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:

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

- Undirected graph \( G = (V, E) \)
- Joint distribution of \((X_1, \ldots, X_p)\), 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:

$$X_j \perp \!
\!
\perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)}, \quad \forall j$$
Goal: Edge recovery from $n$ samples: $\{(X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)})\}_{i=1}^n$
Goal: Edge recovery from $n$ samples: $\{(X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)})\}_{i=1}^n$

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

High-dimensional setting: $p \gg n$, assume $\text{deg}(G) \leq d$

Sources of corruption: non-i.i.d. observations, contamination by noise/missing data
Goal: Edge recovery from $n$ samples: $\{(X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)})\}_{i=1}^n$

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):
  \[
  \hat{\Theta} \in \arg \min_{\Theta \succeq 0} \left\{ \text{trace}(\hat{\Sigma} \Theta) - \log \det(\Theta) + \lambda \sum_{s \neq t} |\Theta_{st}| \right\}
  \]
Some observations

Only sample-based quantity is $\hat{\Sigma}$.

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

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

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

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

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

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

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

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

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

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

- Only sample-based quantity is $\hat{\Sigma}$:
  \[
  \hat{\Theta} \in \text{arg min}_{\Theta \succeq 0} \left\{ \text{trace}(\hat{\Sigma} \Theta) - \log \det(\Theta) + \lambda \sum_{s \neq t} |\Theta_{st}| \right\}
  \]

- Although graphical Lasso is penalized Gaussian MLE, can always be used to estimate $\hat{\Theta}$ from $\hat{\Sigma}$:
  \[
  (\Sigma^*)^{-1} = \text{arg min}_\Theta \left\{ \text{trace}(\Sigma^* \Theta) - \log \det(\Theta) \right\}
  \]

- We extend graphical Lasso to discrete-valued data (undirected case) and linear structural equation models (directed case)
If
\[ \| \hat{\Sigma} - \Sigma^* \|_{\text{max}} \lesssim \sqrt{\frac{\log p}{n}} \] and \[ \lambda \gtrsim \sqrt{\frac{\log p}{n}}, \]
then
\[ \| \hat{\Theta} - \Theta^* \|_{\text{max}} \lesssim \left( \sqrt{\frac{\log p}{n}} + \lambda \right) \]
Theory for graphical Lasso

- If
  \[ \| \hat{\Sigma} - \Sigma^* \|_{\text{max}} \lesssim \sqrt{\frac{\log p}{n}} \text{ and } \lambda \gtrsim \sqrt{\frac{\log p}{n}}, \]
  then
  \[ \| \hat{\Theta} - \Theta^* \|_{\text{max}} \lesssim \left( \sqrt{\frac{\log p}{n}} + \lambda \right) \]

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

- If
  \[
  \| \hat{\Sigma} - \Sigma^* \|_{\max} \lesssim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \lesssim \sqrt{\frac{\log p}{n}},
  \]
  then
  \[
  \| \hat{\Theta} - \Theta^* \|_{\max} \lesssim \left( \sqrt{\frac{\log p}{n}} + \lambda \right)
  \]

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

- Thresholding \( \hat{\Theta} \) at level \( \sqrt{\frac{\log p}{n}} \) yields correct support
Outline

1. Introduction
2. Generalized inverse covariances
3. Linear structural equation models
4. Corrupted data
(Liu et al. ’09, ’12): \((X_1, \ldots, X_p)\) follows nonparanormal distribution if \((f_1(X_1), \ldots, f_p(X_p)) \sim N(0, \Sigma)\), and \(f_j\)'s monotone and differentiable.
(Liu et al. ’09, ’12): \((X_1, \ldots, X_p)\) follows nonparanormal distribution if \((f_1(X_1), \ldots, f_p(X_p)) \sim N(0, \Sigma)\), and \(f_j\)'s monotone and differentiable.

Then \((i, j) \notin E \text{ iff } \Theta_{ij} = 0\).
Non-Gaussian distributions

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

- In general non-Gaussian setting, relationship between entries of \(\Theta = \Sigma^{-1}\) and edges of \(G\) unknown
Assume $X_i$’s take values in a discrete set: \{0, 1, \ldots, m - 1\}
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

Discrete graphical models
An illustrative example

- Binary Ising model:

\[
P_\theta(x_1, \ldots, x_p) \propto \exp \left( \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \right),
\]
Binary Ising model:

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

$$\theta \in \mathbb{R}^{p^{+}\binom{p}{2}}$$,

$$(x_1, \ldots, x_p) \in \{0, 1\}^p$$
An illustrative example

- Ising models with $\theta_s = 0.1$, $\theta_{st} = 2$
An illustrative example

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

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

\[
\Theta_{\text{loop}} = \begin{bmatrix}
51.37 & -5.37 & -0.17 & -5.37 \\
-5.37 & 51.37 & -5.37 & -0.17 \\
-0.17 & -5.37 & 51.37 & -5.37 \\
-5.37 & -0.17 & -5.37 & 51.37
\end{bmatrix}
\]
An illustrative example

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

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

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

- $\Theta$ is graph-structured for chain, but not loop
An illustrative example

Θ graph-structured

Θ not graph-structured

\[ \Theta \text{ graph-structured} \]

\[ \Theta \text{ not graph-structured} \]
An illustrative example

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

$$
\begin{bmatrix}
115 & -2 & 109 & -2 & -114 \\
-2 & 5 & -2 & \text{\color{red}0} & 1 \\
109 & -2 & 114 & -2 & -114 \\
-2 & \text{\color{red}0} & -2 & 5 & 1 \\
-114 & 1 & -114 & 1 & 119
\end{bmatrix}
$$
Assume \((X_1, \ldots, X_p) \in \{0, \ldots, m - 1\}^p\)
Notation

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

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

\textbf{Ex:} When \(m = 2\) and \(U = \{1, 2\}\), \(\phi_U = (x_1, x_2, x_1x_2)\)
Notation

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

Example:
- When \(m = 2\) and \(U = \{1, 2\}\), \(\phi_U = (x_1, x_2, x_1x_2)\)
- When \(U = \{1\}\), \(\phi_U = (I\{x_1 = 1\}, \ldots, I\{x_1 = m - 1\})\)
Assume \((X_1, \ldots, X_p) \in \{0, \ldots, m-1\}^p\)

For any subset \(U \subseteq V\), associate vector \(\phi_U\) of sufficient statistics

**Ex:** When \(m = 2\) and \(U = \{1, 2\}\), \(\phi_U = (x_1, x_2, x_1x_2)\)

**Ex:** When \(U = \{1\}\), \(\phi_U = (I\{x_1 = 1\}, \ldots, I\{x_1 = m-1\})\)

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

Let $S +$ be the set of nodes and separator sets. The augmented covariance matrix is:

$$\text{Cov}(S +)$$

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

- Triangulate $G$

\[
\begin{array}{c}
X_1 \\
X_2 \\
X_3 \\
X_4
\end{array}
\quad
\begin{array}{c}
X_1 \\
X_2 \\
X_3 \\
X_4
\end{array}
\]

$G$  triangulated

Theorem

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

- Triangulate $G$
- Form junction tree with separator sets
Augmented covariance matrices

- Triangulate $G$
- Form junction tree with separator sets
- Let $S^+ = \text{nodes} + \text{separator sets}$
Augmented covariance matrices

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

Theorem

The inverse covariance matrix of $\{\phi_U : U \in S^+\}$ 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

\[
G \quad \text{triangulated} \quad \text{junction tree} \quad \text{augmented matrix}
\]

\[
\Gamma = (\text{Cov}(\phi_{S^+}))^{-1} \propto \begin{bmatrix}
115 & -2 & 109 & -2 & -114 \\
-2 & 5 & -2 & 0 & 1 \\
109 & -2 & 114 & -2 & -114 \\
-2 & 0 & -2 & 5 & 1 \\
-114 & 1 & -114 & 1 & 119 \\
\end{bmatrix}
\]
Example: Binary Ising model

\[ \begin{array}{ccc} X_1 & X_4 & X_1 \\ X_2 & X_3 & X_2 \\ \end{array} \] triangulated junction tree

- Statistics included in \( \phi_{S^+} \) depend on triangulation

\[ \begin{array}{ccc} X_1 & X_2 & X_3 & X_4 & X_{1X3} \\ \end{array} \] augmented matrix

\[ \text{Cov}(\phi_{S^+}) \]
Consequences for trees

- When there exists a triangulation with singleton separator sets, \( S^+ = \{1, \ldots, p\} \)
Consequences for trees

- When there exists a triangulation with singleton separator sets, \( S^+ = \{1, \ldots, p\} \)

**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, $S^+ = \{1, \ldots, p\}$

**Corollary**

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

\[
(Cov(X_1, \ldots, X_p))^{-1}
\]
Proof sketch

Based on *exponential family* representation of pdf:

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

\[
q_\theta(x_1, \ldots, x_p) = \exp \left( \sum_{C \in \mathcal{C}} \langle \theta_C, I_C \rangle - \Phi(\theta) \right)
\]

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

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

- Relationship between \(\Phi^*\) 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:

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

so

\[ H(q) = \sum_{C \in C} H_C(q_C) - \sum_{S \in S} H_S(q_S) \]
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
Plug in sample covariance matrix of \textit{augmented} vector to graphical Lasso, then compute $\text{supp}(\hat{\Theta})$

\[
\hat{\Theta} \in \arg \min_{\Theta \succeq 0} \left\{ \text{trace}(\hat{\Sigma} \Theta) - \log \det(\Theta) + \lambda \sum_{s \neq t} |\Theta_{st}| \right\}
\]
Plug in sample covariance matrix of augmented vector to graphical Lasso, then compute \( \text{supp}(\hat{\Theta}) \)

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

When graph has singleton separators, ordinary graphical Lasso suffices
Structure learning

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

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

- 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 \textit{augmented} vector to graphical Lasso, then compute $\text{supp}(\hat{\Theta})$

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

- When graph has singleton separators, ordinary graphical Lasso suffices

\[\text{Corollary}\]

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

\[ \begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
X_4 \\
X_1X_3
\end{bmatrix} \]

\[ \text{Cov}(\phi_{S^+}) \]

G

triangulated

junction tree

augmented matrix

\[ \begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
X_4 \\
X_2X_4
\end{bmatrix} \]

\[ \text{Cov}(\phi_{S^+}) \]
- Nodewise method: recovers neighborhood $N(s)$ for any fixed $s \in V$
Nodewise method: recovers neighborhood $N(s)$ for any fixed $s \in V$

Form junction tree by fully-connecting all nodes in $V \setminus s$

$G$

$N(s)$

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 $V \setminus s$

![Diagram]

- $G$
- $s \cup N(s)$
- Junction tree
By theorem, inverse covariance matrix over nodes and sufficient statistics of $N(s)$ exposes neighbors of $s$. 

\[ V \{ s \} \]
\[ N(s) \]
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 $V \setminus s$. 
Nodewise algorithm

- For each $s \in V$: 

Regress sufficient statistics of $X_s$ against sufficient statistics of all subsets of $V \setminus s$ of size $\leq d$, using Lasso.

Threshold entries of regression vector to obtain $\hat{N}(s)$.

Combine estimates $\hat{N}(s)$ with AND/OR to recover edges of graph.

Method succeeds w.h.p. for $n \gtrapprox 2d \log p$.

Can incorporate noisy/missing data into Lasso-based regression.
For each \( s \in V \):

- Regress sufficient statistics of \( X_s \) against sufficient statistics of all subsets of \( V \setminus s \) 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 $V \setminus s$ of size $\leq d$, using Lasso
- Threshold entries of regression vector to obtain $\hat{N}(s)$

Combine estimates $\hat{N}(s)$ with AND/OR to recover edges of graph

Method succeeds w.h.p. for $n \gg 2d \log p$
Nodewise algorithm

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

success prob vs. sample size for Erdos–Renyi graph

success prob vs. sample size for grid graph

Erdős–Renyi graph, \( d \approx 3 \)

grid-shaped graph, \( d = 4 \)
Outline

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

Markov property:
\[ X_j \perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)} \]

Linear SEM:
\[ X_j = b_j^T X_{\text{Pa}(j)} + \epsilon_j, \quad \epsilon_j \perp X_{\text{Pa}(j)} \]
Markov property:

\[ X_j \perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)} \]

Linear SEM:

\[ X_j = b_j^T X_{\text{Pa}(j)} + \epsilon_j, \quad \epsilon_j \perp X_{\text{Pa}(j)} \]

\[ X = B^T X + \epsilon, \quad X, \epsilon \in \mathbb{R}^p \text{ and } B \in \mathbb{R}^{p \times p} \text{ strictly upper triangular} \]
Linear structural equation models

Markov property:
\[ X_j \perp X_{\text{Nondesc}(j)} \mid X_{\text{Pa}(j)} \]

Linear SEM:
\[ X_j = b_j^T X_{\text{Pa}(j)} + \epsilon_j, \quad \epsilon_j \perp X_{\text{Pa}(j)} \]

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

**Goal:** Learn support of \( B \) (\( B_{jk} \neq 0 \) iff \( j \to k \) is edge in DAG)
Denote $\text{Cov}(\epsilon) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\Theta = \text{Cov}(X)^{-1}$.
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*

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

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

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

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

$$
\implies \Theta_{jk} \neq 0 \text{ only when } j \to k \text{ is an edge or } j, k \text{ are parents to } \ell
$$
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, \( M(G) = \text{supp}(\Theta) \)
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, \( M(G) = \text{supp}(\Theta) \)

- **Apply graphical Lasso to estimate moralized graph**
Score-based approaches for learning DAG may be sped up with superstructure of skeleton or moralized graph (Perrier et al. ’08, Ordyniak & Szeider ’12)
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
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)
1. Introduction

2. Generalized inverse covariances

3. Linear structural equation models

4. Corrupted data
Systematically corrupted data

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

Examples:

- Additive noise: $Z^{(i)} = X^{(i)} + W^{(i)}$, $W^{(i)} \perp \perp X^{(i)}$.

- Missing data: $Z^{(i)}_j = \{X^{(i)}_j \text{ with prob. } 1 - \alpha \}$ with prob. $\alpha$. 

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

P. Loh (UC Berkeley)
Beyond the graphical Lasso
June 26, 2014
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: \( Z^{(i)} = X^{(i)} + W^{(i)} \), \( W^{(i)} \perp \perp X^{(i)} \)
Systematically corrupted data

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

- Examples:
  - Additive noise: \( Z^{(i)} = X^{(i)} + W^{(i)}, \quad W^{(i)} \perp \perp X^{(i)} \)
  - Missing data:

\[
Z_j^{(i)} = \begin{cases} 
X_j^{(i)} & \text{with prob. } 1 - \alpha \\
\ast & \text{with prob. } \alpha
\end{cases}
\]
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: \( Z^{(i)} = X^{(i)} + W^{(i)}, \quad W^{(i)} \perp \perp X^{(i)} \)
- 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 \( \{(Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_p^{(i)})\}_{i=1}^n \)
**Idea:** Construct surrogate for $\hat{\Sigma}$ based on corrupted samples $\{Z^{(i)}\}_{i=1}^{n}$

Additive noise:

$$\hat{\Sigma} = Z^T Z_n - \Sigma$$

Missing data:

$$\hat{Z}^{(i)}_j = \begin{cases} Z^{(i)}_j & \text{if } Z^{(i)}_j \text{ observed} \\ 0 & \text{otherwise} \end{cases}$$

use

$$\hat{\Sigma} = \hat{Z}^T \hat{Z}_n - \alpha \text{diag}(\hat{Z}^T \hat{Z}_n)$$
Modified graphical Lasso

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

- Additive noise:

$$\hat{\Sigma} = \frac{Z^T Z}{n} - \Sigma_w$$
Modified graphical Lasso

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

Additive noise:

\[
\hat{\Sigma} = \frac{Z^T Z}{n} - \Sigma_w
\]

Missing data: Let

\[
\hat{Z}_{j}^{(i)} = \begin{cases} 
\frac{Z_{j}^{(i)}}{1-\alpha} & \text{if } Z_{j}^{(i)} \text{ observed} \\
0 & \text{otherwise},
\end{cases}
\]

use

\[
\hat{\Sigma} = \frac{\hat{Z}^T \hat{Z}}{n} - \alpha \text{ diag} \left( \frac{\hat{Z}^T \hat{Z}}{n} \right)
\]
If

\[ \| \hat{\Sigma} - \Sigma^* \|_{\text{max}} \lesssim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \lesssim \sqrt{\frac{\log p}{n}}, \]

then

\[ \| \hat{\Theta} - \Theta^* \|_{\text{max}} \lesssim \left( \sqrt{\frac{\log p}{n}} + \lambda \right) \]
Theory for graphical Lasso

- If
  \[ \|\hat{\Sigma} - \Sigma^*\|_{\text{max}} \gtrsim \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \lambda \gtrsim \sqrt{\frac{\log p}{n}}, \]

  then

  \[ \|\hat{\Theta} - \Theta^*\|_{\text{max}} \gtrsim \left( \sqrt{\frac{\log p}{n}} + \lambda \right) \]

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