Undirected Graphical Models

Qing Zhou

UCLA Department of Statistics

Stats 212 Graphical Models
Lecture Notes

Outline

Graphoid

Graphoid axioms (Pearl (1988), §3.1.2.)

CI statement defines a ternary relation: $\langle X, Y \mid Z\rangle$ for $X \perp Y \mid Z$. Suppose X, Y, Z, W are disjoint subsets of random variables from a joint distribution $\mathbb{P}$. Then the CI relation satisfies

(C1) symmetry: $\langle X, Y \mid Z\rangle \Rightarrow \langle Y, X \mid Z\rangle$;

(C2) decomposition: $\langle X, YW \mid Z\rangle \Rightarrow \langle X, Y \mid Z\rangle$;

(C3) weak union: $\langle X, YW \mid Z\rangle \Rightarrow \langle X, Y \mid ZW\rangle$;

(C4) contraction: $\langle X, Y \mid Z\rangle \& \langle X, W \mid ZY\rangle \Rightarrow \langle X, YW \mid Z\rangle$.

If the joint density of $\mathbb{P}$ wrt a product measure is positive and continuous, then

(C5) intersection: $\langle X, Y \mid ZW\rangle \& \langle X, W \mid ZY\rangle \Rightarrow \langle X, YW \mid Z\rangle$.

In the above, $YW := Y \cup W$.

Graphoid

Any ternary relation $\langle A, B \mid C \rangle$ that satisfies (C1) to (C4) is called a *semi-graphoid*. If (C5) also holds, then it is called a *graphoid*.

Examples of graphoid:

1. Conditional independence of $\mathbb{P}$ (positive and continous).
2. Graph separation in undirected graph: $\langle X, Y \mid Z \rangle$ means nodes Z separate X and Y, i.e. $X - Z - Y$.

Graph separation provides an intuitive graphical interpretation for the CI axioms.

Undirected graphs

Definition: A graph $\mathcal{G} = (V, E)$, $V = \{1, \ldots, p\}$ is a set of vertices (or nodes) and $E \subset V \times V$ is a set of edges.

- Undirected edge $i - j$: $(i, j) \in E \Leftrightarrow (j, i) \in E$.
- Associate V to random variables X_i $(i = 1, \ldots, p)$ with joint distribution $\mathbb{P}$. Then $(\mathcal{G}, \mathbb{P})$ is called a graphical model. Often use node i and X_i interchangeably.
- Use graph separation to represent conditional independence among $X_1, \ldots, X_p$.

Undirected graphs

Reference: Lauritzen (1996), chapters 2 and 3.

Terminology for undirected graph $\mathcal{G} = (V, E)$

- i and j are *neighbors* if $(i, j) \in E$; $\mathrm{ne}(i)$ denotes the set of neighbors of i.
- A *path* of length n from i to j is a sequence $a_0 = i, \ldots, a_n = j$ of distinct vertices so that $(a_{k-1}, a_k) \in E$ for all $k = 1, \ldots, n$.
- A subset $C \subset V$ separates a and b if all paths from a to b intersect C.
- C separates A and B if C separates a and b for every $a \in A$ and $b \in B$. Write $A - C - B$.

Markov properties

Markov properties on undirected graphs

Consider undirected graphical model $(\mathcal{G}, \mathbb{P})$. We say $\mathbb{P}$ satisfies

- (P) the pairwise Markov property wrt $\mathcal{G}$ if

$$(i,j) \notin E \Rightarrow i \perp j \mid V \setminus \{i,j\} := [V]_{ij};$$

- (L) the local Markov property wrt $\mathcal{G}$ if

$$(i,j) \notin E \Rightarrow i \perp j \mid \mathrm{ne}(i);$$

- (G) the global Markov property wrt $\mathcal{G}$ if for any disjoint (A, B, C),

$$A - C - B \Rightarrow A \perp B \mid C.$$

Factorization via cliques

- Complete subset and clique: A subset of $C \subset V$ is complete if the subgraph on C is complete. A complete subset that is maximal (wrt $\subset$) is called a clique.
- (F) Factorization: $\mathbb{P}$ factorizes according to $\mathcal{G}$ if for every clique A, there exists $\psi_A(x_A) \geq 0$, such that the joint density of $\mathbb{P}$ has the form

$$f(x) = \prod_{A \in \mathcal{C}} \psi_A(x_A),$$

  where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$.
- Relations: (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P).

Markov properties

Examples.

- Markov chain

$$X_1 - X_2 - X_3 - \cdots - X_{n-1} - X_n$$

Cliques: $\{i, i+1\}, i = 1, \ldots, n-1$.
(F) holds:

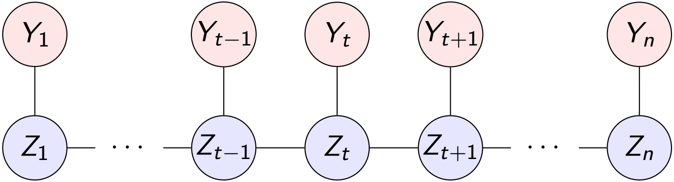
$$\begin{aligned}\mathbb{P}(X_1, \ldots, X_n) &= \mathbb{P}(X_1)\mathbb{P}(X_2 \mid X_1) \cdots \mathbb{P}(X_n \mid X_{n-1}) \\ &= \psi_1(X_1, X_2) \cdots \psi_{n-1}(X_{n-1}, X_n).\end{aligned}$$

Thus, (G) holds: For any $i < j < k$,

$$i - j - k \Rightarrow X_i \perp X_k \mid X_j.$$

Markov properties

- Hidden Markov model $\{(Z_t, Y_t) : t = 1, \ldots, n\}$.



Cliques: $\{Z_t, Z_{t+1}\}, t = 1, \ldots, n-1$, $\{Z_t, Y_t\}, t = 1, \ldots, n$.

(F) holds:
$$\begin{aligned}\mathbb{P}(Y, Z) &= \mathbb{P}(Z_1)\mathbb{P}(Y_1 \mid Z_1)\mathbb{P}(Z_2 \mid Z_1)\mathbb{P}(Y_2 \mid Z_2) \\ &\qquad \cdots \mathbb{P}(Z_n \mid Z_{n-1})\mathbb{P}(Y_n \mid Z_n) \\ &= \prod_{t=1}^{n-1} f_t(Z_t, Z_{t+1}) \prod_{t=1}^{n} g_t(Z_t, Y_t)\end{aligned}$$

Thus, (G) holds: V_{t-i}, Y_t and V_{t+j} are mutually independent conditional on Z_t for $i, j \geq 1$, where $V_k = \{Y_k, Z_k\}$.

Markov properties

When does (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L) $\Leftrightarrow$ (P)?

Theorem 1

If $\mathbb{P}$ has a positive and continuous density f with respect to a product measure, then (F) $\Leftrightarrow$ (P).

- Product measure: (1) $X_j \in \mathbb{R}$, use Lebesgue measure; (2) X_j finite discrete, use counting measure.
- Conclusion implies (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L) $\Leftrightarrow$ (P).
- Counter example. Let $p = 5$, $X_1, X_5 \sim_{iid}$ Bern(0.5), $X_2 = X_1$, $X_4 = X_5$, and $X_3 = X_2 X_4$. This defines $\mathbb{P}$. Let $\mathcal{G}$ be a chain $E = \{(i, i+1) : i = 1, \ldots, 4\}$.
  Then (L) holds but not (G). Because density (probability mass function) is not positive on all possible values of X_i's.
  (L): $X_2 \perp X_4 \mid (X_1, X_3)$ true; (G): $X_2 \perp X_4 \mid X_3$ false!

Markov properties

Conditional independence graph (CIG):

- Definition: A CIG is a graphical model $(\mathcal{G}, \mathbb{P})$ such that (P) holds. That is,

$$(i,j) \notin E \Rightarrow i \perp j \mid V \setminus \{i,j\} := [V]_{ij}.$$

- Sparser graph $\mathcal{G}$ implies more conditional independence (CI) relations.
- One can always choose the minimal $\mathcal{G}$ such that (P) holds to be the CIG, i.e., replace $\Rightarrow$ by $\Leftrightarrow$.
- Estimate the structure of $\mathcal{G}$ to detect CI relations, assuming we have observed iid data from $\mathbb{P}$.

Gaussian graphical models

A CIG with $\mathbb{P} = \mathcal{N}_p(0, \Sigma)$, $\Sigma > 0$ (positive definite).

Lemma 1

Suppose $(X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$ with $\Sigma > 0$ and let $\Theta = (\theta_{jk})_{p \times p} = \Sigma^{-1}$. Then

$$\theta_{jk} = 0 \Leftrightarrow X_j \perp X_k \mid X_{-\{j,k\}}. \tag{1}$$

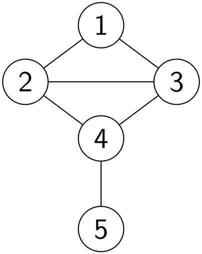
- Θ is called the precision matrix.
- According to (1), construct a graph $\mathcal{G}$ as

$$\theta_{jk} \neq 0 \Leftrightarrow (j, k) \in E, \tag{2}$$

  i.e. (P) holds. Since $\mathbb{P}$ has a continuous and positive density, (L), (G) and (F) hold.
- One can verify (F) directly as well.

Example: Given the following Θ, construct $\mathcal{G}$ by (2).

$$\Theta = \begin{bmatrix} * & * & * & 0 & 0 \\ * & * & * & * & 0 \\ * & * & * & * & 0 \\ 0 & * & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}$$



- Find all S such that $X_1 \perp X_5 \mid S$.
  By (G), find all S that separates nodes 1 and 5:
  $S = \{2,3\}, \{4\}, \{2,4\}, \{3,4\}, \{2,3,4\}$.
- Cliques: $\{1,2,3\}, \{2,3,4\}, \{4,5\}$; directly verify (F).

Partial correlation and neighborhood regression

- Partial correlation between j and k given $[V]_{jk}$:
  $\rho_{jk} = -\theta_{jk}/\sqrt{\theta_{jj}\theta_{kk}}$.
  Correlation calculated from $\Sigma_{(j,k)|[V]_{jk}} = \mathrm{Var}(j, k \mid [V]_{jk})$.
- Neighborhood regression, regress X_j on X_{-j}:

$$X_j = \sum_{i \neq j} \beta_{ij} X_i + \varepsilon_j. \tag{3}$$

  Then $\beta_{kj} = -\theta_{jk}/\theta_{jj}$. (By symmetry $\beta_{jk} = -\theta_{kj}/\theta_{kk}$.)
- Thus, we have

$$(j, k) \notin E \Leftrightarrow \theta_{jk} = 0 \Leftrightarrow \rho_{jk} = 0 \Leftrightarrow \beta_{kj} = \beta_{jk} = 0. \tag{4}$$

Learning GGMs: Given $x_i \sim_{iid} \mathcal{N}_p(0, \Sigma)$, $i = 1, \ldots, n$, estimate

the structure of $\mathcal{G} \Leftrightarrow \text{supp}(\Theta) = \{(j,k) : \theta_{jk} \neq 0\}$.

Also called covariance selection (Dempster 1972).

- Log-likelihood

$$\ell(\Sigma) = -\frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \text{tr}(S\Sigma^{-1}),$$

where $S = \sum_i x_i x_i^\mathsf{T}$ is a $p \times p$ matrix (sufficient statistic).

- $\hat{\Sigma}^{\text{MLE}} = S/n$ (always exists).
- If $n > p$, inverte $\hat{\Sigma}^{\text{MLE}} \Rightarrow \hat{\Theta}^{\text{MLE}} = (\hat{\Sigma}^{\text{MLE}})^{-1}$.
  Then obtain $\widehat{\mathcal{G}}$ by thresholding: $\widehat{E} = \{(j,k) : |\hat{\theta}_{jk}^{\text{MLE}}| > \tau\}$.

Gaussian graphical models

Regularized estimation under ℓ_1 penalty (Yuan and Lin 2007; Friedman et al. 2008; Banerjee et al. 2008)

- Element-wise ℓ_1 norm $\|\Theta\|_1 := \sum_{j<k} |\theta_{jk}|$.
- ℓ_1 regularized estimate $\hat{\Theta} = \text{argmin}_{\Theta > 0} f(\Theta)$,

$$
\begin{aligned}
f(\Theta) &= -\frac{2}{n}\ell(\Theta^{-1}) + \lambda\|\Theta\|_1 \\
&= -\log\det(\Theta) + \text{tr}(\hat{\Sigma}^{\text{MLE}}\Theta) + \lambda\|\Theta\|_1.
\end{aligned}
$$

- f is convex, efficient algorithm.
- Well-defined for $p > n$.
- Sparse solution, $\hat{\theta}_{jk} = 0$ for some (j, k).

Gaussian graphical models

Estimate $\mathcal{G}$ from $\hat{\Theta}$

- $\widehat{E} = \{(j,k) : \hat{\theta}_{jk} \neq 0\}$, but needs very strong assumptions (irrepresentability) for $\mathbb{P}(\widehat{E} = E_0) \to 1$.
- Operator norm error:

$$\|\hat{\Theta} - \Theta_0\|_2 \lesssim \sqrt{d^2 \log p/n}. \tag{5}$$

  d: Maximum degree of G.
- Thresholding $\hat{\Theta}$: $\widehat{E} = \{(j,k) : |\hat{\theta}_{jk}| > \tau\}$. Weaker assumptions (RE, beta-min) for $\mathbb{P}(\widehat{E} = E_0) \to 1$.

Choosing λ by cross-validation, λ^*_{CV}, then $\mathbb{P}(\widehat{E}(\lambda^*_{CV}) \supset E_0) \to 1$ under certain conditions (RE, beta-min).

Gaussian graphical models

Estimate $\mathcal{G}$ by neighborhood regression (Meinshausen and Bühlmann 2006)

- Apply model selection (e.g. lasso) for each neighborhood regression (3) $\Rightarrow \hat{\beta}_{jk}$ $(j, k = 1, \dots, p)$.
- Combine results to define $\widehat{\mathcal{G}}$, e.g.,

$$\widehat{E} = \{(j,k) : \hat{\beta}_{jk} \neq 0, \hat{\beta}_{kj} \neq 0\}.$$

- Approximate $\hat{\Theta}$ if lasso is used in neighborhood regression.

Discrete graphical models

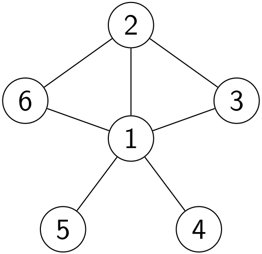
Reference: Hastie et al. (2015), Ch 9.
Ising model:

- $X_i \in \{-1, +1\}, i \in V = [p]$.
- Given an undirected graph $\mathcal{G} = (V, E)$, define a joint distribution

$$\mathbb{P}(x_1, \ldots, x_p; \theta) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{i \in V} \theta_i x_i + \sum_{(j,k) \in E} \theta_{jk} x_j x_k \right\}. \tag{6}$$

- Easy to verify (F) holds $\Rightarrow$ (G), (L), (P).
- Example application: model social networks.

Example: Given the following $\mathcal{G}$, define $\mathbb{P}(x_1, \ldots, x_6)$ as in (6).



- Cliques: $\{1,2,3\}, \{1,2,6\}, \{1,4\}, \{1,5\}$.
- Verify (F) $\Rightarrow$ (G), (L), (P).
- Example CI statements by (G):
  $X_4 \perp X_5 \mid X_1$
  $X_3 \perp X_6 \mid \{X_1, X_2\}$
  $\{X_2, X_3, X_6\} \perp \{X_4, X_5\} \mid X_1$

Discrete graphical models

Generalization:

- $X_i \in \{1, \ldots, m\}, i \in V = [p]$.
- Given an undirected graph $\mathcal{G} = (V, E)$, define a joint distribution

$$\mathbb{P}(x_1, \ldots, x_p; \theta) = \frac{1}{Z(\gamma, \theta)} \exp\left\{ \sum_{i \in V} \sum_{z=1}^{m} \gamma_{iz} I(x_i = z) + \sum_{(j,k) \in E} \theta_{jk} I(x_j = x_k) \right\}.$$

Discrete graphical models

Learning graphs from data:

- Full likelihood-based learning is difficult: $Z(\theta)$ no closed-form.
- More practical to do neighborhood regression. From (6), get $[X_i \mid X_{-i}]$ which leads to a logistic regression model:

$$\log\left[\frac{\mathbb{P}(X_i = 1 \mid X_{-i})}{\mathbb{P}(X_i = -1 \mid X_{-i})}\right] = 2\theta_i + \sum_{j \in \mathrm{ne}(i)} 2\theta_{ij} X_j,$$

  where $\mathrm{ne}(i) = \{j \in V : (i,j) \in E\}$ is the set of neighbors of node i in G.

Discrete graphical models

Learning graphs from data:

- For each $i \in [p]$, apply logistic regression X_i on X_{-i} with variable selection to estimate $\widehat{N}(i)$ (estimated neighbor set).
- For example, ℓ_1-regularized logistic regression or BIC stepwise selection.
- Combine $\{\widehat{N}(i) : i \in V\}$ to construct $\widehat{\mathcal{G}}$.
- Sample size $n = \Omega(d^2 \log p)$ sufficient for $\widehat{\mathcal{G}} = \mathcal{G}$ with high probability.

Faithfulness

Given a graphical model $(\mathcal{G}, \mathbb{P})$ where $\mathbb{P}$ satisfies, say (G).
Then graph separation $\Rightarrow$ condition independence, but not $\Leftarrow$.
If $\mathbb{P}$ is faithful to $\mathcal{G}$ then $\Leftarrow$ holds as well. In this case, we have $\Leftrightarrow$ (perfectness).

Definition 1

For a graphical model $(\mathcal{G}, \mathbb{P})$, we say the distribution $\mathbb{P}$ is faithful to the graph $\mathcal{G}$ if for every triple of disjoint sets $A, B, S \subset V$,

$$A \perp B \mid S \Leftrightarrow S \text{ separates } A \text{ and } B.$$

Faithfulness

How likely is $\mathbb{P}$ faithful?

Gaussian graphical models, $\mathbb{P}$ is Gaussian $\mathcal{N}(0, \Sigma) = \mathcal{N}(0, \Theta^{-1})$.

- Given $\mathcal{G}$, consider all positive-definite Θ such that $\text{supp}(\Theta) = E \cup \{(i, i) : i \in [p]\}$. Then for almost all such Θ, the distribution $\mathcal{N}(0, \Theta^{-1})$ is faithful to $\mathcal{G}$.
- Counterexamples: The parameters in Θ satisfy additional equality constraints that define CI in $\mathbb{P}$ not implied by any separation in $\mathcal{G}$.

Markov blanket

Definition 2 (Markov blanket)

A *Markov blanket* of $i \in V$ is any subset $S \subset V_{-i}$ such that

$$X_i \perp V_{-i} \setminus S \mid S. \tag{7}$$

A *Markov boundary* is a minimal Markov blanket, i.e., none of its proper subset satisfies (7).

- For an undirected graph model $(\mathcal{G}, \mathbb{P})$, $\mathrm{ne}(i)$ is a Markov blanket of i (by local Markov property) and it is a Markov boundary if $\mathbb{P}$ is faithful.
- Neighborhood regression: find Markov boundary (MB) of i.

Markov blanket

The grow-shrink algorithm (Margaritis and Thrun 1999)

Find MB of $i \in V$:

$S \leftarrow \varnothing$.
while there is $j \in V_{-i}$ such that $j \not\perp i \mid S$ **do**
    $S \leftarrow S \cup \{j\}$. ▷ Growing phase
end while
while there is $j \in S$ such that $j \perp i \mid S \setminus \{j\}$ **do**
    $S \leftarrow S \setminus \{j\}$. ▷ Shrinking phase
end while
$\mathrm{MB}(i) \leftarrow S$.

Markov blanket

Notes:

1. After growing phase, S is a Markov blanket.
2. Line 6:
   Suppose j has been removed from S. Consider $k \notin S \cup \{j\}$.
   By (C4) contraction of CI axioms,

   $$i \perp k \mid \{S, j\} \quad \& \quad i \perp j \mid S \quad \Rightarrow \quad i \perp \{k, j\} \mid S.$$

   This means that S is still a Markov blanket of i.
3. Growing phase can be replaced by lasso or ℓ_1-regularized logistic regression.

References I

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Arthur P Dempster. Covariance selection. *Biometrics*, 28(1): 157–175, 1972.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. CRC Press, Boca Raton, FL, 2015.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.

References II

Dimitris Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems (NIPS)*, pages 505–511, 1999.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.