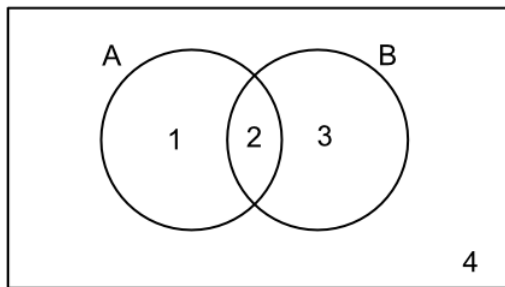


Graphical Models

March 4, 2019

- 1 Probability Review
- 2 Graph Theory
- 3 Undirected Graphical Models

Union Bound



$$P(A^c) = 1 - P(A) \quad (1)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2)$$

$$P(A \cup B) \leq P(A) + P(B) \quad (3)$$

Expectation

The expectation (mean) of a random variable X can be expressed as

$$E(X) = X_{\text{mean}} = \sum_{x \in \mathcal{X}} xP(X = x). \quad (4)$$

The variance and covariance can be defined therefore in terms of the expectation, where

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2, \quad (5)$$

$$\text{Cov}(X) = E(XY) - E(X)E(Y). \quad (6)$$

Conditional expectations and variances follow from the conditional distributions

$$E(X | Y = y) = \sum_{x \in \mathcal{X}} xP(X = x | Y = y). \quad (7)$$

Independence

Two random variables are independent when the probability distribution of one random variable does not affect the other. More concretely, two random variables X and Y are independent, that is, $X \perp Y$, if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y), \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}. \quad (8)$$

If X and Y are continuous with joint density function $f_{X,Y}(x, y)$, then the above condition reduces to finding functions $h(x)$ and $g(y)$ such that

$$f_{X,Y}(x, y) = h(x)g(y). \quad (9)$$

Conditional Independence

Two random variables X and Y are conditionally independent given a third variable Z , denoted as $X \perp Y \mid Z$, if and only if

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z), \quad (10)$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$.

This is equivalent to saying

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z).$$

Note that $X \perp Y \mid Z$ does not imply that $X \perp Y$, and vice versa.

Conditional independence relations

- Symmetry:

$$X \perp Y | Z \implies Y \perp X | Z$$

- Decomposition:

$$X \perp Y, W | Z \implies X \perp Y | Z \quad (\text{and } X \perp W | Z)$$

- Weak union:

$$X \perp Y, W | Z \implies X \perp Y | Z, W \quad (\text{and } X \perp W | Y, Z)$$

- Contraction:

$$X \perp Y | Z \text{ and } X \perp W | Y, Z \implies X \perp Y, W | Z$$

Graph Theory Preliminaries

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of:

- A set of nodes/vertices $\mathcal{V} = \{1, 2, \dots, n\}$. Illustrated by a dot or cross.
- A set of edges $\mathcal{E} = \{(u, v) \mid u, v \in \mathcal{V}\}$, which consists of node pairs. This is illustrated by a line connecting the two nodes in the node pair.
- If the pairs in \mathcal{E} are unordered, that is $(u, v) = (v, u)$, then graph is called undirected (the lines in the graph have no directional arrows). If the pairs are ordered, that is $(u, v) \neq (v, u)$, then the graph is called directed (the lines in the graph has arrows on them).
- We will be dealing with *simple* graphs here, which means we do not allow self loops and multiple edges between the same node pairs. We also will not be considering graphs with a mixture of ordered and unordered edges.

Directed and Undirected Graphs

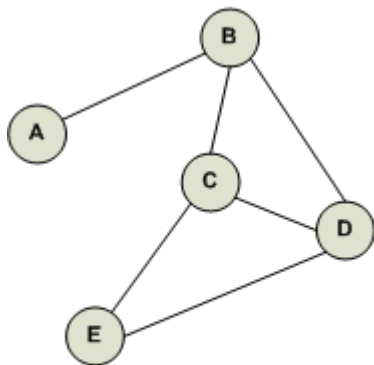


Fig 1. Undirected Graph

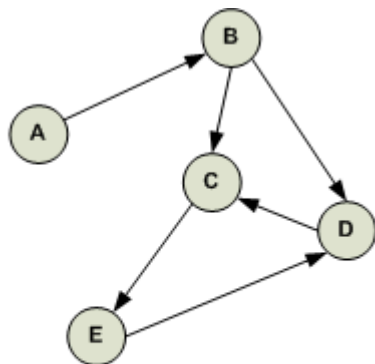
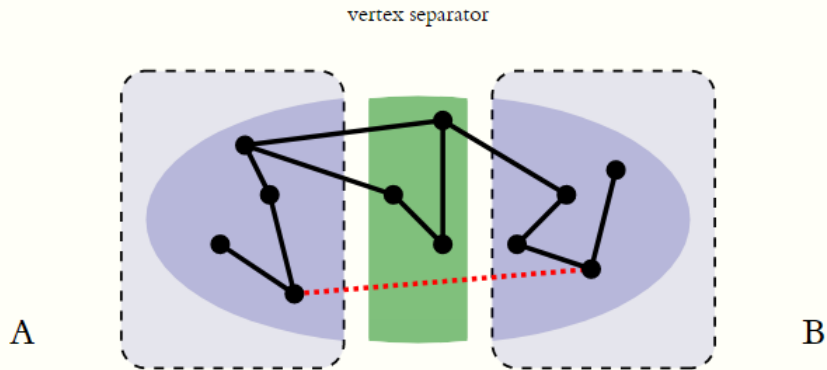
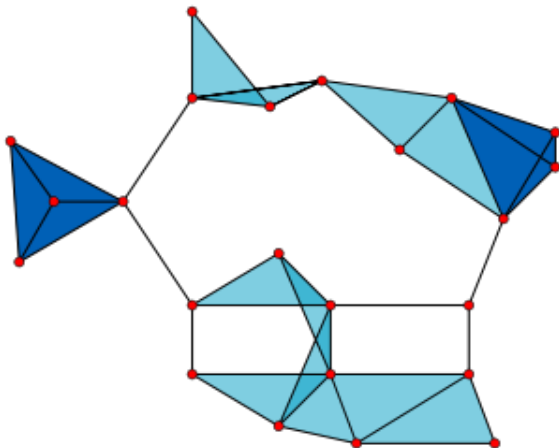


Fig 2. Directed Graph

Separator Node Sets



Graph Cliques



A graphical model is a multivariate probability distribution associated with a graph representation, where the nodes of the graph represent the variables in the distribution and the edges represent the relationships between nodes.

If the associated graph is directed (undirected), then the graphical model is directed (undirected).

In this lecture, we will be discussing undirected graphical models.

Undirected Graphical Models

Which is more important, independence or conditional independence?

One main type of undirected graphical model is known as the Markov random field (MRF).

In a MRF, if no edge exists between two nodes u and v , then X_u must be conditionally independent of X_v given the rest of the variables.

Markov Properties on Undirected Graphs

Fact: (F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P).

(F) P factorizes according to G : $p(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C)$.

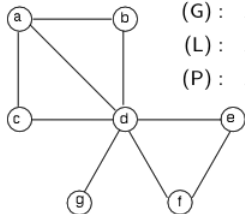
(G) For any disjoint subsets A, B, S of V , $A \perp B | S \Rightarrow X_A \perp X_B | X_S$.

(L) For all $v \in V$, $X_v \perp X_{V \setminus \text{cl}(v)} | X_{\text{bd}(v)}$.

(P) For all pairs of non-adjacent vertices (v, v') in G , $X_v \perp X_{v'} | V \setminus \{v, v'\}$.

Illustration:

G:



(F): $p(x) = \phi_1(x_a, x_b, x_d) \phi_2(x_a, x_c, x_d) \phi_3(x_d, x_e, x_f) \phi_4(x_d, x_g)$

(This is the most general form of p that satisfies (F).)

(G): $X_{\{a,b\}} \perp X_{\{e,g\}} | X_d$, $X_{\{c,g\}} \perp X_{\{b,f\}} | X_{\{a,d\}}$, etc.

(L): $X_a \perp X_{\{e,f,g\}} | X_{\{b,c,d\}}$, $X_g \perp X_{\{a,b,c,e,f\}} | X_d$, etc.

(P): $X_a \perp X_e | X_{\{b,c,d,e,f\}}$, $X_c \perp X_b | X_{\{a,d,e,f,g\}}$, etc.

Multivariate Gaussian Distribution

The Gaussian distribution $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has the density function,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad (11)$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, The edges of the associated graph \mathcal{G} are exactly those corresponding to the nonzero entries of $\boldsymbol{\Omega}$, the precision matrix.

Example

Let $\mathbf{X} = (X_1, X_2, X_3)$ be a multivariate Gaussian distribution with zero mean and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 3 \end{bmatrix}, \quad \mathbf{\Omega} = \begin{bmatrix} 2 & -3 & -1 \\ -3 & 6 & 2 \\ -1 & 2 & 1 \end{bmatrix}.$$

Observe that $X_1 \perp X_3$, or $X_1 \perp X_3 \mid \mathbf{X}_S$, where S is the null set. However, the graph according to the precision matrix is the complete graph, as the precision matrix has no off-diagonal zero entries. So while graph separability implies conditional independence, the converse does not necessarily hold.

Ising Models

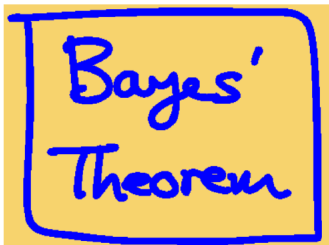
The Ising model $\mathbf{X} = (X_1, \dots, X_n)$, used in the study of ferromagnetism, has state space $\mathcal{X} = \{-1, +1\}^n$ and density function

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp \left\{ \sum_{ij} J_{ij} x_i x_j + \sum_i h_i x_i \right\}, \quad (12)$$

where the associated graph $\mathcal{G} = \mathcal{V}, \mathcal{E}$ has an edge between nodes i and j if and only if $J_{ij} \neq 0$.

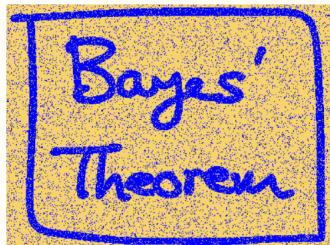
The normalizing constant has the form

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \exp \left\{ \sum_{ij} J_{ij} x_i x_j + \sum_i h_i x_i \right\}. \quad (13)$$



Bayes'
Theorem

A square frame with rounded corners, drawn in blue, containing the handwritten text "Bayes' Theorem" in blue ink on a solid yellow background.



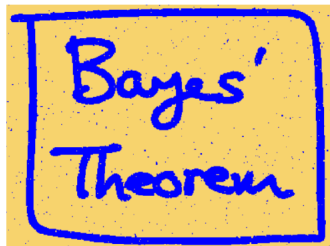
Bayes'
Theorem

A square frame with rounded corners, drawn in blue, containing the handwritten text "Bayes' Theorem" in blue ink on a yellow background with a dense, noisy pattern of small black dots.



Bayes'
Theorem

A square frame with rounded corners, drawn in blue, containing the handwritten text "Bayes' Theorem" in blue ink on a yellow background with a noisy pattern. The text is slightly more legible than in the noisy original.



Bayes'
Theorem

A square frame with rounded corners, drawn in blue, containing the handwritten text "Bayes' Theorem" in blue ink on a solid yellow background, identical to the original image.

The graph learning problem involves learning the underlying graph structure given samples from the distribution.

In the case of the Gaussian graphical model, we want to learn the non-zero structure of the precision matrix given the sample covariance matrix.

Difficult problem especially when dimension n is large.

From samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$, the sample covariance is defined by

$$S = \frac{1}{N-1} \sum_{i=1}^N \left(\mathbf{x}^{(i)} - \sum_{j=1}^N \mathbf{x}^{(j)} / N \right) \left(\mathbf{x}^{(i)} - \sum_{j=1}^N \mathbf{x}^{(j)} / N \right)^T. \quad (14)$$

The graphical lasso uses maximum likelihood and sparsity in the graph to estimate the precision matrix,

$$\hat{\Omega} = \min_{\Omega} -\log |\Omega| + \text{tr}(S\Omega) + \lambda \|\Omega\|_1. \quad (15)$$