GRAPHICAL MODELS:

CONDITION INDEPENDENCE (CI):

$$\boxed{X \perp Y \mid Z}$$

Distribution D: $(x_1, x_2 \dots x_d) \in \{0,1\}^d$

Structure: Rooted Tree.

KL-DIVERGENCE / CROSS-ENTROPY:

Distance between 2 distributions over some space, $\Omega \rightarrow d_1, d_2$

$$\boxed{KL(d_1 \| d_2) = \sum_{s \in \Omega} d_1(s) \log \frac{d_2(s)}{d_1(s)}}$$

$d_1(s)$: Probability s happens under d_1.

Goal: Given distribution P on Σ^d generated from unknown Bayes net (Tree: T^*), find the T and Bayes net P_T such that

$$KL(P, P_T) \leq \varepsilon.$$

$$\Sigma = \{0,1\}$$

Chow - Liu Bound:

For any tree T,

$$\boxed{KL(P \| P_T) = \underbrace{J_P}_{\text{function of } P} - \sum_{(i,j) \text{ is an edge in } T} I(x_i, x_j)}$$

Mutual Information:

$I(x_i; x_j) \equiv$ Measures how much information x_i has about x_j.

$\Downarrow$

can be estimated from samples [Independent of T].

CHOW- LIU ALGORITHM

- $\rightarrow$ Use samples to estimate $I(X_i;X_j)$ for all i,j.
- $\rightarrow$ Form a weighted graph where weights are exactly $I(X_i;X_j)$
- $\rightarrow$ Compute the max. Spanning tree T' in G
- $\rightarrow$ Output $P_{T'}$.

UNDIRECTED GRAPHICAL MODELS:

MARKOV RANDOM FIELDS:

Given $D : (x_1, x_2 \dots x_d)$ $\{0,1\}^d$

G: Dependency graph for G.

D satisfies with respect to G If i,j have no edge, then	Pairwise Markov Property	Local	Global
	$x_i \perp x_j \mid x_{rest}$	$x_i \perp x_j \mid x_{\{\text{neighbors of } i\}}$	$x_i \perp x_j \mid x_{\{\text{any separating set}\}}$ ↓ vertices removing which i,j disconnected.

Global => Local => Pairwise.

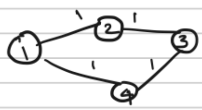
Remark: We'll say D has dependency graph G if it satisfies Markov property with respect to G.

Goal: Find D such that each vertex has degree $\leq k$.

Learning Boltzmann Machines

D on $\{1,-1\}^d$

$$\Pr[X=x] \propto \exp\Big(\sum_{\{i,j\}\in G} W_{ij}\, x_i x_j\Big)$$



$$\Pr[X=x] \propto \exp\big(x_1x_2 + x_2x_3 + x_3x_4 + x_4x_1\big)$$

Distributions as defined above, they satisfy Markov property wrt G.

Gaussian Graphical Models

D on $\mathbb{R}^d$

Dist D is $N(0,\Sigma)$

Σ is the covariance matrix. $X \sim N(0,\Sigma)$

$\Sigma_{ij} = \mathbb{E}[X_i X_j]$

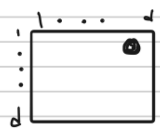
$\Sigma_{ij}=0 \Rightarrow X_i, X_j$ are independent.

(Dempsey 1972):

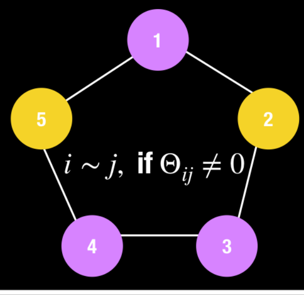
Precision Matrix:

$\Theta = \Sigma^{-1}$.

Thm: Gaussian dist has dependency graph with $\mathrm{Supp}(\Theta)$.



$i,j \in G$ if $\Theta_{ij} \neq 0$.



Example: $(X_1 \mid X_2, X_5)$ **independent of** $(X_3 \mid X_2, X_5)$

Markov property: $\Theta_{ij} = 0 \Rightarrow X_i, X_j$ **are independent conditioned on neighbors of** i.

Structure Learning for GGMs

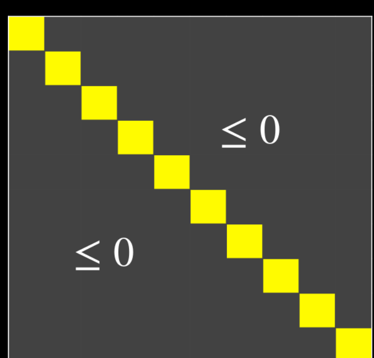
Given samples $X^1, X^2, \ldots, X^n$ from a GGM of degree $d \ll p$, can we efficiently find the dependency graph with $n \ll p$?

(**Think:** $n = O_d(\log p)$.)

Attractive GGMs

GGM is attractive if all covariances are non-negative.

(Equivalently, Θ has non-positive off-diagonals.)



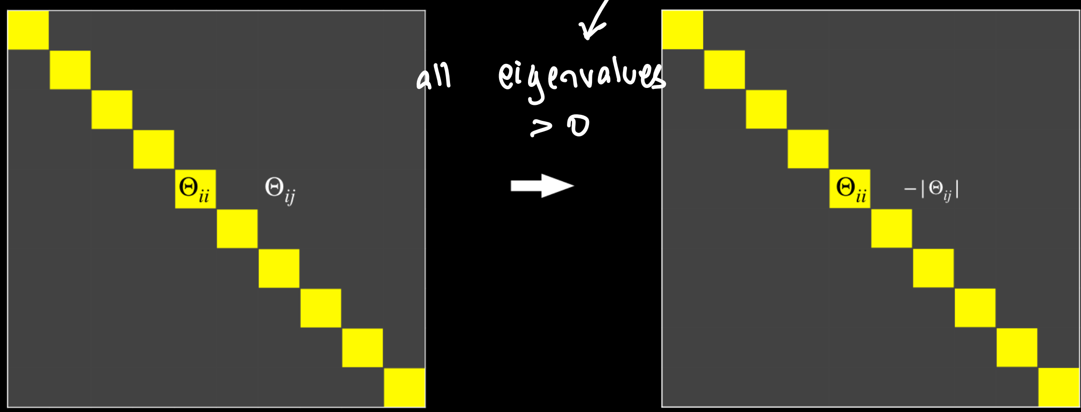
Θ : Precision Matrix

Ex: Gaussian Free Fields

- **Many applications via Gaussian processes**

Walk-Summable GGMs

GGM walk-summable if making off-diagonals of precision matrix negative preserves positive semi-definiteness.



all eigenvalues > 0

Θ : Precision Matrix

Offdiagonals negative $\succeq 0$

GREEDYPRUNE

1. **Recover neighborhood of each vertex in parallel.**
2. **Grow a candidate neighborhood.**
3. **Prune out some vertices.**

GREEDY-GROWING

1. **Set** $S \leftarrow \varnothing$
2. **While S is small enough:**
   1. **Find j to minimize estimate of** $Var(X_1 \mid X_{S \cup j})$.
   2. $S \leftarrow S \cup \{j\}$.

→ That j and x_1 are clearly dependent as j minimizes x_1's variance.

Intuition: Add vertex that gives maximum decrease in conditional variance.

$\hookrightarrow$ For each $var(x_1 | x_j)$ we need $1/\varepsilon^2$ samples to get $(1-\varepsilon)$ accuracy.

If $var(x_1 | x_2, x_3, x_4) \to 3/\varepsilon^2$

GREEDY-PRUNING

1. **For each j in S:**
   1. **If** $Var(X_1 | X_{S \setminus \{j\}}) < (1+\tau) Var(X_1 | X_S)$**, drop j from S.**

Intuition: If dropping a vertex, does not hurt too much, drop it.

Can learn Attractive and Walk-Summable GGMs with $O(d^2 \log p / \kappa^6)$ samples and quadratic run-time.

d - degree (max)

p - total number of parameters

$$\kappa(\Theta) = \min_{i,j:\Theta_{ij} \neq 0} \frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}}$$

SUMMARY:

1. KL-Divergence:

$$KL(d_1 \| d_2) = \sum_{s \in \Omega} d_1(s) \log \frac{d_2(s)}{d_1(s)}$$

2. Chow-Liu Bound:

$$KL(P \| P_T) = J_P - \sum_{(i,j) \text{ is an edge in } T} I(x_i, x_j)$$

3. Chow-Liu Algorithm:

- Find $I(x_i, x_j)$ - graph
- Max Spanning Tree $\to T'$

4. Markov Random Fields:

D (Markov Property) with respect to G : if no edge i, j

$$x_i \perp x_j \mid x_{\{\text{neighbors of } i\}}$$

5. Learning Boltzmann Machines: D on $\{1, -1\}^d$

$$\Pr[X = x] \propto e^{\sum_{(i,j) \in G} w_{ij} x_i x_j}$$

6. Gaussian Graphical models:

$D: N(0, \Sigma)$ $\quad \Sigma_{ij} = E[x_i x_j]$

$\Sigma_{ij} = 0 \Rightarrow$ independent

$$\Theta = \Sigma^{-1}$$

$\text{Support}(\Theta) \to G \quad i,j \in G \text{ if } \Theta_{ij} \neq 0$

7. Attractive: $\Sigma_{ij} > 0$, $\Theta_{ij} \leq 0 \ \forall i \neq j$

8. Walk summable: $\Theta_{ij} \rightarrow -\Theta_{ij} \ \forall i \neq j$ => Still positive semi Definite.

9. Greedy Prune

For each vertex i → find neighborhood in parallel

→ $\arg\min_{j} var(x_i \mid x_{S \cup j})$

→ $var(x_i \mid x_{S - \{j\}}) < (1 + \tau)\, var(x_i \mid x_S)$

→ learns 2 types $O(d^2 \log p / \kappa^6)$ samples, quadratic time

d - max degree

p - number of parameters

$$\kappa(\Theta) = \min_{i,j : \Theta_{ij} \neq 0} \frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}}$$