



Learning Tree Structures from Noisy Data



Konstantinos Nikolakakis[†], Dionysios Kalogerias^{*}, Anand Sarwate[†]

[†] Rutgers University, ^{*} Princeton University

Motivation

- Graphical models are used to model the structure of data
- Real data are almost always collected using noisy sensors
- How well can we estimate the structure from noisy data?

Model Description and Assumptions

- Binary data:** Ising model distribution

$$p(\mathbf{x}) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{(s,t) \in \mathcal{E}} \theta_{st} x_s x_t \right\}, \quad \mathbf{x} \in \{-1, 1\}^p.$$

Hidden Ising Model:

- Hidden layer \mathbf{X} , where $\mathbf{X} \sim p(\cdot) \in \mathcal{P}_T(\alpha, \beta)$
- Observable layer \mathbf{Y} . $\text{BSC}(q)^p$ acts on \mathbf{X} and generates \mathbf{Y} .

- Continuous data:** Gaussian distribution

Hidden Gaussian Model:

- Hidden layer \mathbf{X} , where $\mathbf{X} \sim p(\cdot) \in \mathcal{N}_T^{m,M}$.
- Observable layer $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{N} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

Applications

- Finance:** Dynamics of a market
- Biology:** Epidemic dynamics and neoplastic transitions
- Psychology:** Social and pedestrian dynamics
- Computer Science:** Differential Privacy

The Chow-Liu Algorithm

Algorithm 1 Chow-Liu

Require: Data set $\mathcal{D} = \{\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n)\}$

$$\hat{\mu}_{i,j}^\dagger \leftarrow \frac{1}{n} \sum_k y_i(k) y_j(k), \text{ for all } i, j \in \mathcal{V}$$

$$T_\dagger^{\text{CL}} \leftarrow \text{MaximumSpanningTree}\left(\cup_{i \neq j} \{|\hat{\mu}_{i,j}^\dagger|\}\right)$$

Analysis and Tools

Lower bound for structure recovery

- Correlation decay property
- Concentration of measure inequalities
- Characterization of the error event: $T \neq T_\dagger^{\text{CL}}$

Upper bound for structure reconstruction

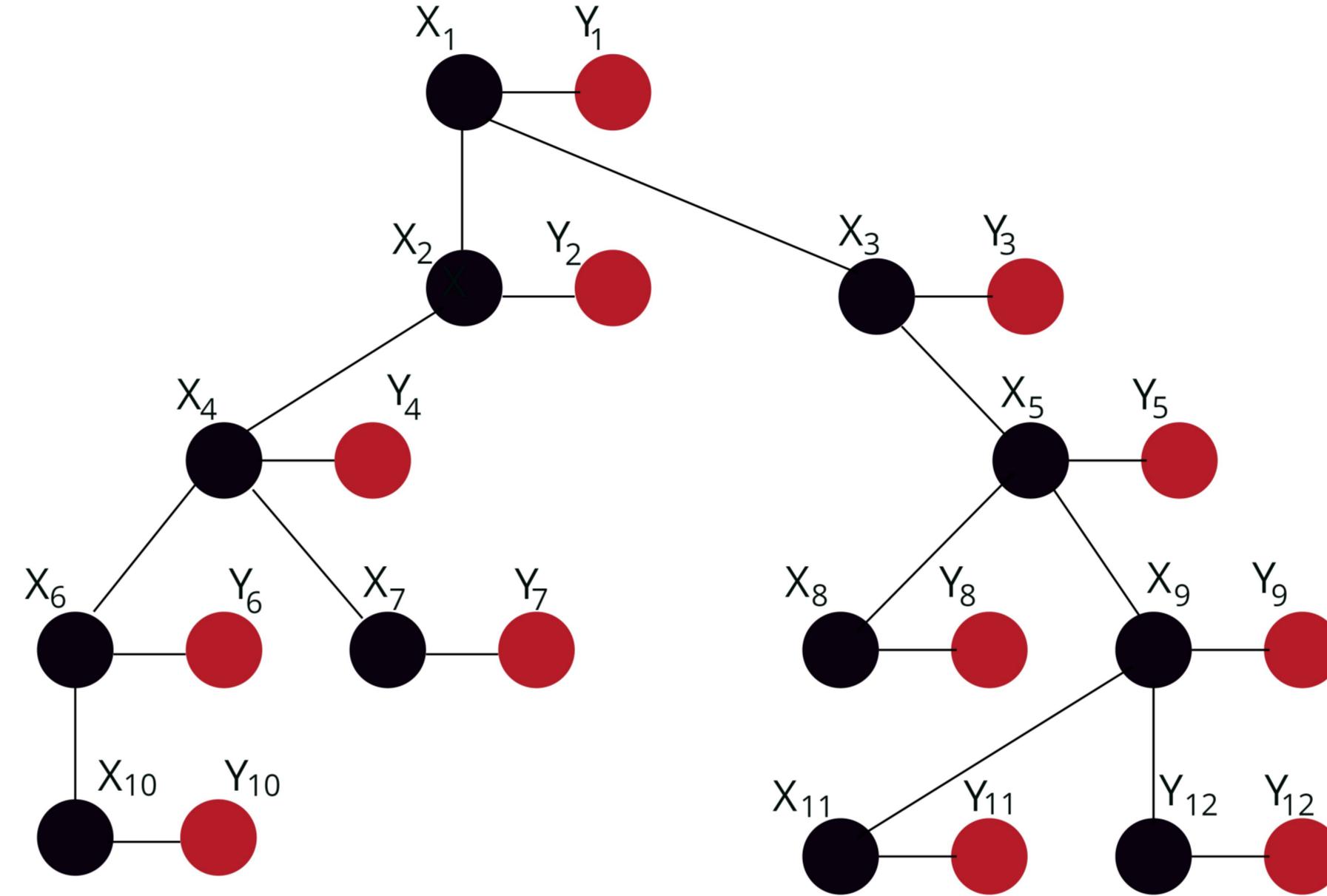
- Minimax bounds, Fano's method
- Strong Data Processing Inequality

Problem Statement

- X_1, X_2, \dots, X_p are hidden variables
- Y_1, Y_2, \dots, Y_p are observable variables

$$T \rightarrow \mathbf{X} \xrightarrow{\text{BSC}^p(q)} \mathbf{Y} \rightarrow \{\mathbf{y}^{1:n}\} \rightarrow T_\dagger^{\text{CL}}$$

- Black nodes: Hidden variables
- Red nodes: Observable variables



Goal: Find the required number of samples for exact structure recovery, assuming that observations from noisy variables are available.

Theorem (Sufficient Number of Samples - Ising Model, BSC Channel)

Fix a number $\delta \in (0, 1)$. If

$$n \geq \frac{32 \left[1 - (1 - 2q)^4 \tanh \beta \right]}{(1 - 2q)^4 (1 - \tanh \beta)^2 \tanh^2 \alpha} \log \frac{2p^2}{\delta},$$

then Chow-Liu's Algorithm returns $T_\dagger^{\text{CL}} = T$ with probability at least $1 - \delta$.

Theorem (Necessary Number of Samples - Ising Model, BSC Channel)

$$\text{If } n < \frac{[1 - (4q(1 - q))^p]^{-1}}{16\alpha \tanh(\alpha)} e^{2\beta} \log(p),$$

then for any estimator ψ , it is true that

$$\inf_{\psi} \sup_{\substack{T \in \mathcal{T} \\ P \in \mathcal{P}_T(\alpha, \beta)}} \mathbb{P}(\psi(\mathbf{Y}_{1:n}) \neq T) > \frac{1}{2}.$$

Theorem (Sufficient Number of Samples - Gaussian Model, AWGN Channel)

Fix a number $\delta \in (0, 1)$. Chow-Liu algorithm recovers the structure, $T = T_\dagger^{\text{CL}}$ with probability at least $1 - \delta$, if

$$n \geq \frac{R^2 [7(1 + \sigma^2)^2 + \rho_M] \log^4 \left(\frac{e^2 p^3}{\delta} \right)}{\rho_m^2 (1 - \rho_M)^2}, \quad R > 0.$$

Measures of Performance

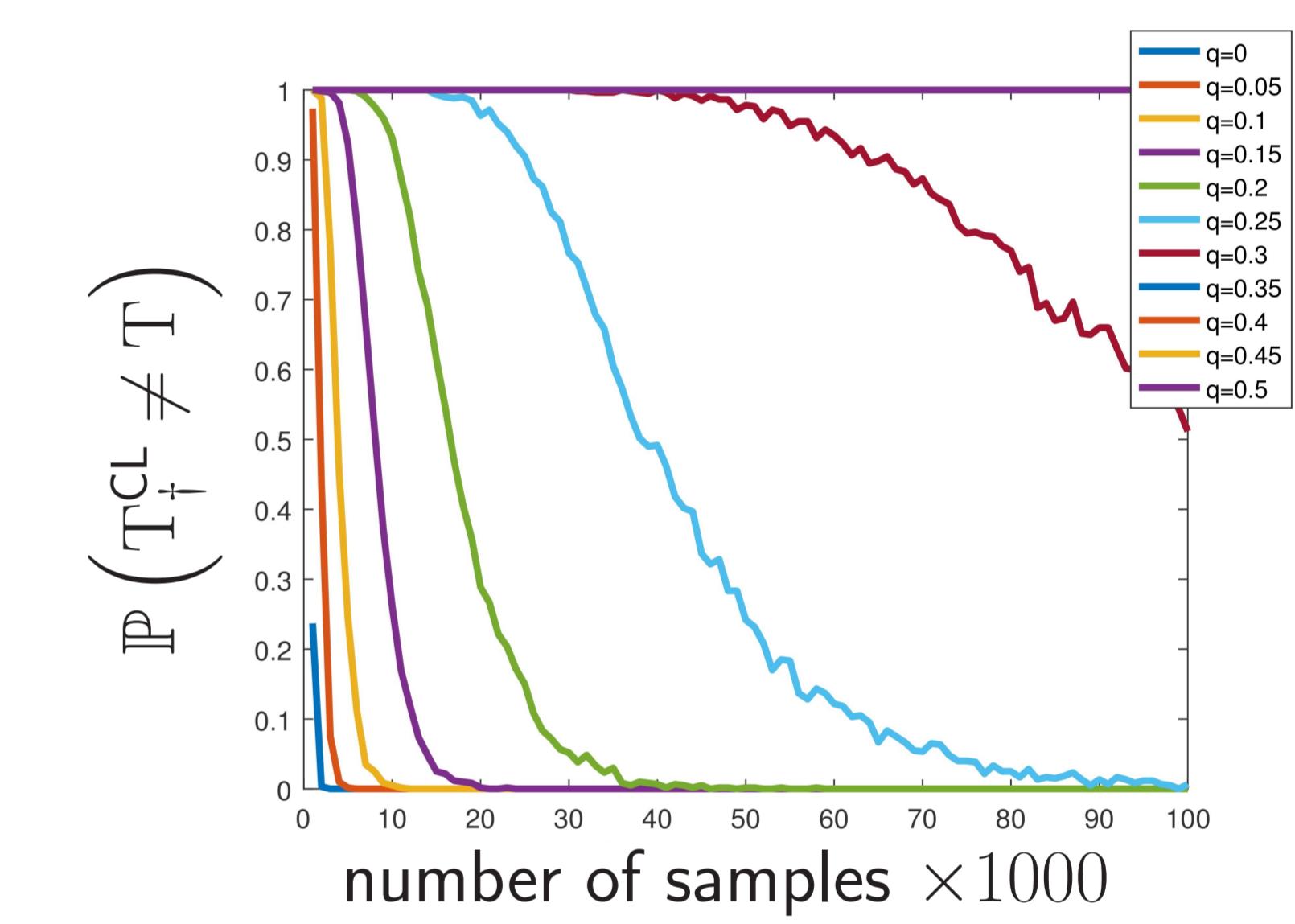
- Probability of incorrect reconstruction:

$$\mathbb{P}(T_\dagger^{\text{CL}} \neq T) = \mathbb{E} [\mathcal{L}^{0-1}(T, T_\dagger^{\text{CL}})]$$

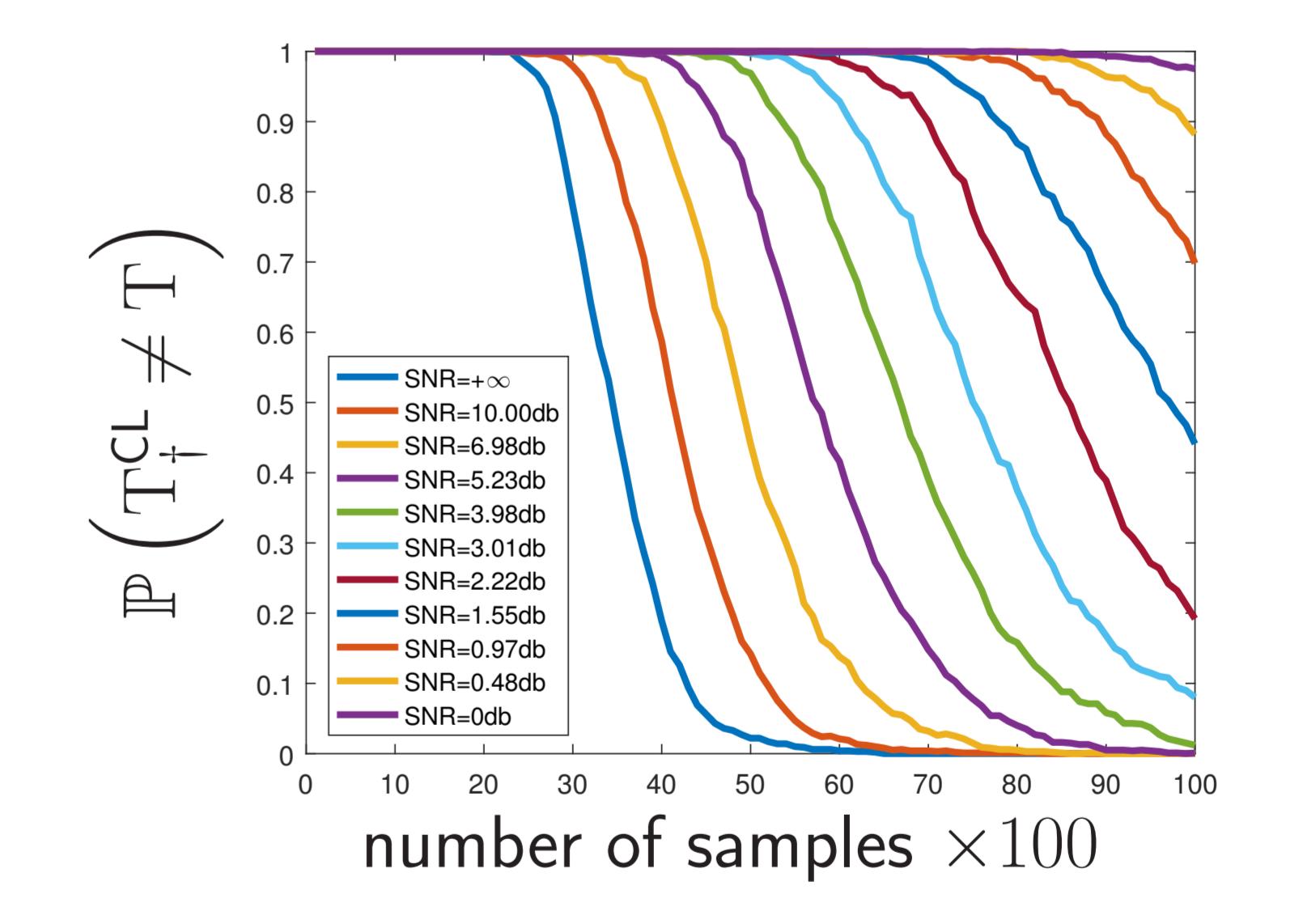
- Mismatched edges:

$$\mathcal{D}_T(T, T_\dagger^{\text{CL}}) \triangleq \frac{|\mathcal{E}_T \Delta \mathcal{E}_{T_\dagger^{\text{CL}}}|}{2}$$

Ising Model, Synthetic Data



Gaussian Model, Synthetic Data



Semi-Synthetic Binary Data

