Non-Parametric Structure Learning on Hidden Tree-Shaped Distributions

[1]Konstantinos E. Nikolakakis, [2]Dionysios S. Kalogerias,
[1]Anand D. Sarwate

[1]Department of Electrical & Computer Engineering, Rutgers University
[2]Department of Electrical & Systems Engineering, University of Pennsylvania

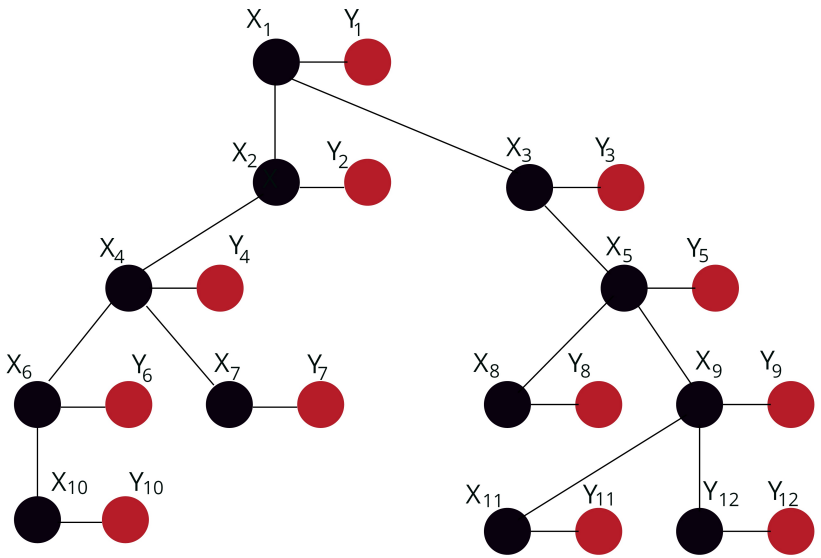
Why learn from noisy data?

- Data acquisition devices or sensors introduce noise
- Local differential privacy
- Communication constrains and quantization error
- Adversarial attacks

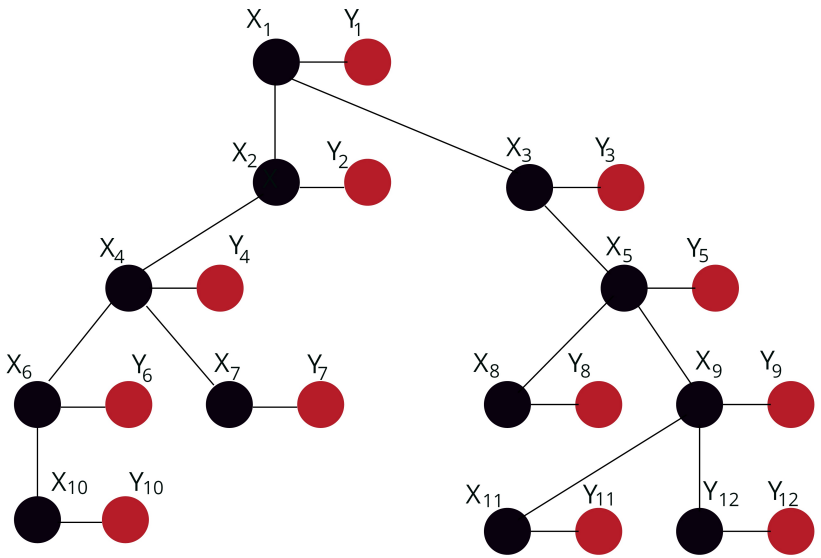
Problem Statement (Learning Hidden Tree Structures)

Observe noisy values for each node of the unknown tree structure

- $X_1, X_2, \ldots, X_p$ are hidden variables (black nodes)
- $Y_1, Y_2, \ldots, Y_p$ are observable variables (red nodes)



Learning a tree structure



Assumptions:

- Distribution of $\mathbf{X}$ is nondegenerate and factorizes according to a tree T.
- $\mathrm{T} = (\mathcal{V}, \mathcal{E})$ is connected.
- $I(X_i; X_j) > 0$ for all $i, j \in \mathcal{V}$.

Chow-Liu Algorithm

Given: Data set $\mathcal{D} = \boldsymbol{Y} \in \mathcal{Y}^{|\mathcal{V}| \times n}$

1. Compute empirical distribution on each edge:

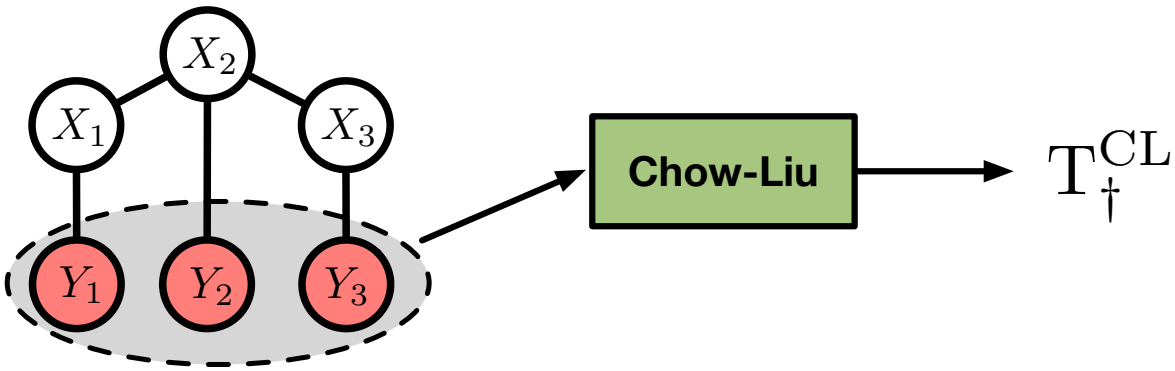
$$\hat{\mathrm{p}}_{i,j}(\ell, m) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{Y_{i,k}=\ell, Y_{j,k}=m\}} \quad \forall i, j \in \mathcal{V}$$

2. Find plug-in estimate of mutual information:

$$\hat{I}(Y_i; Y_j) = \sum_{\ell,m} \hat{\mathrm{p}}_{i,j}(\ell, m) \log_2 \frac{\hat{\mathrm{p}}_{i,j}(\ell, m)}{\hat{\mathrm{p}}_i(\ell)\hat{\mathrm{p}}_j(m)}$$

3. Output $\mathrm{T}_{\dagger}^{\mathrm{CL}} = \mathsf{MST}\left(\{\hat{I}(Z_i; Z_j) : i, j \in \mathcal{V}\}\right)$

Main questions



Given noise corrupted data:

- Is Chow-Liu consistent?
- How does noise affect the sample complexity?

Prior work: Finite sample complexity for Ising and Gaussian Models.

- Tan et al. (2011), Liu et al. (2011), Bresler & Karzand (2018)
- Hidden models: Our work (2019), Goel-Kane-Klivans (2019)

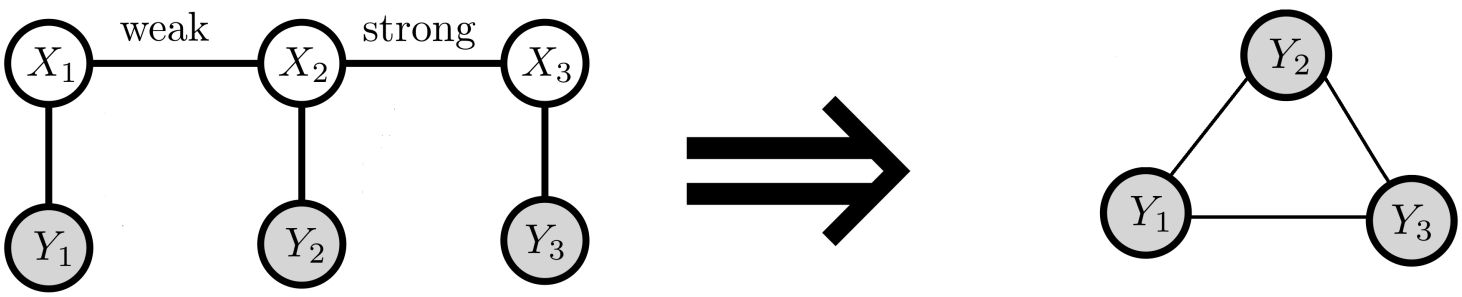
A motivating example: 3-node hidden model

What can go wrong when we have noise?

A motivating example: 3-node hidden model

What can go wrong when we have noise?

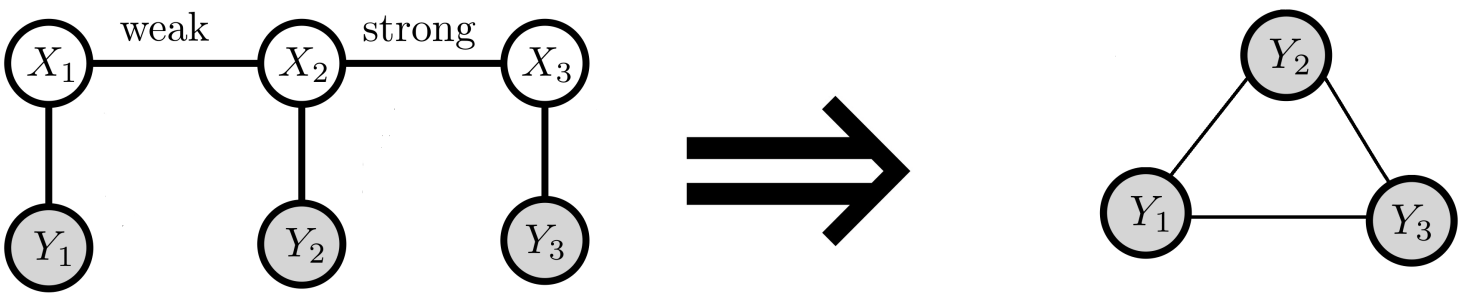
The MRF of the observable is a complete graph!



A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!

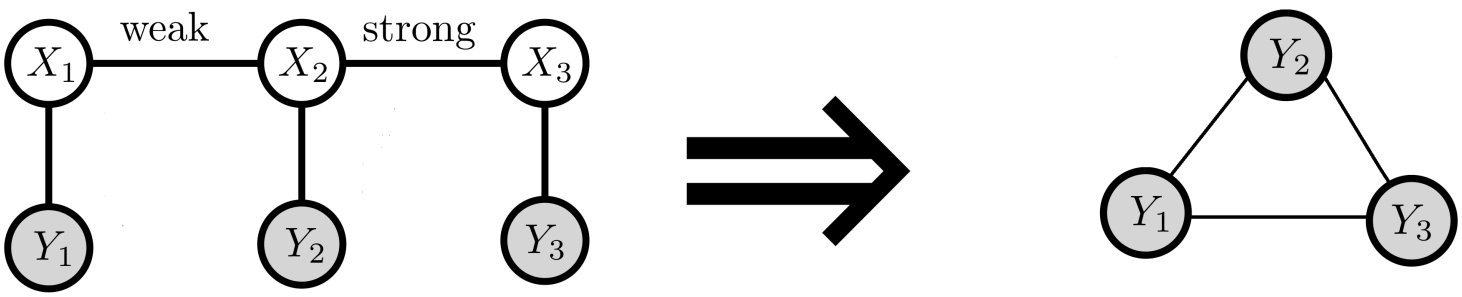


Questions:

A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!



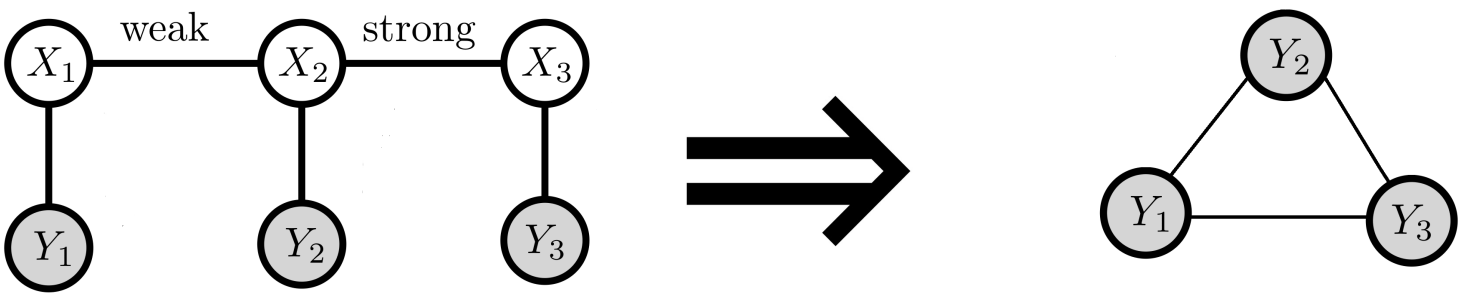
Questions:

- Is Chow-Liu consistent? **NO**

A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!



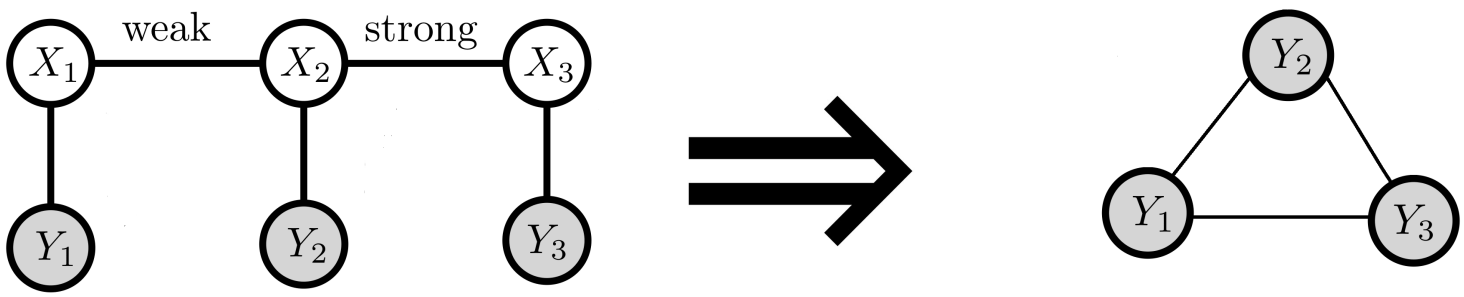
Questions:

- Is Chow-Liu consistent? **NO**
- When does $\lim_{n\to\infty} \mathrm{T}^{\mathrm{CL}} \to \mathrm{T}$ w.p. 1? **A sufficient condition**

A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!

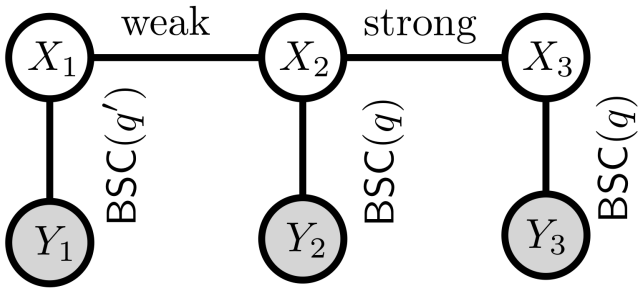


Questions:

- Is Chow-Liu consistent? **NO**
- When does $\lim_{n\to\infty} \mathrm{T}^{\mathrm{CL}} \to \mathrm{T}$ w.p. 1? **A sufficient condition**
- Can we tweak Chow-Liu to fix it? **Sometimes**

A closer look at the example

$$X_1, X_2, X_3 \in \{-1, +1\}, \quad 0 < \underbrace{|\mathbb{E}[X_1 X_2]|}_{\text{weak}} \leq \underbrace{|\mathbb{E}[X_2 X_3]|}_{\text{strong}} < 1$$

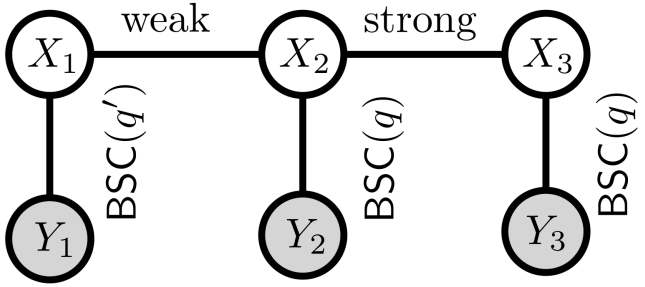


$$I(X_2; X_3) > I(X_1; X_2) \overset{\text{DPI}}{>} I(X_1; X_3)$$

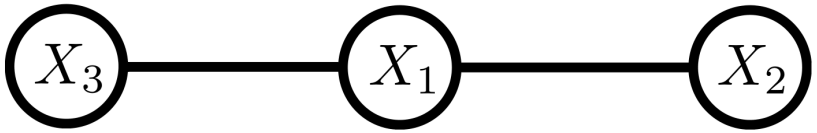
$$\lim_{n\to\infty} \hat{I}(X_i; X_j) \to I(X_i; X_j) \text{ and } \lim_{n\to\infty} \mathcal{E}_{\mathrm{TCL}} \to \{(1,2),(2,3)\} \equiv \mathcal{E}_{\mathrm{T}}$$

- Does a similar condition hold for the observables?
- Could we have $\lim_{n\to\infty} \mathcal{E}_{\mathrm{T}^{\mathrm{CL}}_{\dagger}} \neq \mathcal{E}_{\mathrm{T}}$?

Feasibility Threshold



- If $I(Y_1;Y_2) > I(Y_1;Y_3) > I(Y_2;Y_3)$
- then $\lim_{n\to\infty} \mathcal{E}_{\mathrm{T}^{\mathsf{CL}}_{\dagger}} \neq \mathcal{E}_{\mathrm{T}}$



$$I(Y_1;Y_3) > I(Y_2;Y_3) \iff |\mathbb{E}[Y_1Y_3]| > |\mathbb{E}[Y_2Y_3]| \iff$$
$$|\mathbb{E}[X_1X_2]| > \frac{1-2q}{1-2q'}, \quad q, q' \in [0, 1/2).$$

Unprocessed vs Processed Data

- What if $|\mathbb{E}[X_1 X_2]| > (1-2q)/(1-2q')$?
- We have to pre-process

$$Z_1 \triangleq Y_1/(1-2q'), \quad Z_2 \triangleq Y_2/(1-2q), \quad Z_3 \triangleq Y_3/(1-2q)$$

- Correct order, $I(Z_2; Z_3) > I(Z_1; Z_2) > I(Z_1; Z_3)$
- Then $\lim_{n\to\infty} \mathcal{E}_{\mathrm{T}_\dagger^{\mathsf{CL}}} = \mathcal{E}_{\mathrm{T}}$ with probability 1.

Unprocessed vs Processed Data

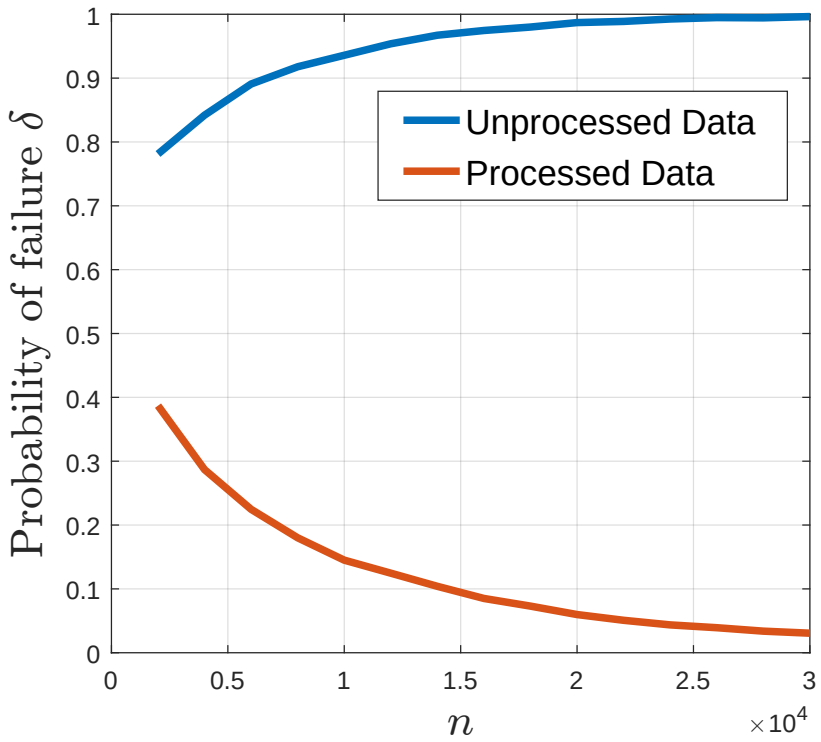
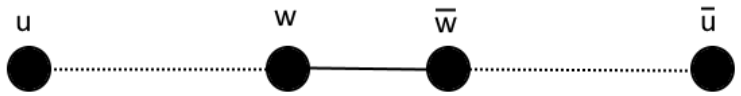


Figure: *Synthetic data,* $q' = 0.2$, $q = 0.25$

Definition

(**The set $\mathcal{EV}^2$**) Let $e \equiv (w, \bar{w}) \in \mathcal{E}_{\mathrm{T}}$ be an edge and $u, \bar{u} \in \mathcal{V}_{\mathrm{T}}$ be a pair of nodes such that $e \in \mathrm{path}_{\mathrm{T}}(u, \bar{u})$ and $|\mathrm{path}_{\mathrm{T}}(u, \bar{u})| \geq 2$. Then

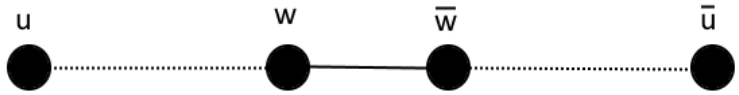
$$\mathcal{EV}^2 \triangleq \{(w, \bar{w}), u, \bar{u} \in \mathcal{E}_{\mathrm{T}} \times \mathcal{V}_{\mathrm{T}} \times \mathcal{V}_{\mathrm{T}} : (w, \bar{w}) \in \mathrm{path}_{\mathrm{T}}(u, \bar{u}) \text{ and } |\mathrm{path}_{\mathrm{T}}(u, \bar{u})| \geq 2\}.$$



Definition

(**The set $\mathcal{EV}^2$**) Let $e \equiv (w, \bar{w}) \in \mathcal{E}_{\mathrm{T}}$ be an edge and $u, \bar{u} \in \mathcal{V}_{\mathrm{T}}$ be a pair of nodes such that $e \in \mathrm{path}_{\mathrm{T}}(u, \bar{u})$ and $|\mathrm{path}_{\mathrm{T}}(u, \bar{u})| \geq 2$. Then

$$\mathcal{EV}^2 \triangleq \{(w, \bar{w}), u, \bar{u} \in \mathcal{E}_{\mathrm{T}} \times \mathcal{V}_{\mathrm{T}} \times \mathcal{V}_{\mathrm{T}} : (w, \bar{w}) \in \mathrm{path}_{\mathrm{T}}(u, \bar{u}) \text{ and } |\mathrm{path}_{\mathrm{T}}(u, \bar{u})| \geq 2\}.$$



Error Characterization of CL algorithm (Bresler & Karzand 2018):

$$\text{If } \mathrm{T}_{\dagger}^{\mathrm{CL}} \neq T \implies \exists\, ((w, \bar{w}), u, \bar{u}) \in \mathcal{EV}^2 : \hat{I}(Y_w; Y_{\bar{w}}) \leq \hat{I}(Y_u; Y_{\bar{u}})$$

Sufficient Condition for Exact Recovery

Exact recovery:

$$\text{If } \forall \left((w,\bar{w}),u,\bar{u}\right) \in \mathcal{EV}^2 : \hat{I}\left(Y_w;Y_{\bar{w}}\right) > \hat{I}\left(Y_u;Y_{\bar{u}}\right) \implies \mathrm{T}_{\dagger}^{\mathrm{CL}} = T$$

$$\hat{I}\left(Y_w;Y_{\bar{w}}\right) > \hat{I}\left(Y_u;Y_{\bar{u}}\right) \iff$$

$$I\left(Y_w;Y_{\bar{w}}\right) - I\left(Y_u;Y_{\bar{u}}\right) >$$

$$\left[\hat{I}\left(Y_u;Y_{\bar{u}}\right) - I\left(Y_u;Y_{\bar{u}}\right)\right] - \left[\hat{I}\left(Y_w;Y_{\bar{w}}\right) - I\left(Y_w;Y_{\bar{w}}\right)\right].$$

Sufficient Condition for Exact Recovery

Exact recovery:

$$\forall\left((w,\bar{w}),u,\bar{u}\right)\in\mathcal{EV}^{2}:\hat{I}\left(Y_{w};Y_{\bar{w}}\right)>\hat{I}\left(Y_{u};Y_{\bar{u}}\right)\iff\mathrm{T}_{\dagger}^{\mathrm{CL}}=T$$

$$\hat{I}\left(Y_{w};Y_{\bar{w}}\right)>\hat{I}\left(Y_{u};Y_{\bar{u}}\right)\iff$$

$$I\left(Y_{w};Y_{\bar{w}}\right)-I\left(Y_{u};Y_{\bar{u}}\right)>$$

$$\left[\hat{I}\left(Y_{u};Y_{\bar{u}}\right)-I\left(Y_{u};Y_{\bar{u}}\right)\right]-\left[\hat{I}\left(Y_{w};Y_{\bar{w}}\right)-I\left(Y_{w};Y_{\bar{w}}\right)\right].$$

Sufficient Condition

If $\left|\hat{I}\left(Y_{\ell};Y_{\bar{\ell}}\right)-I\left(Y_{\ell};Y_{\bar{\ell}}\right)\right|<\frac{1}{2}\min_{(e,u,\bar{u})\in\mathcal{EV}^{2}}\left\{I\left(Y_{w};Y_{\bar{w}}\right)-I\left(Y_{u};Y_{\bar{u}}\right)\right\}$

for all $\ell,\ell'\in\mathcal{V}$ then $\mathrm{T}_{\dagger}^{\mathrm{CL}}=\mathrm{T}$.

Definition

(Information Thresholds $\mathbf{I}^o$, $\mathbf{I}^o_\dagger$)

$$\mathbf{I}^o \triangleq \frac{1}{2} \min_{((w,\bar{w}),u,\bar{u})\in\mathcal{EV}^2} \left[I\left(X_w; X_{\bar{w}}\right) - I\left(X_u; X_{\bar{u}}\right)\right]$$

$$\mathbf{I}^o_\dagger \triangleq \frac{1}{2} \min_{((w,\bar{w}),u,\bar{u})\in\mathcal{EV}^2} \left[I\left(Y_w; Y_{\bar{w}}\right) - I\left(Y_u; Y_{\bar{u}}\right)\right]$$

- Always $\mathbf{I}^o \geq 0$, DPI
- $\mathbf{I}^o_\dagger \lessgtr 0$ generalizes the condition $\frac{1-2q}{1-2q'} \lessgtr |\mathbb{E}[X_1 X_2]|$ to non-parametric models and general channels
- $\mathbf{I}^o_\dagger < 0$ implies that structure learning is infeasible without post-processing

Sample Complexity

- Sufficient condition $\left|\hat{I}\left(Y_\ell; Y_{\bar{\ell}}\right) - I\left(Y_\ell; Y_{\bar{\ell}}\right)\right| < \mathbf{I}^o_\dagger$
- Concentration of measure of mutual information estimates
- Union bound over the pairs $\ell, \ell' \in \mathcal{V}$

Theorem

Fix $\delta \in (0,1)$. There exist constants $C > 0$ and $c \in (1,2]$ independent of δ such that, if $\mathbf{I}^o_\dagger > 0$ and

$$\frac{n}{\log_2^2 n} \geq \frac{72 \log\left(\frac{p}{\delta}\right)}{\left(\mathbf{I}^o_\dagger - Cn^{\frac{1-c}{c}}\right)^2} \quad \text{and} \quad \mathbf{I}^o_\dagger > Cn^{\frac{1-c}{c}},$$

then CL with input $\mathcal{D} = \mathbf{Y}^{1:n}$ returns $\mathrm{T}^{CL}_\dagger = \mathrm{T}$ w.p. at least $1-\delta$.

Almost logarithmic order: $\mathcal{O}\left(\log^{1+\zeta}(p/\delta)\right)$, for all $\zeta > 0$

Experiments: Noiseless Binary Data

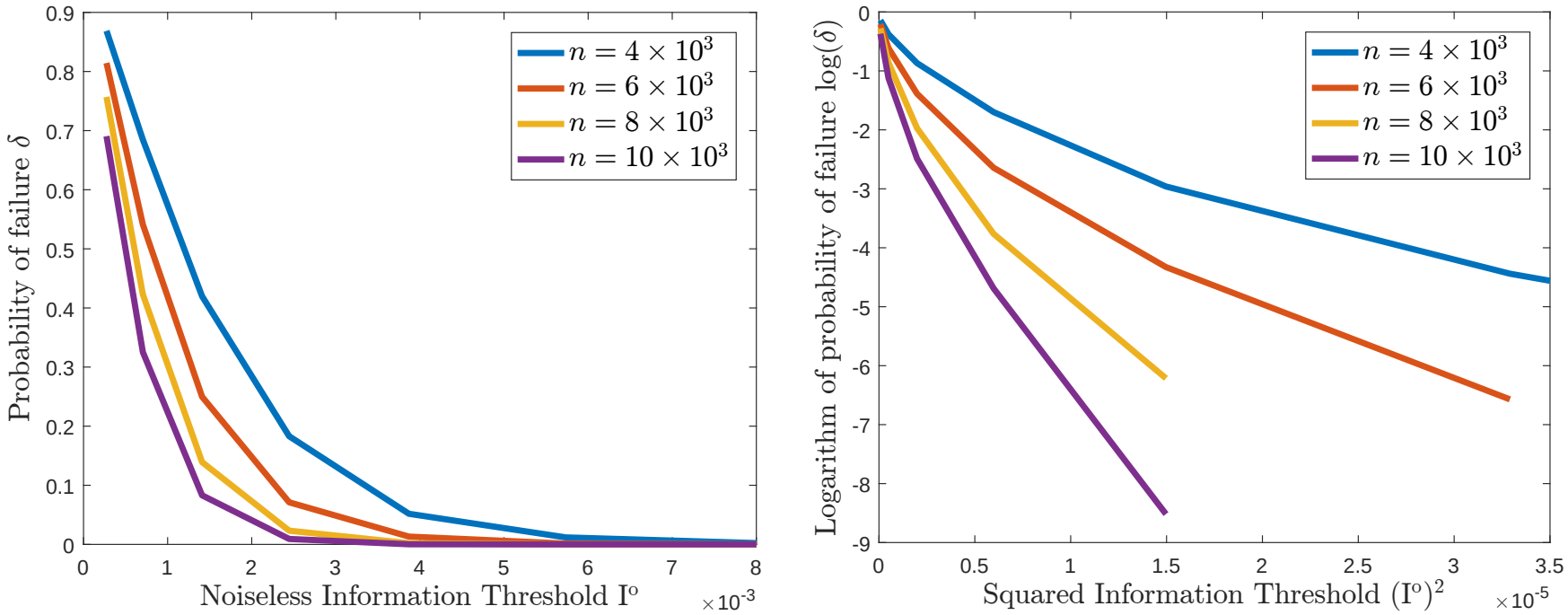


Figure: *Left:* $\hat{\mathbb{P}}\left(\mathrm{T}^{\mathrm{CL}} \neq \mathrm{T}\right)$ *vs* $\mathbf{I}^o$, *Right:* $\log \hat{\mathbb{P}}\left(\mathrm{T}^{\mathrm{CL}} \neq \mathrm{T}\right)$ *vs* $(\mathbf{I}^o)^2$

Experiments: Noisy Binary Data (BSC)

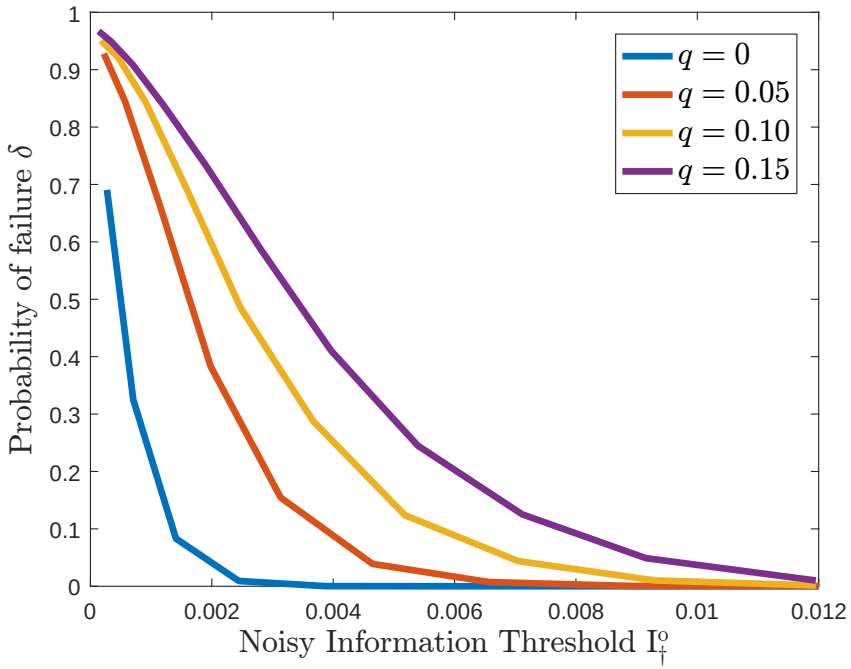


Figure: $\hat{\mathbb{P}}\left(\mathrm{T}^{\mathrm{CL}}_{\dagger} \neq \mathrm{T}\right)$ *vs* $\mathbf{I}^{o}_{\dagger}$

Further Questions and Future Directions

- What is the relationship of $\mathbf{I}^o$ and $\mathbf{I}^o_\dagger$? Connection with SDPI
- How to estimate $\mathbf{I}^o_\dagger$ from training data?
- How to preserve privacy while structure learning remains feasible?
- Find robust methods for pre-processing against adversarial attacks

Thank you!