* Structural Learning Recap.

Approach 1: ML for DAG

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

[Chow-Liu Algorithm] $\longleftrightarrow$ Max-weight spanning
If search over all trees, then optimization
can be solved efficiently exactly.

Approach 2. ML for Ising models

$$\min_{\theta \in \mathbb{R}^{n \times n}} \log Z_G(\theta) - \langle \hat{M}, \theta \rangle + \lambda \|\theta\|_{L1}$$

$\|\theta\|_{L1} = \sum_{ij} |\theta_{ij}|$

$\hat{M} = \begin{bmatrix} \hat{M}_{11} & \hat{M}_{12} & \cdots \end{bmatrix}$ $\quad \theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots \\ & \ddots & \ddots \end{bmatrix}$

$\hat{M}_{11} = \frac{1}{N}\sum_{j=1}^{N} X_1^{(j)}$ $\quad \hat{M}_{12} = \frac{1}{N}\sum_{j=1}^{N} X_1^{(j)} X_2^{(j)}$

Complex to evaluated

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

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 $X_{-i} \triangleq \{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots x_n\}$ (subset of this set)

Claim: A condition distribution of x_1 given the rest is

$$(*) \quad \frac{P(X_1 = 1 | \overbrace{X_2 \cdots X_n}^{X_{-1}})}{P(X_1 = 0 | X_2 \cdots X_n)} = \exp\left\{ \theta_{11} + \sum_{j \in \partial 1} \theta_{1j} x_j \right\}$$

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

$$(*) \longrightarrow P(X_1 = 1 | X_{-1}) = \frac{e^{\theta_{11} + \sum_j \theta_{1j} x_j}}{1 + e^{\theta_{11} + \sum_j \theta_{1j} x_j}}$$

$$= \text{Sigmoid}\left(\underbrace{\theta_{11} + \textstyle\sum_j \theta_{1j} x_j}_{\text{linear in } \theta_1} \right).$$

Proof (*) >

$$P(X_1 = 1 | X_2^n) = \frac{P(X_1 = 1 \wedge X_2^n)}{P(X_2^n)}$$

$$= \frac{1/Z(\theta) \cdot \exp\left\{ \theta_{11} \cdot 1 + \sum_{j=2}^{n} \theta_{1j} x_j + \sum_{j=2}^{n} \theta_{1j} x_j \cdot 1 + \sum_{\substack{i,j > 1 \\ (i,j) \in E}} \theta_{ij} x_i x_j \right\}}{P(X_2^n)}$$

$j \in \partial 1$

$$P(X_1 = 0 | X_2^n) =$$

$$P(X_1=1 \mid X_2^n) = \frac{\exp\{\theta_{11} + \sum_j X_j \theta_{1j}\}}{1+\exp\{\theta_{11} + \sum_j X_j \theta_{1j}\}}$$

* Review of logistic regression

- We are given labelled examples of features $z^{(l)} = (z_1^{(l)}, \dots z_L^{(l)}) \in \mathbb{R}^L$

  Data $= \{(z^{(l)}, y^{(l)})\}_{l=1}^N$ $\quad l \in \{1, \dots, N\}$ $\quad$ label $y^{(l)} \in \{0,1\}$

- Suppose a parametric model with $w \in \mathbb{R}^L$

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

$$P(y=0 \mid z) = \frac{1}{1+\exp(\cdot \quad)}$$

- Maximum likelihood

log-likelihood $\mathcal{L}(\{(z^{(l)}, y^{(l)})\}_{l=1}^N, w \in \mathbb{R}^L) \triangleq \sum_{l=1}^N \log P_w(y^{(l)} \mid z^{(l)})$

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

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

utility/objective over w

* this is a concave maximization over $w \in \mathbb{R}^L$
  use gradient ascent to solve efficiently.

* Logistic Regression for neighborhood selection

structural learning.

$$p(X_1 = 1 \mid X_{\setminus 1}) = \frac{e^{\theta_{11} + \sum_{j \in \partial 1} \theta_{1j} X_j}}{1 + e^{\theta_{11} + \sum_{j \in \partial 1} \theta_{1j} X_j}}$$

$\partial 1$ ↑ G

logistic regression

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

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 $|\theta_{ij}| \le 1$, and degree $\le K$

this motivates following **Sparse logistic regression.**

$$\min_{\theta_{1\cdot} \in \mathbb{R}^n} \; -\mathcal{L}\left(\{ \underbrace{(1, x_2^{(l)}, \dots, x_n^{(l)})}_{\text{feature}}, \underbrace{x_1^{(l)}}_{\text{label}} \}_{l=1}^N, \theta_{1\cdot} \right)$$

$(\theta_{11}, \theta_{12}, \dots \theta_{1n})$

s.t. $\| \theta_{1\cdot} \|_{L_1} \le K$ or equivalently

$\| \theta_{1\cdot} \|_{L_1} = \sum_j |\theta_{1j}|$

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

$$\Rightarrow \quad \theta_{1\cdot} \approx (0, \underbrace{\theta_{13}, 0, 0, \dots 0, \theta_{1,18}}_{\text{Graph structure.}}, \dots)$$

Theorem [Klivans, Meka 2017]

If max degree $\le k$

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

for some $\varepsilon > 0$

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

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

$\|\cdot\|_\infty$: $\max_{ij} |\theta_{ij} - \theta^*_{ij}| \le \varepsilon$, with high probability

and runtime $O(n^2 \cdot \text{poly}\log \frac{n}{\varepsilon})$

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

learn structure exactly.

---

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

* This principle can be used for more general graphical models

- Gaussian Graphical Models.

$$P(x) = \frac{1}{Z_J} \exp\left\{ -\frac{1}{2} x^T J x \right\}$$

$J = \Sigma^{-1}$

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

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

① Compute conditional distribution

$$P(X_1 \mid X_2^n) = \frac{1}{Z_1} \cdot \exp\left\{ -\frac{1}{2} \left(J_{11} x_1^2 + 2 \underbrace{\left(\sum_{j \ne 1} J_{1j} X_j \right)}_{\beta} \cdot X_1 \right\} \right.$$

β: 2 $\times$ neighborhood of node 1

$$= \frac{1}{\sqrt{2\pi \cdot \frac{1}{J_{11}}}} \cdot \exp\left\{ -\frac{1}{2} \frac{\left(X_1 - \frac{\beta}{J_{11}}\right)^2}{1/J_{11}} \right\}$$

② Apply Maximum Likelihood principle to get X_1 from X_2^n

$$\min -\frac{1}{N}\sum_{l=1}^{N} \log P(X_1^{(l)} \mid X_2^{(l)} \dots X_n^{(l)})$$

$$= \frac{J_{11}}{2N} \cdot \sum_{l=1}^{N} \left(X_1^{(l)} - \sum_{j\neq 1} \frac{J_{1j}}{-J_{11}} X_j^{(l)} \right)^2 - \frac{1}{2}\log J_{11}$$

ratio. label features. L_2-loss

w_j : modal parameter

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

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

---

* Graphical Lasso.

Motivation: — Get J directly. without separating each neighborhood.

Step 1. Compute the empirial covariance matrix

$$S_{ij} = \frac{1}{N}\sum_{l=1}^{N} x_i^{(l)} x_j^{(l)}$$

Step 2. maximize likelihood over Information matrix $J \in \mathbb{R}^{n\times n}$

$$\hat{J} \in \arg\max_{J} \{ \log|J| - \mathrm{Tr}(S\cdot J) - \lambda \|J\|_{1,1} \}$$

determinant; $\mathrm{Tr}(A) = \sum_i A_{ii}$; likelihood; regularizer.

s.t. $J \succ 0$

Others : linear regression → semi-definite program.