Beyond the graphical Lasso: Structure learning via inverse covariance estimation

Po-Ling Loh

UC Berkeley
Department of Statistics

ICML Workshop on Covariance Selection and Graphical Model Structure Learning

June 26, 2014

Joint work with Martin Wainwright (UC Berkeley) & Peter Bühlmann (ETH Zürich)

Outline

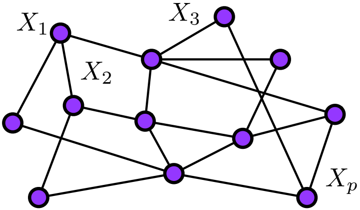
1. Introduction
2. Generalized inverse covariances
3. Linear structural equation models
4. Corrupted data

Outline

1. Introduction
2. Generalized inverse covariances
3. Linear structural equation models
4. Corrupted data

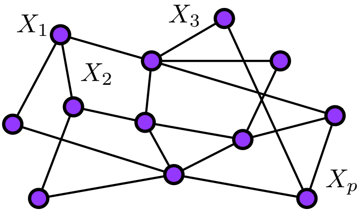
Undirected graphical models

- Undirected graph $G = (V, E)$
- Joint distribution of $(X_1, \ldots, X_p)$, where $|V| = p$



Undirected graphical models

- Undirected graph $G = (V, E)$
- Joint distribution of $(X_1, \ldots, X_p)$, where $|V| = p$

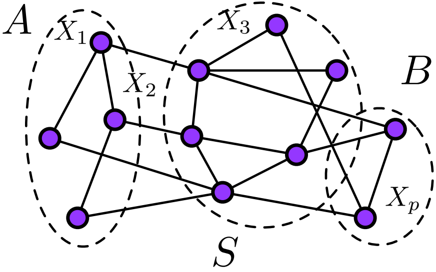


- Markov property:

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

Undirected graphical models

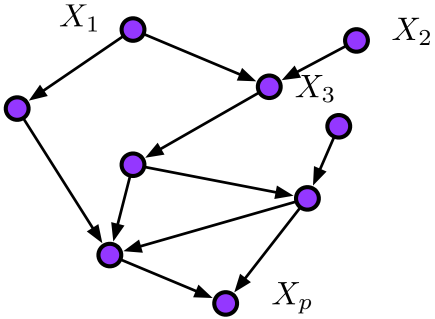
- Undirected graph $G = (V, E)$
- Joint distribution of $(X_1, \dots, X_p)$, where $|V| = p$



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

Directed graphical models

- Directed acyclic graph $G = (V, E)$



- Markov property:

$$X_j \perp\!\!\!\perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)}, \qquad \forall j$$

Structure learning

- **Goal:** Edge recovery from n samples: $\left\{(X_1^{(i)}, X_2^{(i)}, \dots, X_p^{(i)})\right\}_{i=1}^n$

Structure learning

- **Goal:** Edge recovery from n samples: $\left\{(X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)})\right\}_{i=1}^n$
- High-dimensional setting: $p \gg n$, assume $\deg(G) \leq d$

Structure learning

- **Goal:** Edge recovery from n samples: $\left\{(X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)})\right\}_{i=1}^n$
- High-dimensional setting: $p \gg n$, assume $\deg(G) \leq d$
- Sources of corruption: non-i.i.d. observations, contamination by noise/missing data

Structure learning

- **Goal:** Edge recovery from n samples: $\left\{(X_1^{(i)}, X_2^{(i)}, \dots, X_p^{(i)})\right\}_{i=1}^n$
- High-dimensional setting: $p \gg n$, assume $\deg(G) \leq d$
- Sources of corruption: non-i.i.d. observations, contamination by noise/missing data

- **Note:** Structure learning generally harder for directed graphs (topological order unknown)

Graphical Lasso

- When $(X_1, \ldots, X_p) \sim N(0, \Sigma)$, well-known fact:

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

Graphical Lasso

- When $(X_1, \ldots, X_p) \sim N(0, \Sigma)$, well-known fact:

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

- Establishes statistical consistency of graphical Lasso (Yuan & Lin '07):

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

Some observations

Some observations

- Only sample-based quantity is $\widehat{\Sigma}$:

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

Some observations

- Only sample-based quantity is $\widehat{\Sigma}$:

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

- Although graphical Lasso is *penalized Gaussian MLE*, can *always* be used to estimate $\widehat{\Theta}$ from $\widehat{\Sigma}$:

$$(\Sigma^*)^{-1} = \arg\min_{\Theta} \left\{ \text{trace}(\Sigma^*\Theta) - \log\det(\Theta) \right\}$$

Some observations

- Only sample-based quantity is $\widehat{\Sigma}$:

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

- Although graphical Lasso is *penalized Gaussian MLE*, can *always* be used to estimate $\widehat{\Theta}$ from $\widehat{\Sigma}$:

$$(\Sigma^*)^{-1} = \arg\min_{\Theta} \left\{ \text{trace}(\Sigma^*\Theta) - \log\det(\Theta) \right\}$$

- **We extend graphical Lasso to discrete-valued data (undirected case) and linear structural equation models (directed case)**

Theory for graphical Lasso

- If
$$\|\widehat{\Sigma} - \Sigma^*\|_{\max} \precsim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \succsim \sqrt{\frac{\log p}{n}},$$
then
$$\|\widehat{\Theta} - \Theta^*\|_{\max} \precsim \left(\sqrt{\frac{\log p}{n}} + \lambda\right)$$

Theory for graphical Lasso

- If
$$\|\widehat{\Sigma} - \Sigma^*\|_{\max} \precsim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \succsim \sqrt{\frac{\log p}{n}},$$
then
$$\|\widehat{\Theta} - \Theta^*\|_{\max} \precsim \left(\sqrt{\frac{\log p}{n}} + \lambda \right)$$

- Deviation condition holds w.h.p. for various ensembles (e.g., sub-Gaussian)

Theory for graphical Lasso

- If
$$\|\widehat{\Sigma} - \Sigma^*\|_{\max} \precsim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \succsim \sqrt{\frac{\log p}{n}},$$
then
$$\|\widehat{\Theta} - \Theta^*\|_{\max} \precsim \left(\sqrt{\frac{\log p}{n}} + \lambda\right)$$
- Deviation condition holds w.h.p. for various ensembles (e.g., sub-Gaussian)
- Thresholding $\widehat{\Theta}$ at level $\sqrt{\frac{\log p}{n}}$ yields correct support

Outline

Non-Gaussian distributions

- (Liu et al. '09, '12): $(X_1, \dots, X_p)$ follows *nonparanormal distribution* if $(f_1(X_1), \dots, f_p(X_p)) \sim N(0, \Sigma)$, and f_j's monotone and differentiable

Non-Gaussian distributions

- (Liu et al. '09, '12): $(X_1, \dots, X_p)$ follows *nonparanormal distribution* if $(f_1(X_1), \dots, f_p(X_p)) \sim N(0, \Sigma)$, and f_j's monotone and differentiable
- Then $(i, j) \notin E$ iff $\Theta_{ij} = 0$

Non-Gaussian distributions

- (Liu et al. '09, '12): $(X_1, \dots, X_p)$ follows *nonparanormal distribution* if $(f_1(X_1), \dots, f_p(X_p)) \sim N(0, \Sigma)$, and f_j's monotone and differentiable
- Then $(i,j) \notin E$ iff $\Theta_{ij} = 0$

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

Discrete graphical models

- Assume X_i's take values in a discrete set: $\{0, 1, \ldots, m-1\}$

Discrete graphical models

- Assume X_i's take values in a discrete set: $\{0, 1, \ldots, m-1\}$

Our results:

- Establish relationship between **augmented** inverse covariance matrices and edge structure
- New algorithms for structure learning in discrete graphs

An illustrative example

- Binary Ising model:

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

An illustrative example

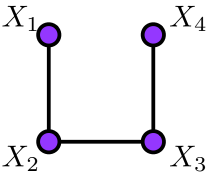
- Binary Ising model:

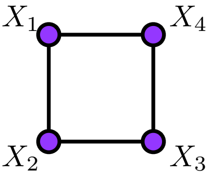
$$\mathbb{P}_\theta(x_1, \dots, x_p) \;\propto\; \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\right),$$

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

An illustrative example

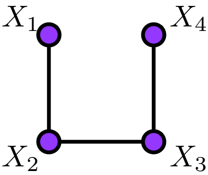
- Ising models with $\theta_s = 0.1, \quad \theta_{st} = 2$



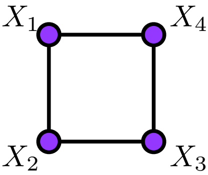


An illustrative example

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



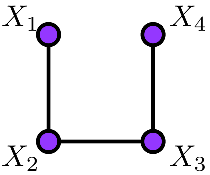
$$\Theta_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix}$$



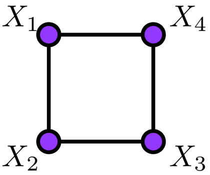
$$\Theta_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix}$$

An illustrative example

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



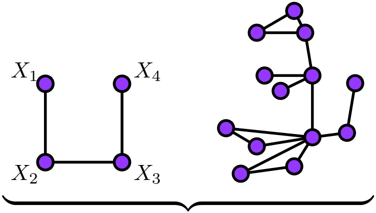
$$\Theta_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix}$$



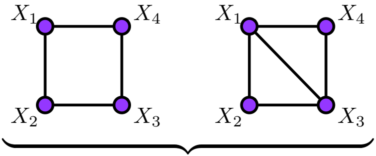
$$\Theta_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix}$$

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

An illustrative example

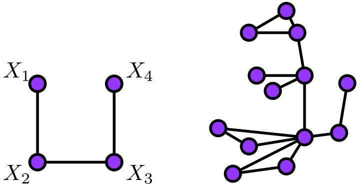


Θ graph-structured

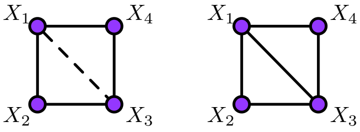


Θ **not** graph-structured

An illustrative 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:

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

Notation

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

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

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 $U = \{1, 2\}$, $\phi_U = (x_1, x_2, x_1 x_2)$

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 $U = \{1\}$, $\phi_U = (\mathbb{I}\{x_1 = 1\}, \ldots, \mathbb{I}\{x_1 = m-1\})$

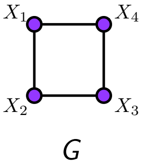
Notation

- 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 $U = \{1\}$, $\phi_U = (\mathbb{I}\{x_1 = 1\}, \ldots, \mathbb{I}\{x_1 = m-1\})$

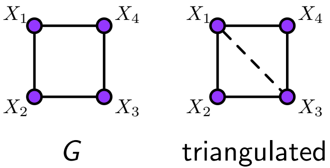
- **In general:** Clique $C \in \mathcal{C}$ has $(m-1)^{|C|}$ indicators of nonzero states

Augmented covariance matrices



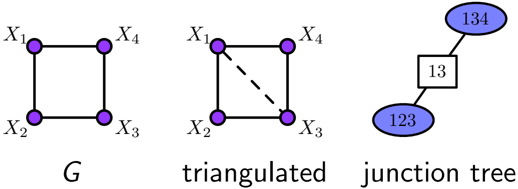
Augmented covariance matrices

- Triangulate G



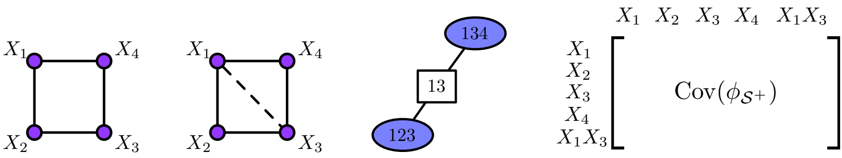
Augmented covariance matrices

- Triangulate G
- Form junction tree with separator sets



Augmented covariance matrices

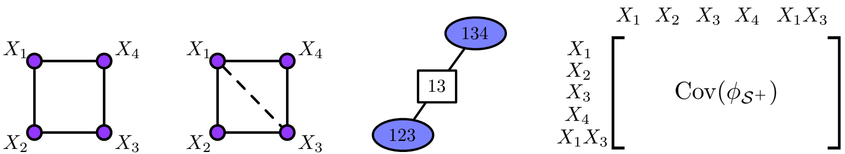
- Triangulate G
- Form junction tree with separator sets
- Let $\mathcal{S}^+$ = nodes + separator sets



G triangulated junction tree augmented matrix

Augmented covariance matrices

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

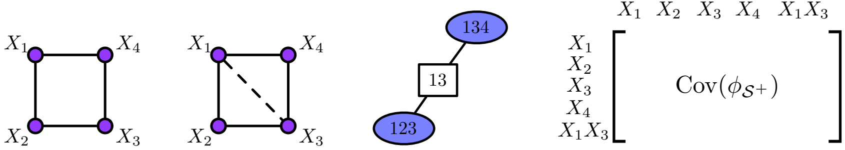


G triangulated junction tree augmented matrix

Theorem

The inverse covariance matrix of $\{\phi_U : U \in \mathcal{S}^+\}$ from any junction tree triangulation is graph-structured: $\Gamma_{A,B} \neq 0$ iff A, B are contained in a common clique

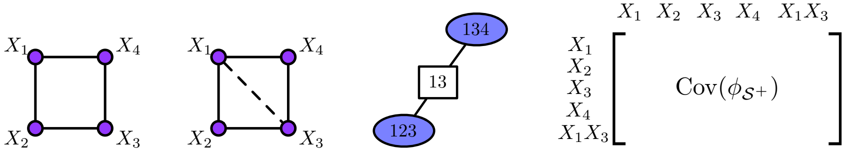
Example: Binary Ising model



G triangulated junction tree augmented matrix

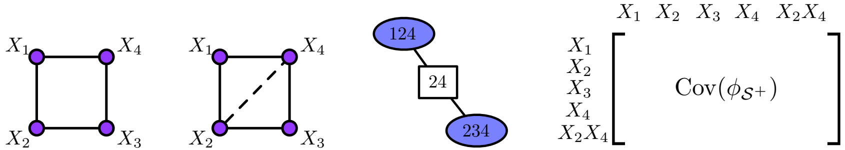
$$\Gamma = (\text{Cov}(\phi_{\mathcal{S}^+}))^{-1} \propto \begin{bmatrix} 115 & -2 & 109 & -2 & -114 \\ -2 & 5 & -2 & 0 & 1 \\ 109 & -2 & 114 & -2 & -114 \\ -2 & 0 & -2 & 5 & 1 \\ -114 & 1 & -114 & 1 & 119 \end{bmatrix}$$

Example: Binary Ising model



G triangulated junction tree augmented matrix

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



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

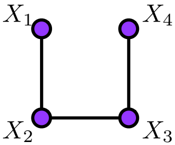
When G has only singleton separators, inverse covariance matrix of sufficient statistics on nodes is graph-structured

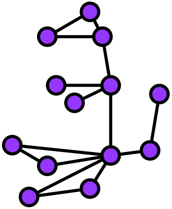
Consequences for trees

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

Corollary

When G has only singleton separators, inverse covariance matrix of sufficient statistics on nodes is graph-structured





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

Proof sketch

- Based on *exponential family* representation of pdf:

$$q_\theta(x_1, \dots, x_p) = \exp\left(\sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C \rangle - \Phi(\theta)\right)$$

Proof sketch

- Based on *exponential family* representation of pdf:

$$q_\theta(x_1, \ldots, x_p) = \exp\left(\sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C \rangle - \Phi(\theta)\right)$$

- $(\text{cov}_\theta[\mathbb{I}(X)])^{-1} = \nabla^2 \Phi^*(\mu)$, where

$$\Phi^*(\mu) := \sup_{\theta \in \mathbb{R}^D} \{\langle \mu, \theta \rangle - \Phi(\theta)\}$$

Proof sketch

- Based on *exponential family* representation of pdf:

$$q_\theta(x_1, \dots, x_p) = \exp\left(\sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C \rangle - \Phi(\theta)\right)$$

- $(\mathrm{cov}_\theta[\mathbb{I}(X)])^{-1} = \nabla^2 \Phi^*(\mu)$, where

$$\Phi^*(\mu) := \sup_{\theta \in \mathbb{R}^D} \{\langle \mu, \theta \rangle - \Phi(\theta)\}$$

- Relationship between Φ^* and entropy:

$$-\Phi^*(\mu) = H(q_{\theta(\mu)}(x)) = -\sum_x q_{\theta(\mu)}(x) \log q_{\theta(\mu)}(x)$$

Proof sketch

- Junction tree theorem:

$$q(x_1, \ldots, x_p) = \frac{\prod_{C \in \mathcal{C}} q_C(x_C)}{\prod_{S \in \mathcal{S}} q_S(x_S)},$$

so

$$H(q) = \sum_{C \in \mathcal{C}} H_C(q_C) - \sum_{S \in \mathcal{S}} H_S(q_S)$$

Proof sketch

- Junction tree theorem:

$$q(x_1, \ldots, x_p) = \frac{\prod_{C \in \mathcal{C}} q_C(x_C)}{\prod_{S \in \mathcal{S}} q_S(x_S)},$$

so

$$H(q) = \sum_{C \in \mathcal{C}} H_C(q_C) - \sum_{S \in \mathcal{S}} H_S(q_S)$$

- Then take Hessian

Structure learning

- Plug in sample covariance matrix of **augmented** vector to graphical Lasso, then compute $\text{supp}(\widehat{\Theta})$

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

Structure learning

- Plug in sample covariance matrix of **augmented** vector to graphical Lasso, then compute $\text{supp}(\widehat{\Theta})$

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

- When graph has singleton separators, ordinary graphical Lasso suffices

Structure learning

- Plug in sample covariance matrix of **augmented** vector to graphical Lasso, then compute $\text{supp}(\widehat{\Theta})$

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

- When graph has singleton separators, ordinary graphical Lasso suffices

Corollary

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

Structure learning

- Plug in sample covariance matrix of **augmented** vector to graphical Lasso, then compute $\text{supp}(\widehat{\Theta})$

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

- When graph has singleton separators, ordinary graphical Lasso suffices

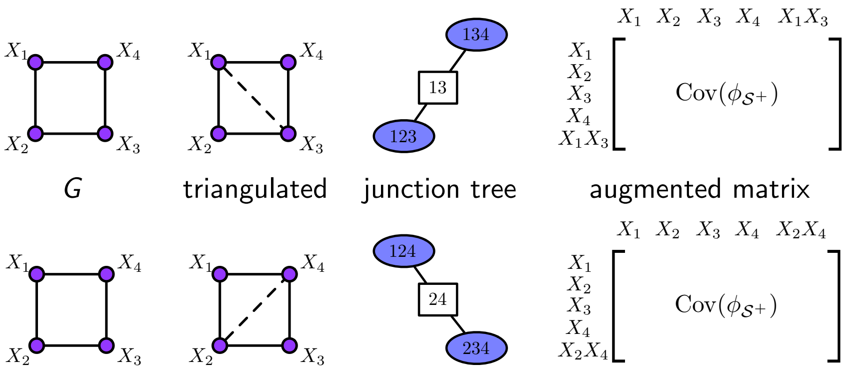
Corollary

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

- Group graphical Lasso for $m > 2$, similar theoretical guarantees

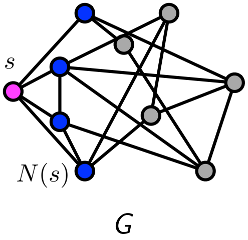
Problem

- **However**, augmented vector depends on structure of graph ...



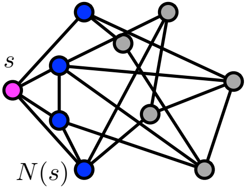
Beyond graphical Lasso

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

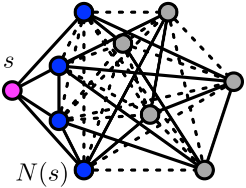


Beyond graphical Lasso

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



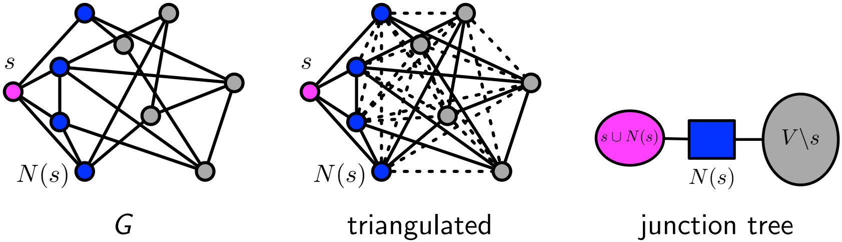
G



triangulated

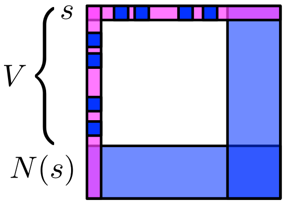
Beyond graphical Lasso

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



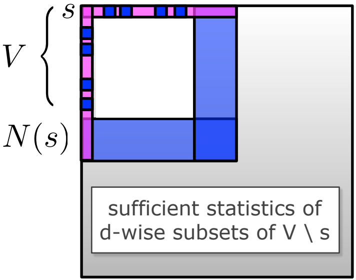
Inference for general graphs

- By theorem, inverse covariance matrix over nodes and sufficient statistics of $N(s)$ exposes neighbors of s



Inference for general graphs

- By theorem, inverse covariance matrix over nodes and sufficient statistics of $N(s)$ exposes neighbors of s



- Same result holds for matrix augmented by **all** d-subsets of $V \backslash s$

Nodewise algorithm

- For each $s \in V$:

Nodewise algorithm

- For each $s \in V$:
  - Regress sufficient statistics of X_s against sufficient statistics of all subsets of $V \backslash s$ of size $\leq d$, using Lasso

Nodewise algorithm

- For each $s \in V$:
  - Regress sufficient statistics of X_s against sufficient statistics of all subsets of $V \backslash s$ of size $\leq d$, using Lasso
  - Threshold entries of regression vector to obtain $\widehat{N(s)}$

Nodewise algorithm

- For each $s \in V$:
  - Regress sufficient statistics of X_s against sufficient statistics of all subsets of $V \backslash s$ of size $\leq d$, using Lasso
  - Threshold entries of regression vector to obtain $\widehat{N(s)}$
- Combine estimates $\widehat{N(s)}$ with AND/OR to recover edges of graph

Nodewise algorithm

- For each $s \in V$:
  - Regress sufficient statistics of X_s against sufficient statistics of all subsets of $V \backslash s$ of size $\leq d$, using Lasso
  - Threshold entries of regression vector to obtain $\widehat{N(s)}$
- Combine estimates $\widehat{N(s)}$ with AND/OR to recover edges of graph

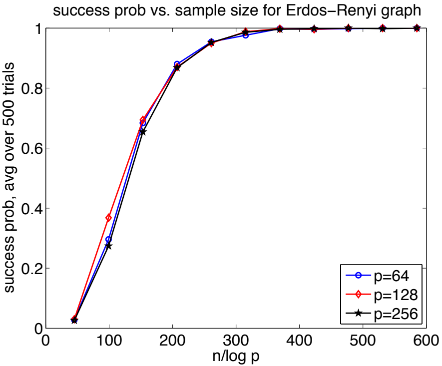
- Method succeeds w.h.p. for $n \gtrsim 2^d \log p$

Nodewise algorithm

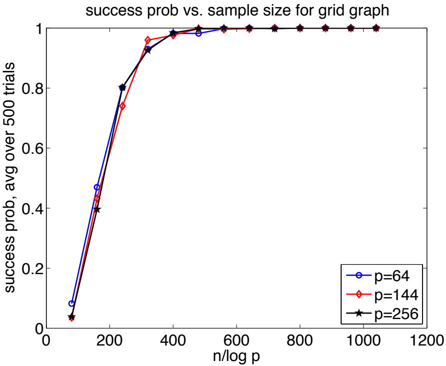
- For each $s \in V$:
  - Regress sufficient statistics of X_s against sufficient statistics of all subsets of $V \backslash s$ of size $\leq d$, using Lasso
  - Threshold entries of regression vector to obtain $\widehat{N(s)}$
- Combine estimates $\widehat{N(s)}$ with AND/OR to recover edges of graph

- Method succeeds w.h.p. for $n \gtrsim 2^d \log p$
- Can incorporate noisy/missing data into Lasso-based regression

Simulations for nodewise method



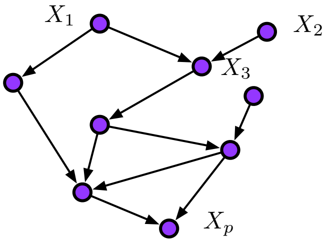
Erdös-Renyi graph, $d \approx 3$



grid-shaped graph, $d = 4$

Outline

Linear structural equation models



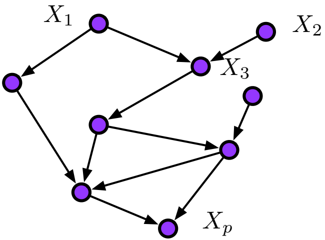
Markov property:

$$X_j \perp\!\!\!\perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)}$$

linear SEM:

$$X_j = b_j^T X_{\text{Pa}(j)} + \epsilon_j, \qquad \epsilon_j \perp\!\!\!\perp X_{\text{Pa}(j)}$$

Linear structural equation models



Markov property:

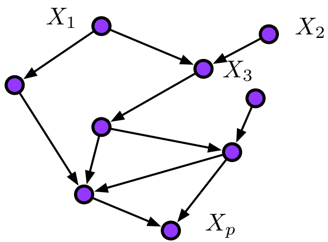
$$X_j \perp\!\!\!\perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)}$$

linear SEM:

$$X_j = b_j^T X_{\text{Pa}(j)} + \epsilon_j, \qquad \epsilon_j \perp\!\!\!\perp X_{\text{Pa}(j)}$$

- $X = B^T X + \epsilon, \quad X, \epsilon \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p \times p}$ strictly upper triangular

Linear structural equation models



Markov property:

$$X_j \perp\!\!\!\perp X_{\mathrm{Nondesc}(j)} \mid X_{\mathrm{Pa}(j)}$$

linear SEM:

$$X_j = b_j^T X_{\mathrm{Pa}(j)} + \epsilon_j, \qquad \epsilon_j \perp\!\!\!\perp X_{\mathrm{Pa}(j)}$$

- $X = B^T X + \epsilon, \quad X, \epsilon \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p \times p}$ strictly upper triangular
- **Goal:** Learn support of B ($B_{jk} \neq 0$ iff $j \to k$ is edge in DAG)

Inverse covariance matrix

- Denote $\text{Cov}(\epsilon) = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ and $\Theta = \text{Cov}(X)^{-1}$

Inverse covariance matrix

- Denote $\text{Cov}(\epsilon) = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ and $\Theta = \text{Cov}(X)^{-1}$

Theorem

The inverse covariance matrix of X is given by

$$\Theta_{jk} = -\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell}, \qquad \forall j < k$$

$$\Theta_{jj} = \sigma_j^{-2} + \sum_{\ell > j} \sigma_\ell^{-2} B_{jk}^2, \qquad \forall j$$

Inverse covariance matrix

- Denote $\text{Cov}(\epsilon) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\Theta = \text{Cov}(X)^{-1}$

Theorem

The inverse covariance matrix of X is given by

$$\Theta_{jk} = -\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell}, \qquad \forall j < k$$

$$\Theta_{jj} = \sigma_j^{-2} + \sum_{\ell > j} \sigma_\ell^{-2} B_{jk}^2, \qquad \forall j$$

- $\implies \Theta_{jk} \neq 0$ only when $j \to k$ is an edge or j, k are parents to ℓ

Consequence for structure learning

- **Faithfulness assumption:**

$$-\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell} = 0$$

only if $B_{jk} = 0$ and $B_{j\ell} B_{k\ell} = 0$ for all $\ell > k$

Consequence for structure learning

- **Faithfulness assumption:**

$$-\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell} = 0$$

  only if $B_{jk} = 0$ and $B_{j\ell} B_{k\ell} = 0$ for all $\ell > k$

- Under faithfulness assumption, $\mathcal{M}(G) = \text{supp}(\Theta)$

Consequence for structure learning

- **Faithfulness assumption:**

$$-\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell} = 0$$

  only if $B_{jk} = 0$ and $B_{j\ell} B_{k\ell} = 0$ for all $\ell > k$

- Under faithfulness assumption, $\mathcal{M}(G) = \text{supp}(\Theta)$

- **Apply graphical Lasso to estimate moralized graph**

Graphical Lasso for preprocessing

- Score-based approaches for learning DAG may be sped up with superstructure of skeleton or moralized graph (Perrier et al. '08, Ordyniak & Szeider '12)

Graphical Lasso for preprocessing

- Score-based approaches for learning DAG may be sped up with superstructure of skeleton or moralized graph (Perrier et al. '08, Ordyniak & Szeider '12)
- For linear SEM, first apply graphical Lasso to learn moralized graph

Graphical Lasso for preprocessing

- Score-based approaches for learning DAG may be sped up with superstructure of skeleton or moralized graph (Perrier et al. '08, Ordyniak & Szeider '12)
- For linear SEM, first apply graphical Lasso to learn moralized graph
- Can also accommodate systematically corrupted data (next section)

Outline

1. Introduction
2. Generalized inverse covariances
3. Linear structural equation models
4. Corrupted data

Systematically corrupted data

- Observe corrupted samples $\{(Z_1^{(i)}, Z_2^{(i)}, \dots, Z_p^{(i)})\}_{i=1}^n$, where $Z^{(i)}$ is noisy version of $X^{(i)}$

Systematically corrupted data

- Observe corrupted samples $\{(Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_p^{(i)})\}_{i=1}^n$, where $Z^{(i)}$ is noisy version of $X^{(i)}$
- Examples:
  - Additive noise: $Z^{(i)} = X^{(i)} + W^{(i)}, \qquad W^{(i)} \perp\!\!\!\perp X^{(i)}$

Systematically corrupted data

- Observe corrupted samples $\{(Z_1^{(i)}, Z_2^{(i)}, \dots, Z_p^{(i)})\}_{i=1}^n$, where $Z^{(i)}$ is noisy version of $X^{(i)}$
- Examples:
  - Additive noise: $Z^{(i)} = X^{(i)} + W^{(i)}, \qquad W^{(i)} \perp\!\!\!\perp X^{(i)}$
  - Missing data:

$$Z_j^{(i)} = \begin{cases} X_j^{(i)} & \text{with prob. } 1-\alpha \\ \star & \text{with prob. } \alpha \end{cases}$$

Systematically corrupted data

- Observe corrupted samples $\{(Z_1^{(i)}, Z_2^{(i)}, \dots, Z_p^{(i)})\}_{i=1}^n$, where $Z^{(i)}$ is noisy version of $X^{(i)}$
- Examples:
  - Additive noise: $Z^{(i)} = X^{(i)} + W^{(i)}, \qquad W^{(i)} \perp\!\!\!\perp X^{(i)}$
  - Missing data:

$$Z_j^{(i)} = \begin{cases} X_j^{(i)} & \text{with prob. } 1 - \alpha \\ \star & \text{with prob. } \alpha \end{cases}$$

- **Goal:** Structure learning based on $\{(Z_1^{(i)}, Z_2^{(i)}, \dots, Z_p^{(i)})\}_{i=1}^n$

Modified graphical Lasso

- **Idea:** Construct surrogate for $\widehat{\Sigma}$ based on corrupted samples $\{Z^{(i)}\}_{i=1}^n$

Modified graphical Lasso

- **Idea:** Construct surrogate for $\widehat{\Sigma}$ based on corrupted samples $\{Z^{(i)}\}_{i=1}^n$
- Additive noise:

$$\widehat{\Sigma} = \frac{Z^T Z}{n} - \Sigma_w$$

Modified graphical Lasso

- **Idea:** Construct surrogate for $\widehat{\Sigma}$ based on corrupted samples $\{Z^{(i)}\}_{i=1}^n$
- Additive noise:

$$\widehat{\Sigma} = \frac{Z^T Z}{n} - \Sigma_w$$

- Missing data: Let

$$\widehat{Z}_j^{(i)} = \begin{cases} \frac{Z_j^{(i)}}{1-\alpha} & \text{if } Z_j^{(i)} \text{ observed} \\ 0 & \text{otherwise,} \end{cases}$$

use

$$\widehat{\Sigma} = \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \, \mathrm{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n} \right)$$

Theory for graphical Lasso

- If

$$\|\widehat{\Sigma} - \Sigma^*\|_{\max} \precsim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \succsim \sqrt{\frac{\log p}{n}},$$

then

$$\|\widehat{\Theta} - \Theta^*\|_{\max} \precsim \left(\sqrt{\frac{\log p}{n}} + \lambda \right)$$

Theory for graphical Lasso

- If

$$\|\widehat{\Sigma} - \Sigma^*\|_{\max} \precsim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \succsim \sqrt{\frac{\log p}{n}},$$

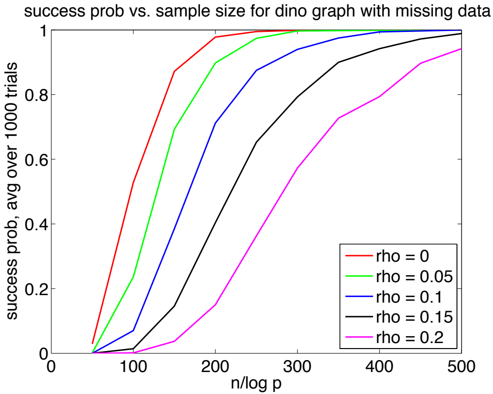
  then

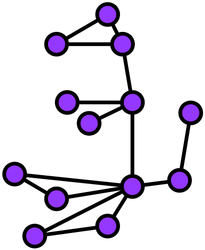
$$\|\widehat{\Theta} - \Theta^*\|_{\max} \precsim \left(\sqrt{\frac{\log p}{n}} + \lambda \right)$$

- Can establish deviation condition w.h.p. for modified estimators with corrupted data

Simulation study

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





Summary

- Significance of inverse covariance matrix for non-Gaussian data
  - For discrete variables, inverse of augmented covariance matrix is graph structured
  - For linear SEMs, support of inverse covariance is moralized graph

Summary

- Significance of inverse covariance matrix for non-Gaussian data
  - For discrete variables, inverse of augmented covariance matrix is graph structured
  - For linear SEMs, support of inverse covariance is moralized graph
- Use graphical Lasso to estimate (augmented) inverse
- Nodewise method for general discrete-valued graphs

Summary

- Significance of inverse covariance matrix for non-Gaussian data
  - For discrete variables, inverse of augmented covariance matrix is graph structured
  - For linear SEMs, support of inverse covariance is moralized graph
- Use graphical Lasso to estimate (augmented) inverse
- Nodewise method for general discrete-valued graphs
- Modifications for corrupted data

Open questions

- Computationally tractable method for structure learning in general discrete graphs
- Robustness results: Inverse covariance matrix of approximately Gaussian and/or approximately tree-structured graphs
- More general analysis of inverse covariances via exponential family representation

References

- P. Loh and P. Bühlmann (2013). High-dimensional learning of linear causal networks via inverse covariance estimation. ArXiv paper.
- P. Loh and M.J. Wainwright (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics.*
- P. Loh and M.J. Wainwright (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics.*