No voodoo here! Learning discrete graphical models via inverse covariance estimation

Po-Ling Loh

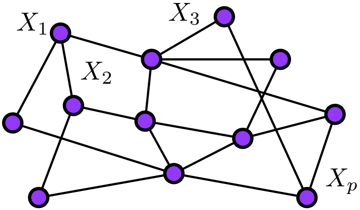
UC Berkeley
Department of Statistics

NIPS 2012
December 5, 2012

Joint work with Martin Wainwright

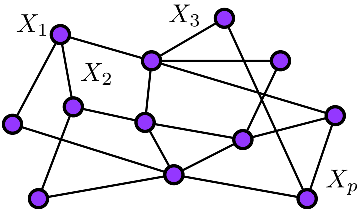
Introduction: Graphical models

- Graph $G = (V, E)$
- Represents joint distribution of $(X_1, \ldots, X_p)$, where $|V| = p$



Introduction: Graphical models

- Graph $G = (V, E)$
- Represents joint distribution of $(X_1, \ldots, X_p)$, where $|V| = p$

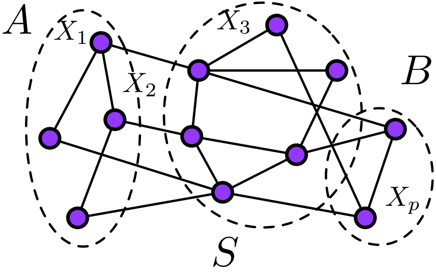


- Absent edges indicate conditional independence:

$$(s, t) \notin E \implies X_s \perp\!\!\!\perp X_t \mid X_{\setminus\{s,t\}}$$

Introduction: Graphical models

- Graph $G = (V, E)$
- Represents joint distribution of $(X_1, \ldots, X_p)$, where $|V| = p$



- More generally, $X_A \perp\!\!\!\perp X_B \mid X_S$ whenever $S \subseteq V$ separates A from B

Introduction: Graphical models

- Wide applications in computer vision, civil engineering, political science, epidemiology ...
- **Goal:** Edge recovery from n samples: $\left\{(X_1^{(i)}, X_2^{(i)}, \dots, X_p^{(i)})\right\}_{i=1}^n$

- Wide applications in computer vision, civil engineering, political science, epidemiology ...
- **Goal:** Edge recovery from n samples: $\left\{(X_1^{(i)}, X_2^{(i)}, \dots, X_p^{(i)})\right\}_{i=1}^n$
- High-dimensional setting: $p \gg n$
- Samples may be non-i.i.d. or corrupted by noise/missing data

Structure learning for Gaussians

- When $(X_1, \dots, X_p) \sim N(0, \Sigma)$, Hammersley-Clifford implies

$$(\Sigma^{-1})_{st} = 0 \iff (s,t) \notin E$$

Structure learning for Gaussians

- When $(X_1, \ldots, X_p) \sim N(0, \Sigma)$, Hammersley-Clifford implies

$$(\Sigma^{-1})_{st} = 0 \iff (s,t) \notin E$$

- Numerous methods for edge recovery by estimating $\Theta = \Sigma^{-1}$:
  - Nodewise regression with Lasso (Meinshausen & Bühlmann '06)
  - Global estimation of Θ with penalized MLE (Yuan & Lin '07)
  - Nonparanormal (Liu et al. '09, '12)

Non-Gaussian distributions

- In non-Gaussian setting, relationship between entries of $\Theta = \Sigma^{-1}$ and edges of G is unclear

Non-Gaussian distributions

- In non-Gaussian setting, relationship between entries of $\Theta = \Sigma^{-1}$ and edges of G is unclear

Main contributions:

- Establish relationship between **augmented** inverse covariance matrices and edge structure in discrete graphical models

Non-Gaussian distributions

- In non-Gaussian setting, relationship between entries of $\Theta = \Sigma^{-1}$ and edges of G is unclear

Main contributions:

- Establish relationship between **augmented** inverse covariance matrices and edge structure in discrete graphical models
- Propose two new algorithms for structure learning in discrete graphs

A curious example

- Binary Ising model:

$$\mathbb{P}_\theta(x_1, \ldots, x_p) \;\propto\; \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\right),$$

A curious example

- Binary Ising model:

$$\mathbb{P}_\theta(x_1, \ldots, x_p) \;\propto\; \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\right),$$

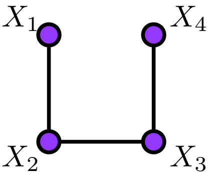
$$\theta \in \mathbb{R}^{p + \binom{p}{2}}, \qquad (x_1, \ldots, x_p) \in \{0, 1\}^p$$

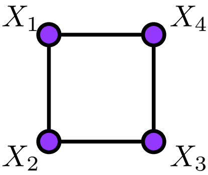
A curious example

- Ising models with $\theta_s = 0.1, \quad \theta_{st} = 2$

A curious example

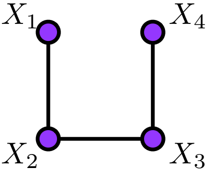
- Ising models with $\theta_s = 0.1, \quad \theta_{st} = 2$



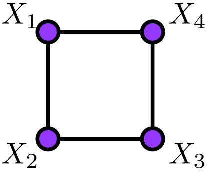


A curious example

- Ising models with $\theta_s = 0.1, \quad \theta_{st} = 2$



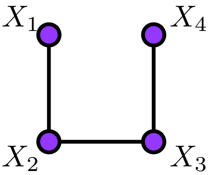
$$\Theta_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix}$$



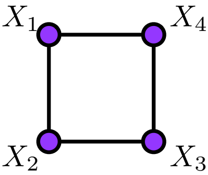
$$\Theta_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix}$$

A curious example

- Ising models with $\theta_s = 0.1, \quad \theta_{st} = 2$



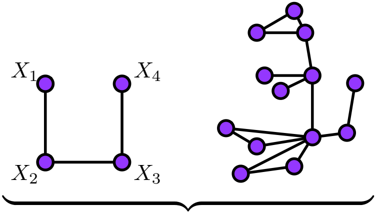
$$\Theta_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix}$$



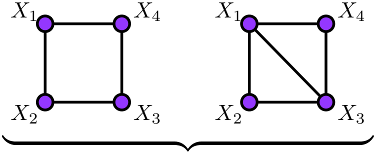
$$\Theta_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix}$$

- Θ is graph-structured for chain, but not loop

A curious example

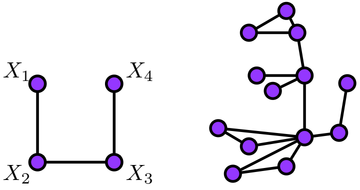


Θ graph-structured

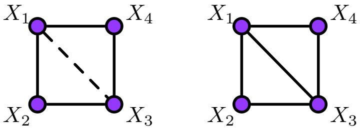


Θ **not** graph-structured

A curious example



Θ graph-structured



Θ **not** graph-structured

- However, letting $\Gamma_{\text{aug}} = \text{Cov}(X_1, X_2, X_3, X_4, X_1X_3)^{-1}$ for loop:

$$\Gamma_{\text{aug}} \propto \begin{bmatrix} 115 & -2 & 109 & -2 & -114 \\ -2 & 5 & -2 & 0 & 1 \\ 109 & -2 & 114 & -2 & -114 \\ -2 & 0 & -2 & 5 & 1 \\ -114 & 1 & -114 & 1 & 119 \end{bmatrix}$$

Some notation

- Assume $(X_1, \dots, X_p) \in \{0, \dots, m-1\}^p$

Some notation

- Assume $(X_1, \ldots, X_p) \in \{0, \ldots, m-1\}^p$
- For any subset $U \subseteq V$, associate vector ϕ_U of sufficient statistics

Some notation

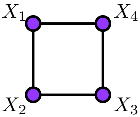
- Assume $(X_1, \ldots, X_p) \in \{0, \ldots, m-1\}^p$
- For any subset $U \subseteq V$, associate vector ϕ_U of sufficient statistics
- **Ex:** When $m = 2$ and $U = \{1, 2\}$, $\phi_U = (x_1, x_2, x_1 x_2)$

Some notation

- Assume $(X_1, \ldots, X_p) \in \{0, \ldots, m-1\}^p$
- For any subset $U \subseteq V$, associate vector ϕ_U of sufficient statistics

- **Ex:** When $m = 2$ and $U = \{1, 2\}$, $\phi_U = (x_1, x_2, x_1 x_2)$
- **Ex:** When $U = \{1\}$, $\phi_U = (\mathbb{I}\{x_1 = 1\}, \ldots, \mathbb{I}\{x_1 = m-1\})$

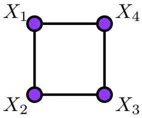
Augmenting covariance matrix



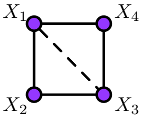
G

Augmenting covariance matrix

- Triangulate G



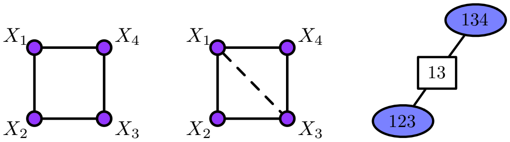
G



triangulated

Augmenting covariance matrix

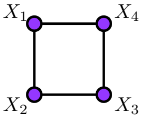
- Triangulate G
- Form junction tree with separator sets



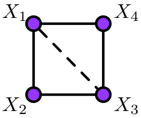
G triangulated junction tree

Augmenting covariance matrix

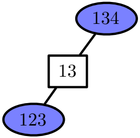
- Triangulate G
- Form junction tree with separator sets
- Let $\mathcal{S}^+$ = nodes + separator sets



G



triangulated



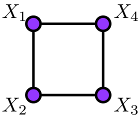
junction tree

$$
\begin{array}{c c}
 & \begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & X_1X_3 \end{array} \\
\begin{array}{c} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_1X_3 \end{array} & \left[\begin{array}{c} \\ \\ \mathrm{Cov}(\phi_{\mathcal{S}^+}) \\ \\ \\ \end{array} \right]
\end{array}
$$

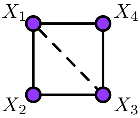
augmented matrix

Augmenting covariance matrix

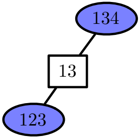
- Triangulate G
- Form junction tree with separator sets
- Let $\mathcal{S}^+$ = nodes + separator sets



G



triangulated



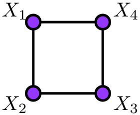
junction tree

$$\begin{array}{c|c} & X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_1X_3 \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_1X_3 \end{array} & \left[\begin{array}{c} \\ \\ \mathrm{Cov}(\phi_{\mathcal{S}^+}) \\ \\ \\ \end{array} \right] \end{array}$$

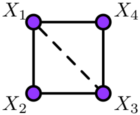
augmented matrix

Theorem

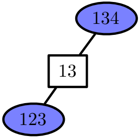
The inverse covariance matrix of $\{\phi_U : U \in \mathcal{S}^+\}$ from any junction tree is graph-structured



G



triangulated



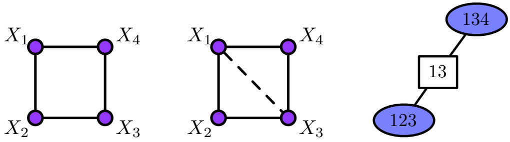
junction tree

$$\begin{array}{c|c} & X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_1X_3 \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_1X_3 \end{array} & \left[\begin{array}{c} \\ \\ \mathrm{Cov}(\phi_{\mathcal{S}^+}) \\ \\ \\ \end{array} \right] \end{array}$$

augmented matrix

$$\Gamma = (\mathrm{Cov}(\phi_{\mathcal{S}^+}))^{-1} \propto \begin{bmatrix} 115 & -2 & 109 & -2 & -114 \\ -2 & 5 & -2 & 0 & 1 \\ 109 & -2 & 114 & -2 & -114 \\ -2 & 0 & -2 & 5 & 1 \\ -114 & 1 & -114 & 1 & 119 \end{bmatrix}$$

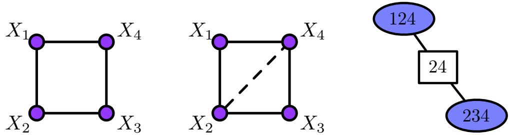
Example: Binary Ising model



$$\begin{array}{cc} & \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_1X_3 \end{matrix} \\ \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_1X_3 \end{matrix} & \begin{bmatrix} \\ \qquad \mathrm{Cov}(\phi_{\mathcal{S}^+}) \qquad \\ \\ \end{bmatrix} \end{array}$$

G triangulated junction tree augmented matrix

- Statistics included in $\phi_{\mathcal{S}^+}$ depend on triangulation



$$\begin{array}{cc} & \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_2X_4 \end{matrix} \\ \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_2X_4 \end{matrix} & \begin{bmatrix} \\ \qquad \mathrm{Cov}(\phi_{\mathcal{S}^+}) \qquad \\ \\ \end{bmatrix} \end{array}$$

Consequences for trees

- When $\exists$ triangulation with singleton separator sets, $\mathcal{S}^+ = \{1, \dots, p\}$

Consequences for trees

- When $\exists$ triangulation with singleton separator sets, $\mathcal{S}^+ = \{1, \ldots, p\}$

Corollary

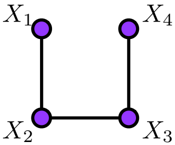
When G is a tree, inverse covariance matrix of sufficient statistics on nodes is graph-structured

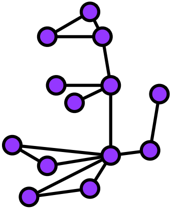
Consequences for trees

- When $\exists$ triangulation with singleton separator sets, $\mathcal{S}^+ = \{1, \ldots, p\}$

Corollary

When G is a tree, inverse covariance matrix of sufficient statistics on nodes is graph-structured





$$(\mathrm{Cov}(X_1, \ldots, X_p))^{-1}$$

Structure learning

Structure learning

- Graphical Lasso for Gaussian graphical models:

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0} \left\{ \operatorname{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda \sum_{s \neq t} |\Theta_{st}| \right\}$$

- Based on penalized MLE for Gaussians

Structure learning

- Graphical Lasso for Gaussian graphical models:

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0} \left\{ \operatorname{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda \sum_{s \neq t} |\Theta_{st}| \right\}$$

- Based on penalized MLE for Gaussians

- **When does graphical Lasso succeed for non-Gaussians?**

Graphical Lasso

Corollary

For binary Ising models with singleton separators, the graphical Lasso succeeds w.h.p. when $n \gtrsim d^2 \log p$

Graphical Lasso

Corollary

For binary Ising models with singleton separators, the graphical Lasso succeeds w.h.p. when $n \gtrsim d^2 \log p$

- Graphical Lasso **completely unrelated** to MLE for non-Gaussians
- Population results imply graphical Lasso is **inconsistent in general**

Graphical Lasso

Corollary

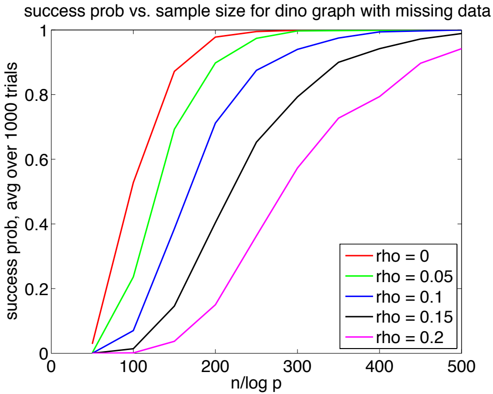
For binary Ising models with singleton separators, the graphical Lasso succeeds w.h.p. when $n \gtrsim d^2 \log p$

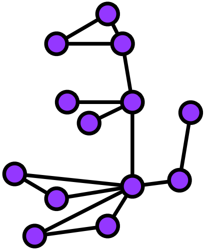
- Graphical Lasso **completely unrelated** to MLE for non-Gaussians
- Population results imply graphical Lasso is **inconsistent in general**

- Group graphical Lasso for $m > 2$
- Easily accommodates additive noise/missing data (modify $\widehat{\Sigma}$)

Simulation study

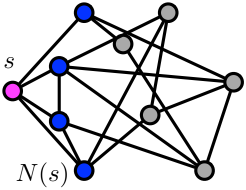
- Graphical Lasso for dinosaur graph: probability of success for recovering 15 edges vs. rescaled sample size (with missing data)





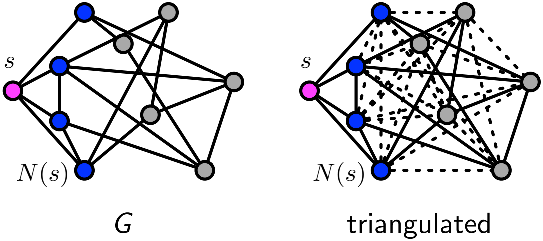
Inference methods for non-trees

- Nodewise method: recovers neighborhood $N(s)$ for any fixed $s \in V$



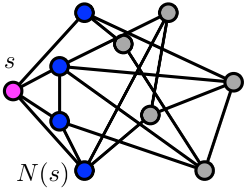
G

- Nodewise method: recovers neighborhood $N(s)$ for any fixed $s \in V$
- Form junction tree by fully-connecting all nodes in $V \backslash s$

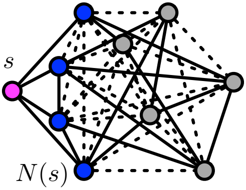


Inference methods for non-trees

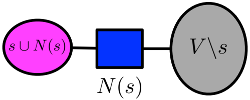
- Nodewise method: recovers neighborhood $N(s)$ for any fixed $s \in V$
- Form junction tree by fully-connecting all nodes in $V \backslash s$



G



triangulated



junction tree

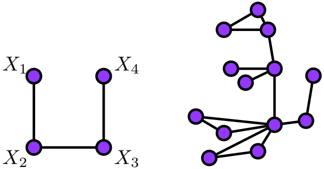
More details at poster!

Summary

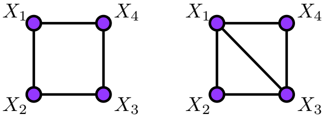
- Established relationship between augmented inverse covariance matrices and edge structure in discrete graphical models

Summary

- Established relationship between augmented inverse covariance matrices and edge structure in discrete graphical models
- Demystified relationship between

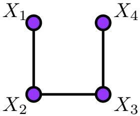


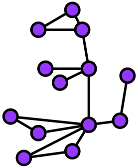
vs.



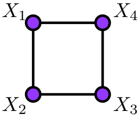
Summary

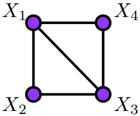
- Established relationship between augmented inverse covariance matrices and edge structure in discrete graphical models
- Demystified relationship between





vs.

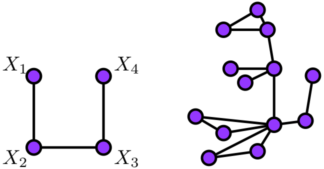




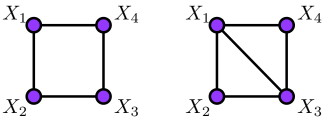
- Proposed structure learning methods for arbitrary discrete graphs

Summary

- Established relationship between augmented inverse covariance matrices and edge structure in discrete graphical models
- Demystified relationship between



vs.



- Proposed structure learning methods for arbitrary discrete graphs
- Methods are theoretically rigorous and easily adapted to corrupted observations