Learning Tree Structures from Noisy Data

Konstantinos Nikolakakis[1], Dionysios Kalogerias[2], Anand Sarwate[3]

[1] Rutgers University, [2] Princeton University, [3] Rutgers 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?

Problem Statement

Noisy values for each node of the unknown tree structure T are observed,

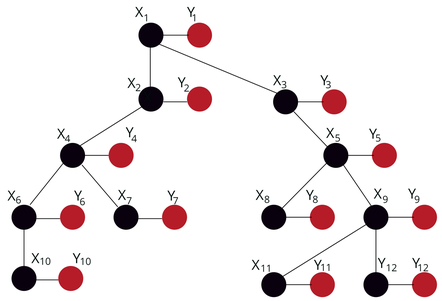
- $X_1, X_2, \ldots, X_p$ are hidden variables
- $Y_1, Y_2, \ldots, Y_p$ are observable variables

$$\mathrm{T} \to \mathbf{X} \overset{\mathrm{BSC}^p(q)}{\longrightarrow} \mathbf{Y} \to \mathrm{T}_{\dagger}^{\mathrm{CL}}$$

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

Example of a hidden tree structure model.

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



Example Applications

- Finance: Dynamics of a market
- Computer Science: Differential Privacy, Image reconstruction
- Biology: Epidemic dynamics and Neoplastic transitions

Measures of Performance

- Probability of incorrect reconstruction:

$$\mathbb{P}\left(\mathrm{T}_{\dagger}^{\mathsf{CL}} \neq \mathrm{T}\right) = \mathbb{E}\left[\mathcal{L}^{0-1}\left(\mathrm{T}, \mathrm{T}_{\dagger}^{\mathsf{CL}}\right)\right]$$

- Mismatched edges:

$$\mathcal{D}_{\mathcal{T}}(\mathrm{T}, \mathrm{T}_{\dagger}^{\mathsf{CL}}) \triangleq \frac{|\mathcal{E}_{\mathrm{T}} \triangle \mathcal{E}_{\mathrm{T}_{\dagger}^{\mathsf{CL}}}|}{2}$$

Model Description and Underlying Assumptions

- **Binary data:** Ising model distribution

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

Hidden Ising Model:

- Hidden layer $\mathbf{X}$, where $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha, \beta)$
- Observable layer $\mathbf{Y}$. $\mathrm{BSC}(q)^p$ with cross-over q acts compentwise independently on $\mathbf{X}$ and generates $\mathbf{Y}$.

- **Continuous data:** Gaussian distribution

$\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, the nonzero entries of Σ^{-1} indicate the existence of the corresponding edges.

Hidden Gaussian Model:

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

Theorem (Sufficient number of samples - Ising Model)

Let $\mathbf{Y}$ be the output of a $\mathrm{BSC}(q)^p$, with input variable $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha, \beta)$. Fix a number $\delta \in (0,1)$. If the number of samples n of $\mathbf{Y}$ satisfies the inequality

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

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

Theorem (Necessary number of samples - Ising Model)

Let $\mathbf{Y}$ be the output of a $\mathrm{BSC}(q)^p$, with input variable $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha, \beta)$. If the given number of samples satisfies

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

then for any (measurable) estimator ψ, it is true that

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

Theorem (Sufficient number of samples - Gaussian Model)

Let $\mathbf{Y}$ be the output of a Gaussian channel, $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{X} \sim \mathrm{p}(\cdot) \in \boldsymbol{\mathcal{N}}_{\mathrm{T}}^{m,M}$ and $\mathbf{N} \sim \mathcal{N}(0, \sigma^2)$. Fix a number $\delta \in (0,1)$. Chow-Liu algorithm recovers the structure, $\mathrm{T} = \mathrm{T}_{\dagger}^{\mathsf{CL}}$ with probability at least $1-\delta$, if the number of samples satisfies the inequality

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

and R is a positive constant.

The Chow-Liu Algorithm

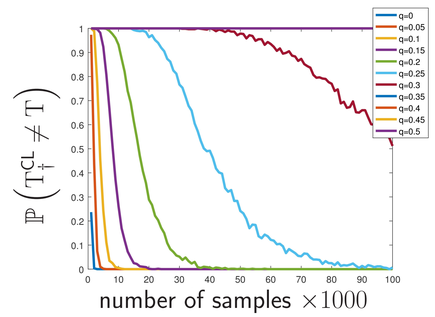
Algorithm 1 Chow-Liu

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

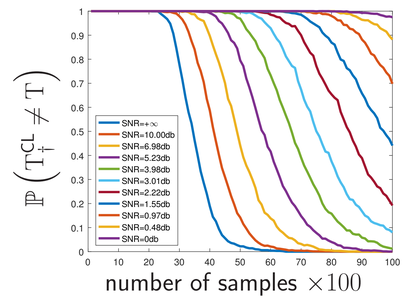
$\hat{\mu}_{i,j}^{\dagger} \leftarrow \frac{1}{n}\sum_k y_i(k) y_j(k)$, for all $i, j \in \mathcal{V}$

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

Ising Model, Synthetic Data



Gaussian Model, Synthetic Data



Semi-Synthetic Binary Data