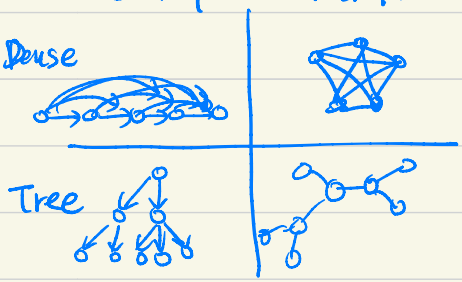
* structure Learning Recap.

Approach 1. ML for DAG

Derived under B.N.

$$\text{SCORE}(G) = \sum_{i=1}^{n} I_{\hat{p}}(X_i ; X_{\pi_i}) \leftarrow \{P(x_i | x_{\pi_i})\}$$

π_i: parents of i in G



tractable for G ∈ Tree $\leftrightarrow$ max weight spanning tree [Chow-Liu algorithm]

Approach 2. ML for Ising Models.

$$\min_{\theta \in \mathbb{R}^{mn}} \underbrace{\log Z_G(\theta) - \langle \hat{M}, \theta \rangle}_{\text{- log-likelihood}} + \underbrace{\lambda \|\theta\|_{L1}}_{\text{regularization}}$$

intractable for general graphs even for sparse graphs.

$O(|\mathcal{X}|^n)$

Approach 3. Local Independence Tests. with max-~~degree~~ $\leq K$

Enumerate over all nodes.

Select a neighborhood S_i.

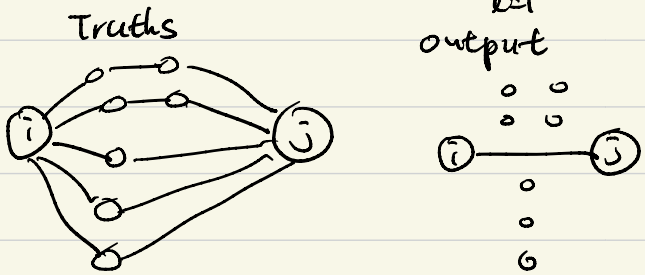
with best SCORE(S_i) $\cong$ conditional independence

$O((n)^{k+1})$

Approach 4. pairwise tests.

$O(n^2)$

$$\text{SCORE}((i,j)) = \frac{1}{N} \sum_{l=1}^{N} x_i^{(l)} x_j^{(l)}$$: correlation



* None of the algorithms above work in large scale practical problems.

Q. Can we design an algorithm that is efficient & works well in practice?

Consider Binary Random Variables $\mathcal{X} = \{0, 1\}$

Undirected graphical models.

$$P(X_1, \cdots, X_n) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{i=1}^{n} \theta_{ii} X_i + \sum_{(i,j) \in E} \theta_{ij} X_i X_j \right\}$$

Strategy is to predict X_i using X_{-i}, but with a subset.

Claim: A conditional distribution of X_1 given the rest is

$$(*) \qquad \frac{P(X_1 = 1 \mid X_2 \cdots X_n)}{P(X_1 = 0 \mid X_2 \cdots X_n)} = \exp\left\{ \theta_{11} + \sum_{j \in \partial 1} \theta_{1j} X_j \right\}$$

this is a logistic regression problem, where we try to predict X_1 from X_2^n can be solved efficiently.

$$(*) \rightarrow P(X_1 = 1 \mid X_2 \cdots X_n) = \frac{e^{\theta_{11} + \sum_j \theta_{1j} X_j}}{1 + e^{\theta_{11} + \sum \theta_{1j} X_j}}$$

$$= \text{sigmoid}\left(\theta_{11} + \sum_j \theta_{1j} X_j \right)$$

sigmoid: logistic function.

linear function $\theta_1 \cdot$ " $(\theta_{11}, \theta_{12}, \cdots, \theta_{1n})$

proof of (*)

$$P(X_1 = 1 \mid X_2^n) = \frac{P(X_1 = 1, X_2^n)}{P(X_2^n)}$$

$$= \frac{\frac{1}{Z(\theta)} \cdot \exp\left\{ \theta_{11} + \sum_{j \in \partial 1} \theta_{1j} X_j + \sum_{\substack{i,j > 1 \\ i,j \in E}} \theta_{ij} X_i X_j \right\}}{P(X_2^n)}$$

$$P(X_1 = 0 \mid X_2^n) = \frac{\frac{1}{Z(\theta)} \cdot \exp\left\{ \sum_{\substack{i,j > 1 \\ i,j \in E}} \theta_{ij} X_i X_j \right\}}{P(X_2^n)}$$

$$P(X_i=1 \mid X_1^n) = \frac{\exp\{\theta_{i1} + \sum \theta_{ij} X_j\}}{1+\exp\{\theta_{i1} + \sum \theta_{ij} X_j\}}$$

* Review of Logistic Regression.

we are given labelled samples of features $z^{(l)} = (z_1^{(l)}, \dots, z_L^{(l)}) \in \mathbb{R}^L$ and labels $Y^{(l)} \in \{0,1\}$

Data: $\{(z^{(l)}, Y^{(l)})\}_{l=1}^N$

we suppose a parametric model with $w \in \mathbb{R}^L$

$$P(Y=1 \mid z) = \frac{\exp(\sum_{k=1}^L w_k z_k)}{1+\exp(\sum_{k=1}^L w_k z_k)}$$

$$P(Y=0 \mid z) = \frac{1}{1+\exp(\sum_{k=1}^L w_k z_k)}$$

Apply maximum likelihood.

log-likelihood $\mathcal{L}(\{z^{(l)}, Y^{(l)}\}_{l=1}^N, w) = \frac{1}{N} \sum_{l=1}^N \log P_w(Y^{(l)} \mid z^{(l)})$

$$= \frac{1}{N} \sum_{l=1}^N \left\{ Y^{(l)} P_w(Y=1 \mid z^{(l)}) + (1-Y^{(l)}) P_w(Y=0 \mid z^{(l)}) \right\}$$

$$= \frac{1}{N} \sum_{l=1}^N \left\{ Y^{(l)} \cdot \left(\sum_{k=1}^L w_k \cdot z_k^{(l)}\right) - \log\left(1+\exp\left(\sum_{k=1}^L w_k z_k^{(l)}\right)\right) \right\}$$

this is a concave maximization in $w \in \mathbb{R}^L$.

use Gradient Ascent to find $\hat{w}$.

Initialize $w^{(0)} = \mathbb{1}$

Repeat $w^{(t+1)} = w^{(t)} + \frac{1}{t} \cdot \nabla_w \mathcal{L}(\{Y^{(l)}, z^{(l)}\}_{l=1}^N, w^{(t)})$

$$\frac{1}{N} \sum_{l=1}^N Y^{(l)} \cdot z^{(l)} - \frac{\exp(\langle w^{(t)}, z^{(l)} \rangle)}{1+\exp(\langle w^{(t)}, z^{(l)} \rangle)} \cdot z^{(l)}$$

$$\left[\!\left[\right\}_L - \square \cdot \left[\!\left[\right.\right.\right.$$

$$= w^{(t)} + \frac{1}{t} \cdot \frac{1}{N} \sum_{l=1}^N z^{(l)} \left(Y^{(l)} - P_{w^{(t)}}(Y=1 \mid z^{(l)}) \right)$$

$Y^{(l)}$: sample label; $P_{w^{(t)}}(Y=1 \mid z^{(l)})$: current prediction

* Logistic regression for neighborhood selection.

structural learning

$$P(X_1=1 \mid X_2 \cdots X_n) = \frac{e^{\theta_{11} + \sum_j \theta_{1j} X_j}}{1 + e^{\theta_{11} + \sum_{j \neq 1} \theta_{1j} X_j}}$$

θ_{11} ↑ G

logistic regression

$$P(Y=1 \mid Z) = \frac{e^{\sum w_k z_k}}{1 + e^{\sum w_k z_k}}$$

for node 1, if we know G, then features are $(1, X_{\partial 1})$
label is X_1.

as we do not know G, suppose ground truths $|\theta_{ij}| \leq 1$.
and degree $\leq K$.

this motivates the following algorithm.

$$\min_{\theta_{1\cdot}} \; -\mathcal{L}\left(\{X_1^{(l)}, \underbrace{(1, X_2^{(l)}, \cdots, X_n^{(l)})}_{\text{feature.}}\}_{l=1}^{N}, \theta_{1\cdot}\right)$$

$\theta_{1\cdot} = (\theta_{11}, \theta_{12}, \cdots \theta_{1n}) \in \mathbb{R}^n$

$\theta_{11}, \theta_{12}, \cdots, \theta_{1n}$

$\theta_{11} + \theta_{12}X_2 + \theta_3 X_3 \cdots$ (feature 1.)

s.t. $\|\theta_{1\cdot}\|_{L_1} \leq K$

$\exists \lambda^*(K)$ gives same solution

OR equivalently,

$$(**) \quad \min_{\theta_{1\cdot}} \; -\mathcal{L}(\quad, \theta_{1\cdot}) + \lambda \|\theta_{1\cdot}\|_{L_1}$$

$\hat{\theta}_{1\cdot} \rightarrow (0\ \hat{\theta}_{13}\ 0\ \cdots\ 0\ \hat{\theta}_{17}\ 0\ 0\ 0)$

structure

values ≃ factors

Theorem [Klivans Meka 2017]

If max degree $\le K$

$P(x_i = 1 \mid x_{-i}) \in [\delta, 1-\delta]$ for some $\delta > 0$

number of samples $N \ge C \cdot \log n \cdot \frac{1}{\varepsilon^2} \times K \times C_\delta$

Then (**) achieves $\|\hat{\theta} - \theta^*\|_\infty \le \varepsilon$

$\max_{ij} |\hat{\theta}_{ij} - \theta^*_{ij}| \le \varepsilon$

run-time $O(n^2 \operatorname{poly} \log \frac{n}{\varepsilon})$.

if all $|\theta^*_{ij}| > 2\varepsilon$, then threshold $|\hat{\theta}_{ij}| \gtrless \varepsilon$

learn structure exactly.

* This principle can be used for more general graphical models example of Gaussian Graphical models.

$$P(x) = \frac{1}{Z} \exp\{-\tfrac{1}{2} x^T J x\}$$

w.l.o.g $h = 0$

$\mathcal{X} = \mathbb{R}$

① Compute conditional distribution.

$$P(x_1 \mid x_2^n) = \frac{1}{Z_1} \cdot \exp\{-\tfrac{1}{2}(J_{11} x_1^2 + 2 x_1 \underbrace{(\sum_{j \ge 1} J_{1j} \cdot x_j)}_{\beta})\}$$

$$= \frac{1}{\sqrt{2\pi \cdot \frac{1}{J_{11}}}} \cdot \exp\left\{-\frac{1}{2} \frac{(x_1 + \frac{\beta}{J_{11}})^2}{\frac{1}{J_{11}}}\right\}$$

② Apply maximum likelihood to estimate X_1 from X_2^n

$$\min -\frac{1}{N} \sum_{\ell=1}^{N} \log P(x_1^{(\ell)} \mid x_2^{(\ell)} \dots x_n^{(\ell)}) = \frac{J_{11}}{2N} \sum_{\ell=1}^{N} \left(x_1^{(\ell)} - \sum_{j \ne 1} \frac{J_{1j}}{-J_{11}} x_j^{(\ell)} \right)^2$$

J_{11} can be made small

ratio

label

w_j model

x_j features

J_{11}, J_{1j} be small

$-\frac{1}{2} \log J_{11}$

as we care about structure,
we re-parametrize it by $w \in \mathbb{R}^{n-1}$

[Hw4]

$$\min_{w \in \mathbb{R}^{n-1}} \frac{1}{N} \sum_{l=1}^{N} \left(X_1^{(l)} - \sum_{j=2}^{n} w_j \cdot X_j^{(l)} \right)^2 + \lambda \|w\|_1$$

* Motivation: [Get J directly
without separating the neighborhoods

Graphical Lasso.

Step 1: Compute the Covariance Matrix $S_{ij} = \frac{1}{N} \sum_{l=1}^{N} X_i^{(l)} X_j^{(l)}$

Step 2: Maximize likelihood for information matrix J.

$$\hat{J} = \arg\max_{J} \{ \underbrace{\log|J| - \mathrm{Tr}(S \cdot J)}_{\text{likelihood}} - \underbrace{\lambda \|J\|_{L_1}}_{\text{regularize}} \}$$

$\log|J|$: $|J|$ = determinant

$\mathrm{Tr}(S \cdot J)$: $\mathrm{Tr}(A) = \sum_i A_{ii}$

$\|J\|_{L_1} = \sum_{ij} |J_{ij}|$

s.t $J \succ 0$

*This concave maximization