Graphical Model Selection

Binan Gu

Department of Mathematical Sciences, New Jersey Institute of Technology

New Jersey Institute of Technology
Fall 2020 Machine Learning Talk III

Motivation: Data Representation

Ising Model

Motivation: Data Representation

Ising Model

Given undirected $G = (V, E)$ and Bernoulli variables $X = (X_1, \ldots, X_p) \in \{-1, +1\}^p$ on V,

Motivation: Data Representation

Ising Model

Given undirected $G = (V, E)$ and Bernoulli variables $X = (X_1, \ldots, X_p) \in \{-1, +1\}^p$ on V, the Ising model is the family of distributions

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

where θ is connection strength and $A(\theta)$ a normalization constant.

Motivation: Data Representation

Ising Model

Given undirected $G = (V, E)$ and Bernoulli variables $X = (X_1, \ldots, X_p) \in \{-1, +1\}^p$ on V, the Ising model is the family of distributions

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

where θ is connection strength and $A(\theta)$ a normalization constant. In practice, $A(\theta)$ becomes computationally taxing when p is big.

Motivation: Data Representation

Reformulation of Multivariate Gaussian Variables

Motivation: Data Representation

Reformulation of Multivariate Gaussian Variables

Let X be a variable with distribution

$$\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) = \frac{1}{(2\pi)^{\frac{s}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu)}.$$

Motivation: Data Representation

Reformulation of Multivariate Gaussian Variables

Let X be a variable with distribution

$$\mathbb{P}_{\mu,\mathbf{\Sigma}}(x) = \frac{1}{(2\pi)^{\frac{s}{2}} \det(\mathbf{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}.$$

With $\gamma = -\mathbf{\Sigma}^{-1}\mu$, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, we obtain an "Ising" form

Motivation: Data Representation

Reformulation of Multivariate Gaussian Variables

Let X be a variable with distribution

$$\mathbb{P}_{\mu,\mathbf{\Sigma}}(x) = \frac{1}{(2\pi)^{\frac{s}{2}} \det(\mathbf{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}.$$

With $\gamma = -\mathbf{\Sigma}^{-1}\mu$, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, we obtain an “Ising” form

$$\mathbb{P}_{\gamma,\mathbf{\Theta}}(x) = \exp\left\{ \underbrace{\sum_{s=1} \gamma_s x_s}_{\text{diagonal}} \right.$$

Motivation: Data Representation

Reformulation of Multivariate Gaussian Variables

Let X be a variable with distribution

$$\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) = \frac{1}{(2\pi)^{\frac{s}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu)}.$$

With $\gamma = -\boldsymbol{\Sigma}^{-1}\mu$, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, we obtain an “Ising” form

$$\mathbb{P}_{\gamma,\boldsymbol{\Theta}}(x) = \exp\left\{ \underbrace{\sum_{s=1} \gamma_s x_s}_{\text{diagonal}} - \underbrace{\frac{1}{2} \sum_{s,t=1}^{p} \theta_{st} x_s x_t}_{\text{off-diagonal}} - \right.$$

Motivation: Data Representation

Reformulation of Multivariate Gaussian Variables

Let X be a variable with distribution

$$\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) = \frac{1}{(2\pi)^{\frac{s}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu)}.$$

With $\gamma = -\boldsymbol{\Sigma}^{-1}\mu$, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, we obtain an “Ising” form

$$\mathbb{P}_{\gamma,\boldsymbol{\Theta}}(x) = \exp\left\{ \underbrace{\sum_{s=1} \gamma_s x_s}_{\text{diagonal}} - \underbrace{\frac{1}{2} \sum_{s,t=1}^{p} \theta_{st} x_s x_t}_{\text{off-diagonal}} - A(\boldsymbol{\Theta}) \right\}$$

where $A(\boldsymbol{\Theta}) = -\frac{1}{2} \log\det\left[\frac{\boldsymbol{\Theta}}{2\pi}\right]$.

Sparsity of the Precision Matrix Θ

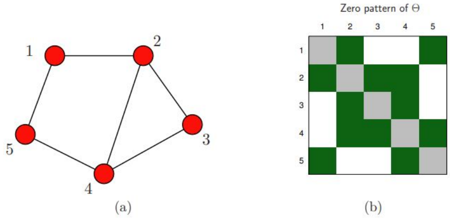

Figure 9.3 *(a) An undirected graph G on five vertices. (b) Associated sparsity pattern of the precision matrix $\mathbf{\Theta}$. White squares correspond to zero entries.*

Sparsity of the Precision Matrix Θ

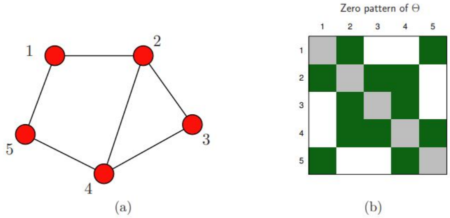

Figure 9.3 *(a) An undirected graph G on five vertices. (b) Associated sparsity pattern of the precision matrix* $\boldsymbol{\Theta}$. *White squares correspond to zero entries.*

- The entire graph represents the joint distribution.

Sparsity of the Precision Matrix Θ

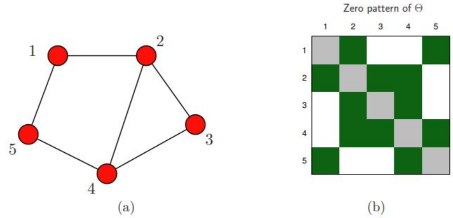

Figure 9.3 *(a) An undirected graph G on five vertices. (b) Associated sparsity pattern of the precision matrix* $\boldsymbol{\Theta}$*. White squares correspond to zero entries.*

- The entire graph represents the joint distribution.
- Dependence structure is represented by edges, e.g.

$$X_1 \perp X_4 \mid X_2, X_3, X_5,$$

also known as *conditional independence*.

Conditional Dependence Structure

Given p-dimensional $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Consider $Y = X_p$ and $Z = (X_1, \ldots, X_{p-1})$. Thus

$$\mu = \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

Conditional Dependence Structure

Given p-dimensional $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Consider $Y = X_p$ and $Z = \left(X_1, \ldots, X_{p-1}\right)$. Thus

$$\mu = \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

and the conditional distribution

$$Y \mid Z = z \sim \mathcal{N}\left(\mu_Y + (z - \mu_Z)^T \underbrace{\mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY}}_{\text{regression coef.}}, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY} \right).$$

Conditional Dependence Structure

Given p-dimensional $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Consider $Y = X_p$ and $Z = (X_1, \ldots, X_{p-1})$. Thus

$$\mu = \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

and the conditional distribution

$$Y \mid Z = z \sim \mathcal{N}\left(\mu_Y + (z - \mu_Z)^T \underbrace{\mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}}_{\text{regression coef.}}, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY} \right).$$

Write $\beta_{Y|Z} = \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}$.

Conditional Dependence Structure

Given p-dimensional $X \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$. Consider $Y = X_p$ and $Z = \left(X_1, \ldots, X_{p-1}\right)$. Thus

$$\mu = \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

and the conditional distribution

$$Y \mid Z = z \sim \mathcal{N}\left(\mu_Y + (z - \mu_Z)^T \underbrace{\boldsymbol{\Sigma}_{ZZ}^{-1}\sigma_{ZY}}_{\text{regression coef.}}, \sigma_{YY} - \sigma_{ZY}^T \boldsymbol{\Sigma}_{ZZ}^{-1}\sigma_{ZY}\right).$$

Write $\beta_{Y|Z} = \boldsymbol{\Sigma}_{ZZ}^{-1}\sigma_{ZY}$. If $\beta_{Y|Z_j} = 0$, then Y and Z_j are conditionally independent given the rest.

Symmetry of Grpahical Models

One can do this for arbitrary Y and thus form a matrix β such that each entry captures the conditional dependence of variable X_i and X_j.

Symmetry of Grpahical Models

One can do this for arbitrary Y and thus form a matrix β such that each entry captures the conditional dependence of variable X_i and X_j.

Symmetry

Consider $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. Then,

$$\mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{ZZ} & \theta_{ZY} \\ \theta_{ZY}^T & \theta_{YY} \end{pmatrix}$$

where in particular,

$$\theta_{ZY} = -\theta_{YY}\mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} = -\theta_{YY}\beta_{Y|Z}$$

Symmetry of Grpahical Models

One can do this for arbitrary Y and thus form a matrix β such that each entry captures the conditional dependence of variable X_i and X_j.

Symmetry

Consider $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. Then,

$$\mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{ZZ} & \theta_{ZY} \\ \theta_{ZY}^T & \theta_{YY} \end{pmatrix}$$

where in particular,

$$\theta_{ZY} = -\theta_{YY}\mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} = -\theta_{YY}\beta_{Y|Z}$$

which means $\mathbf{\Theta}$ symmetrically and completely determines conditional dependence structure. Regression analysis doesn't honor this symmetry (normal equation solutions).

The Graph Selection Problem

Problem Description

Given some data sampled from a graphical model whose underlying structure is unknown, how do we use this data to select the correct graphical representation?

The Graph Selection Problem

Problem Description

Given some data sampled from a graphical model whose underlying structure is unknown, how do we use this data to select the correct graphical representation?

Mathematical Formulation of *covariance selection*

For a collection $\{x_1, \ldots, x_N\}$ sampled from random variables $X \in \mathbb{R}^p$ where $p \gg N$, can we estimate $\mathbf{\Theta}$ which, in turn, gives us the graphical structure of X?

The Graph Selection Problem

Problem Description

Given some data sampled from a graphical model whose underlying structure is unknown, how do we use this data to select the correct graphical representation?

Mathematical Formulation of *covariance selection*

For a collection $\{x_1, \ldots, x_N\}$ sampled from random variables $X \in \mathbb{R}^p$ where $p \gg N$, can we estimate $\mathbf{\Theta}$ which, in turn, gives us the graphical structure of X?

Two Approaches

The Graph Selection Problem

Problem Description

Given some data sampled from a graphical model whose underlying structure is unknown, how do we use this data to select the correct graphical representation?

Mathematical Formulation of *covariance selection*

For a collection $\{x_1, \ldots, x_N\}$ sampled from random variables $X \in \mathbb{R}^p$ where $p \gg N$, can we estimate $\mathbf{\Theta}$ which, in turn, gives us the graphical structure of X?

Two Approaches

1. Conditional Inference.

The Graph Selection Problem

Problem Description

Given some data sampled from a graphical model whose underlying structure is unknown, how do we use this data to select the correct graphical representation?

Mathematical Formulation of *covariance selection*

For a collection $\{x_1, \ldots, x_N\}$ sampled from random variables $X \in \mathbb{R}^p$ where $p \gg N$, can we estimate $\mathbf{\Theta}$ which, in turn, gives us the graphical structure of X?

Two Approaches

1. Conditional Inference.
2. Penalized Likelihood.

The Graph Selection Problem

Problem Description

Given some data sampled from a graphical model whose underlying structure is unknown, how do we use this data to select the correct graphical representation?

Mathematical Formulation of *covariance selection*

For a collection $\{x_1, \ldots, x_N\}$ sampled from random variables $X \in \mathbb{R}^p$ where $p \gg N$, can we estimate $\mathbf{\Theta}$ which, in turn, gives us the graphical structure of X?

Two Approaches

1. Conditional Inference.
2. Penalized Likelihood.

Conditional Inference

Consider Gaussian variables $X = (X_1, \ldots, X_p)$ embedded in $G = (V, E)$.

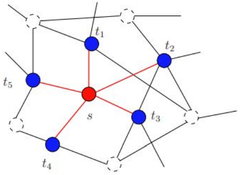

Figure 9.6 *The dark blue vertices form the neighborhood set $\mathcal{N}(s)$ of vertex s (drawn in red); the set $\mathcal{N}^+(s)$ is given by the union $\mathcal{N}(s) \cup \{s\}$. Note that $\mathcal{N}(s)$ is a cut set in the graph that separates $\{s\}$ from $V \backslash \mathcal{N}^+(s)$. Consequently, the variable X_s is conditionally independent of $X_{V \backslash \mathcal{N}^+(s)}$ given the variables $X_{\mathcal{N}(s)}$ in the neighborhood set. This conditional independence implies that the optimal predictor of X_s based on all other variables in the graph depends only on $X_{\mathcal{N}(s)}$.*

Conditional Inference

Consider Gaussian variables $X = (X_1, \ldots, X_p)$ embedded in $G = (V, E)$. For $s \in V$, define its complement and neighborhood

$$X_{\setminus\{s\}} = \{X_t, t \in V \setminus \{s\}\} \in \mathbb{R}^{p-1}$$
$$\mathcal{N}(s) = \{t \in V \mid (s,t) \in E\}$$

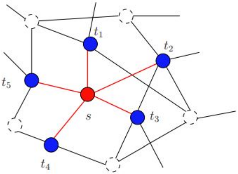

Figure 9.6 *The dark blue vertices form the neighborhood set $\mathcal{N}(s)$ of vertex s (drawn in red); the set $\mathcal{N}^+(s)$ is given by the union $\mathcal{N}(s) \cup \{s\}$. Note that $\mathcal{N}(s)$ is a cut set in the graph that separates $\{s\}$ from $V \setminus \mathcal{N}^+(s)$. Consequently, the variable X_s is conditionally independent of $X_{V \setminus \mathcal{N}^+(s)}$ given the variables $X_{\mathcal{N}(s)}$ in the neighborhood set. This conditional independence implies that the optimal predictor of X_s based on all other variables in the graph depends only on $X_{\mathcal{N}(s)}$.*

Conditional Inference

Distributional Equivalence

$$\left(X_s \mid X_{\setminus\{s\}}\right) \overset{d}{=} \left(X_s \mid X_{\mathcal{N}(s)}\right)$$

If one wants to predict X_s given the rest, you only need to look "around" X_s, i.e. the best predictor is a function of $X_{\mathcal{N}(s)}$.

$$X_s = X_{\setminus\{s\}}^T \beta_s + W_{\setminus\{s\}}.$$

To estimate $\mathbf{\Theta}$ with $\hat{\mathbf{\Theta}}$ is to approximate the true edge set E with an estimate $\hat{E}$.

Parallel Graphical Lasso

Key steps

1. For each vertex $s = 1, 2, \ldots, p$, do PARALLEL!
   1.1 Some type of regression, say, *lasso*,

$$\hat{\beta}^s \in \arg\min_{\beta_s \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(x_{i,s} - x_{i,V\setminus\{s\}}^T \beta^s \right)^2 + \lambda \left\| \beta_s \right\|_1 \right\}$$

   1.2 Compute the estimate $\hat{\mathcal{N}}(s) = \mathrm{supp}\left(\hat{\beta}_s\right)$, i.e. nodes where $\hat{\beta}^s$ is nonzero.
2. Combine the estimates $\hat{\mathcal{N}}(s)$ for every $s \in V$.

Theoretical Guarantee

For graphical *lasso* to ensure $\hat{G} = G$ with high probability,

$$\left\| \hat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right\|_2 \lesssim \sqrt{\frac{d^2 \log p}{N}}$$

where d is maximum degree of any node. We see that if $N = \Omega\left(d^2 \log N\right)$, we have recovery of the covariance structure and thus a graphical model. The proof relies on concentration.

Take-aways

Advantanges

- Parallelizable
- Fast without the use of extensive packages

Disadvantanges

- No confidence interval of parameter estimation

References

Hastie, Trevor and Tibshirani, Robert and Wainwright, Martin. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. 2015.