Structure Learning of Sparse Random Ising Models

Vincent Tan

Joint work with Anima Anandkumar (UCI) and Alan Willsky (MIT)

Department of ECE,
University of Wisconsin-Madison

ITA (Feb 11, 2011)

Outline

1. Graphical Models

Outline

1. Graphical Models
2. Problem Definition

Outline

1. Graphical Models
2. Problem Definition
3. Necessary Conditions on Sample Complexity

Outline

1. Graphical Models
2. Problem Definition
3. Necessary Conditions on Sample Complexity
4. Sufficient Conditions on Sample Complexity
    - Correlation Thresholding
    - Conditional Mutual Information Thresholding

Outline

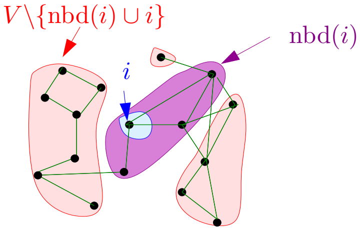
1. Graphical Models
2. Problem Definition
3. Necessary Conditions on Sample Complexity
4. Sufficient Conditions on Sample Complexity
    - Correlation Thresholding
    - Conditional Mutual Information Thresholding
5. Conclusion

Graphical Models: Introduction

- Graph $G = (V, E)$ represents a multivariate prob. distribution of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ indexed by $V = \{1, \ldots, d\}$
- Node $i \in V$ corresponds to random variable X_i
- Edge set E corresponds to conditional independencies

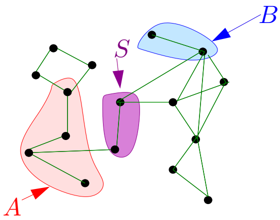
Graphical Models: Introduction

- Graph $G = (V, E)$ represents a multivariate prob. distribution of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ indexed by $V = \{1, \ldots, d\}$
- Node $i \in V$ corresponds to random variable X_i
- Edge set E corresponds to conditional independencies



$X_i \perp\!\!\!\perp \mathbf{X}_{V\setminus\{\mathrm{nbd}(i)\cup i\}} | \mathbf{X}_{\mathrm{nbd}(i)}$

Local Markov Property



$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$

Global Markov Property

High Dimensional Learning of Graphical Models

High Dimensional Learning of Graphical Models

- Given k training samples $\mathbf{x}^k := \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ drawn from a graphical model P, Markov on $G_n = (V, E)$ (graph with n nodes)
- Information about model class (e.g., Gaussian, discrete, Ising....)
- Would like an estimate $\hat{G}_n = \hat{G}_n(\mathbf{x}^k)$ that is consistent

High Dimensional Learning of Graphical Models

- Given k training samples $\mathbf{x}^k := \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ drawn from a graphical model P, Markov on $G_n = (V, E)$ (graph with n nodes)
- Information about model class (e.g., Gaussian, discrete, Ising....)
- Would like an estimate $\hat{G}_n = \hat{G}_n(\mathbf{x}^k)$ that is consistent

Definition: Structural Consistency

$$\lim_{\substack{k,n\to\infty \\ k=O(f(n))}} \Pr\left(\hat{G}_n(\mathbf{x}^k) \neq G_n\right) = 0$$

High Dimensional Learning of Graphical Models

- Given k training samples $\mathbf{x}^k := \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ drawn from a graphical model P, Markov on $G_n = (V, E)$ (graph with n nodes)
- Information about model class (e.g., Gaussian, discrete, Ising....)
- Would like an estimate $\hat{G}_n = \hat{G}_n(\mathbf{x}^k)$ that is consistent

Definition: Structural Consistency

$$\lim_{\substack{k,n\to\infty \\ k=O(f(n))}} \Pr\left(\hat{G}_n(\mathbf{x}^k) \neq G_n\right) = 0$$

- Desideratum 1: k is grows very slowly with n

High Dimensional Learning of Graphical Models

- Given k training samples $\mathbf{x}^k := \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ drawn from a graphical model P, Markov on $G_n = (V, E)$ (graph with n nodes)
- Information about model class (e.g., Gaussian, discrete, Ising....)
- Would like an estimate $\hat{G}_n = \hat{G}_n(\mathbf{x}^k)$ that is consistent

Definition: Structural Consistency

$$\lim_{\substack{k,n\to\infty \\ k=O(f(n))}} \Pr\left(\hat{G}_n(\mathbf{x}^k) \neq G_n\right) = 0$$

- Desideratum 1: k is grows very slowly with n
- Desideratum 2: Low computational complexity

High Dimensional Learning of Graphical Models

- Given k training samples $\mathbf{x}^k := \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ drawn from a graphical model P, Markov on $G_n = (V, E)$ (graph with n nodes)
- Information about model class (e.g., Gaussian, discrete, Ising....)
- Would like an estimate $\hat{G}_n = \hat{G}_n(\mathbf{x}^k)$ that is consistent

Definition: Structural Consistency

$$\lim_{\substack{k,n\to\infty \\ k=O(f(n))}} \Pr\left(\hat{G}_n(\mathbf{x}^k) \neq G_n\right) = 0$$

- Desideratum 1: k is grows very slowly with n
- Desideratum 2: Low computational complexity
- Motivation: High-dimensional data (microarray, social networks)

Related Work on Learning Graphical Models

Efficient Algorithms for Structure Learning

- ML for trees: Max-weight spanning tree (Chow & Liu 68)
  - Error exponents (T., Anandkumar, Tong, Willsky IT-'11)
  - Forest models (T., Anandkumar, Willsky JMLR-'11)
  - Latent models (Choi, T., Anandkumar, Willsky JMLR-'11)

Related Work on Learning Graphical Models

Efficient Algorithms for Structure Learning

- ML for trees: Max-weight spanning tree (Chow & Liu 68)
  - Error exponents (T., Anandkumar, Tong, Willsky IT-'11)
  - Forest models (T., Anandkumar, Willsky JMLR-'11)
  - Latent models (Choi, T., Anandkumar, Willsky JMLR-'11)
- Hardness: NP hard to learn non-trees (Karger & Srebro '01)
- Difficulty: Correlation decay necessary (Bento & Montanari '10)

Related Work on Learning Graphical Models

Efficient Algorithms for Structure Learning

- ML for trees: Max-weight spanning tree (Chow & Liu 68)
    - Error exponents (T., Anandkumar, Tong, Willsky IT-'11)
    - Forest models (T., Anandkumar, Willsky JMLR-'11)
    - Latent models (Choi, T., Anandkumar, Willsky JMLR-'11)
- Hardness: NP hard to learn non-trees (Karger & Srebro '01)
- Difficulty: Correlation decay necessary (Bento & Montanari '10)
- Conditional independence tests for bounded degree graphs (Abbeel et al. '06, Bresler et al. '09)
- Convex optimization: ℓ_1 regularization (Dudik et al. '04, Lee et al. '06, Meinshausen & Buehlmann '06, Ravikumar et al. '10)
- Information-theoretic lower bounds (Santhanam & Wainwright '08)

This Talk: Learning Random Graphical Models

- We consider the case where the underlying graph G is random

This Talk: Learning Random Graphical Models

- We consider the case where the underlying graph G is random
- Relax the assumption that graph comes from a particular set

This Talk: Learning Random Graphical Models

- We consider the case where the underlying graph G is random
- Relax the assumption that graph comes from a particular set
- "Real-world" networks can be modeled by random graphs

This Talk: Learning Random Graphical Models

- We consider the case where the underlying graph G is random
- Relax the assumption that graph comes from a particular set
- "Real-world" networks can be modeled by random graphs
- Our work is a first-step in understanding the fundamental limits in learning random graphical models

- Ising model
- Markov on Erdős-Rényi ensemble $G_n \sim \mathcal{G}(n, \frac{c}{n})$

Crisis = Danger + Opportunity

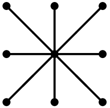
Crisis = Danger + Opportunity

Why difficult or dangerous?

Crisis = Danger + Opportunity

Why difficult or dangerous?

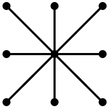
- Random graphs consists of some nodes that have large degrees
- Existing algorithms may not be consistent because they typically assume the max degree grows slowly.



Crisis = Danger + Opportunity

Why difficult or dangerous?

- Random graphs consists of some nodes that have large degrees
- Existing algorithms may not be consistent because they typically assume the max degree grows slowly.

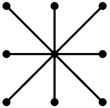


Saving grace(s)...

Crisis = Danger + Opportunity

Why difficult or dangerous?

- Random graphs consists of some nodes that have large degrees
- Existing algorithms may not be consistent because they typically assume the max degree grows slowly.



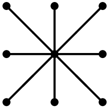
Saving grace(s)...

- $G_n \sim \mathcal{G}(n, \frac{c}{n})$ is locally tree-like

Crisis = Danger + Opportunity

Why difficult or dangerous?

- Random graphs consists of some nodes that have large degrees
- Existing algorithms may not be consistent because they typically assume the max degree grows slowly.



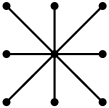
Saving grace(s)...

- $G_n \sim \mathcal{G}(n, \frac{c}{n})$ is locally tree-like
- Correlation decay: Influences of “faraway” nodes on node i are negligible, model behaves locally as a tree distribution

Crisis = Danger + Opportunity

Why difficult or dangerous?

- Random graphs consists of some nodes that have large degrees
- Existing algorithms may not be consistent because they typically assume the max degree grows slowly.



Saving grace(s)...

- $G_n \sim \mathcal{G}(n, \frac{c}{n})$ is locally tree-like
- Correlation decay: Influences of "faraway" nodes on node i are negligible, model behaves locally as a tree distribution
- Tree-based algorithms (Chow-Liu, Thresholding) may work

Setup: Ising Models on Random Graphs

- We have k vector valued samples $\mathbf{x}^k$. Each sample in $\{\pm 1\}^n$

Setup: Ising Models on Random Graphs

- We have k vector valued samples $\mathbf{x}^k$. Each sample in $\{\pm 1\}^n$
- Graph is drawn from the Erdős-Rényi ensemble, i.e., $G_n \sim \mathcal{G}(n, \frac{c}{n})$

Setup: Ising Models on Random Graphs

- We have k vector valued samples $\mathbf{x}^k$. Each sample in $\{\pm 1\}^n$
- Graph is drawn from the Erdős-Rényi ensemble, i.e., $G_n \sim \mathcal{G}(n, \frac{c}{n})$
- Each edge in G_n has an appearance probability of $\frac{c}{n}$, independent of all other edges

Setup: Ising Models on Random Graphs

- We have k vector valued samples $\mathbf{x}^k$. Each sample in $\{\pm 1\}^n$
- Graph is drawn from the Erdős-Rényi ensemble, i.e., $G_n \sim \mathcal{G}(n, \frac{c}{n})$
- Each edge in G_n has an appearance probability of $\frac{c}{n}$, independent of all other edges
- Ising model on $G = (V, E)$:

$$P(\mathbf{x}|G) \propto \exp\left(\sum_{(i,j)\in E} J_{i,j} x_i x_j \right), \qquad \mathbf{x} \in \{\pm 1\}^n$$

Assumptions:

- Ferromagnetism: $J_{i,j} \in [J_{\min}, J_{\max}] \subset (0, \infty)$ for all $(i,j) \in E$

Setup: Ising Models on Random Graphs

- We have k vector valued samples $\mathbf{x}^k$. Each sample in $\{\pm 1\}^n$
- Graph is drawn from the Erdős-Rényi ensemble, i.e., $G_n \sim \mathcal{G}(n, \frac{c}{n})$
- Each edge in G_n has an appearance probability of $\frac{c}{n}$, independent of all other edges
- Ising model on $G = (V, E)$:

$$P(\mathbf{x}|G) \propto \exp\left(\sum_{(i,j)\in E} J_{i,j} x_i x_j\right), \qquad \mathbf{x} \in \{\pm 1\}^n$$

Assumptions:

- Ferromagnetism: $J_{i,j} \in [J_{\min}, J_{\max}] \subset (0, \infty)$ for all $(i,j) \in E$
- Correlation Decay:

$$c \tanh(J_{\max}) < 1$$

A Strong Converse

- Converse result: Lower bound on sample complexity
- Any algorithm fails if number of samples k does not exceed the prescribed lower bound.

A Strong Converse

- Converse result: Lower bound on sample complexity
- Any algorithm fails if number of samples k does not exceed the prescribed lower bound.
- Recall that $G_n \sim \mathcal{G}(n, \frac{c}{n})$
- Assume that $c \leq n/2$, i.e., graph does not need to be sparse

A Strong Converse

- Converse result: Lower bound on sample complexity
- Any algorithm fails if number of samples k does not exceed the prescribed lower bound.
- Recall that $G_n \sim \mathcal{G}(n, \frac{c}{n})$
- Assume that $c \leq n/2$, i.e., graph does not need to be sparse

Theorem (Converse)

There exists an $\varepsilon > 0$ such that if

$$k \leq \varepsilon c \log n,$$

then,

$$\lim_{k,n\to\infty} \Pr\left(\hat{G}_n(\mathbf{x}^k) \neq G_n\right) = 1$$

for any estimator $\hat{G}_n(\,\cdot\,)$.

Proof Idea for the Strong Converse

Moral of the story: Need $k = \Omega(c \log n)$ samples for consistent recovery

Proof Idea for the Strong Converse

Moral of the story: Need $k = \Omega(c \log n)$ samples for consistent recovery

- Follows closely the converse technique in Bresler et al. '09.
- Main modification: Underlying graph not deterministic so counting argument needs to be modified

Proof Idea for the Strong Converse

Moral of the story: Need $k = \Omega(c \log n)$ samples for consistent recovery

- Follows closely the converse technique in Bresler et al. '09.
- Main modification: Underlying graph not deterministic so counting argument needs to be modified
- Focus on graphs "with the highest likelihoods"
- Note from

$$k \leq \varepsilon c \log n,$$

  that number of samples k is required to grow linearly with the average degree c

Correlation Thresholding

- Intuition: Edges with strong correlations should be included in the model

Correlation Thresholding

- Intuition: Edges with strong correlations should be included in the model
- Compute for each pair of variables $u, v \in V$, the empirical correlation

$$\hat{C}^k_{u,v} := \frac{1}{k}\sum_{i=1}^{k} x_u^{(i)} x_v^{(i)}$$

Correlation Thresholding

- Intuition: Edges with strong correlations should be included in the model
- Compute for each pair of variables $u, v \in V$, the empirical correlation

$$\hat{C}^k_{u,v} := \frac{1}{k} \sum_{i=1}^{k} x_u^{(i)} x_v^{(i)}$$

- Set $(u, v) \in \hat{G}_n$ iff

$$\hat{C}^k_{u,v} \geq \delta(J_{\min}, J_{\max})$$

Correlation Thresholding: Theoretical Properties

- Assume correlation decay: $c \tanh J_{\max} < 1$ (c constant)

Correlation Thresholding: Theoretical Properties

- Assume correlation decay: $c \tanh J_{\max} < 1$ (c constant)
- Assume homogeneity: $2 \tanh^2 J_{\max} < \tanh J_{\min}$

Correlation Thresholding: Theoretical Properties

- Assume correlation decay: $c \tanh J_{\max} < 1$ (c constant)
- Assume homogeneity: $2 \tanh^2 J_{\max} < \tanh J_{\min}$

Theorem (Structural Consistency of CorrThres)

For a.e. graph G_n, we have

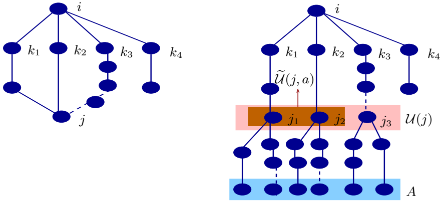
$$\lim_{\substack{k,n\to\infty \\ k=\Omega(\log n)}} \Pr\left(\mathsf{CorrThres}(\{\hat{C}^k_{u,v}\}_{(u,v)\in V^2};\delta);\neq G_n\right) = 0$$

Correlation Thresholding: Why / How does it work?

- Correlations are higher on edges than non-edges for nearly homogeneous Ising models on $\mathcal{G}(n, \frac{c}{n})$

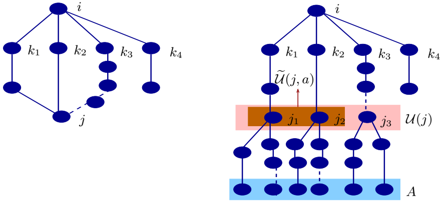
Correlation Thresholding: Why / How does it work?

- Correlations are higher on edges than non-edges for nearly homogeneous Ising models on $\mathcal{G}(n, \frac{c}{n})$
- Self-avoiding walk tree (SAW) construction



Correlation Thresholding: Why / How does it work?

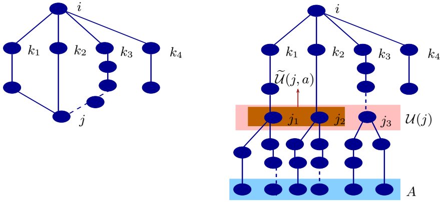
- Correlations are higher on edges than non-edges for nearly homogeneous Ising models on $\mathcal{G}(n, \frac{c}{n})$
- Self-avoiding walk tree (SAW) construction



- Prove result for exact statistics; generalization to sample statistics using large deviations

Correlation Thresholding: Why / How does it work?

- Correlations are higher on edges than non-edges for nearly homogeneous Ising models on $\mathcal{G}(n, \frac{c}{n})$
- Self-avoiding walk tree (SAW) construction



- Prove result for exact statistics; generalization to sample statistics using large deviations
- Can homogeneity assumption be removed?

Separation Property

- Fact: For sets $A, B, S \in V$, if S separates A, B, then

$$I(X_A; X_B | X_S) = 0$$

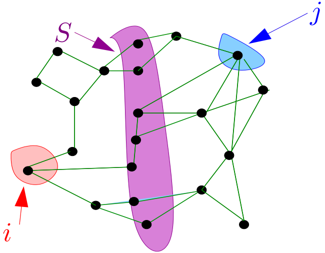
The global Markov property.

Separation Property

- Fact: For sets $A, B, S \in V$, if S separates A, B, then

$$I(X_A; X_B | X_S) = 0$$

The global Markov property.



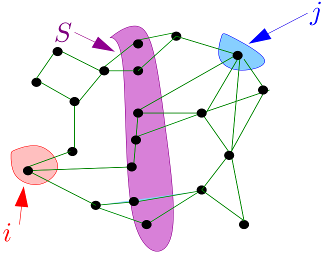
$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

Separation Property

- Fact: For sets $A, B, S \in V$, if S separates A, B, then

$$I(X_A; X_B | X_S) = 0$$

The global Markov property.



$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

- S separates i and j.

Conditional Mutual Information Thresholding

- Compute the empirical conditional mutual information; if small, we may have found a candidate separator

Conditional Mutual Information Thresholding

- Compute the empirical conditional mutual information; if small, we may have found a candidate separator
- What should be the cardinality of the conditioning set?

Conditional Mutual Information Thresholding

- Compute the empirical conditional mutual information; if small, we may have found a candidate separator
- What should be the cardinality of the conditioning set? Roughly 2!

Conditional Mutual Information Thresholding

- Compute the empirical conditional mutual information; if small, we may have found a candidate separator
- What should be the cardinality of the conditioning set? Roughly 2!
- Rule: $(i,j) \in \hat{G}_n$ if and only if

$$\min_{S \subset V \setminus \{i,j\}, |S| \leq 2} \hat{I}(X_i, X_j | X_S) \leq \tau_{k,n}$$

Conditional Mutual Information Thresholding

- Compute the empirical conditional mutual information; if small, we may have found a candidate separator
- What should be the cardinality of the conditioning set? Roughly 2!
- Rule: $(i,j) \in \hat{G}_n$ if and only if

$$\min_{S \subset V \setminus \{i,j\}, |S| \le 2} \hat{I}(X_i, X_j | X_S) \le \tau_{k,n}$$

- $\tau_{k,n}$ is the threshold
- Depends on number of variables n and sample size k

Conditional Mutual Information Thresholding

Again assume correlation decay condition: $c \tanh J_{\max} < 1$

Conditional Mutual Information Thresholding

Again assume correlation decay condition: $c \tanh J_{\max} < 1$

Theorem (Structural Consistency of CMIT)

For a.e. graph G_n, we have

$$\lim_{\substack{k,n\to\infty \\ k=\omega(\log n)}} \Pr\left(\mathrm{CMIT}(\mathbf{x}^k;\tau_{k,n}); \neq G_n\right) = 0$$

Conditional MI Thresholding: Why / How does it work?

- Challenge: Separators in graphical models may be large, i.e.,

$$\hat{I}(X_i; X_j | X_S) = f(\hat{P}_{i,j,S})$$

depends on the type over many variables

Conditional MI Thresholding: Why / How does it work?

- Challenge: Separators in graphical models may be large, i.e.,

$$\hat{I}(X_i;X_j|X_S) = f(\hat{P}_{i,j,S})$$

depends on the type over many variables

- Approximate separation?

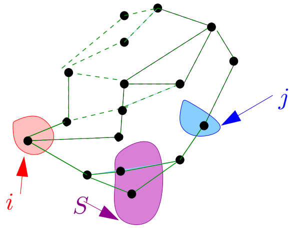
Conditional MI Thresholding: Why / How does it work?

- Challenge: Separators in graphical models may be large, i.e.,

$$\hat{I}(X_i; X_j | X_S) = f(\hat{P}_{i,j,S})$$

depends on the type over many variables

- Approximate separation?
- In such random graphical models, the size of an approximate separator is ≤ 2 asymptotically



- Ignore effects of long paths separating i and j

Conclusion

- Proposed a framework for learning random graphical models

Conclusion

- Proposed a framework for learning random graphical models
- Necessary condition on sample complexity

$$k = \Omega(c \log n)$$

Conclusion

- Proposed a framework for learning random graphical models
- Necessary condition on sample complexity

$$k = \Omega(c \log n)$$

- Tractable, simple algorithms with provable theoretical properties

$$\begin{aligned} \text{CorrThres :} \quad & k = \Omega(\log n) \\ \text{CMIT :} \quad & k = \omega(\log n) \end{aligned}$$

Conclusion

- Proposed a framework for learning random graphical models
- Necessary condition on sample complexity

$$k = \Omega(c \log n)$$

- Tractable, simple algorithms with provable theoretical properties

$$\begin{aligned} \text{CorrThres}: \quad & k = \Omega(\log n) \\ \text{CMIT}: \quad & k = \omega(\log n) \end{aligned}$$

- http://arxiv.org/abs/1011.0129