Graphical Models for High-Dimensional Data

Ahmed Gashgash, Kody Law, Omer Gokalp, Sudeep Salgia, Taylor Ferebee

Cornell University

10 December 2019

Overview

Basics

Overview

- Graphical Models
- Two ways to connect
  - Factorization
  - Conditional Independence
- Hammersley-Clifford equivalence

Graphical Models

- Types: directed, undirected, or hybrid
- Undirected graphical models (Markov random fields)
  - Undirected graph $G = (V, E)$
  - **Vertices** $V = \{1, 2, ..., d\}$
    - Each vertex j is associated with a random variable $X_j \in \chi_j$
  - **Edges** E, edge (j, k) - an unordered pair of distinct vertices $j, k \in V$
- $\mathcal{P}$ - distribution of the d-dimensional vector $X = (X_1, ..., X_d)$

Graphical Model Connections

- Want to analyze the connections between the structure of $\mathcal{P}$, and the structure of the underlying graph G
- Two ways to connect
  1. Factorization
  2. Conditional independence properties
- Hammersley-Clifford theorem says both approaches are the same

Factorization

- A **clique** C is a subset of vertices that are all joined by edges - $(j, k) \in E$ for all distinct vertices $j, k \in C$ - let $\mathcal{C}$ be the set of all cliques in G
- **Maximal clique** - clique that is not a subset of any other clique
- The random vector $(X_1, ..., X_d)$ *factorizes according to the graph* G if its density function p is represented as

$$p(x_1, ..., x_d) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

for some collection of *clique compatibility function* $\psi_C : \chi^C \to [0, \infty)$

Factorization

$$p(x_1, \ldots, x_7) \propto \psi_{123}(x_1, x_2, x_3)\, \psi_{345}(x_3, x_4, x_5)\, \psi_{46}(x_4, x_6)\, \psi_{57}(x_5, x_7).$$

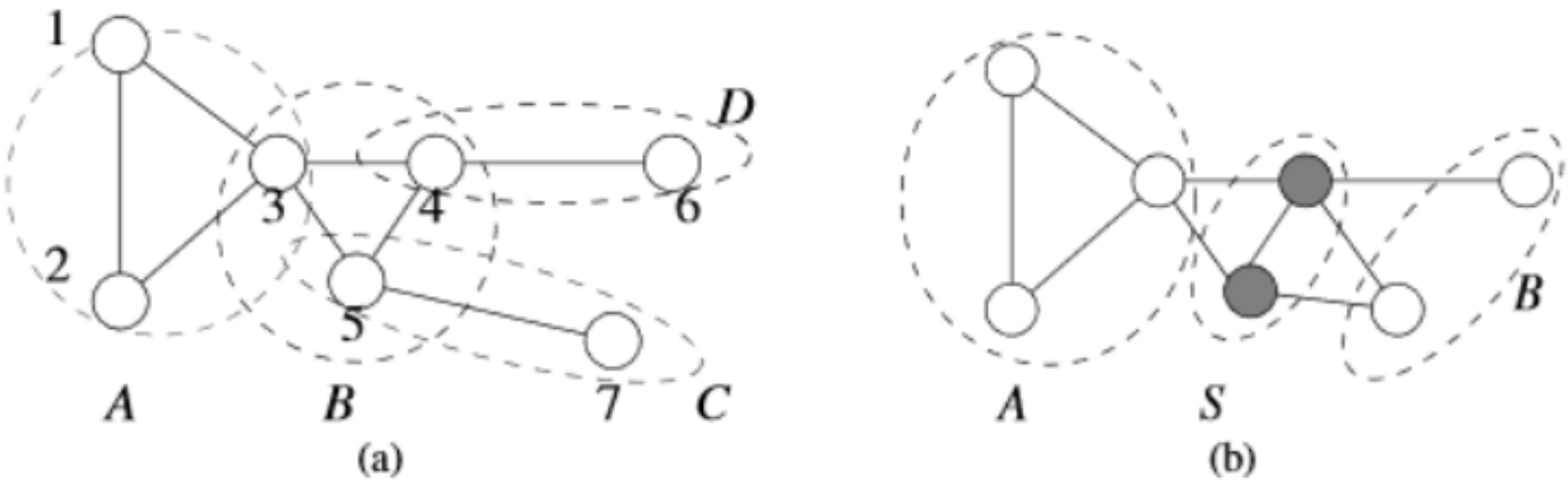


Figure 11.1 Illustration of basic graph-theoretic properties. (a) Subsets A and B are 3-cliques, whereas subsets C and D are 2-cliques. All of these cliques are maximal. Each vertex is a clique as well, but none of these singleton cliques are maximal for this graph. (b) Subset S is a vertex cutset, breaking the graph into two disconnected subgraphs with vertex sets A and B, respectively.

Factorization Examples

- Markov chain factorization
  - Let p_1 denote the marginal distribution of X_1, and for $j \in 1, 2, ..., d-1$, let $p_{j+1|j}$ denote the condition distribution of X_{j+1} given X_j
  - The standard way of factoring the distribution of a Markov chain is $p(x_1, ..., x_d) = p_1(x_1)p_{2|1}(x_2|x_1)...p_{d|d-1}(x_d|x_{d-1})$
    - $\psi_1(x_1) = p_1(x_1)$ at vertex 1, and $\psi_j(x_j) = 1$ for all $j = 2, ..., d$
    - $\psi_{j,j+1}(x_j, x_{j+1}) = p_{j+1|j}(x_{j+1}|x_j)$ for $j = 1, ..., d-1$

Factorization Examples

- Multivariate Gaussian factorization
  - Any non-degenerate Gaussian distribution with zero mean can be parameterized in terms of its inverse covariance matrix, or *precision matrix*, $\Theta^* = \Sigma^{-1}$
  - $p(x_1, ..., x_d, \Theta^*) = \frac{\sqrt{\det(\Theta^*)}}{(2\pi)^{d/2}} e^{-\frac{1}{2}x^T \Theta^* x}$
  - $e^{-\frac{1}{2}x^T \Theta^* x} = \exp(\frac{1}{2} \sum_{(j,k)\in E} \Theta^*_{jk} x_j x_k) = \prod_{(j,k)\in E} e^{-\frac{1}{2}\Theta^*_{jk} x_j x_k}$
    - $e^{-\frac{1}{2}\Theta^*_{jk} x_j x_k} = \psi_{jk}(x_j, x_k)$
  - Any zero-mean Gaussian distribution can be factorized in terms of functions on edges, or cliques of size two, even if the underlying graph has higher-order cliques.

Conditional Independence

- A *vertex cutset* S is a subset of vertices whose removal from the graph breaks it into 2+ disjoint pieces
- Removing S from V leads to the vertex-induced subgraph $G(V \setminus S)$, which consists of the vertex set $V \setminus S$, and the residual edge set $E(V \setminus S) := \{(j,k) \in E | j,k \in V \setminus S\}$ - S is a vertex cutset if $G(V \setminus S)$ has 2+ disconnected non-empty components.
- A random vector $X = (X_1, ..., X_d)$ is *Markov with respect to a graph* G if for all vertex cutsets S breaking the graph into A and B disjoint pieces, $X_a \perp X_B | X_S$ holds

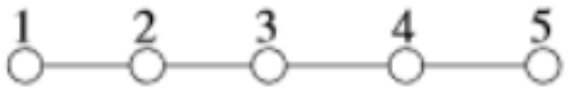
Conditional Independence Examples

- Markov chain conditional independence
  - A chain graph on the vertex set $V = \{1, ..., d\}$ contains the edges $(j, j+1)$ for $j = 1, 2, ..., d-1$
  - Each vertex $j \in \{2, 3, ..., d-1\}$ is a non-trivial cutset, breaking graph into the "past" $P = \{1, 2, ..., j-1\}$ and "future" $P = \{j+1, ..., d\}$
  - Past X_P and future X_F are conditionally independent given the present X_j

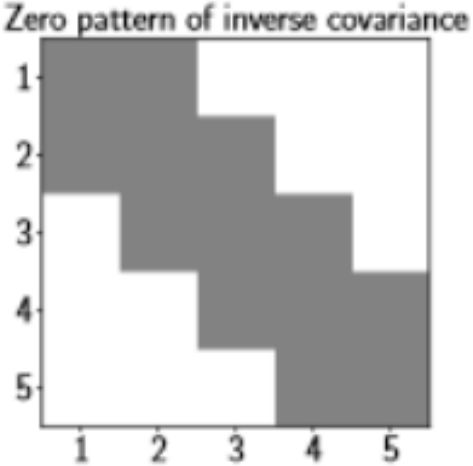
Conditional Independence Examples

- Neighborhood-based cutsets
  - For any vertex $j \in V$, its *neighborhood set* is the subset of vertices $N(j) := \{k \in V | (j,k) \in E\}$ that are joined to j by an edge
    - $N(j)$ is always a vertex cutset, and is non-trivial if j is not connected to every other vertex
    - Graph separated into $A = \{j\}$ and $B = V \setminus (N(j) \cup \{j\})$

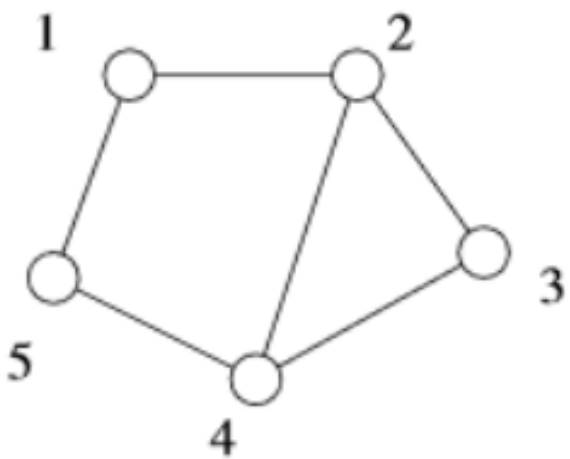
Conditional Independence Examples



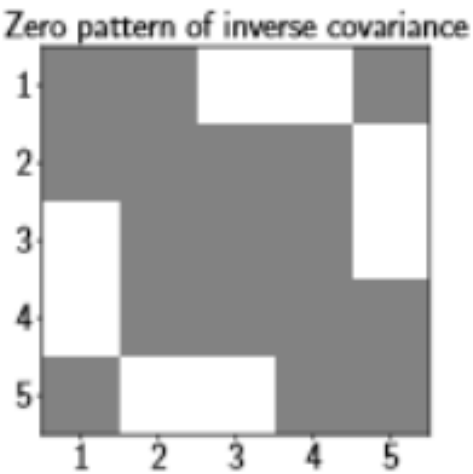
(a)



(b)



(c)



(d)

Hammersley-Clifford equivalence

- For a given undirected graph and any random vector $X = (X_1, ..., X_d)$ with strictly positive density p, both properties are equivalent:
  1. X factories according to the structure of G
  2. X is Markov w.r.t G
- In effect, both factorization and conditional independence are equivalent for any strictly positive distribution.

Hammersley-Clifford Proof: Factorization to conditional independence

- Suppose factorization holds, and let S be an arbitrary vertex cutset of the graph such that non-empty subsets A and B are separated by S.
- Let the clique $\mathbb{C}$ be split into subsets $\mathbb{C} = \mathbb{C}_A \cup \mathbb{C}_S \cup \mathbb{C}_B$, where $\mathbb{C}_j := \{C \in \mathbb{C} | C \cap j \neq 0\}$ for $j \in \{A, B\}$, and $\mathbb{C}_S := \{C \in \mathbb{C} | C \subseteq S \neq 0\}$
- $p(x_A, x_S, x_B) =$
  $\frac{1}{Z} \left[\prod_{C \in \mathbb{C}_A} \psi_C(x_C)\right] \left[\prod_{C \in \mathbb{C}_S} \psi_C(x_C)\right] \left[\prod_{C \in \mathbb{C}_B} \psi_C(x_C)\right]$

Hammersley-Clifford Proof: Factorization to conditional independence

- $\prod_{C \in \mathbb{C}_j} \psi_C(x_C) = \psi_j(x_j, x_S)$ for $j \in \{A, B\}$
- $\prod_{C \in \mathbb{C}_S} \psi_C(x_C) = \psi_S(x_S)$
- $Z_A(x_s) := \sum_{x_A} \psi_A(x_A, x_S)$ and $Z_B(x_s) := \sum_{x_B} \psi_B(x_B, x_S)$, which then leads to:
  - $p(x_S) = \frac{Z_A(x_s) Z_B(x_s)}{Z} \psi_S(x_S)$
  - $p(X_A, x_S) = \frac{Z_B(x_s)}{Z} \psi_A(x_A, x_S) \psi_S(x_S)$ and $p(X_B, x_S) = \frac{Z_A(x_s)}{Z} \psi_B(x_B, x_S) \psi_S(x_S)$
- $\frac{p(x_A, x_S, x_B)}{p(x_S)} = \frac{\frac{1}{Z} \psi_A(x_A, x_S) \psi_S(x_S) \psi_B(x_B, x_S)}{\frac{Z_A(x_s) Z_B(x_s)}{Z} \psi_S(x_S)} = \frac{\psi_A(x_A, x_S) \psi_B(x_B, x_S)}{Z_A(x_s) Z_B(x_s)}$

Hammersley-Clifford Proof: Factorization to conditional independence

- $\frac{p(x_A,x_S)}{p(x_S)} = \frac{\frac{Z_B(x_S)}{Z}\psi_A(x_A,x_S)\psi_S(x_S)}{\frac{Z_A(x_S)Z_B(x_S)}{Z}\psi_S(x_S)} = \frac{\psi_A(x_A,x_S)}{Z_A(x_s)}$
- $\frac{p(x_B,x_S)}{p(x_S)} = \frac{\frac{Z_A(x_S)}{Z}\psi_B(x_B,x_S)\psi_S(x_S)}{\frac{Z_A(x_S)Z_B(x_S)}{Z}\psi_S(x_S)} = \frac{\psi_B(x_B,x_S)}{Z_A(x_s)}$
- Combining both above equations leads to $p(x_A, x_B|x_S) = \frac{p(x_A,x_B,x_S)}{p(x_S)} = \frac{p(x_A,x_S)}{p(x_S)}\frac{p(x_B,x_S)}{p(x_S)} = p(x_A|x_S)p(x_B|x_S)$
- Thus, we show that factorization implies conditional independence - $X_A \perp X_B|X_S$

Hammersley-Clifford Proof: Conditional independence to factorization

- See Grimmett (1973) and Besag (1974) for proofs of the converse

Estimation of Graphical Models

Overview

- Motivation
- Graphical Lasso
- Neighborhood Based Methods
- Consistency results

Motivation

Typically, graphical model applications are associated with an inverse problem that is generally about a collection of samples $\{x_i\}_{i=1}^n$ where each $x_i = (x_{i1}, ..., x_{id})$ is a d-dimensional vector, hypothesized to have been drawn from some graph structured probability distribution.

We want to estimate certain aspects of the underlying graphical model, including *graphical parameters* and *graphical model* selection.

Motivating Examples: Graphical Model Selection, or Inverse Covariance Selection in Gaussian MRFs

Goal: Given an estimate of the precision matrix $\hat{\Theta}$ of Θ^*, recover the edge set E of the underlying graph G

$P[\hat{E} \neq E]$, where $\hat{E}$ is the estimate of edge set based upon $\hat{\Theta}^*$

Goal: Given information about the precision matrix, what is the probability that we have recovered a fraction of the edge set?

$P[\hat{E} \leq 1 - \delta]$, where $\delta \in (0, 1)$ is user set tolerance parameter

Goal: To better understand relationships among the samples across items, scales or groups, estimate the inverse covariance matrix itself.

$$\|\hat{\Theta} - \Theta^*\|_{op} \text{ or } \|\hat{\Theta} - \Theta^*\|_F$$

ℓ_1-regularized maximum likelihood

When the graph G is expected to have relatively few edges, a natural form of regularization is to impose an ℓ_1-constraint of the entries of $\mathbf{\Theta}$. We combine this with the negative log-likelihood to arrive to the **Graphical Lasso Estimator**:

$$\hat{\mathbf{\Theta}} \in arg \min_{\mathbf{\Theta} \in S^{d \times d}} \left\{ \langle\langle \mathbf{\Theta}, \hat{\mathbf{\Sigma}} \rangle\rangle - \log det \mathbf{\Theta} + \lambda_n \|\mathbf{\Theta}\|_{1,off} \right\}$$

$\|\mathbf{\Theta}\|_{1,off}$ refers to the $\ell_1$norm applied to off diagonal entries of $\mathbf{\Theta}$

Frobenius Norm Bounds for Graphical Lasso

Suppose that the inverse covariance matrix $\mathbf{\Theta}^*$ has at most m non-zero entries per row, and we solve the graphical Lasso with regularization parameters $\lambda_n = 8\sigma^2(\sqrt{\frac{\log d}{n}} + \delta)$ for some $\delta \in (0, 1]$. Then as long as $6(\|\mathbf{\Theta}^*\|_2 + 1)^2\lambda_n\sqrt{md} < 1$, the graphical Lasso estimate $\hat{\mathbf{\Theta}}$ satifies

$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2 \leq \frac{9}{(\|\mathbf{\Theta}^*\|_2 + 1)^4} md\lambda_n^2$$

with probability at least $1 - 8e^{-\frac{1}{16}n\delta^2}$.

Proof Outline The proof is broken into two parts:

- Establish that restricted strong convexity holds over the Frobenius norm ball $\mathbb{B}_F(1)$ and apply the result for bounds of general models (Corollary 9.20).
- Verify that the regularization parameter is valid for the bound stated by localizing the error matrix.

Edge Selection

- The previous proposition is quite crude. We want to also guarantee that the edge structure of the underlying graph is preserved.
- The problem of edge selection is very similar to the problem of variable selection in sparse linear models. The next proposition gives some insight into graphical Lasso and graph structure.

Proposition 11.10

Consider a zero mean d-dimensional Gaussian distribution based on an $\alpha-$incoherent inverse covariance matrix $\mathbf{\Theta}^*$.
Given a sample size lower bounded: $n > c_0(1+8\alpha^{-1})^2 m^2 \log d$
Solve graphical Lasso with regularization parameter:
$\lambda = \frac{c_1}{\alpha}\sqrt{\frac{\log d}{n}} + \delta$ for some $\delta \in (0,1]$. Then with probability at least $1 - c_2 e^{-c_3 n\delta^2}$:

- The graphical Lasso solution leads to no false inclusions
- It satisfies the sup-norm bound

$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{max} \leq c_4\{(1+8\alpha^{-1})\sqrt{\frac{\log d}{n}} + \lambda_n\}$$

Comments on Assumptions

- Dependency Condition: Relevant Covariates are not overly dependent
- Incoherence Condition: Large number of irrelevant covariates have *limited* influence over relevant covariates
- Strong concentration bounds namely, sample quantities quickly converge to expectations

Neighborhood Based Regression

General idea: We want to be able to detect the conditional independence relationships in a neighborhood for any given vertex.

Neighborhood regression: For a given vertex $j \in V$, we use the random variables $X_{\setminus\{j\}} := \{X_k \mid k \in V \setminus \{j\}\}$

Lasso-based neighborhood regression

- For each node $j \in V$:
  - Extract the column vector $X_j \in \mathbb{R}^n$ and the submatrix $\boldsymbol{X}_{\backslash\{j\}} \in \mathbb{R}^{n \times (d-1)}$
  - Solve the Lasso problem

$$\hat{\theta} = argmin\Big\{ \frac{1}{2n} \|X_j - \boldsymbol{X}_{\backslash\{j\}}\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\}$$

  - Return the neighborhood estimate $\hat{\boldsymbol{N}}(j) = \{k \in V \backslash \{j\} | \hat{\theta}_k \neq 0\}$

Combine the neighborhood estimates to form an edge estimate $\hat{E}$, using either the OR rule of the AND rule

OR Rule: $(j,k) \in \hat{E}_{OR}$ if either $k \in \hat{\boldsymbol{N}}(j)$ or $j \in \hat{\boldsymbol{N}}(j)$
AND Rule: $(j,k) \in \hat{E}_{AND}$ if $k \in \hat{\boldsymbol{N}}(j)$ and $j \in \hat{\boldsymbol{N}}(j)$

Consistency

$*$ *For further discussion and information on the consistency results, refer to Chapter 7 discussions on primal-witness dual technique*

- We can guarantee graph selection consistency of the previously stated procedure with AND & OR rules.
- Note this is for a GMRF with covariance matrix $\mathbf{\Sigma}^* = (\mathbf{\Theta}^*)^{-1}$ with maximum degree m and scaled diagonals ≤ 1
- With probability greater than $1 - c_2 \exp\{-c_3 n \min\{\delta^2, \frac{1}{m}\}\}$ the estimated edge set has the following properties:
  - No false inclusions, it includes no false edges so that $\hat{E} \subseteq E$
  - All significant edges are captured: it includes all edges (j, k) for which $|\mathbf{\Theta} *_{jk}| \geq 7b\lambda_n$

Graphical Models in Exponential Form

Graphical Models in Exponential Form

✓This section evaluates graph estimation problem for a broader class, which can be factorized in exponential form.

- Factorization of Exponential Forms and Model Examples
- A General Form of Neighborhood Regression
- Graph Selection for Ising Models

Factorization and Model Examples

- Given a graph $G = (E, V)$ we denote vector of parameters of a vertex $j \in V$ as $\mathbf{\Theta}_j^*$ and $\mathbf{\Theta}_{jk}^*$ is matrix denotes the parameters of edge $(j, k) \in E$. We may do the following pairwise factorization with these parameters in exponential form.

Factorization of Exponential Forms:

$$p_{\mathbf{\Theta}^*}(x_1, \ldots x_d) \propto exp\left\{ \sum_{j \in V} \phi_j(x_j; \mathbf{\Theta}_j^*) + \sum_{(j,k) \in E} \phi_{jk}(x_j, x_k; \mathbf{\Theta}_{jk}^*) \right\} \tag{1}$$

Comparison with Gaussian Model

- The difference between our previous Gaussian model and this broader exponential class is Gaussian's parameters are scalars. We may write their potential functions as follows:

$$\phi(x_j; \theta_j^*) = \theta_j^* x_j, \qquad \phi_{jk}(x_j, x_k; \theta_{jk}^*) = \theta_{jk}^* x_j x_k \tag{2}$$

- In exponential forms, we may use Lebesgue measure(e.g Gaussian) over $\mathbb{R}^d$ or counting measure(e.g Ising Model) on the binary hypercube $\{0, 1\}^d$ depending on the model we use.

Potts Model

- Each X_s is a random variable and its value is an element of discrete set $\{0, \ldots M-1\}$.
- $\mathbf{\Theta}_j^* = \{\mathbf{\Theta}_{j;a}^*, a = 1, \ldots, M-1\}$ is an $(M-1)$ vector.
- $\mathbf{\Theta}_{jk}^* = \{\mathbf{\Theta}_{jk;a,b}^*, a = 1, \ldots, M-1\}$ is an $(M-1) \times (M-1)$ matrix.
- The potential functions are as follows:

$$\phi_j(x_j; \mathbf{\Theta}_j^*) = \sum_{a=1}^{M-1} \mathbf{\Theta}_{j;a}^* 1\{x_j = a\} \tag{3a}$$

$$\phi_j(x_j, x_k; \mathbf{\Theta}_{jk}^*) = \sum_{a=1}^{M-1} \sum_{b=1}^{M-1} \mathbf{\Theta}_{jk;ab}^* 1\{x_j = a, x_k = b\} \tag{3b}$$

Poisson Graphical Model

- Collection of random variables(e.g type of count data) $(X_1, \ldots, X_d)$ taking values from the set $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$.

- To build a graphical model for this collection, we may think of conditional distribution of each variable with its given neighbors is Poisson r.v. with mean:

$$\mu_j = exp\left(\theta_j^* + \sum_{k \in \mathcal{N}(j)} \theta_{jk}^* x_k\right)$$

Poisson Graphical Model

- This setup leads us to a Markov random field with potential functions:

$$\phi(x_j; \theta_j^*) = \theta_j^* x_j - \log(x!) \quad \text{for all } j \in V, \tag{4a}$$

$$\phi_{jk}(x_j, x_k; \theta_{jk}^*) = \theta_{jk}^* x_j x_k \quad \text{for all } (j,k) \in E. \tag{4b}$$

- Counting measure on $\mathbb{Z}_+$ is used for density functions.

- Moreover, we may have mixed graphical models such as Gaussian mixtures.

A General Form of Neighborhood Regression

- $\{x_i\}_{i=1}^n$ is a collection of i.i.d. samples from a graphical model in exponential form where x_i is a d-vector.
- Form a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where x_i^T is the i^th row.
- Let $X_j \in \mathbb{R}^n$ denotes j^th column for $j = 1, \ldots, d$.
- Neighborhood regression predicts X_j using the columns of submatrix $\mathbf{X}_{\setminus\{j\}} \in \mathbb{R}^{n \times (d-1)}$

A General Form of Neighborhood Regression

- Conditional likelihood depends only on the vector of parameters that involve node j

$$\mathbf{\Theta}_{j+} := \left\{ \mathbf{\Theta}_j, \mathbf{\Theta}_{jk}, k \in V \setminus \{j\} \right\} \tag{5}$$

- In true model $\mathbf{\Theta}^*$ we are guaranteed that $\mathbf{\Theta}^*_{jk} = 0$, $(j,k) \notin E$.
- $\|\cdot\|$ denotes a matrix norm, then we have the following:

$$\widehat{\mathbf{\Theta}}_{j+} = \arg\min_{\mathbf{\Theta}_{j+}} \left\{ \underbrace{-\frac{1}{n}\sum_{i=n}^{n} \log p_{\mathbf{\Theta}_{j+}}(x_{ij}|x_{i|\{j\}})}_{\mathcal{L}_n(\mathbf{\Theta}_{j+};x_j,x_{\setminus\{j\}})} + \lambda_n \sum_{k \in V\setminus\{j\}} \|\mathbf{\Theta}_{jk}\| \right\} \tag{6}$$

Graph Selection for Ising Models

- The factorization of Ising Model, which is a distribution over binary variables, takes the following form:

$$p_{\theta^*}(x_1, \ldots, x_d) \propto exp\left\{ \sum_{j \in V} \theta^* x_j + \sum_{(j,k) \in E} \theta^*_{jk} x_j x_k \right\} \qquad (7)$$

- We have only one parameter per edge, then ℓ_1- penalty suffices to encourage sparsity in neighborhood regression.

- Then define a subset of coefficients associated with the following for any $j \in V$

$$\theta_{j+} := \left\{ \theta_j, \theta_{jk}, k \in V \setminus \{j\} \right\}$$

Graph Selection for Ising Models

- The neighborhood regression reduced to a form of logistic regression, which uses logistic function $f(t) = \log(1 + e^t)$:

$$\widehat{\theta}_{j+} = \arg \min_{\theta_{j+} \in \mathbb{R}^d} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} f\left(\theta_j x_{ij} + \sum_{k \in V \setminus \{j\}} \theta_{jk} x_{ij} x_{ik}\right)}_{\mathcal{L}_n(\theta_{j+} : x_j, x_{\setminus \{j\}})} + \lambda_n \sum_{k \in V \setminus \{j\}} |\theta_{jk}| \right\} \tag{8}$$

- What are our conditions to recover the correct neighborhood set $\mathcal{N}(j)$ with the estimate we get above?

Graph Selection for Ising Models

- Let θ^*_{j+} denote the minimizer of population objective function:

$$\overline{\mathcal{L}}(\theta_{j+}) = \mathbb{E}[\mathcal{L}_n(\theta_{j+}; X_j, \mathbf{X}_{\setminus\{j\}}]$$

- $\mathbf{J} := \nabla^2 \overline{\mathcal{L}}(\theta^*_{j+})$ where $\mathbf{J}$ is a d-dimensional matrix

- Given $\alpha \in (0, 1]$ $\mathbf{J}$ satisfies α-incoherence condition at node $j \in V$ if

$$\max_{k \notin S} \|J_{kS}(\mathbf{J}_{SS})^{-1}\|_1 \leq 1 - \alpha, \text{ where } S = \mathcal{N}(j) \qquad (9)$$

- We also assume that the smallest eigenvalue of $\mathbf{J}_{SS}$ is lower bounded by some $c_{min} > 0$.

Graph Selection for Ising Models

Theorem 11.15

- n i.i.d. samples with $n > c_0 m^2 \log d$
- We consider the estimator given in 8 with $\lambda_n = \frac{32}{\alpha}\sqrt{\frac{\log d}{n}} + \delta$ for some $\delta \in [0, 1]$
- Then with probability at least $1 - c_1 e^{-c_2(n\delta^2 + \log d)}$ our estimate $\widehat{\theta}_{j+}$ has the following properties:

a) $\widehat{S} = supp(\widehat{\theta})$ is contained within the set $\mathcal{N}(j)$

b) ℓ_∞-$bound \|\widehat{\theta}_{j+} - \theta^*_{j+}\| \leq \frac{c_3}{c_{min}}\sqrt{m}\lambda_n$.

Graphs with Corrupted or Hidden Variables

Graphs with Corrupted Data

- So far we assumed samples $\{x_i\}_{i=1}^n$ are observed perfectly
- Samples could be corrupted by measurement noise or missing entries
- In this section we'll discuss some methods for addressing this problem
- We assume the Gaussian case for simplicity

Gaussian Graph Estimation with Corrupted Data

- Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote data matrix corresponding to original samples
- We observe $\mathbf{Z} \in \mathbb{R}^{n \times d}$ such that $\mathbf{Z} = \mathbf{X} + \mathbf{V}$ where $\mathbf{V}$ is some type of measurement noise
- Naiive approach: apply the standard Gaussian graph estimator to $\mathbf{Z} \rightarrow$ inconsistent estimates

Gaussian Graph Estimation with Corrupted Data

Consider the graphical lasso, a naiive approach is to solve the convex program:

$$\hat{\Theta}_{NAI} = arg \min_{\Theta \in S^d} \left\{ \langle\langle \Theta, \hat{\Sigma}_z \rangle\rangle + \log det\Theta + \lambda_n \|\Theta\|_{1,off} \right\}$$

- where $\hat{\Sigma}_z = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ is the corrupted sample covariance matrix
- The addition of noise does not preserve Markov properties
- In general, the estimate $\hat{\Theta}_{NAI}$ will not lead to a consistent estimate of the edge set or the precision matrix
- Replace $\hat{\Sigma}_z$ with an unbiased estimate of $\text{cov}(x)$ based on $\mathbf{Z}$

Example: Unbiased covariance estimate for additive corruptions

Let $\mathbf{Z} = \mathbf{X} + \mathbf{V}$:

- each row v_i in $\mathbf{V}$ is iid from a zero mean distribution with covariance Σ_v
- a natural estimate of $\Sigma_x := \text{cov}(x)$ is :

$$\hat{\Gamma} := \frac{1}{n}\mathbf{Z}^T\mathbf{Z} - \Sigma_v$$

- as long as the noise matrix $\mathbf{V}$ is independent of $\mathbf{X} \rightarrow \hat{\Gamma}$ is an unbiased estimate of Σ_x

Correcting the Gaussian graphical Lasso

More generally:

- any unbiased estimate $\hat{\Gamma}$ of Σ_x defines a form of the corrected graphical Lasso estimator:

$$\tilde{\Theta} = arg \min_{\Theta \in S_+^{d \times d}} \left\{ \langle\langle \Theta, \hat{\Gamma} \rangle\rangle + \log det \Theta + \lambda_n \|\Theta\|_{1,off} \right\}$$

- as with the usual graphical Lasso, this is a strictly convex program
- Solution exists as long as $\lambda_n > \|\hat{\Gamma} - \Sigma_x\|_{max}$

Correcting Neighborhood Regression

- As stated before, neighborhood regression involves solving a linear regression problem
- The observation vector $X_j \in \mathbb{R}^n$ at a given node j plays the role of the response variable and the remaining (d-1) variables play the role of the predictors

Correcting Neighborhood Regression

Throughout this section let:

- $\mathbf{X}$ denote the $n \times (d-1)$ matrix with $\{X_k, k \in V \setminus \{j\}\}$ as its columns
- $y = X_j$ denotes the response vector

Therefore an instance of a corrupted linear regression model is:

$$y = \mathbf{X}\theta^* + w, \qquad Z \sim \mathbb{Q}(.|\mathbf{X})$$

where the distribution $\mathbb{Q}$ varies according to the nature of corruption

Correcting Neighborhood Regression

As before, the naiive approach is to solve :

$$\min_{\theta} \frac{1}{n} \|y - \mathbf{Z}\theta\|_2^2$$

- will lead to an inconsistent estimate of the neighborhood regression vector θ^*
- Least squares error can be corrected
- What quantities need to be "corrected" to obtain a consistent form of linear regression ?

Correcting Neighborhood Regression

Consider the following population-level objective function:

$$\overline{\mathcal{L}}(\theta) = \frac{1}{2}\theta^T \Gamma \theta - \langle \theta, \gamma \rangle$$

where $\Gamma := cov(x)$ & $\gamma := cov(x, y)$. The true regression vector is the unique global minimizer of $\overline{\mathcal{L}}$

Strategy: solve a penalized version where (Γ, γ) is replaced by data-dependent estimates $(\hat{\Gamma}, \hat{\gamma})$ leading to :

$$\mathcal{L}_n(\theta) = \frac{1}{2}\theta^T \hat{\Gamma} \theta - \langle \theta, \hat{\gamma} \rangle$$

Correcting Neighborhood Regression

- We previously described a suitable unbiased estimator $\hat{\Gamma}$ for the case of additive corruptions
- Exercise 11.12 discusses a suitable unbiased estimator $\hat{\gamma}$ for the cross variance vector γ

Combining the above we are led to study the following corrected lasso estimator :

$$\min_{\|\theta\|_1 \leq \sqrt{\frac{n}{\log d}}} \left\{ \frac{1}{2} \theta^T \hat{\Gamma} \theta - \langle \theta, \hat{\gamma} \rangle + \lambda_n \|\theta\|_1 \right\}$$

Exercise 11.11 shows that an l_1 penalty and l_1 constraint are both needed when the objective function is non-convex

Property of the Corrected Lasso

Under suitable conditions - ones that still permit non-convexity, any local optimum is relatively close to the true regression vector. (**RE condition:**) We impose a restricted eigenvalue condition on $\hat{\Gamma}$, we assume there exists $\kappa > 0$ s.t :

$$\langle \Delta, \hat{\Gamma}\Delta \rangle \geq \kappa \|\Delta\|_2^2 - c_0 \frac{\log d}{n} \|\Delta\|_1^2 \qquad \forall \Delta \in \mathbb{R}^d$$

Also assume θ^* of the population objective has sparsity s and l_2 norm at most 1, also $n \geq s \log d$. These ensure that θ^* is feasible for the non-convex lasso.

Property of the Corrected Lasso

Theorem

Under the RE condition, suppose the pair $(\hat{\Gamma}, \hat{\gamma})$ satisfy the deviation condition:

$$\|\hat{\Gamma}\theta^* - \hat{\gamma}\|_{max} \leq \phi(\mathbb{Q}, \sigma_w)\sqrt{\frac{\log d}{n}}$$

for a prefactor $\phi(\mathbb{Q}, \sigma_w)$ depending on the conditional distribution $\mathbb{Q}$ and the noise standard deviation σ_w. Then for any regularization paramater $\lambda_n \geq 2(2c_0 + \phi(\mathbb{Q}, \sigma_w))\sqrt{\frac{\log d}{n}}$. Any local optimum to the corrected lasso program satisfies:

$$\|\tilde{\theta} - \theta^*\|_2 \leq \frac{2}{\kappa}\sqrt{s}\lambda_n$$

Graphs with Hidden Variables

- Sometimes, a given set of random variables might not be represented as a sparse graphical model on its own
- However, when augmented with an additional set of 'hidden' variables, the augmented system offers a sparse representation
- An extreme example is that of conditional independence on a hidden variable
- The fundamental idea underlying this is that given set is almost determined by a much smaller set of random variables which allows such a modelling

Estimation of the Graphical Model

- Consider a family of $d + r$ random variables, $X := (X_1, \ldots, X_d, X_{d+1}, \ldots, X_{d+r})$ that admits a representation as a sparse graphical model with $d + r$ vertices
- Assume that only the subvector $X_O := (X_1, \ldots, X_d)$ is observed while $X_H := (X_{d+1}, \ldots X_{d+r})$ stays hidden.
- The goal is to recover useful information about the underlying graph given the partial information
- The problem in the most general setup can get intractable, however, has an attractive matrix theoretic formulation in the Gaussian case

Estimation of Gaussian Graphical Model

- Under the Gaussian setup, the objective as before is to estimate the precision matrix
- The observed samples of X_O give us information about the covariance matrix Σ^*_{OO}
- The precision matrix $\Theta^\diamond$ of the full vector $X = (X_O, X_H)$ is sparse which follows from the Hammersley-Clifford Theorem

Thus, we have

$$\Theta^\diamond = \begin{bmatrix} \Theta^\diamond_{OO} & \Theta^\diamond_{OH} \\ \Theta^\diamond_{HO} & \Theta^\diamond_{HH} \end{bmatrix}$$

Estimation of Gaussian Graphical Models

Using block matrix inversion formula, one can write,

$$(\Sigma^*_{OO})^{-1} = \underbrace{\Theta^{\diamond}_{OO}}_{\Gamma^*} - \underbrace{\Theta^{\diamond}_{OH}(\Theta^{\diamond}_{HH})^{-1}\Theta^{\diamond}_{HO}}_{\Lambda^*}$$

where

- $\Gamma^* := \Theta^{\diamond}_{OO}$ is sparse
- $\Lambda^* := \Theta^{\diamond}_{OH}(\Theta^{\diamond}_{HH})^{-1}\Theta^{\diamond}_{HO}$ has a rank atmost $\min(r, d)$

If $r \ll d$, then the inverse covariance matrix of the observed variables can be written as a sum of a sparse and a low-rank matrix

Estimation of Gaussian Graphical Models

Assume that we have been given n i.i.d. samples $x_i \in \mathbb{R}^d$ and $x_i \sim \mathcal{N}(0, \Sigma^*_{OO})$. Then we can consider an observation model of the form

$$\mathbf{Y} = \Gamma^* - \Lambda^* + \mathbf{W}$$

where

- $\mathbf{Y} := (\hat{\Sigma})^{-1}$ where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$
- $\mathbf{W} \in \mathbb{R}^{d \times d}$ is the noise stochastic matrix modelling the difference in population and sample covariance matrices

It is also assumed that $n > d$ to ensure that $\mathbf{Y}$ exists and is well-defined with high probability

Estimation of Gaussian Graphical Models

Estimation of the components can be done using simple two-step estimator, by first estimating the sparse component and then using it to deduce the low rank component. Thus, for some given threshold $\nu_n > 0$ we have

$$\hat{\Gamma} := T_{\nu_n}((\hat{\Sigma})^{-1})) \text{ and } \hat{\Lambda} := \hat{\Gamma} - (\hat{\Sigma})^{-1})$$

where $T_{\nu_n}(\nu) = \nu \mathbb{1}\{|\nu| > \nu_n\}$

Estimation of Gaussian Graphical Models

The two-step estimator, though simple, need not be very consistent since it tries to estimate Γ^* as if $\Gamma^* - \Lambda^*$ is also sparse. To ensure better performance, we need additional assumptions on the pair (Γ^*, Λ^*).

We assume that Λ^* satisfies a α-spikiness constraint, that is $\|\Lambda^*\|_{\max} \leq \frac{\alpha}{d}$. Intuitively, this ensures that the elements in the low-rank part are not too big and hence estimate of Γ^* is rather consistent.

In addition, we assume that $|||\sqrt{\Theta^*}|||_\infty = \max_{j=1,2,\ldots,d} \sum_{k=1}^{d} \sqrt{|\Theta^*_{jk}|} \leq \sqrt{M}$ for some given M. This essentially helps in designing the threshold used for the sparse component.

Estimation of Gaussian Graphical Models

Theorem

Consider a precision matrix Θ^* *that can be decomposed as the difference* $\Gamma^* - \Lambda^*$ *where* Γ^* *has atmost* s *non-zero entries per row and* Λ^* *is* α*-spiky. Given* $n > d$ *i.i.d. samples from* $\mathcal{N}(0, (\Theta^*_{OO})^{-1})$ *and* $\delta \in (0, 1]$*, then estimators* $(\hat{\Gamma}, \hat{\Lambda})$ *obtained by choosing* $\nu_n := M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{\alpha}{d}$ *satisfy the bounds*

$$\|\hat{\Gamma} - \Gamma^*\|_{\max} \leq 2M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{2\alpha}{d}$$

$$\|\hat{\Lambda} - \Lambda^*\|_2 \leq M\left(2\sqrt{\frac{d}{n}} + \delta\right) + s\|\hat{\Gamma} - \Gamma^*\|_{\max}$$

with probability atleast $1 - c_1 e^{-c_2 n\delta^2}$

Estimation of Gaussian Graphical Models

We provide a sketch of the proof here. First it can be shown that $\mathbf{Y} := (\hat{\Sigma})^{-1})$ is a good estimate of Θ^*. In this regard, it can be shown that

$$\|\mathbf{Y} - \Theta^*\|_2 \leq M\left(2\sqrt{\frac{d}{n}} + \delta\right)$$

$$\|\mathbf{Y} - \Theta^*\|_{\max} \leq M\left(4\sqrt{\frac{\log d}{n}} + \delta\right)$$

with probability atleast $1 - c_1 e^{-c_2 n\delta^2}$. The main idea here is to use the identity

$$(\hat{\Sigma})^{-1} - \Theta^* = \sqrt{\Theta^*}\left\{n^{-1}V^T V - I_d\right\}\sqrt{\Theta^*}$$

Estimation of Gaussian Graphical Models

Then using the above results, we can obtain the required bounds as follows

$$\begin{aligned}\|\hat{\Gamma} - \Gamma^*\|_{\max} &\leq \|\mathbf{Y} - \Theta^*\|_{\max} + \|\mathbf{Y} - T_{\nu_n}(\mathbf{Y})\|_{\max} + \|\Lambda^*\|_{\max} \\ &\leq M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \nu_n + \frac{\alpha}{d} \\ &\leq 2M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{2\alpha}{d}\end{aligned}$$

and

$$\begin{aligned}\|\hat{\Lambda} - \Lambda^*\|_2 \leq \|\mathbf{Y} - \Theta^*\|_2 + \|\hat{\Gamma} - \Gamma^*\|_2 &\leq \|\mathbf{Y} - \Theta^*\|_2 + s\|\hat{\Gamma} - \Gamma^*\|_{\max} \\ &\leq M\left(2\sqrt{\frac{d}{n}} + \delta\right) + s\|\hat{\Gamma} - \Gamma^*\|_{\max}\end{aligned}$$