}}\left|\frac{\sqrt{k}\left(T_n(s,t)-R_n(s,t)\right)-W_R(s,t)}{(s\vee t)^{\eta_0}}\right|\right.\\ &\left.\times\left|\int_0^L\left(\max\left\{s_n(x)g_{\frac{n}{k}}(x),yg_{\frac{n}{k}}(x)\right\}\right)^{\eta_0}dx^{-1/\theta}\right|>\frac{\varepsilon}{2}\right).\end{aligned}\tag{S4.11}$$

It follows from Lemma 3 that $\sup_{0<s,t<\widetilde{L}}\left|\frac{\sqrt{k}(T_n(s,t)-R_n(s,t))-W_R(s,t)}{(s\vee t)^{\eta_0}}\right|\to 0$ holds in (S4.11) with $\widetilde{L}>0$ and $\eta_0\in[0,\frac{1}{2})$. Since $\sup_{0<x\le L}\frac{|(\theta x^{2+1/\theta})^{\eta_0}|}{x^{\eta_0}}<L\theta^{\eta_0}$ satisfies the condition in Lemma 5, we have

$$\begin{aligned}\lim_{n\to\infty}\left|\int_0^L(s_n(x)g_{\frac{n}{k}}(x))^{\eta_0}dx^{-1/\theta}\right|=&\int_0^L(\theta x^{2+1/\theta})^{\eta_0}dx^{-1/\theta}\\ =&\frac{\theta^{\eta_0}}{\eta_0+2\eta_0\theta-1}L^{(\eta_0-1)/\theta+2\eta_0}.\end{aligned}\tag{S4.12}$$

Similarly, we have

$$\lim_{n\to\infty}\left|\int_0^L(yg_{\frac{n}{k}}(x))^{\eta_0}dx^{-1/\theta}\right|=y^{\eta_0}\int_0^L(\theta x^{1+1/\theta})^{\eta_0}dx^{-1/\theta}$$

$$=\frac{(\theta y)^{\eta_0}}{\theta\eta_0+\eta_0-1}L^{-\frac{1}{\theta}(1-\eta_0)+\eta_0}. \tag{S4.13}$$

Thus, there exists $n_3(L) > n_2(L)$ such that for $n > n_3(L)$, $\widetilde{K}_{31} < \varepsilon/2$ holds.

Now, the term $\widetilde{K}_{32}$ is bounded by

$$\begin{aligned}\widetilde{K}_{32} \leq & P\left(\sup_{0<s\leq\theta L^{2+1/\theta},0<t<\infty,1/2\leq y\leq 2}\frac{|W_R(s,t)|}{s^{\eta_0}}\left|\int_0^L (s_n(x)g_{\frac{n}{k}}(x))^{\eta_0}dx^{-1/\theta}\right| > \frac{\varepsilon}{4}\right)\\ & + P\left(\sup_{0<s\leq L,0<t<\infty,1/2\leq y\leq 2}\frac{|W_R(s,t)|}{s^{\eta_0}}\left|\int_0^L x^{\eta_0}dx^{-1/\theta}\right| > \frac{\varepsilon}{4}\right)\\ := & \widetilde{K}_{32,1}+\widetilde{K}_{32,2}.\end{aligned} \tag{S4.14}$$

Similar to (S4.12) and (S4.13), we first consider the term $\widetilde{K}_{32,1}$,

$$\begin{aligned}&\widetilde{K}_{32,1}\\ \leq & P\left(\sup_{0<s\leq\theta L^{2+1/\theta},0<t<\infty,1/2\leq y\leq 2}\frac{|W_R(s,t)|}{s^{\eta_0}}\left|\int_0^L (s_n(x)g_{\frac{n}{k}}(x))^{\eta_0}-(\theta x^{2+1/\theta})^{\eta_0}dx^{-1/\theta}\right| \geq \frac{\varepsilon}{8}\right)\\ & + P\left(\sup_{0<s\leq\theta L^{2+1/\theta},0<t<\infty,1/2\leq y\leq 2}\frac{|W_R(s,t)|}{s^{\eta_0}}\left|\int_0^L (\theta x^{2+1/\theta})^{\eta_0}dx^{-1/\theta}\right| \geq \frac{\varepsilon}{8}\right)\\ := & \widetilde{K}_{32,1,1}+\widetilde{K}_{32,1,2}.\end{aligned}$$

Similar to $\widetilde{K}_{31}$, by combining the fact that $\sup_{0<s\leq L,0<y<\infty}\frac{|W_R(s,y)|}{s^{\eta_0}} < \infty$ for any $L > 0$ and $\lim_{n\to\infty}\left|\int_0^L (s_n(x)g_{\frac{n}{k}}(x))^{\eta_0}-(\theta x^{2+1/\theta})^{\eta_0}dx^{-1/\theta}\right| = 0$ from Lemma 5 under Assumption 2, there exists $n_4(L) > n_2(L)$ such that for

$n > n_4(L)$, $\widetilde{K}_{32,1,1}$ is smaller than $\varepsilon/8$. For the term $\widetilde{K}_{32,1,2}$, we write it as

$$\widetilde{K}_{32,1,2} \leq P\left(\sup_{0<s\leq\theta L^{2+1/\theta},0<t<\infty} \frac{|W_R(s,t)|}{s^{\eta_0}} > \frac{\varepsilon(\eta_0+2\eta_0\theta-1)}{8L^{(\eta_0-1)/\theta+2\eta_0}}\right).$$

For any fixed ε, Lemma 4 ensures that there is a positive $L(\varepsilon)$ such that for all $L < L(\epsilon)$, we have $\tilde{K}_{32,1,2} < \epsilon/8$ with the aforementioned $\eta_0 \in (\frac{1}{2\theta+1}, \frac{1}{2})$. Combining the above results, we have

$$\widetilde{K}_{32,1} < \varepsilon/4. \tag{S4.15}$$

For the term $\widetilde{K}_{32,2}$ in(S4.14), we have for $n > n_2$ and $\eta_0 \in (\frac{1}{\theta}, \frac{1}{2})$,

$$\widetilde{K}_{32,2} \leq P\left(\sup_{0<x\leq L,1/2<y\leq 2} \frac{|W_R(x,y)|}{s^{\eta_0}} > \frac{\varepsilon(\theta\eta_0-1)}{4L^{\eta_0-1/\theta}}\right). \tag{S4.16}$$

Similar to $\widetilde{K}_{32,1,2}$, we have for $L < L(\varepsilon)$, $\widetilde{K}_{32,2} < \varepsilon/4$. It follows from (S4.15) and (S4.16) that $\widetilde{K}_{32} < \varepsilon/2$ holds. Thus, $P(K_{3,n}(L_0) > \varepsilon) < \varepsilon$ holds for any L_0 and $n > \max(n_2(L_0), n_3(L_0), n_4(L_0))$.

To summarize, we let $L_0 = L_0(\varepsilon) > \max(L_1, L_2)$ and $n_0(L_0) = \max_{1\leq j\leq 4} n_j(L_0)$, and thus (S4.8) holds, which implies the desired result. ■

To prove Theorem 2, we first prove a non-stochastic limit relationship as shown in Lemma 7.

Lemma 7. *Suppose that Assumptions 1 and 3 are satisfied. Therefore,*

there holds that

$$\sup_{1/2\le y\le 2}\sqrt{k}\left|\int_0^\infty R_n(s_n(x),y)g_{\frac{n}{k}}(x)-R(x,y)\theta x^{1+1/\theta}dx^{-1/\theta}\right|\to 0,\ \ as\ n\to\infty.$$

Proof of Lemma 7: Note that

$$\begin{aligned}
&\sqrt{k}\left|\int_0^\infty R_n(s_n(x),y)g_n(x)-R(x,y)\theta x^{1+1/\theta}dx^{-1/\theta}\right|\\
\le&\sqrt{k}\left|\int_0^\infty \big[R_n(s_n(x),y)-R(x,y)\big]g_n(x)dx^{-1/\theta}\right|\\
&+\sqrt{k}\left|\int_0^\infty R(x,y)\big[g_n(x)-\theta x^{1+1/\theta}\big]dx^{-1/\theta}\right|\\
:=&H_1+H_2.
\end{aligned}$$

It follows from Equation (4.7) in Assumption 1 with $\tau_1,\tau_2<0$ and Assumption 3 with $\gamma<\min\left\{\frac{2\tau_1}{2\tau_1-1},\frac{2(\tau_1+\tau_2)}{2(\tau_1+\tau_2)-1}\right\}=\frac{2\tau_1}{2\tau_1-1}$ that

$$\begin{aligned}
H_1\le&\sqrt{k}\sup_{0<x<\infty,1/2\le y\le 2}|R_n(x,y)/R(x,y)-1|\left|\int_0^\infty g_n(x)R(x,y))dx^{-1/\theta}\right|\\
\le&O(\sqrt{k}(n/k)^{\tau_1})\left(\left|\int_0^\infty (g_n(x)-\theta x^{1+1/\theta})R(x,y)dx^{-1/\theta}\right|\right.\\
&\qquad\left.+\left|\int_0^\infty \theta x^{1+1/\theta}R(x,y)dx^{-1/\theta}\right|\right)\\
\le&O(\sqrt{k}(n/k)^{\tau_1+\tau_2})\left|\int_0^\infty R(x,y)dx^{-1/\theta}\right|\\
&+O(\sqrt{k}(n/k)^{\tau_1})\left|\int_0^\infty \theta x^{1+1/\theta}R(x,y)dx^{-1/\theta}\right|
\end{aligned}$$

$$\xrightarrow{\mathrm{P}} 0. \tag{S4.17}$$

For H_2, by Equation (4.8) in Assumption 1 with $\tau_2 < 0$ and Assumption 3 with $\gamma < \frac{-2\tau_2}{-2\tau_2+1}$, we have

$$\begin{aligned} H_2 \leq & \sqrt{k} \sup_{0<x<\infty} \left| g_{\frac{n}{k}}(x) - \theta x^{1+1/\theta} \right| \left| \int_0^{\infty} R(x,y) dx^{-1/\theta} \right| \\ = & O\left(\sqrt{k} \left(\frac{n}{k} \right)^{\tau_2} \right) \left| \int_0^{\infty} R(x,y) dx^{-1/\theta} \right| \to 0, \end{aligned} \tag{S4.18}$$

as $\int_0^{\infty} R(x,y) dx^{-1/\theta} < \infty$ holds. By combining (S4.17) and (S4.18), the desired result is derived. This completes the proof. ■

Next, we proceed to establish the asymptotic normality of $\widehat{\Gamma}_{jm}$.

Proof of Theorem 2: It follows from Assumption 1 with $\tau_3 < -1$ and Assumption 3 with $\gamma < 1$ that

$$\frac{n}{\sqrt{k}} \left(E(F_m(X_m) | F_j(X_j) > 1 - k/n) - \Gamma_{jm} \right) = O\left(n^{\frac{(1-\gamma)((1+\tau_3)}{2}} \right) = o_p(1). \tag{S4.19}$$

Note that $\widehat{\Gamma}_{jm} = e_n \widetilde{\Gamma}_{jm}^{1-ke_n/n}$ holds with probability one, with $e_n = (n/k)(1 - F_j(X^n_{(n-k),j}))$. Then, we have

$$\frac{n}{\sqrt{k}} \left[\widehat{\Gamma}_{jm} - \Gamma_{jm} \right] - \Theta$$

$$
\begin{aligned}
=&\frac{n}{\sqrt{k}}\big[\widehat{\Gamma}_{jm} - E(F_m(X_m)|F_j(X_j) > 1 - k/n)\big] - \Theta \\
&+ \frac{n}{\sqrt{k}}\left[E(F_m(X_m)|F_j(X_j) > 1 - k/n) - \Gamma_{jm}\right] \\
=&\left\{\frac{n}{\sqrt{k}}\big[e_n\widetilde{\Gamma}_{jm}^{1-ke_n/n} - e_nE(F_m(X_m)|F_j(X_j) > 1 - ke_n/n)\big] + \int_0^\infty W_R(s,1)ds^{-1/\theta}\right\} \\
&+ \left\{\frac{n}{\sqrt{k}}\big[e_nE(F_m(X_m)|F_j(X_j) > 1 - ke_n/n) - E(F_m(X_m)|F_j(X_j) > 1 - k/n)\big]\right. \\
&\quad \left. -(1/\theta - 1)W_R(\infty,1)\int_0^\infty R(s,1)s^{1+1/\theta}ds^{-1/\theta}\right\} + o_p(1) \\
:=&I_1 + I_2 + o_p(1),
\end{aligned}
$$

where the second equality follows from the fact (S4.19).

We first deal with the term I_1. Similar to the treatment of the Equation (26) in Cai et al. (2015), it follows from Lemma 3 that $T_n(\infty, e_n) = 1$ and then $\sqrt{k}(e_n - 1) \xrightarrow{\mathrm{d}} -W_R(\infty, 1)$. This leads to $\lim_{n\to\infty} P(|e_n - 1| > k^{-1/4}) = 0$. Therefore, with probability approaching one,

$$
\begin{aligned}
|I_1| \le &\sup_{|y-1|<k^{-1/4}} \left|\frac{n}{\sqrt{k}}\big[y\widetilde{\Gamma}_{jm}^{1-ky/n} - yE(F_m(X_m)|F_j(X_j) > 1 - ky/n)\big]\right. \\
&\qquad \left. + \int_0^\infty W_R(s,y)ds^{-1/\theta}\right| \\
&+ \sup_{|y-1|<k^{-1/4}} \left|\int_0^\infty W_R(s,y) - W_R(s,1)ds^{-1/\theta}\right|.
\end{aligned}
$$

where the first part directly follows from Lemma 6 and the proof of the second part is outlined in the following. For and $\varepsilon > 0$, $0 < \delta < 1$, and the

aforementioned $\eta \in (\frac{1}{\theta}, \frac{1}{2})$, we have

$$\begin{aligned}
&P\left(\sup_{|y-1|<k^{-1/4}}\left|\int_0^{\infty} W_R(s,y)-W_R(s,1)ds^{-1/\theta}\right|>\varepsilon\right)\\
\leq& P\left(\sup_{|y-1|<k^{-1/4}}\left|\int_0^{\delta} W_R(s,y)-W_R(s,1)ds^{-1/\theta}\right|>\varepsilon\right)\\
&+P\left(\sup_{|y-1|<k^{-1/4}}\left|\int_{\delta}^{\infty} W_R(s,y)-W_R(s,1)ds^{-1/\theta}\right|>\varepsilon\right)\\
=&P\left(\sup_{0<s\leq 1,\frac{1}{2}\leq y\leq 2}\frac{|W_R(s,y)|}{s^{\eta}}>\frac{\varepsilon(\eta-1/\theta)}{4/\theta}\delta^{1/\theta-\eta}\right)\\
&+P\left(\sup_{s>0,|y-1|<k^{-1/4}}|W_R(s,y)-W_R(s,1)|\delta^{-1/\delta}>\frac{\varepsilon}{2}\right)\\
:=&I_{11}+I_{12}.
\end{aligned}$$

It follows from Lemma 4 that for any fixed ε, there is a positive $\omega(\varepsilon) = \varepsilon^{1/(\eta-1/\theta)}$, such that for all $\omega < \omega(\varepsilon)$, $I_{11} < \varepsilon$ holds. Furthermore, it is shown that $\sup_{s>0,|y-1|<k^{-1/4}} |W_R(s,y) - W_R(s,1)|\delta^{-1/\delta} > \frac{\epsilon}{2} \to 0$ almost surely in the proof of Proposition 3 in Cai et al. (2015), and thus we omit it here. Therefore, we have $I_{12} < \epsilon$, and also $I_1 \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$.

Next, we consider the term I_2. It follows from (S4.5) and Lemma 7 that

$$\begin{aligned}
&\frac{e_n n}{k}E\left(F_m(X_m)|F_j(X_j)>1-\frac{ke_n}{n}\right)\\
=&\int_0^{\infty} R_n(s_n(x),e_n)g_{\frac{n}{k}}(x)dx^{-1/\theta}
\end{aligned}$$

$$= \int_0^\infty R(x, e_n)\theta x^{1+1/\theta} dx^{-1/\theta} + o_p(1/\sqrt{k}), \tag{S4.20}$$

and

$$\begin{aligned} &\frac{n}{k} E\left(F_m(X_m) | F_j(X_j) > 1 - \frac{k}{n} \right) \\ =& \int_0^\infty R_n(s_n(x), 1) g_{\frac{n}{k}}(x) dx^{-1/\theta} \\ =& \int_0^\infty R(x, 1)\theta x^{1+1/\theta} dx^{-1/\theta} + o_p(1/\sqrt{k}). \end{aligned} \tag{S4.21}$$

By the definition of the function $R(\cdot, \cdot)$, $\int_0^\infty R(x, y) dx^{-1/\theta} = y^{1-1/\theta} \int_0^\infty R(x, 1) dx^{-1/\theta}$ holds for $y > 0$. This combines with (S4.20) and (S4.21) and we apply Cramér delta method,

$$\begin{aligned} &\frac{n}{\sqrt{k}} \left[e_n E(F_m(X_m) | F_j(X_j) > 1 - k e_n/n) - E(F_m(X_m) | F_j(X_j) > 1 - k/n) \right] \\ =& \sqrt{k} \left(\int_0^\infty R(x, e_n)\theta x^{1+1/\theta} dx^{-1/\theta} - \int_0^\infty R(x, 1)\theta x^{1+1/\theta} dx^{-1/\theta} + o_p(1/\sqrt{k}) \right) \\ =& \sqrt{k}(e_n^{1-1/\theta} - 1) \int_0^\infty R(x, 1)\theta x^{1+1/\theta} dx^{-1/\theta} + o_p(1) \\ \xrightarrow{\mathrm{P}}& (1/\theta - 1) W_R(\infty, 1) \int_0^\infty R(x, 1) x^{1+1/\theta} dx^{-1/\theta}, \end{aligned}$$

as $n \to \infty$. Thus, we have $I_2 \xrightarrow{\mathrm{P}} 0$. This completes the proof of Theorem 2.

■

S5 Proof of Theorem 3

Denote the event as $\xi_{1,t} = \{\max_{j,m\in\mathcal{C}_t} : |\widehat{\Gamma}_{jm} - \Gamma_{jm}| \leq \frac{\eta_{\min}}{2}\}$ and $\xi_1 = \cap_{t=0}^{T-2}\xi_{1,t}$, where we use the notation ξ_1^c to denote the complementary of the event ξ_1. Note that

$$
\begin{aligned}
&P(\widehat{\mathcal{L}} \neq \mathcal{L}) \\
=&P(\widehat{\mathcal{L}} \neq \mathcal{L}, \xi_1) + P(\widehat{\mathcal{L}} \neq \mathcal{L}, \xi_1^c) \\
\leq&P(\widehat{\mathcal{L}} \neq \mathcal{L}, \xi_1) + P(\xi_1^c) \\
\leq&P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0, \xi_1) + P(\widehat{\mathcal{A}}_1 \neq \mathcal{A}_1, \xi_1 | \widehat{\mathcal{A}}_0 = \mathcal{A}_0) + \ldots \\
&+ P(\widehat{\mathcal{A}}_{T-1} \neq \mathcal{A}_{T-1}, \xi_1 | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{T-2} = \mathcal{A}_{T-2}) + P(\xi_1^c). \qquad \text{(S5.22)}
\end{aligned}
$$

Clearly, if $\widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots$, and $\widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1}$ hold, then $\widehat{T} = T$.

We now bound the terms in (S5.22) by induction, and first deal with $P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0, \xi_1)$ as follows,

$$
\begin{aligned}
&P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0, \xi_1) \\
\leq&P\big(\text{there exists some } m \in \mathcal{A}_0 \text{ such that } |\max_{j\in\mathcal{C}_0} \widehat{\Gamma}_{jm} - 1| \leq \epsilon_0,\ \xi_1\big) \\
&+ P\big(\text{there exists some } m \in \mathcal{V}\backslash\mathcal{A}_0 \text{ such that } \min_{j\in\mathcal{C}_0} |\widehat{\Gamma}_{jm} - 1| > \epsilon_0,\ \xi_1\big) \\
=&\max_{m\in\mathcal{A}_0} P\big(|\max_{j\in\mathcal{C}_0} \widehat{\Gamma}_{jm} - 1| \leq \epsilon_0,\ \max_{j\in\mathcal{C}_0} |\widehat{\Gamma}_{jm} - \Gamma_{jm}| \leq \frac{\eta_{\min}}{2}, \Gamma_{jm} < 1\big) + 0 \\
=&0, \qquad \text{(S5.23)}
\end{aligned}
$$

where the first equality follows from the event ξ_1 and the choice of $\epsilon_0 = \frac{\eta_{\min}}{2}$, and the last equality is obtained by the fact that $\Gamma_{jm} < 1 - \eta_{\min}$ if $j \notin \text{an}_m$ and ϵ_0 is chosen such that $\Gamma_{jm} + \frac{\eta_{\min}}{2} < 1 - \frac{\eta_{\min}}{2} = 1 - \epsilon_0$.

Without loss of generality, we assume all the first t layers have been correctly recovered at the t-th step such that $\{\widehat{\mathcal{A}}_0, ..., \widehat{\mathcal{A}}_{t-1}\} = \{\mathcal{A}_0, ..., \mathcal{A}_{t-1}\}$. Then, at the $t+1$-th step, we turn to bound the following term,

$$
\begin{aligned}
&P\big(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t, \xi_1 | \widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\big) \\
\leq & P\big(\text{there exists some } m \in \mathcal{A}_t \text{ such that } |\widehat{F}_{jm} - 1| \leq \epsilon_t \text{ for some } j \in \mathcal{C}_t, \xi_1 \\
& \quad |\widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\big) \\
& + P\big(\text{there exists some } m \in \mathcal{C}_t \backslash \mathcal{A}_t \text{ such that } \min_{j \in \mathcal{C}_t} |\widehat{F}_{jm} - 1| > \epsilon_t, \ \xi_1 \\
& \quad\quad |\widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\big) \\
= & 0,
\end{aligned}
$$

where the inequality follows from a similar treatment as that of (S5.23).

Clearly, the upper bound of (S5.22) reduces to

$$
\begin{aligned}
P(\widehat{\mathcal{L}} \neq \mathcal{L}) \leq & P(\xi_1^c) \leq \sum_{t=0}^{T-2} P(\xi_{1,t}^c) = \sum_{t=0}^{T-2} P\left(\max_{j,m \in \mathcal{C}_t} |\widehat{\Gamma}_{jm} - \Gamma_{jm}| > \frac{\eta_{\min}}{2}\right) \\
\leq & Tp^2 \max_{j,m \in \mathcal{C}_t} P\left(|\widehat{\Gamma}_{jm} - \Gamma_{jm}| > \frac{\eta_{\min}}{2}\right) \\
\leq & C_0 T p^2 \sqrt{k}/n,
\end{aligned}
$$

for some positive constant C_0 and the last inequality follows from the Markov's inequality by Theorem 2. This completes the proof of Theorem 3. ■

In practice, the tuning parameter ϵ_t are chosen by the stability selection method (Sun et al., 2013), which is originally proposed to choose the tuning parameter for variable selection. Note that Assumptions 1 and 2 in Sun et al. (2013) can be verified in our case with $\lambda = 1/\epsilon_t$. For example, we illustrate that Assumption 1 in Sun et al. (2013) holds in terms of ϵ_0. Since Theorem 3 implies that the layer recovery consistency result holds if $\epsilon_0 \in (r_0, \eta_{\min}/2]$ for some $r_0 > 0$, we have $P(\widehat{\mathcal{A}}_{\epsilon_0{}^*} = \mathcal{A}_0) \geq 1 - a_n$ for some $a_n \to 0$ as $n \to \infty$ if $\epsilon_0{}^* \in (r_0, \eta_{\min}/2]$. Furthermore, this yields that $P(\cap_{\lambda_0 r_0 \leq \epsilon_0 \leq \epsilon_0^*}\{\widehat{\mathcal{A}}_{\epsilon_0} = \mathcal{A}_0\}) \geq 1 - c_0(\lambda_0)$ with $c_0(\lambda_0) = P(\text{there exists some } m \in \mathcal{A}_0 \text{ such that } \max_{j \in \widehat{\mathcal{C}}_1} |\widehat{\Gamma}_{jm} - 1| < \epsilon_0, \xi_1) \to 0$ as λ_0 diverges. Therefore, the required Assumption 1 in Sun et al. (2013) is verified.

S6 Proof of Proposition 1

Since a very appealing CIT measure aimed to the sub-exponential distributions is proposed in Azadkia and Chatterjee (2021), our goal is to extend the CIT measure to accommodate the heavy-tailed distributions. Denote $\widetilde{\mathbf{X}}_i^n$ as the nearest neighbor (NN) of $\mathbf{X}_i^n$ among $\mathbf{X}_1^n, \ldots, \mathbf{X}_{i-1}^n, \mathbf{X}_{i+1}^n, \ldots, \mathbf{X}_n^n$. Note

that Lemma 14.1 in Azadkia and Chatterjee (2021) establishes the upper bound for $E(\|\mathbf{X}_i^n - \widetilde{\mathbf{X}}_i^n\|)$ for each $i = 1, \ldots, n$, under the sub-exponential distribution assumption. Next, we first provide the corresponding result for the variables from the heavy-tailed distributions, given in Lemma 8 below.

Lemma 8. *Under the heavy-tailed distribution assumption* (2.2)*, there exists a positive constant* C_6 *which may rely on* c_j *in* (2.2) *and* $|\mathcal{S}_t|$ *such that*

$$E(\min\{\|\mathbf{X}_1^n - \widetilde{\mathbf{X}}_1^n\|, 1\}) \leq \begin{cases} C_6 n^{-1}(\log n)^3, & |\mathcal{S}_t| \in \{1, 2\}, \\ C_6 n^{-1/p}(\log n)^{|\mathcal{S}_t|}, & |\mathcal{S}_t| \geq 3. \end{cases}$$

Proof of Lemma 8: The proof of Lemma 8 is similar to that of Lemma 14.1 in Azadkia and Chatterjee (2021), and here we only point out the differences. Since the heavy-tailed assumption (2.2) implies that $P(X_j > t) \sim \sum_{k \in \mathrm{An}_j} \pi_{jk} h(t) t^{-\theta}$ as $t \to \infty$, for $\varepsilon \in (n^{-1/|\mathcal{S}_t|}, 1)$, any $t > 0$, and some constants $C_7, C_8 > 0$, we have

$$P(\|\mathbf{X}_1^n - \widetilde{\mathbf{X}}_1^n\| \geq \varepsilon) \leq C_7 \frac{1}{t^\theta} + (1-\delta)^{n-1} + C_8 t^{|\mathcal{S}_t|} \varepsilon^{-|\mathcal{S}_t|} \delta \leq \frac{C_5 (\log n)^{|\mathcal{S}_t|+1}}{n \varepsilon^{|\mathcal{S}_t|}},$$

with $\delta = C_9 \frac{\log n}{n^{2|\mathcal{S}_t|}}$ and $t = C_9 n \varepsilon^{|\mathcal{S}_t|}$ for some large enough $C_9 > 0$ and C_7 is determined by π_{jk} and $h(t)$. The completes the proof. ■

Note that we replace Lemma 14.1 in Azadkia and Chatterjee (2021)

with our modified Lemma 8 here, and thus we can apply other results in Azadkia and Chatterjee (2021) under the heavy-tailed distribution assumption in the proof of Proposition 1.

For convenience, we give some notations below. Let $\mathbf{X}_m^n = (X_{1,m}^n, \ldots, X_{n,m}^n)^T$ for $m = 1, \ldots, p$ and $\mathbf{X}_{\mathcal{S}_t\setminus\{j\}}^n = (\mathbf{X}_{1,\mathcal{S}_t\setminus\{j\}}^{n,T}, \ldots, \mathbf{X}_{n,\mathcal{S}_t\setminus\{j\}}^{n,T})^T$. Recall that the CIT measure $Q_{m,j,t}$ is defined in (3.6), we rewrite it as

$$Q_{m,j,t} := \frac{Y(X_m, X_j|\mathbf{X}_{\mathcal{S}_t\setminus\{j\}})}{Z(X_m, X_j|\mathbf{X}_{\mathcal{S}_t\setminus\{j\}})} := \frac{Y}{Z}.$$

For its estimator $\widehat{Q}_{m,j,t}$ given in Azadkia and Chatterjee (2021), we have

$$\begin{aligned}\widehat{Q}_{m,j,t} =& \frac{\sum_{i=1}^n \{\min(R_i, R_{NM(i)}) - \min(R_i, R_{NN(i)})\}}{\sum_{i=1}^n \{R_i - \min(R_i, R_{NN(i)})\}} \\ :=& \frac{Y_n(X_m, X_j|\mathbf{X}_{\mathcal{S}_t\setminus\{j\}})}{Z_n(X_m, X_j|\mathbf{X}_{\mathcal{S}_t\setminus\{j\}})} := \frac{Y_n}{Z_n},\end{aligned}$$

where R_i is the rank of $\mathbf{X}_{i,\mathcal{S}_t\setminus\{j\}}^n$, and $NN(i)$ and $NM(i)$ are the indices of NN of $\mathbf{X}_{i,\mathcal{S}_t\setminus\{j\}}^{n,T}$ and $\mathbf{X}_{i,\mathcal{S}_t}^{n,T}$, respectively, with

$$\begin{aligned}R_i &= \sum_{i_1=1}^n I(\mathbf{X}_{i_1,\mathcal{S}_t\setminus\{j\}}^{n,T} \leq \mathbf{X}_{i,\mathcal{S}_t\setminus\{j\}}^{n,T}), \\ NN(i) &= \big\{i_1 \neq i : \mathbf{X}_{i_1,\mathcal{S}_t\setminus\{j\}}^{n,T} \text{ is the NN of } \mathbf{X}_{i,\mathcal{S}_t\setminus\{j\}}^{n,T}\big\}, \\ NM(i) &= \big\{i_1 \neq i : \mathbf{X}_{i_1,\mathcal{S}_t}^{n,T} \text{ is the NN of } \mathbf{X}_{i,\mathcal{S}_t}^{n,T}\big\}.\end{aligned}$$

In practice, the asymptotic variance σ^2 can be estimated by nearest neighbor graphs guided by Lemma 3.2 and 3.3 in Shi et al. (2024) with $\sigma^2 = \frac{4}{5} + \frac{2}{5}(q_{|\mathcal{S}_t|} + q_{|\mathcal{S}_t|-1}) + \frac{4}{5}(o_{|\mathcal{S}_t|} + o_{|\mathcal{S}_t|-1})$ where q_h and o_h denote as the sample proportion of the nearest neighbor and the v structure in a h-dimensional vector, respectively. Denote $\mathcal{G}_{NN}$ as a directed nearest neighbor graph with n nodes and its edge sets $\mathcal{E}_{NN}$. The graph $\mathcal{G}_{NN}$ contains a directed edge from i_1 to i_2 if $\mathbf{J}_{i_1}$ is a nearest neighbor of $\mathbf{J}_{i_2}$ with $\mathbf{J} = \{\mathbf{J}_i \in \mathbb{R}^p, i = 1, \ldots, n\}$. Thus, σ^2 can be estimated with

$$\widehat{\sigma}^2 = \frac{4}{5} + \frac{2}{5}(\widehat{q}_{|\mathcal{S}_t|} + \widehat{q}_{|\mathcal{S}_t|-1}) + \frac{4}{5}(\widehat{o}_{|\mathcal{S}_t|} + \widehat{o}_{|\mathcal{S}_t|-1})$$

where $\widehat{q}_h = \widehat{E}(\frac{1}{n}\#\{(i_1, i_2) : i_1 \to i_2 \text{ and } i_2 \to i_1 \in \mathcal{E}_{NN}\})$, and $\widehat{o}_h = \widehat{E}(\frac{1}{n}\#\{(i_1, i_2, i_3) : i_1 \to i_3 \leftarrow i_2 \in \mathcal{E}_{NN}\})$ with $h \in \{|\mathcal{S}_t|, |\mathcal{S}_t| - 1\}$, which are the empirical versions of q_h and o_h, respectively. It is important to remark that the bootstrap method Efron (1992) can also be used to estimate the variance σ^2 with the extensive computational cost.

Proof of Proposition 1: Note that for any $\eta > 0$,

$$\begin{aligned} P(|\widehat{Q}_{m,j,t} - Q_{m,j,t}| \leq \eta) =& P\big(\big|\frac{Y_n - Y}{Z_n} - \frac{Y}{Z}\frac{Z_n - Z}{Z_n}\big| \leq \eta\big) \\ \geq& 1 - P\big(\big|\frac{Y_n - Y}{Z_n}\big| > \frac{\eta}{2}\big) - P\big(\big|\frac{Y}{Z}\frac{Z_n - Z}{Z_n}\big| > \frac{\eta}{2}\big) \end{aligned}$$

$$:=1-V_1-V_2. \tag{S6.24}$$

We first deal with the term V_1. It follows from Lemma 13.1 in Azadkia and Chatterjee (2021) that

$$\begin{aligned}
&Y_n(X_m, X_j|\mathbf{X}_{\mathcal{S}_t\setminus\{j\}}) = Y_n(X_m, \mathbf{X}_{\mathcal{S}_t}) - Y_n(X_m, \mathbf{X}_{\mathcal{S}_t\setminus\{j\}}),\\
&Y(X_m, X_j|\mathbf{X}_{\mathcal{S}_t\setminus\{j\}}) = Y(X_m, \mathbf{X}_{\mathcal{S}_t}) - Y(X_m, \mathbf{X}_{\mathcal{S}_t\setminus\{j\}}),
\end{aligned}$$

where

$$\begin{aligned}
Y(X_m, \mathbf{X}_{\mathcal{S}_t}) &= \int Var\left(P(X_m \geq t|\mathbf{X}_{\mathcal{S}_t})\right) dF_m(t),\\
Y_n(X_m, \mathbf{X}_{\mathcal{S}_t}) &= \frac{1}{n}\sum_{i=1}^{n} \min\{F_{n,m}(X_{i,m}^n), F_{n,m}(X_{N(i),m}^n)\} - G_{n,m}(X_{i,m}^n)^2,
\end{aligned} \tag{S6.25}$$

with $F_{n,m}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_{i,m}^n \leq t\}$, $G_{n,m}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_{i,m}^n \geq t\}$, and $N(i) = \{j : X_{j,m}^n$ is the nearest-neighbor of $X_{n,m}^n$ for$1 \leq j \leq n$ and $j \neq i\}$. Therefore, there holds that

$$\begin{aligned}
V_1 =& 1 - P(|Y_n - Y| \leq \frac{\eta}{2}|Z_n|)\\
=& 1 - P(|Y_n(X_m, \mathbf{X}_{\mathcal{S}_t}) - Y(X_m, \mathbf{X}_{\mathcal{S}_t})\\
&-[Y_n(X_m, \mathbf{X}_{\mathcal{S}_t\setminus\{j\}}) - Y(X_m, \mathbf{X}_{\mathcal{S}_t\setminus\{j\}})]\big| \leq \frac{\eta}{2}|Z_n|)
\end{aligned}$$

$$
\begin{aligned}
\leq & P(|Y_n(X_m, \mathbf{X}_{\mathcal{S}_t}) - Y(X_m, \mathbf{X}_{\mathcal{S}_t})| \geq \frac{\eta}{4}|Z_n|) \\
& + P(\big|Y_n(X_m, \mathbf{X}_{\mathcal{S}_t \setminus \{j\}}) - Y(X_m, \mathbf{X}_{\mathcal{S}_t \setminus \{j\}})\big| \geq \frac{\eta}{4}|Z_n|) \\
:= & V_{11} + V_{12}.
\end{aligned}
$$

Next, we first focus on the term V_{11}, and omit the proof details for V_{12} as they are similar. We define $Y'_n(X_m, \mathbf{X}_{\mathcal{S}_t})$ as the statistic by replacing the empirical distribution function in (S6.25) with the true one such that $Y'_n(X_m, \mathbf{X}_{\mathcal{S}_t}) = \frac{1}{n}\sum_{i=1}^n \min\{F_m(X^n_{i,m}), F_m(X^n_{N(i),m})\} - G_m(X^n_{i,m})^2$. Therefore, we have

$$
\begin{aligned}
V_{11} = & P\big(\big|Y_n(X_m, \mathbf{X}_{\mathcal{S}_t}) - E[Y_n(X_m, \mathbf{X}_{\mathcal{S}_t})] + E[Y_n(X_m, \mathbf{X}_{\mathcal{S}_t})] - E[Y'_n(X_m, \mathbf{X}_{\mathcal{S}_t})] \\
& + E[Y'_n(X_m, \mathbf{X}_{\mathcal{S}_t})] - Y(X_m, \mathbf{X}_{\mathcal{S}_t})\big| \geq \frac{\eta}{4}|Z_n|\big) \\
\leq & C_{11}\exp(-C_{12} n\eta^2 |Z_n|^2)
\end{aligned}
$$

holds for some positive constants $C_{11}, C_{12} > 0$ by Lemma 14.2 in Azadkia and Chatterjee (2021). It follows from Lemma 13.2 in Azadkia and Chatterjee (2021) that

$$
|Z_n| = \big|\frac{1}{n}\sum_{i=1}^n (\widehat{F}_{n,m}(X^n_{i,m}) - \min\{\widehat{F}_{n,m}(X^n_{i,m}), \widehat{F}_{n,m}(X^n_{N(i),m})\})\big|.
$$

Thus, the upper bound of $|Z_n|$ is 2 by applying triangle inequality. Mean-

while, the lower bound of $|Z_n|$ is obtained with $o(1/\sqrt{n}) + C_{10}$ by the convergence rate of the empirical cdf for some constant $C_{10} > 0$. This leads to that

$$V_{11} \leq C_{11} \exp(-C_{12} n\eta^2).$$

Similar to the treatment as in V_1, we consider the term V_2 and apply Lemmas 13.2 and 13.3 in Azadkia and Chatterjee (2021),

$$V_2 \leq C_{13} \exp(-C_{14} n\eta^2),$$

for constants $C_{13}, C_{14} > 0$. Then, it yields from (S6.24) that

$$\sup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} P(|\widehat{Q}_{m,j,t} - Q_{m,j,t}| > \eta) \leq V_1 + V_2 \leq C_{15} \exp(-C_{16} n\eta^2),$$

for constants $C_{15}, C_{16} > 0$. This completes the proof. ■

S7 Proof of Theorem 4

Note that

$$\begin{aligned} P(\widehat{\mathcal{G}} \neq \mathcal{G}) =& P(\widehat{\mathcal{G}} \neq \mathcal{G}, \widehat{\mathcal{L}} = \mathcal{L}) + P(\widehat{\mathcal{G}} \neq \mathcal{G}, \widehat{\mathcal{L}} \neq \mathcal{L}) \\ \leq& P(\widehat{\mathcal{G}} \neq \mathcal{G}, \widehat{\mathcal{L}} = \mathcal{L}) + P(\widehat{\mathcal{L}} \neq \mathcal{L}), \end{aligned} \tag{S7.26}$$

where the second term converges to 0 if $n = \Omega(T^{2/(2-\gamma)} p^{4/(2-\gamma)})$ by Theorem 3. Next, we focus on the first term in the right hand of (S7.26).

Specifically, we estimate directed edges by testing whether $m \in \mathcal{A}_t$ and $j \in \mathcal{S}_t$ are dependent while conditioning on $\mathcal{S}_t \backslash \{j\}$ for $t = 1, 2, \ldots, T-1$ if the event $\widehat{\mathcal{L}} = \mathcal{L}$ is given. The hypothesis test is given as in (3.5) and we denote $\mathcal{M}_{m,j}$ as the the event that we make an error while performing one CIT procedure for nodes m, j with its conditioning set $\mathcal{S}_t \backslash \{j\}$. Therefore, we obtain $\mathcal{M}_{m,j} = \mathcal{M}^I_{m,j} \cup \mathcal{M}^{II}_{m,j}$, where

$$\text{Type I error: } \mathcal{M}^I_{m,j} = \big\{\sqrt{n}|\widehat{Q}_{m,j,t}|/\widehat{\sigma}^2 > \Phi^{-1}(1-\alpha/2) \text{ and } Q_{m,j,t} = 0\big\},$$

$$\text{Type II error: } \mathcal{M}^{II}_{m,j} = \big\{\sqrt{n}|\widehat{Q}_{m,j,t}|/\widehat{\sigma}^2 \leq \Phi^{-1}(1-\alpha/2) \text{ and } Q_{m,j,t} \neq 0\big\}.$$

Therefore, we have

$$\begin{aligned} P(\widehat{\mathcal{G}} \neq \mathcal{G}, \widehat{\mathcal{L}} = \mathcal{L}) \leq & p^2 P(\cup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} \mathcal{M}_{m,j}) \\ \leq & p^2 \sup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} \big[P(\mathcal{M}^I_{m,j}) + P(\mathcal{M}^{II}_{m,j})\big]. \end{aligned}$$

Then, it suffices to bound both $P(\mathcal{M}^I_{m,j})$ and $P(\mathcal{M}^{II}_{m,j})$.

We choose $\alpha = \alpha_n = 2(1 - \Phi(\sqrt{n}\phi_n/\sigma^2))$ where ϕ_n satisfies Assumption 5 and obtain that

$$\begin{aligned} \sup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} P(\mathcal{M}^I_{m,j}) = & \sup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} P(\sqrt{n}|\widehat{Q}_{m,j,t} - Q_{m,j,t}|/\widehat{\sigma}^2 > \Phi^{-1}(1-\alpha/2)) \\ \leq & \sup_{m \in \mathcal{A}_t, j \in \mathcal{S}_t} P(|\widehat{Q}_{m,j,t} - Q_{m,j,t}| > \phi_n \widehat{\sigma}^2/\sigma^2), \end{aligned} \tag{S7.27}$$

and

$$\sup_{m\in\mathcal{A}_t,j\in\mathcal{S}_t} P(\mathcal{M}^{II}_{m,j}) = \sup_{m\in\mathcal{A}_t,j\in\mathcal{S}_t} P(\sqrt{n}|\widehat{Q}_{m,j,t}|/\widehat{\sigma}^2 \leq \Phi^{-1}(1-\alpha/2)) \\ \leq \sup_{m\in\mathcal{A}_t,j\in\mathcal{S}_t} P(|\widehat{Q}_{m,j,t} - Q_{m,j,t}| > \phi_n - \phi_n\widehat{\sigma}^2/\sigma^2). \tag{S7.28}$$

Combining (S7.27) and (S7.28), we apply Proposition 1,

$$P(\widehat{\mathcal{G}} \neq \mathcal{G}, \widehat{\mathcal{L}} = \mathcal{L}) \leq p^2 \sup_{m\in\mathcal{A}_t,j\in\mathcal{S}_t} P(\mathcal{M}_{mj}) \leq \exp\big(2\log p - n\phi_n^2(\widehat{\sigma}^2/\sigma^2)^2\big). \tag{S7.29}$$

Note that the convergence rate of the variance estimator $\widehat{\sigma}^2$ has not been directly established in the literature and only the consistency is derived in Devroye (1988) and Henze (1988). Since the expected number of the nearest neighbor pairs q_h and o_h in σ^2 is based on the the nearest neighbor distance, whose convergence rate is given in Biau and Devroye (2015), we can directly obtain that $\widehat{\sigma}^2/\sigma^2 = 1 + O_p(1/n^\xi)$ with $\xi \leq \min\big\{\frac{1}{2}, \frac{2}{|\mathcal{S}_t|}\big\}$, where ξ is related to the convergence rate of the variance estimator $\widehat{\sigma}^2$.

By Theorem 3, $P(\widehat{\mathcal{L}} \neq \mathcal{L}) \to 0$ if $2\log p \leq \frac{2-\gamma}{2}\log n < \frac{1}{2}\log n$ for $T \geq 1$. Combining with (S7.29) and the convergence rate of $\widehat{\sigma}^2$, we obtain

$$P(\widehat{\mathcal{G}} \neq \mathcal{G}, \widehat{\mathcal{L}} = \mathcal{L}) \leq \exp\big(\frac{1}{2}\log n - n^{1-2c}(1 + O_p(1/n^\xi))^2\big) \xrightarrow{\mathrm{P}} 0,$$

as $n \to \infty$ if $\frac{1-\xi}{2} < c < \frac{1}{2}$ for $\xi \in (0, \frac{1}{2}]$ holds in Assumption 5. This result combines with (S7.26) and yields that $P(\widehat{\mathcal{G}} \neq \mathcal{G}) \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$. This completes the proof of Theorem 4. ■

S8 Theoretical results for Real-valued coefficients

In this section, we relax the positive assumption of β_{jm} in (2.1), by allowing $\beta_{jm} \in \mathbb{R}$ for $j, m \in \mathcal{V}$. Different from (2.2), to describe both the upper and lower tails of the noise term, we define

$$P(\varepsilon_j > x) \sim c_{j,+} h(x) x^{-\theta}, \quad P(\varepsilon_j < -x) \sim c_{j,-} h(x) x^{-\theta}, \tag{S8.30}$$

as $x \to \infty$, $c_{j,+}, c_{j,-} > 0$ and $h(\cdot) \in \mathrm{RV}_0$. Similar to Gnecco et al. (2021), the causal tail coefficient matrix $\boldsymbol{\lambda} = (\lambda_{jm})_{j,m=1}^{p} \in \mathbb{R}^{p \times p}$ is given as,

$$\lambda_{jm} = \lim_{u \to 1^-} E(\mu(F_m(X_m)) | \mu(F_j(X_j)) > u), \tag{S8.31}$$

where we consider the effect of both tails and assume that the limit exists for the function $\mu : x \to |2x - 1|$. Then, λ_{jm} in (S8.31) is decomposed as

$$\begin{aligned} \lambda_{jm} = & \lim_{u \to 1^-} \frac{1}{2} E(\mu(F_m(X_m)) | F_j(X_j) > u) + \lim_{u \to 0^+} \frac{1}{2} E(\mu(F_m(X_m)) | F_j(X_j) < u) \\ := & \lambda_{jm,+} + \lambda_{jm,-}, \end{aligned} \tag{S8.32}$$

where $\lambda_{jm,+}$ and $\lambda_{jm,-}$ represent the extremely large and extremely small cases for X_j. Now, we give the following Theorem S1 to recover the topological layers of a heavy-tailed DAG for the real-valued coefficients based on $\boldsymbol{\lambda}$, similar to Theorem 1.

Theorem S1. *We consider the heavy-tailed linear SCM model in* (2.1) *with* $\beta_{jm} \in \mathbb{R}$. *Assume that* $\pi_{jm} \neq 0$ *if* m *is an ancestor of* j. *Given* $\mathcal{A}_0, \ldots, \mathcal{A}_{t-1}$, *we let* $\mathcal{C}_0 = \mathcal{V}$ *and* $\mathcal{C}_t = \mathcal{V} \backslash \bigcup_{d=0}^{t-1} \mathcal{A}_d$, *and then there holds that*

$$\mathcal{A}_t = \{m \in \mathcal{C}_t : \max_{j \in \mathcal{C}_t} \lambda_{jm} < 1\}.$$

We define the lower tail dependence between two random variables X_j and X_m as $L(x,y) = \lim_{t \to 0^+} \frac{1}{t} P(F_j(X_j) \leq tx, F_m(X_m) \leq ty)$ for $(x,y) \in [0,\infty]^2 \backslash \{(\infty,\infty)\}$ and $j, m \in \{1, \ldots, p\}$. Different from the right-hand heavy-tailed assumption in (2.2), the left-hand lower tail assumption in (S8.30) implies that $\lim_{t \to -\infty} F_j(tx)/F_j(t) = x^{-\theta}$ for $x > 0$ and we define $U_j^- = (1/F_j)^{\leftarrow}$.

Assumption S1. *There exist* $\tau_1', \tau_2' < 0$ *and* $\tau_3' < -1$ *such that as* $t \to \infty$,

$$\sup_{0<x<\infty, 1/2 \leq y \leq 2} |tP(F_j(X_j) \leq x/t, F_m(X_m) \leq y/t)/L(x,y) - 1| = O(t^{\tau_1'}),$$

$$\sup_{0<x<\infty} |g_t^-(x) + \theta x^{1+1/\theta}| = O(t^{\tau_2'}),$$

$$\Big|\frac{1}{2} E\left(1 - 2F_m(X_m) | F_j(X_j) < 1/t\right) - \lambda_{jm,-}\Big| = O(t^{(\tau_3'-1)/2}),$$

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

Assumption S2. *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 S3. *As* $n \to \infty$, $k = O(n^\gamma)$ *for some* γ *satisfying* $0 < \gamma < \min\left\{\frac{2\tau_1'}{2\tau_1'-1}, \frac{2\tau_2'}{2\tau_2'-1}, \frac{2\rho'}{2\rho'+\theta(\rho'-1)}\right\}$.

With slight modifications from Assumptions 1–3 in the main text, we propose Assumptions S1–S3 to deal with the case for the lower tail when the extreme small values occur.

Theorem S2. *Assume the heavy-tailed SCM* (2.1) *is encoded with the real-valued coefficients. Let* $\widehat{\lambda}_{jm}$ *be the estimator of* λ_{jm} *given in* (S8.31). *Under the conditions of Theorem 2 where* $\left|\frac{1}{2}E\left(2F_m(X_m) - 1|F_j(X_j) > 1 - 1/t\right) - \lambda_{jm,+}\right| = O(t^{(\tau_3-1)/2})$ *replaces* (4.9) *in Assumption 1, suppose that Assumptions S1–S3 also hold. Therefore, we have*

$$\frac{n}{\sqrt{k}}\left[\widehat{\lambda}_{jm} - \lambda_{jm}\right] \xrightarrow{\mathrm{d}} \widetilde{\Theta},$$

where $\widetilde{\Theta} = (1/\theta - 1)W_R(\infty, 1)\int_0^\infty [R(s,1) - L(s,1)]\,ds - 2\int_0^\infty W_R(s,1)ds^{-1/\theta}$.

Similar to Theorem 2, the asymptotic normality for $\widehat{\lambda}_{jm}$ is derived. Note that the limiting distribution Θ in Theorem 2 is slightly different from $\widetilde{\Theta}$ in Theorem S2, where both the left-hand lower tail dependence $L(\cdot,\cdot)$ and the

right-hand upper tail dependence $R(\cdot,\cdot)$ are introduced in the first term of $\widetilde{\Theta}$, as well as the constant 2 in the second term of $\widetilde{\Theta}$. This fact coincides with the decomposition of λ_{jm} in (S8.32).

Theorem S3. *Assume the heavy-tailed SCM* (2.1) *is encoded with real-valued coefficients. Let* $\widehat{\mathcal{G}}$ *be the estimated DAG based on* $\widehat{\boldsymbol{\lambda}}$*. Under the conditions of Theorem S2, we assume that Assumptions 4–5 hold. If* $n = \Omega(p^{4/(2-\gamma)})$*, we have*

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

Theorem S3 shows that the heavy-tailed DAG, encoded with a SCM (2.1) with real-valued coefficients, can be consistently recovered by the matrix $\widehat{\boldsymbol{\lambda}}$. It requires similar conditions to Theorem 4 and relaxes the positive assumption of direct causal effects, indicating the general applicability of our proposed method.

S9 Proof of Theorem S1

Lemma 9. *(Theorem 3 in Gnecco et al. (2021)) Under the heavy-tailed linear SCM model in* (2.1) *and the coefficient* $\beta_{jm} \in \mathbb{R}$*. We assume* $\pi_{jm} \neq 0$ *if* m *is an ancestor of* j*. For any two variables* X_j *and* X_m*, there holds that*

$$\lambda_{jm} = \frac{1}{2} + \frac{\sum_{l \in An_j \cap An_m} c_{jl,+} |\pi_{jl}|^\theta}{4 \sum_{l \in An_j} c_{jl,+} |\pi_{jl}|^\theta} + \frac{\sum_{l \in An_j \cap An_m} c_{jl,-} |\pi_{jl}|^\theta}{4 \sum_{l \in An_j} c_{jl,-} |\pi_{jl}|^\theta}, \tag{S9.33}$$

where $c_{jl,+} = c_{l,+}, c_{jl,-} = c_{l,-}$ *for* $\pi_{jl} > 0$ *and* $c_{jl,+} = c_{l,-}, c_{jl,-} = c_{l,+}$ *for* $\pi_{jl} < 0$

Further, Table S2 gives the corresponding values of λ_{jm} *when: (a)* $j \in An_m$*; (b)* $m \in An_j$*; (c)* $An_j \cap An_m = \emptyset$*; or (d)* $An_j \cap An_m \neq \emptyset$ *neither* $j \notin An_m$ *nor* $m \notin An_j$*.*

Table S2: Values of λ_{jm} and λ_{mj} under different scenarios, where $-$ indicates impossible scenarios under the heavy-tailed linear SCM model (2.1).

	$\lambda_{mj} = 1$	$\lambda_{mj} \in (1/2, 1)$	$\lambda_{mj} = 1/2$
$\lambda_{jm} = 1$	$-$	(a)	$-$
$\lambda_{jm} \in (1/2, 1)$	(b)	(d)	$-$
$\lambda_{jm} = 1/2$	$-$	$-$	(c)

Based on the result of Lemma 9, the proof of Theorem S1 is identical to that of Theorem 1, by replacing Γ_{jm} with λ_{jm} and referring to Lemma 9 rather than Lemma 1. Therefore, we omit the details here.

S10 Proof of Theorem S2

Note that

$$\begin{aligned}\widehat{\lambda}_{jm} - \lambda_{jm} =& (\widehat{\lambda}_{jm,+} + \widehat{\lambda}_{jm,-}) - (\lambda_{jm,+} + \lambda_{jm,-}) \\ =& (\widehat{\lambda}_{jm,+} - \lambda_{jm,+}) + (\widehat{\lambda}_{jm,-} - \lambda_{jm,-}). \end{aligned} \tag{S10.34}$$

To derive the asymptotic normality property of $\widehat{\lambda}_{jm}$, we need to establish the asymptotic normality of $\widehat{\lambda}_{jm,+}$ and $\widehat{\lambda}_{jm,-}$, respectively. We first focus

on $\widehat{\lambda}_{jm,+}$.

Lemma 10. *For both upper and lower tails, we have*

$$\lim_{u\to 1^-}\frac{1}{2(1-u)}E(2F_m(X_m)-1|F_j(X_j)>u)=\int_0^\infty R(1-F_m(s),1)dF_m(s), \tag{S10.35}$$

$$\lim_{u\to 0^+}\frac{1}{2u}E(1-2F_m(X_m)|F_j(X_j)<u)=\int_0^\infty L(F_m(s),1)dF_m(s). \tag{S10.36}$$

Proof of Lemma 10: Since the proof of Lemma 10 is similar to that of Lemma 2, we only show the differences and omit the details here. Note that

$$\begin{aligned}\frac{1}{2}E(2F_m(X_m)-1|F_j(X_j)>u)=&\frac{1}{2}\int_0^\infty 2P(X_m>s|F_j(X_j)>u)dF_m(s)\\ =&\int_0^\infty P(X_m>s|F_j(X_j)>u)dF_m(s),\end{aligned}$$

and

$$\begin{aligned}\frac{1}{2}E(1-2F_m(X_m)|F_j(X_j)<u)=&\frac{-2}{2}\int_{-\infty}^0 P(X_m<s|F_j(X_j)<u)dF_m(s)\\ =&\int_0^\infty P(X_m<s|F_j(X_j)<u)dF_m(s).\end{aligned}$$

By referring to the proof of Lemma 2, the desired result is obtained. This completes the proof of Lemma 10. ■

Similar to (S4.4), we define $\widetilde{\lambda}_{jm,+}^{1-ky/n}$ as a random function over the interval $[1/2, 2]$,

$$\widetilde{\lambda}_{jm,+}^{1-ky/n} = \frac{1}{2ky}\sum_{i=1}^{n}(2\widehat{F}_m(X_{i,m}^n) - 1)\mathbf{1}\big\{X_{i,j}^n > U_j(\frac{n}{ky})\big\}.$$

The asymptotic behaviour of $\widetilde{\lambda}_{jm,+}^{1-ky/n}$ is given in Lemma 11 by applying the results in Lemmas 3–5.

Lemma 11. *Suppose that Assumption 2 holds and for $\theta > 1$, we have*

$$\sup_{1/2\le y\le 2}\Big|\frac{n}{\sqrt{k}}\big(\widetilde{\lambda}_{jm,+}^{1-ky/n} - \frac{1}{2}E\big(2F_m(X_m) - 1|F_j(X_j) > 1 - \frac{ky}{n}\big)\big) + \frac{1}{y}\int_0^\infty W_R(x,y)dx^{-1/\theta}\Big| \xrightarrow{\mathrm{P}} 0.$$

Combining Lemmas 10–11 with Lemma 7, we similarly establish the asymptotic result for $\widetilde{\lambda}_{jm,+}$ as follows

$$\frac{n}{\sqrt{k}}(\widetilde{\lambda}_{jm,+} - \lambda_{jm,+}) \xrightarrow{\mathrm{d}} \Theta, \tag{S10.37}$$

where Θ is given in Theorem 2.

Next, we deal with $\widehat{\lambda}_{jm,-}$. Define the estimator as

$$\begin{aligned}\widehat{\lambda}_{jm,-}^{k/n} = \widehat{\lambda}_{jm,-} =& \frac{1}{2k}\sum_{i=1}^{n}(1 - 2\widehat{F}_m(X_{i,m}^n))\mathbf{1}\{X_{i,j}^n < X_{(k),j}^n\},\\ =& \frac{1}{2k}\sum_{i=1}^{n}(1 - 2\widehat{F}_m(X_{i,m}^n))\mathbf{1}\big\{X_{i,j}^n < U_j(\frac{n}{(n-k)e_n^-})\big\},\end{aligned}$$

since $e_n^- = \frac{n}{(n-k)}(1 - F_j(X^n_{(k),j})) \xrightarrow{\mathrm{P}} 1$ as $n \to \infty$. Furthermore, we denote $\widetilde{\lambda}^{1-k\cdot/n}_{jm,-}$ as a random function over the interval $[1/2, 2]$,

$$\widetilde{\lambda}^{ky/n}_{jm,-} = \frac{1}{2ky}\sum_{i=1}^{n}(1 - 2\widehat{F}_m(X^n_{i,m}))\mathbf{1}\big\{X^n_{i,j} < U_j\big(\frac{n}{(n-k)y}\big)\big\}.$$

We write $L_n(x_j, x_m) = (n/k)P(F_j(X_j) \le kx_j/n, F_m(X_m) \le kx_m/n)$ and its pseudo estimator is given as $V_n(x_j, x_m) = \frac{1}{k}\sum_{i=1}^{n}\mathbf{1}\{F_j(X_j) \le kx_j/n, F_m(X_m) \le kx_m/n\}$. Now, we first give the asymptotic behaviour of the pseudo estimator $V_n(\cdot,\cdot)$ in the following Lemma 12. Different from the right-hand heavy-tailed assumption in (2.2), the left-hand lower tail assumption in (S8.30) implies that $\lim_{t\to-\infty} F_j(tx)/F_j(t) = x^{-\theta}$ and we define $U_j^- = (1/F_j)^{\leftarrow}$. Denote $s_n^-(x) = (n/k)F_m\big(U_m^-\big(\frac{n}{k}\big)x^{-1/\theta}\big)$ and its derivative $g^-_{\frac{n}{k}}(x) = \frac{n}{k}U_m^-\big(\frac{n}{k}\big)f_m\big(U_m\big(\frac{n}{k}\big)x^{-1/\theta}\big)$ for $x < 0$. Note that $s_n^-(x) \to x$ as $n \to \infty$ implies that $g^-_{\frac{n}{k}}(x) \to -\theta x^{1+1/\theta}$.

Next, we first establish the asymptotic result for the pseudo estimator for the lower tail, similar to Lemma 3 in Cai et al. (2015).

Lemma 12. *Assume that this limit* $L(x, y) = \lim_{t\to0^+}\frac{1}{t}P(F_j(X_j) \le tx, F_m(X_m) \le ty)$ *exists for* $(x, y) \in [0,\infty]^2\backslash\{(\infty,\infty)\}$ *and* $j, m \in \{1, \ldots, p\}$. *For any* $\eta \in [0, \frac{1}{2})$ *and positive* ℓ, *with probability one, there holds that*

$$\sup_{x,y\in(0,\ell]}\Big|\frac{\sqrt{k}(V_n(x, y) - L_n(x, y)) - W_R(x, y)}{x^\eta}\Big| \to 0,$$

$$\sup_{x\in(0,\ell]}\Big|\frac{\sqrt{k}(V_n(x,\infty)-x)-W_R(x,\infty)}{x^{\eta}}\Big|\to 0,$$
$$\sup_{y\in(0,\ell]}\Big|\frac{\sqrt{k}(V_n(\infty,y)-y)-W_R(x,\infty)}{y^{\eta}}\Big|\to 0.$$

Since the lower tail dependence of the random vector (X_j, X_m) is equivalent to the upper tail dependence of the random vector $(-X_j, -X_m)$, we have

$$\begin{aligned}L(x,y)&=\lim_{t\to 0^+}\frac{1}{t}P(F_j(X_j)\le tx, F_m(X_m)\le ty)\\&=\lim_{t\to\infty}tP(F_j(-X_j)\le x/t, F_m(-X_m)\le y/t),\end{aligned}$$

which is the right-hand upper tail dependence between two random variables $-X_j$ and $-X_m$ by definition. Therefore, the proof of Lemma 12 is similar to that of Proposition 3.1 in Einmahl et al. (2006) , and thus omitted here.

Lemma 13. *Suppose that Assumption S2 holds and for $\theta > 1$, we have*

$$\sup_{1/2\le y\le 2}\Big|\frac{n}{\sqrt{k}}\big(\widetilde{\lambda}_{jm,-}^{ky/n}-\frac{1}{2}E\big(1-2F_m(X_m)|F_j(X_j)<\frac{ky}{n}\big)\big)+\frac{1}{y}\int_0^{\infty}W_R(x,y)dx^{-1/\theta}\Big|\xrightarrow{\mathrm{P}}0.$$

Proof of Lemma 13: Note that

$$E(1-2F_m(X_m)|F_j(X_j)<ky/n) = -2\int_{-\infty}^{0} P(X_m \le s|F_j(X_j)<ky/n)dF_m(s)$$
$$= -\frac{2}{ky/n}\int_{-\infty}^{0} P(F_m(X_m)\le F_m(s), F_j(X_j)\le ky/n)dF_m(s).$$

The above equality multiplies $y/2$ in both sides and we obtain that

$$\begin{aligned}
&\frac{y}{2}E(1-2F_m(X_m)|F_j(X_j)<ky/n)\\
=&-\frac{n}{k}\int_{-\infty}^{0} P(F_m(X_m)\le F_m(s), F_j(X_j)\le ky/n)dF_m(s)\\
=&-\frac{n}{k}\int_{-\infty}^{0} P(F_m(X_m)\le \frac{k}{n}\frac{nF_m(s)}{k}, F_j(X_j)\le \frac{ky}{n})dF_m(s)\\
=&-\int_{-\infty}^{0} L_n(\frac{nF_m(s)}{k}, y)dF_m(s)\\
=&\int_{0}^{\infty} L_n(\frac{n}{k}F_m(s), y)dF_m(s)\\
=&U_m^-(\frac{n}{k})\int_0^\infty L_n(s_n^-(x), y)f_m(U_m^-(\frac{n}{k})x^{-1/\theta})dx^{-1/\theta}, \qquad \text{(S10.38)}
\end{aligned}$$

where the last step follows from the fact that

$$\begin{aligned}
&\int_0^\infty L_n(s_n^-(x), y)f_m(U_m^-\left(\frac{n}{k}\right)x^{-1/\theta})dx^{-1/\theta}\\
=&\int_0^\infty L_n(\frac{n}{k}F_m(U_m^-(\frac{n}{k})x^{-1/\theta}), y)f_m(U_m^-(\frac{n}{k})x^{-1/\theta})dx^{-1/\theta}\\
=&\frac{1}{U_m^-(\frac{n}{k})}\int_0^\infty L_n(\frac{n}{k}F_m(s), y)dF_m(s).
\end{aligned}$$

Similar to (S10.38), we also have

$$y\widetilde{\lambda}_{jm,-}^{ky/n} = U_m^-(\frac{n}{k})\int_0^\infty V_n(s_n^-(x),y)f_m(U_m^-(\frac{n}{k})x^{-1/\theta})dx^{-1/\theta}.$$

Similar to in Lemma 6, for any $L > 0$,

$$\begin{aligned}
&\sup_{1/2\le y\le 2}\Big|\frac{n}{\sqrt{k}}(y\widetilde{\lambda}_{jm,-}^{ky/n} - \frac{y}{2}E(1-2F_m(X_m)|F_j(X_j) < \frac{ky}{n})) \\
&\qquad + \int_0^\infty W_R(x,y)dx^{-1/\theta}\Big| \\
=&\sup_{1/2\le y\le 2}\Big|\int_0^\infty \sqrt{k}\big(V_n(s_n^-(x),y) - L_n(s_n^-(x),y)\big)g_{\frac{n}{k}}^-(x)dx^{-1/\theta} \\
&\qquad + \int_0^\infty W_R(x,y)dx^{-1/\theta}\Big| \\
\le&\sup_{1/2\le y\le 2}\Big|\int_L^\infty W_R(x,y)dx^{-1/\theta}\Big| \\
&+\sup_{1/2\le y\le 2}\Big|\int_L^\infty \sqrt{k}\big(V_n(s_n^-(x),y) - L_n(s_n^-(x),y)\big)g_{\frac{n}{k}}^-(x)dx^{-1/\theta}\Big| \\
&+\sup_{1/2\le y\le 2}\Big|\int_0^L \sqrt{k}\big(V_n(s_n(x),y) - L_n(s_n(x),y)\big)g_{\frac{n}{k}}^-(x) - W_R(x,y)dx^{-1/\theta}\Big| \\
:=&K_1^-(L) + K_{2,n}^-(L) + K_{3,n}^-(L).
\end{aligned}$$

Similar to (S4.6) in Lemma 6, the three terms $K_1^-(L), K_{2,n}^-(L), K_{3,n}^-(L)$ can be bounded by using the same technique in the proof of Lemma 6 by using Assumption S2 and by referring to Lemmas 4, 5, and 12. Therefore, we omit the details here. This completes the proof of Lemma 13. ■

Lemma 14. *Suppose that Assumptions 1 and 3 are satisfied. Therefore, there holds that*

$$\sup_{1/2\le y\le 2}\sqrt{k}\Big|\int_0^{\infty} L_n(s_n^-(x),y)g_{\frac{n}{k}}^-(x)+L(x,y)\theta x^{1+1/\theta}dx^{-1/\theta}\Big|\to 0,\ \ as\ n\to\infty.$$

Similar to Lemma 7, we establish the non-stochastic limit relationship result in Lemma 14. This combines with Lemma 13 and the proof of Theorem 2, and thus we derive the asymptotic normality of $\widehat{\lambda}_{jm,-}$ similarly. Here, we only provide the proof sketch and list the main differences below.

Proof of Lemma 14: Different from (S4.20) and (S4.21), it follows from (S10.38) and Lemma 14 that for lower tails

$$\begin{aligned}&\frac{e_n n}{2k}E(1-2F_m(X_m)|F_j(X_j)<\frac{ke_n}{n})\\ =&\int_0^{\infty} L_n(s_n^-(x),e_n)g_{\frac{n}{k}}^-(x)dx^{-1/\theta}=-\int_0^{\infty} L(x,e_n)\theta x^{1+1/\theta}dx^{-1/\theta}+o_p(1/\sqrt{k}),\end{aligned}$$

and

$$\begin{aligned}&\frac{n}{2k}E(1-2F_m(X_m)|F_j(X_j)<\frac{k}{n})\\ =&\int_0^{\infty} L_n(s_n^-(x),1)g_{\frac{n}{k}}^-(x)dx^{-1/\theta}=-\int_0^{\infty} L(x,1)\theta x^{1+1/\theta}dx^{-1/\theta}+o_p(1/\sqrt{k}).\end{aligned}$$

This yields that

$$\begin{aligned}&\frac{n}{\sqrt{k}}\big[\frac{e_n^-}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{ke_n^-}{n})-\frac{1}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{k}{n})\big]\\ =&\sqrt{k}\big[\int_0^{\infty} L(x,1)\theta x^{1+1/\theta}dx^{-1/\theta}-\int_0^{\infty} L(x,e_n^-)\theta x^{1+1/\theta}dx^{-1/\theta}+o_p(1/\sqrt{k})\big]\end{aligned}$$

$$
\begin{aligned}
&=\sqrt{k}(1-(e_n^-)^{1-1/\theta})\int_0^\infty L(x,1)\theta x^{1+1/\theta}dx^{-1/\theta}+o_p(1)\\
&\xrightarrow{\mathrm{P}}(1-1/\theta)W_R(\infty,1)\int_0^\infty L(x,1)x^{1+1/\theta}dx^{-1/\theta}. \qquad \text{(S10.39)}
\end{aligned}
$$

Combining with Lemma 13, (S10.39), and Assumption S1, thus we have

$$
\begin{aligned}
&\frac{n}{\sqrt{k}}(\widehat{\lambda}_{jm,-}-\lambda_{jm,-})\\
=&\frac{n}{\sqrt{k}}(\widehat{\lambda}_{jm,-}-\frac{1}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{k}{n}))\\
&+\frac{n}{\sqrt{k}}(\frac{1}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{k}{n})-\lambda_{jm,-})\\
=&\frac{n}{\sqrt{k}}\big\{e_n^-\widetilde{\lambda}_{jm,-}^{ke_n^-/n}-\frac{e_n^-}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{ke_n^-}{n})\big\}\\
&+\frac{n}{\sqrt{k}}\big\{\frac{e_n^-}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{ke_n^-}{n})\\
&\qquad\qquad-\frac{1}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{k}{n})\big\}\\
&+\frac{n}{\sqrt{k}}\big\{\frac{1}{2}E(1-2F_m(X_m)|F_j(X_j)<\frac{k}{n})-\lambda_{jm,-}\big\}\\
\xrightarrow{\mathrm{d}}&-\int_0^\infty W_R(x,1)dx^{-1/\theta}+(1-1/\theta)W_R(\infty,1)\int_0^\infty L(x,1)x^{1+1/\theta}dx^{-1/\theta}.
\end{aligned}
\qquad \text{(S10.40)}
$$

As a consequence, it follows from (S10.34), (S10.37), and (S10.40) that

$$
\begin{aligned}
&\frac{n}{\sqrt{k}}(\widehat{\lambda}_{jm}-\lambda_{jm})\\
\xrightarrow{\mathrm{d}}&(1/\theta-1)W_R(\infty,1)\int_0^\infty(R(s,1)-L(s,1))ds-2\int_0^\infty W_R(s,1)ds^{-1/\theta}.
\end{aligned}
$$

This completes the proof of Theorem S2. ■

Proof of Theorem S3: It follows from Theorem S2 that the topological layers of a heavy-tailed DAG can be exactly recovered based on $\widehat{\boldsymbol{\lambda}}$, and the proof is similar to that of Theorem 3 with slight modifications. Then, this combines with the proof of Theorem 4, which leads to Theorem S3, and thus we omit it here. ■

S11 Additional Figures and Tables

To select an optimal value of k for estimating causal tail dependence $\boldsymbol{\Gamma}$, we choose $k = \lfloor n^{\gamma} \rfloor$ with different choices of $\gamma \in \{0.2, 0.25, \cdots, 0.7\}$ where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. We consider a hub graph and generate the data from the Student-t distribution with degree of freedom $\{1, 2, 3\}$ and Cauchy distribution with scale $\{1, 3, 9\}$ and location 3, respectively, to evaluate the performance metrics of TopHeat in Figure S1. It is worth pointing out better performance requires larger k for the Student-t distribution with smaller degree of freedom, and also for the Cauchy distribution with different scales. This fact reflects that the best choice for γ largely depends on the tail heaviness of the noise terms, which also corresponds to Assumption 3 where $k/n \to 0$ and $k \to \infty$ as $n \to \infty$. In practice, we take $k = \lfloor n^{0.5} \rfloor$ since it is located within the best range of γ

under different settings.

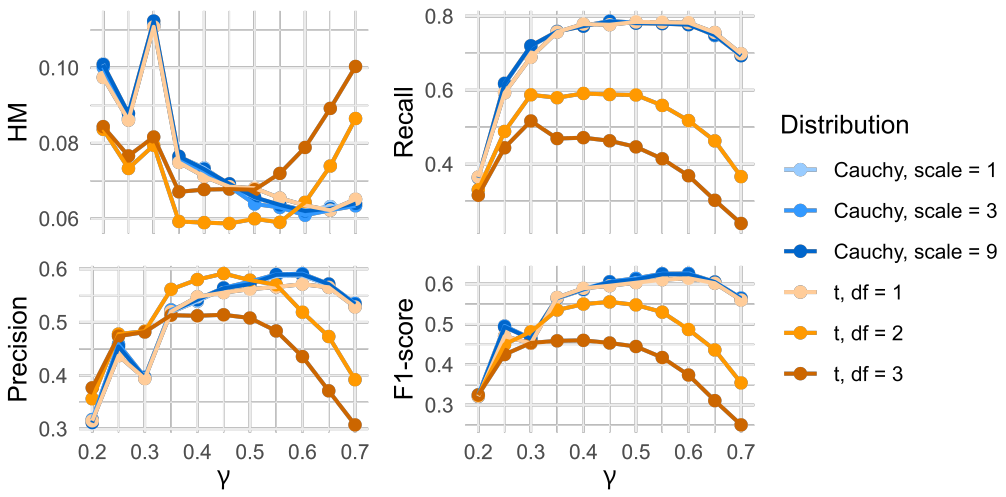
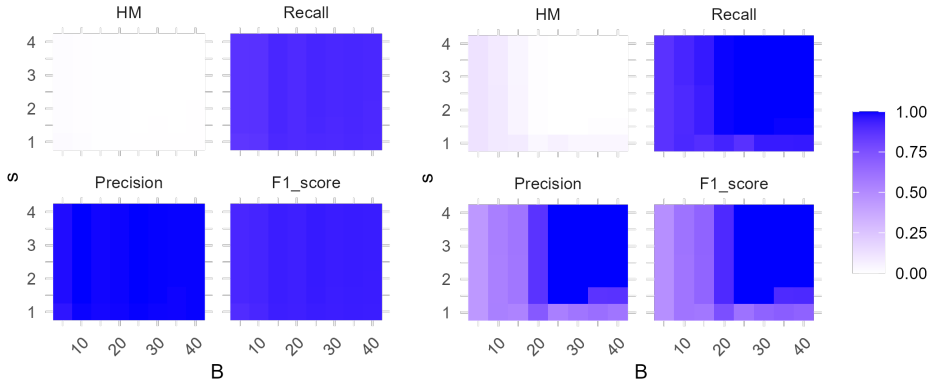


Figure S1: The figure presents the results of HM, Recall, Precision, and F1-score of TopHeat for $k = \lfloor n^{\gamma} \rfloor$ with different values of $\gamma \in \{2.0, 2.5, \cdots, 7\}$. Each point records the average of 50 repeated experiments and 6 settings with $n \in \{500, 1000\}$ and $p \in \{5, 10, 50\}$.

To determine the parameters (a, B) in the stability selection method, we choose the tuning parameter ϵ_t from the grids $\{10^{-2+0.05s}, s = 0, 1, \ldots, 35\}$. It is illustrated in Figure S2 that the effects of different values of a and B on the estimation accuracy is negligible when p and n are relatively small, while smaller a and larger B lead to better performance when p and n are larger. This coincides with the fact that the threshold a and the number of repetition B vary from p and n, which is supported by the discussion in Section S5. Therefore, 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 simulated examples.



(a) $(n, p) = (500, 5)$ (b) $(n, p) = (5000, 50)$

Figure S2: The figure displays a heatmap of performance metrics for the TopHeat algorithm, with the varying repetitions $\mathrm{B} \in \{5, 10, \cdots, 40\}$ and the thresholds $a \in \{10^{-s}, s = 1, 1.5, \cdots, 4\}$ in stability selection. Figure S2(a) in the left panel and Figure S2(b) in the right panel correspond to the estimation accuracy results of a hub graph with $(n, p) = (500, 5)$ and $(n, p) = (5000, 50)$, respectively.

During the CIT procedure of the TopHeat algorithm, a random error is considered as the conditional variable when $|\widehat{\mathcal{S}}_t| = 1$ and we choose it follows from a standard normal distribution. 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 error distribution, including Gaussian, Student-t, Cauchy, and Uniform. However, Gaussian-distributed errors yield more robust performance compared with other distributions.

Furthermore, we investigate the performance metrics of TopHeat for different significant levels α for CIT methods in Figure S4. While considering the Student-t distribution with 1 degree of freedom and the averaged

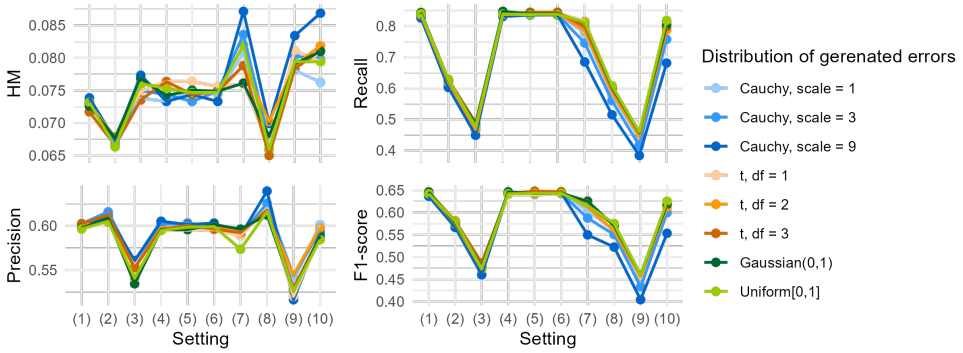


Figure S3: The figure illustrates the estimation results of TopHeat with random errors drawn from various distributions, depicted by different colored lines. The horizontal axis delineates 10 distinct settings: (1) t distribution, df=1, linear; (2) t distribution, df=2, linear; (3) t distribution, df=3, linear; (4) Cauchy, scale=1, linear; (5) Cauchy, scale=3, linear; (6) Cauchy, scale=9, linear; (7) t distribution, df=1, nonlinear; (8) t distribution, df=2, nonlinear; (9) t-distribution, df=3, nonlinear; (10) Cauchy, scale=1, nonlinear. Each point represents the average of 50 repeated experiments across 6 combinations with $n \in \{500, 1000\}$ and $p \in \{5, 10, 50\}$.

metrics of three different graphs, we observe that smaller α leads to better performance for a small graph with $p = 5$, and larger α yields higher accuracy of graph estimation for large graphs when $p \in 20, 50$. The results indicate that a uniform significant level only controls node-wise false discoveries, but fails in the graph-wise manner, which is supported by Li and Maathuis (2021). Theoretically, this finding agrees with Assumption 5, where α is correlated with ϕ_n. In the sequel, we also consider that the significance level α should be smaller and tend towards zero as p and n approach infinity, and choose $(\alpha, p) \in \{(10^{-2}, 5), (10^{-5}, 20), (10^{-10}, 50)\}$ in simulation studies.

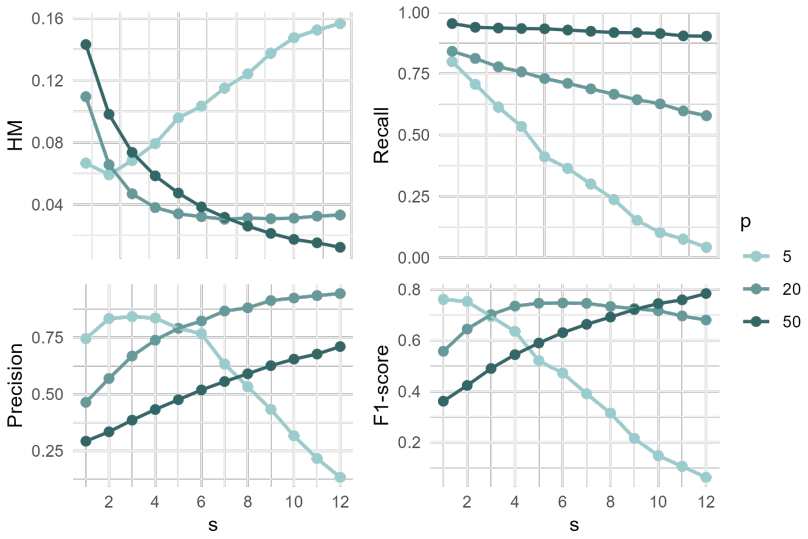


Figure S4: The figure shows the estimation of performance metrics of TopHeat with different α with $\alpha = 10^{-s}, s \in \{1, \cdots, 12\}$. Each point records the average of 30 repeated experiments under different settings.

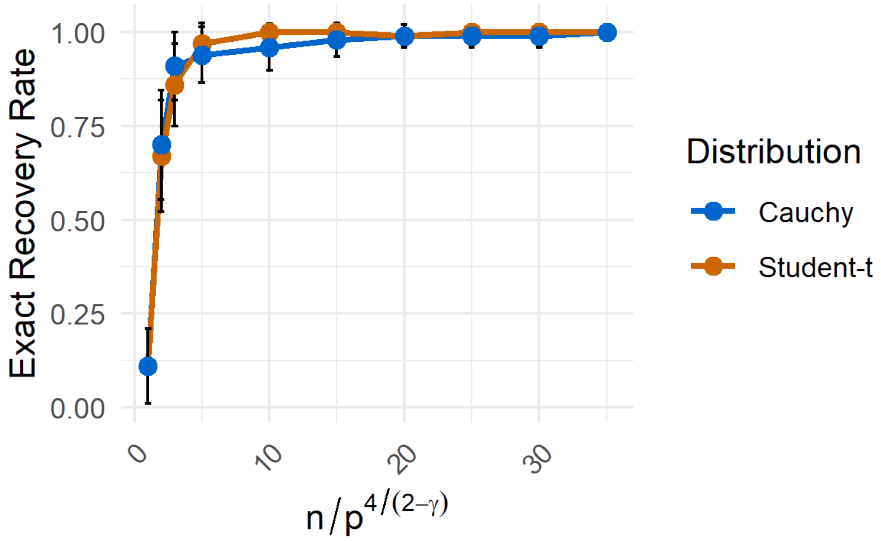


Figure S5: The figure shows the averaged exact recovery rates of the proposed TopHeat method and their standard errors over 100 repetations.

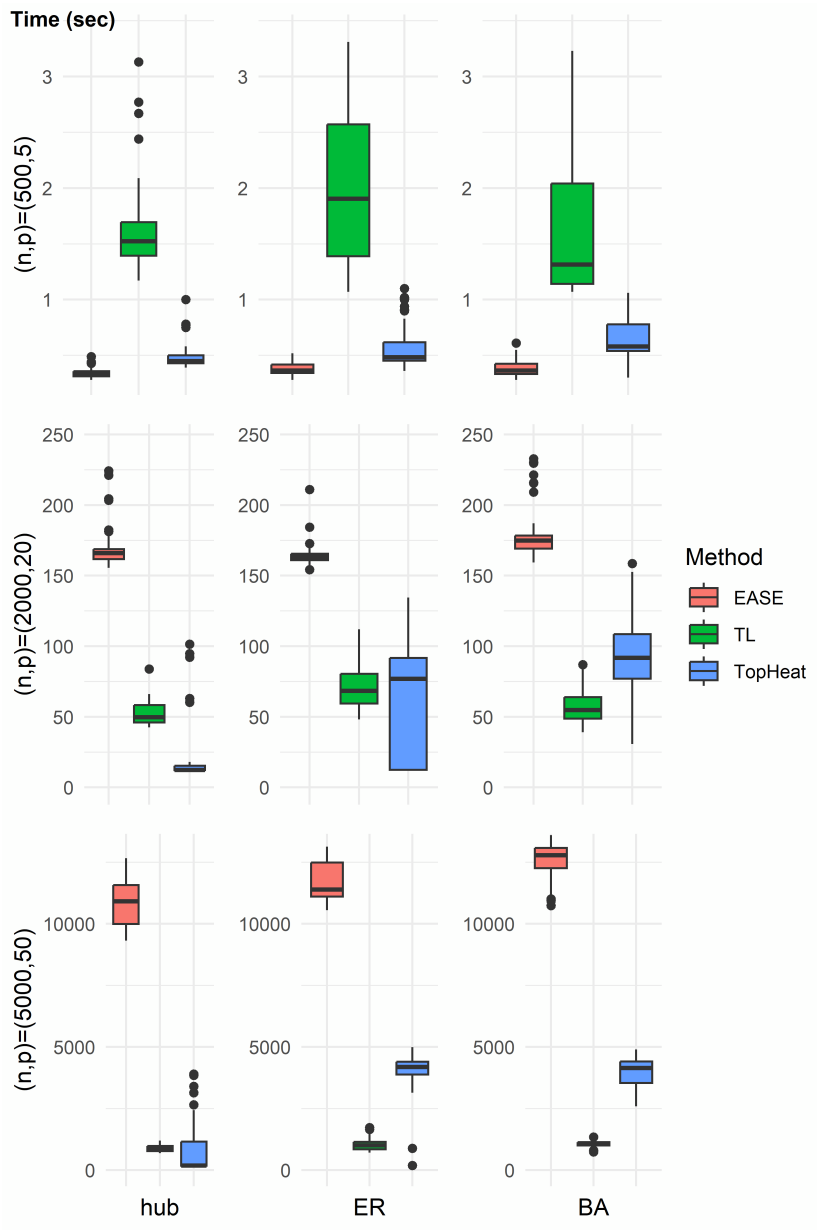


Figure S6: The time complexity of TopHeat, EASE, and TL in terms of the running time for different graphs. Each column corresponds to one graph setting for the Student-t distribution under different n and p, and three types of graphs in Examples 1–3 are considered.

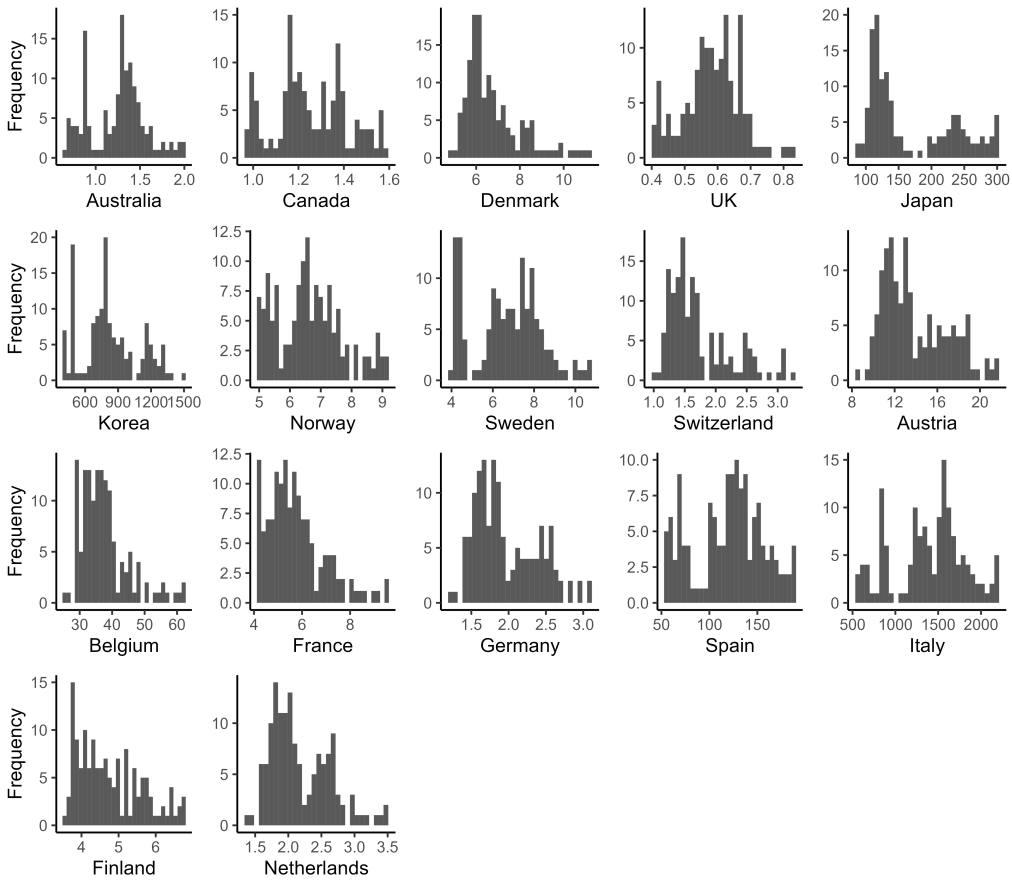


Figure S7: The histograms of exchange rates for 17 currencies.

Table S3: The averaged performance metrics of various methods, as well as their standard errors in parentheses, are presented for a BA graph in Example 2 with the Student-t distribution.

Model	(n, p)	Methods	HM (%)	Recall	Precision	F1-score
linear	$(500, 5)$	TopHeat	7.80(0.93)	0.76(0.03)	0.85(0.03)	0.79(0.03)
		EASE	8.70(0.98)	0.79(0.04)	0.81(0.03)	0.77(0.03)
		TL	7.40(1.07)	0.64(0.05)	0.84(0.05)	0.70(0.05)
		Directed-LiNGAM	2.50(0.59)	0.98(0.01)	0.92(0.02)	0.95(0.01)
		ICA-LiNGAM	4.50(1.01)	0.94(0.02)	0.87(0.03)	0.90(0.02)
		HD-LiNGAM	51.20(1.58)	0.47(0.04)	0.19(0.02)	0.27(0.02)
		Rank PC	28.40(1.20)	0.33(0.03)	0.31(0.03)	0.32(0.03)
	$(2000, 20)$	TopHeat	1.87(0.70)	0.84(0.13)	0.78(0.08)	0.79(0.10)
		EASE	2.88(0.27)	0.56(0.05)	0.82(0.05)	0.66(0.04)
		TL	2.36(0.19)	0.55(0.04)	0.92(0.02)	0.66(0.04)
		Directed-LiNGAM	8.43(1.60)	0.98(0.02)	0.40(0.05)	0.56(0.05)
		ICA-LiNGAM	9.41(1.68)	0.95(0.07)	0.37(0.05)	0.52(0.05)
		HD-LiNGAM	50.36(0.67)	0.46(0.07)	0.05(0.01)	0.08(0.01)
		Rank PC	9.09(0.31)	0.11(0.03)	0.11(0.03)	0.11(0.03)
	$(5000, 50)$	TopHeat	0.86(0.11)	0.94(0.02)	0.75(0.02)	0.82(0.02)
		EASE	1.51(0.03)	0.42(0.01)	0.71(0.01)	0.53(0.01)
		TL	1.00(0.07)	0.53(0.03)	0.94(0.01)	0.65(0.03)
		Directed-LiNGAM	9.31(0.45)	0.99(0.00)	0.19(0.01)	0.31(0.01)
		ICA-LiNGAM	11.03(0.48)	0.88(0.03)	0.15(0.01)	0.26(0.01)
		HD-LiNGAM	50.05(0.06)	0.49(0.02)	0.02(0.00)	0.04(0.00)
		Rank PC	3.75(0.02)	0.06(0.00)	0.06(0.00)	0.06(0.00)
nonlinear	$(500, 5)$	TopHeat	8.30(1.01)	0.71(0.04)	0.86(0.03)	0.76(0.03)
		EASE	9.20(0.98)	0.74(0.04)	0.80(0.03)	0.75(0.03)
		TL	20.20(0.20)	0.01(0.01)	0.03(0.02)	0.02(0.01)
		Directed-LiNGAM	40.30(2.38)	0.29(0.03)	0.23(0.03)	0.25(0.03)
		ICA-LiNGAM	21.20(2.28)	0.60(0.04)	0.54(0.04)	0.56(0.04)
		HD-LiNGAM	23.80(0.74)	0.28(0.02)	0.37(0.02)	0.32(0.02)
		Rank PC	28.40(1.20)	0.33(0.03)	0.31(0.03)	0.32(0.03)
	$(2000, 20)$	TopHeat	1.92(0.23)	0.84(0.04)	0.78(0.02)	0.79(0.03)
		EASE	2.86(0.10)	0.56(0.02)	0.83(0.02)	0.65(0.01)
		TL	5.00(0.00)	0.00(0.00)	0.03(0.02)	0.01(0.00)
		Directed-LiNGAM	13.61(0.66)	0.41(0.02)	0.19(0.02)	0.25(0.02)
		ICA-LiNGAM	9.06(0.72)	0.58(0.03)	0.35(0.03)	0.43(0.03)
		HD-LiNGAM	8.03(0.08)	0.17(0.01)	0.18(0.01)	0.18(0.01)
		Rank PC	9.09(0.09)	0.11(0.01)	0.11(0.01)	0.11(0.01)
	$(5000, 50)$	TopHeat	0.86(0.11)	0.94(0.02)	0.75(0.02)	0.82(0.02)
		EASE	1.43(0.03)	0.46(0.01)	0.73(0.01)	0.56(0.01)
		TL	2.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	7.14(0.42)	0.50(0.02)	0.17(0.02)	0.25(0.02)
		ICA-LiNGAM	6.71(0.41)	0.55(0.02)	0.20(0.02)	0.28(0.02)
		HD-LiNGAM	3.36(0.03)	0.15(0.01)	0.15(0.01)	0.15(0.01)
		Rank PC	3.75(0.02)	0.06(0.00)	0.06(0.00)	0.06(0.00)

Table S4: The averaged performance metrics of various methods, as well as their standard errors in parentheses, are presented for the ER graph in Example 3 with the Student-t distribution.

Model	(n, p)	Methods	HM (%)	Recall	Precision	F1-score
linear	$(500, 5)$	TopHeat	6.70(0.93)	0.68(0.05)	0.75(0.05)	0.68(0.04)
		EASE	5.10(0.68)	0.82(0.04)	0.77(0.04)	0.77(0.04)
		TL	5.10(0.84)	0.59(0.06)	0.75(0.06)	0.63(0.06)
		Directed-LiNGAM	7.20(0.73)	0.92(0.04)	0.62(0.04)	0.72(0.04)
		ICA-LiNGAM	8.60(0.85)	0.87(0.04)	0.59(0.04)	0.68(0.04)
		HD-LiNGAM	51.50(1.31)	0.42(0.05)	0.11(0.02)	0.17(0.02)
		Rank PC	19.00(1.76)	0.28(0.05)	0.26(0.05)	0.27(0.05)
	$(2000, 20)$	TopHeat	0.69(0.13)	0.86(0.03)	0.90(0.02)	0.86(0.03)
		EASE	0.65(0.06)	0.79(0.02)	0.97(0.01)	0.87(0.01)
		TL	1.73(0.12)	0.39(0.04)	0.85(0.04)	0.50(0.04)
		Directed-LiNGAM	9.23(0.46)	0.99(0.01)	0.24(0.01)	0.38(0.01)
		ICA-LiNGAM	10.53(0.45)	0.96(0.02)	0.21(0.01)	0.34(0.01)
		HD-LiNGAM	49.85(0.11)	0.53(0.02)	0.03(0.00)	0.05(0.00)
		Rank PC	4.51(0.23)	0.30(0.02)	0.24(0.02)	0.26(0.02)
	$(5000, 50)$	TopHeat	0.13(0.01)	0.94(0.01)	0.94(0.01)	0.94(0.01)
		EASE	0.64(0.02)	0.39(0.01)	0.94(0.01)	0.55(0.01)
		TL	0.71(0.04)	0.38(0.03)	0.85(0.03)	0.50(0.03)
		Directed-LiNGAM	9.49(0.37)	0.99(0.00)	0.10(0.00)	0.18(0.01)
		ICA-LiNGAM	10.91(0.31)	0.90(0.03)	0.08(0.00)	0.15(0.01)
		HD-LiNGAM	50.02(0.03)	0.49(0.02)	0.01(0.00)	0.02(0.00)
		Rank PC	1.57(0.07)	0.24(0.01)	0.24(0.01)	0.24(0.01)
nonlinear	$(500, 5)$	TopHeat	7.30(0.92)	0.61(0.05)	0.69(0.05)	0.62(0.05)
		EASE	6.40(0.80)	0.75(0.05)	0.74(0.05)	0.72(0.04)
		TL	11.50(0.99)	0.10(0.04)	0.14(0.05)	0.11(0.04)
		Directed-LiNGAM	22.40(1.49)	0.22(0.04)	0.18(0.03)	0.20(0.04)
		ICA-LiNGAM	16.60(1.57)	0.44(0.05)	0.36(0.05)	0.39(0.05)
		HD-LiNGAM	21.50(1.01)	0.25(0.04)	0.21(0.03)	0.22(0.03)
		Rank PC	19.00(1.76)	0.28(0.05)	0.26(0.05)	0.27(0.05)
	$(2000, 20)$	TopHeat	0.65(0.11)	0.85(0.03)	0.91(0.02)	0.87(0.03)
		EASE	1.06(0.09)	0.64(0.02)	0.97(0.01)	0.76(0.02)
		TL	2.66(0.11)	0.03(0.01)	0.14(0.05)	0.05(0.02)
		Directed-LiNGAM	5.68(0.25)	0.24(0.02)	0.16(0.01)	0.19(0.02)
		ICA-LiNGAM	2.62(0.20)	0.64(0.03)	0.52(0.03)	0.57(0.03)
		HD-LiNGAM	5.76(0.11)	0.31(0.02)	0.18(0.01)	0.22(0.01)
		Rank PC	4.51(0.23)	0.30(0.02)	0.24(0.02)	0.26(0.02)
	$(5000, 50)$	TopHeat	0.13(0.01)	0.93(0.01)	0.94(0.01)	0.94(0.01)
		EASE	0.58(0.03)	0.45(0.01)	0.96(0.01)	0.60(0.01)
		TL	1.01(0.03)	0.00(0.00)	0.02(0.02)	0.00(0.00)
		Directed-LiNGAM	2.32(0.10)	0.26(0.01)	0.15(0.01)	0.19(0.01)
		ICA-LiNGAM	1.39(0.10)	0.60(0.02)	0.42(0.02)	0.49(0.02)
		HD-LiNGAM	2.48(0.02)	0.24(0.01)	0.13(0.01)	0.16(0.01)
		Rank PC	1.57(0.07)	0.24(0.01)	0.24(0.01)	0.24(0.01)

Table S5: 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 Cauchy distribution.

Model	(n, p)	Methods	HM (%)	Recall	Precision	F1-score
linear	$(500, 5)$	TopHeat	4.00(0.70)	0.86(0.03)	0.95(0.02)	0.89(0.02)
		EASE	11.80(0.67)	0.58(0.02)	0.82(0.03)	0.66(0.02)
		TL	6.30(1.25)	0.68(0.06)	0.72(0.06)	0.70(0.06)
		Directed-LiNGAM	5.10(0.71)	0.98(0.01)	0.83(0.02)	0.89(0.01)
		ICA-LiNGAM	4.70(0.63)	0.99(0.00)	0.84(0.02)	0.90(0.01)
		HD-LiNGAM	30.00(0.00)	1.00(0.00)	0.40(0.00)	0.57(0.00)
		Rank PC	33.60(0.62)	0.27(0.01)	0.23(0.01)	0.24(0.01)
	$(2000, 20)$	TopHeat	2.55(0.45)	0.85(0.03)	0.77(0.04)	0.80(0.03)
		EASE	4.89(0.06)	0.14(0.01)	0.59(0.03)	0.22(0.01)
		TL	2.00(0.31)	0.61(0.06)	0.67(0.07)	0.64(0.06)
		Directed-LiNGAM	11.52(0.51)	0.99(0.00)	0.32(0.01)	0.48(0.01)
		ICA-LiNGAM	12.70(0.50)	1.00(0.00)	0.29(0.01)	0.45(0.01)
		HD-LiNGAM	45.00(0.00)	1.00(0.00)	0.10(0.00)	0.18(0.00)
		Rank PC	9.04(0.32)	0.13(0.03)	0.13(0.03)	0.13(0.03)
	$(5000, 50)$	TopHeat	0.74(0.19)	0.93(0.02)	0.83(0.04)	0.87(0.03)
		EASE	2.08(0.02)	0.06(0.00)	0.44(0.03)	0.10(0.00)
		TL	0.53(0.11)	0.74(0.05)	0.79(0.06)	0.77(0.05)
		Directed-LiNGAM	13.15(0.59)	1.00(0.00)	0.14(0.01)	0.25(0.01)
		ICA-LiNGAM	14.02(0.50)	1.00(0.00)	0.13(0.01)	0.23(0.01)
		HD-LiNGAM	48.00(0.00)	1.00(0.00)	0.04(0.00)	0.08(0.00)
nonlinear	$(500, 5)$	TopHeat	5.70(0.77)	0.79(0.03)	0.94(0.02)	0.84(0.02)
		EASE	12.50(0.75)	0.55(0.03)	0.77(0.04)	0.62(0.03)
		TL	20.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	54.90(2.10)	0.13(0.02)	0.09(0.02)	0.10(0.02)
		ICA-LiNGAM	15.40(3.07)	0.73(0.05)	0.70(0.05)	0.71(0.05)
		HD-LiNGAM	25.00(0.00)	0.25(0.00)	0.33(0.00)	0.29(0.00)
		Rank PC	33.60(0.62)	0.27(0.01)	0.23(0.01)	0.24(0.01)
	$(2000, 20)$	TopHeat	2.27(0.38)	0.85(0.03)	0.78(0.04)	0.81(0.03)
		EASE	4.80(0.06)	0.14(0.01)	0.61(0.02)	0.23(0.01)
		TL	5.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	23.92(0.72)	0.38(0.01)	0.09(0.01)	0.14(0.01)
		ICA-LiNGAM	35.63(0.90)	0.12(0.02)	0.02(0.00)	0.04(0.01)
		HD-LiNGAM	9.21(0.00)	0.05(0.00)	0.06(0.00)	0.05(0.00)
		Rank PC	9.04(0.32)	0.13(0.03)	0.13(0.03)	0.13(0.03)
	$(5000, 50)$	TopHeat	0.71(0.17)	0.93(0.02)	0.83(0.03)	0.87(0.03)
		EASE	2.06(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	21.57(0.23)	0.52(0.01)	0.05(0.00)	0.09(0.00)
		ICA-LiNGAM	24.51(0.80)	0.34(0.05)	0.04(0.01)	0.06(0.01)
		HD-LiNGAM	3.88(0.00)	0.02(0.00)	0.02(0.00)	0.02(0.00)

Table S6: The averaged performance metrics of various methods, as well as their standard errors in parentheses, are presented for a BA graph in Example 2 with the Cauchy distribution.

Model	(n, p)	Methods	HM (%)	Recall	Precision	F1-score
linear	$(500, 5)$	TopHeat	7.10(0.94)	0.75(0.03)	0.90(0.03)	0.80(0.03)
		EASE	10.00(0.81)	0.71(0.03)	0.79(0.02)	0.73(0.02)
		TL	7.30(1.14)	0.64(0.06)	0.78(0.06)	0.68(0.06)
		Directed-LiNGAM	5.20(0.62)	0.97(0.01)	0.82(0.02)	0.89(0.01)
		ICA-LiNGAM	6.10(0.80)	0.97(0.02)	0.80(0.02)	0.87(0.02)
		HD-LiNGAM	51.20(1.58)	0.47(0.04)	0.19(0.02)	0.27(0.02)
		Rank PC	29.00(1.18)	0.31(0.03)	0.29(0.03)	0.30(0.03)
	$(2000, 20)$	TopHeat	1.97(0.22)	0.85(0.04)	0.77(0.02)	0.79(0.03)
		EASE	2.87(0.10)	0.57(0.02)	0.81(0.01)	0.66(0.01)
		TL	2.32(0.21)	0.56(0.04)	0.88(0.04)	0.66(0.04)
		Directed-LiNGAM	11.92(0.53)	0.99(0.00)	0.31(0.01)	0.47(0.01)
		ICA-LiNGAM	14.14(0.68)	0.93(0.03)	0.27(0.01)	0.41(0.02)
		HD-LiNGAM	50.36(0.23)	0.46(0.02)	0.05(0.00)	0.08(0.00)
		Rank PC	9.01(0.10)	0.11(0.01)	0.11(0.01)	0.11(0.01)
	$(5000, 50)$	TopHeat	1.02(0.10)	0.92(0.03)	0.69(0.02)	0.78(0.02)
		EASE	1.54(0.03)	0.42(0.01)	0.70(0.01)	0.52(0.01)
		TL	0.79(0.07)	0.63(0.03)	0.91(0.03)	0.73(0.03)
		Directed-LiNGAM	14.60(0.58)	0.99(0.00)	0.13(0.00)	0.22(0.01)
		ICA-LiNGAM	16.04(0.48)	0.89(0.02)	0.10(0.00)	0.19(0.01)
		HD-LiNGAM	50.05(0.06)	0.49(0.02)	0.02(0.00)	0.04(0.00)
		Rank PC	3.76(0.02)	0.06(0.00)	0.06(0.00)	0.06(0.00)
nonlinear	$(500, 5)$	TopHeat	8.20(0.98)	0.69(0.03)	0.89(0.03)	0.76(0.03)
		EASE	7.90(0.89)	0.76(0.03)	0.86(0.03)	0.78(0.03)
		TL	20.00(0.00)	0.00(0.00)	0.01(0.01)	0.01(0.01)
		Directed-LiNGAM	39.50(2.66)	0.32(0.04)	0.25(0.03)	0.28(0.03)
		ICA-LiNGAM	21.50(2.16)	0.59(0.04)	0.53(0.04)	0.55(0.04)
		HD-LiNGAM	21.40(0.85)	0.34(0.02)	0.45(0.03)	0.39(0.02)
		Rank PC	29.00(1.18)	0.31(0.03)	0.29(0.03)	0.30(0.03)
	$(2000, 20)$	TopHeat	1.91(0.22)	0.85(0.04)	0.79(0.02)	0.79(0.03)
		EASE	2.86(0.11)	0.57(0.02)	0.81(0.02)	0.66(0.02)
		TL	5.00(0.00)	0.00(0.00)	0.01(0.01)	0.00(0.00)
		Directed-LiNGAM	13.32(0.66)	0.43(0.02)	0.20(0.02)	0.27(0.02)
		ICA-LiNGAM	9.45(0.81)	0.58(0.03)	0.36(0.03)	0.43(0.03)
		HD-LiNGAM	7.85(0.12)	0.19(0.01)	0.20(0.01)	0.19(0.01)
		Rank PC	9.01(0.10)	0.11(0.01)	0.11(0.01)	0.11(0.01)
	$(5000, 50)$	TopHeat	1.02(0.10)	0.92(0.03)	0.70(0.02)	0.78(0.02)
		EASE	1.43(0.03)	0.47(0.01)	0.73(0.01)	0.56(0.01)
		TL	2.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	7.12(0.42)	0.51(0.02)	0.17(0.02)	0.25(0.02)
		ICA-LiNGAM	6.64(0.43)	0.55(0.02)	0.21(0.02)	0.29(0.03)
		HD-LiNGAM	3.35(0.03)	0.15(0.01)	0.16(0.01)	0.15(0.01)
		Rank PC	3.76(0.02)	0.06(0.00)	0.06(0.00)	0.06(0.00)

Table S7: The averaged performance metrics of various methods, as well as their standard errors in parentheses, are presented for the ER graph in Example 3 with the Cauchy distribution.

Model	(n, p)	Methods	HM (%)	Recall	Precision	F1-score
linear	$(500, 5)$	TopHeat	7.50(0.89)	0.60(0.06)	0.69(0.05)	0.60(0.05)
		EASE	6.00(0.88)	0.74(0.05)	0.73(0.05)	0.72(0.05)
		TL	5.20(0.89)	0.57(0.06)	0.71(0.06)	0.61(0.06)
		Directed-LiNGAM	10.10(1.03)	0.92(0.04)	0.55(0.04)	0.67(0.04)
		ICA-LiNGAM	12.60(0.93)	0.94(0.03)	0.50(0.03)	0.63(0.03)
		HD-LiNGAM	51.50(1.31)	0.42(0.05)	0.11(0.02)	0.17(0.02)
		Rank PC	19.00(1.74)	0.29(0.05)	0.28(0.05)	0.29(0.05)
	$(2000, 20)$	TopHeat	0.56(0.09)	0.92(0.01)	0.90(0.02)	0.91(0.01)
		EASE	0.53(0.06)	0.83(0.02)	0.98(0.01)	0.89(0.01)
		TL	1.80(0.13)	0.40(0.04)	0.78(0.04)	0.50(0.04)
		Directed-LiNGAM	13.57(0.51)	1.00(0.00)	0.17(0.01)	0.29(0.01)
		ICA-LiNGAM	15.93(0.48)	0.96(0.01)	0.15(0.01)	0.25(0.01)
		HD-LiNGAM	49.85(0.11)	0.53(0.02)	0.03(0.00)	0.05(0.00)
		Rank PC	4.44(0.21)	0.29(0.02)	0.24(0.02)	0.26(0.02)
	$(5000, 50)$	TopHeat	0.13(0.01)	0.95(0.01)	0.94(0.01)	0.94(0.01)
		EASE	0.63(0.02)	0.40(0.01)	0.95(0.01)	0.56(0.01)
		TL	0.64(0.04)	0.43(0.03)	0.85(0.02)	0.55(0.03)
		Directed-LiNGAM	14.63(0.52)	0.99(0.00)	0.07(0.00)	0.13(0.00)
		ICA-LiNGAM	16.07(0.44)	0.90(0.02)	0.06(0.00)	0.11(0.01)
		HD-LiNGAM	50.02(0.03)	0.49(0.02)	0.01(0.00)	0.02(0.00)
		Rank PC	1.59(0.07)	0.23(0.01)	0.23(0.01)	0.23(0.01)
nonlinear	$(500, 5)$	TopHeat	6.80(0.84)	0.61(0.05)	0.73(0.05)	0.63(0.05)
		EASE	7.00(0.73)	0.69(0.05)	0.69(0.04)	0.66(0.04)
		TL	12.10(0.95)	0.05(0.02)	0.08(0.04)	0.06(0.03)
		Directed-LiNGAM	22.90(1.52)	0.24(0.04)	0.19(0.03)	0.21(0.04)
		ICA-LiNGAM	16.80(1.69)	0.46(0.05)	0.36(0.04)	0.40(0.05)
		HD-LiNGAM	21.50(1.13)	0.26(0.04)	0.21(0.03)	0.22(0.03)
		Rank PC	19.00(1.74)	0.29(0.05)	0.28(0.05)	0.29(0.05)
	$(2000, 20)$	TopHeat	0.49(0.07)	0.90(0.01)	0.94(0.01)	0.92(0.01)
		EASE	0.97(0.08)	0.67(0.02)	0.97(0.01)	0.78(0.02)
		TL	2.68(0.11)	0.02(0.01)	0.10(0.04)	0.03(0.01)
		Directed-LiNGAM	5.54(0.24)	0.27(0.02)	0.17(0.01)	0.21(0.02)
		ICA-LiNGAM	2.95(0.25)	0.62(0.03)	0.49(0.03)	0.55(0.03)
		HD-LiNGAM	5.82(0.11)	0.30(0.02)	0.17(0.01)	0.22(0.01)
		Rank PC	4.44(0.21)	0.29(0.02)	0.24(0.02)	0.26(0.02)
	$(5000, 50)$	TopHeat	0.12(0.01)	0.95(0.01)	0.94(0.01)	0.94(0.01)
		EASE	0.58(0.02)	0.46(0.01)	0.96(0.01)	0.61(0.01)
		TL	1.02(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)
		Directed-LiNGAM	2.31(0.10)	0.26(0.01)	0.15(0.01)	0.19(0.01)
		ICA-LiNGAM	1.29(0.08)	0.60(0.02)	0.43(0.02)	0.50(0.02)
		HD-LiNGAM	2.49(0.03)	0.24(0.01)	0.12(0.01)	0.16(0.01)
		Rank PC	1.58(0.07)	0.24(0.01)	0.24(0.01)	0.24(0.01)

Bibliography

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

Biau, G. and L. Devroye (2015). *Lectures on the nearest neighbor method*, Volume 246. Springer.

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.

Devroye, L. (1988). The expected size of some graphs in computational geometry. *Computers and Mathematics with Applications 15*(1), 53–64.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics*, pp. 569–593. Springer.

Einmahl, J. H. J. (1986). *Multivariate empirical processes*. Ph. D. thesis, [Nijmegen]: Centrum voor Wiskunde en Informatica.

Einmahl, J. H. J., L. de Haan, and D. Li (2006). Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *The Annals of Statistics 34*(4), 1987–2014.

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

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type

coincidences. *The Annals of Statistics 16*(2), 772–783.

Li, J. and M. H. Maathuis (2021). GGM knockoff filter: False discovery rate control for Gaussian graphical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(3), 534–558.

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

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

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

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.