Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

Recitation 3

1 Gaussian Graphical Models: Schur's Complement

Consider a sequence of jointly Gaussian-distributed variables:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathtt{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \underbrace{\begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}}_{\boldsymbol{\Lambda}} \right) = \mathtt{N}^{-1}\left(\begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix}, \underbrace{\begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix}}_{\mathbf{J}} \right).$$

Assume we want to do marginalization over $\mathbf{x}_2$. It's easy when we have $\mathbf{x}$ represented in covariance form. Computing marginals just involves reading off entries from $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, e.g.

$$\mathbf{x}_1 \sim \mathtt{N}(\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_{11}).$$

In contrast, computing marginals using the information form is more complicated:

$$\mathbf{x}_1 \sim \mathtt{N}^{-1}(\mathbf{h}', \mathbf{J}'),$$

where $\mathbf{J}' = \boldsymbol{\Lambda}_{11}^{-1}$.
Since we're working with information form, it would be nice if we can represent $\mathbf{J}'$ in terms of $\mathbf{J}_{11}$, $\mathbf{J}_{12}$, $\mathbf{J}_{21}$, and $\mathbf{J}_{22}$. It turns out that

$$\boldsymbol{\Lambda}_{11}^{-1} = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21} \tag{1}$$

Similarly, when we want to do conditioning, the information form representation is easy[1]. But for the canonical form, we'll need to look at $\boldsymbol{\Lambda}' = \mathbf{J}_{11}^{-1}$, and it turns out that

$$\mathbf{J}_{11}^{-1} = \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21} \tag{2}$$

The expressions (1) and (2) are called the *Schur's complement*.
In this section, we'll provide two ways to prove the validity of the Schur's Complement.

Method 1

$$\because \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

$$\therefore \boldsymbol{\Lambda}_{11}\mathbf{J}_{11} + \boldsymbol{\Lambda}_{12}\mathbf{J}_{21} = \mathbf{I} \tag{3}$$

$$\boldsymbol{\Lambda}_{21}\mathbf{J}_{11} + \boldsymbol{\Lambda}_{22}\mathbf{J}_{21} = \mathbf{0} \tag{4}$$

Left multiply $\boldsymbol{\Lambda}_{22}^{-1}$ to equation (4), we get:

$$\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\mathbf{J}_{11} + \mathbf{J}_{21} = \mathbf{0} \Rightarrow \mathbf{J}_{21} = -\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\mathbf{J}_{11} \tag{5}$$

Plug (5) into (3), we have

$$\boldsymbol{\Lambda}_{11}\mathbf{J}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\mathbf{J}_{11} = \mathbf{I}$$

$$\therefore (\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21})\mathbf{J}_{11} = \mathbf{I}$$

$$\therefore \mathbf{J}_{11}^{-1} = \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$$

Following similar arguments, we can prove $\boldsymbol{\Lambda}_{11}^{-1} = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$ as well.

[1] For details about the distribution for $\mathbf{x}_1|\mathbf{x}_2$, please refer to the notes of lecture 6.

Method 2 The second method is more systematic. It is based on two facts:
1. Diagonal matrix is easy to invert;
2. We can always use row manipulations (which corresponds to left-multiplying a matrix) and column manipulations (which corresponds to right-multiplying a matrix) to convert a full-rank matrix into a diagonal one.

$$\text{Notice } \underbrace{\begin{bmatrix} \mathbf{I} & -\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21} & \mathbf{I} \end{bmatrix}}_{\mathbf{B}} = \underbrace{\begin{bmatrix} \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{22} \end{bmatrix}}_{\mathbf{D}}$$

$$\therefore \mathbf{A\Lambda B} = \mathbf{D} \Rightarrow \mathbf{\Lambda} = \mathbf{A}^{-1}\mathbf{D}\mathbf{B}^{-1} \Rightarrow \mathbf{\Lambda}^{-1} = (\mathbf{A}^{-1}\mathbf{D}\mathbf{B}^{-1})^{-1} = \mathbf{B}\mathbf{D}^{-1}\mathbf{A}$$

Remember $\mathbf{J} = \mathbf{\Lambda}^{-1}$

$$\therefore \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{6}$$

It can be easily verified in (6) that

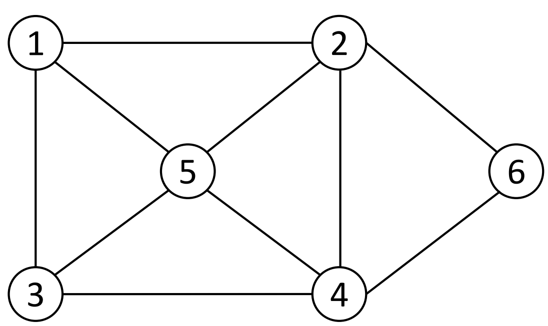
$$\mathbf{J}_{11} = (\mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21})^{-1}$$

$$\therefore \mathbf{J}_{11}^{-1} = \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}$$

Again, $\mathbf{\Lambda}_{11}^{-1} = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$ can be proved in a similar manner.

2 Conversion Among Different Graphical Models

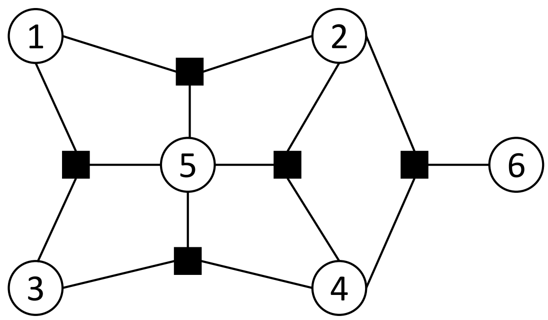
Exercise 1: Consider the following undirected graph:



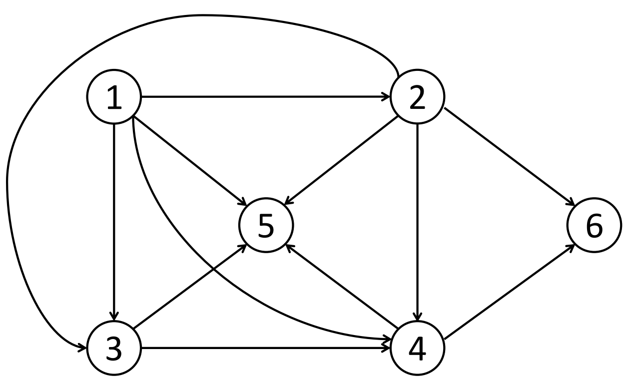
(1) Is this graph chordal?
Solution: No. Nodes 1-2-4-3-1 form a loop of size 4 that has no chord.

(2) Convert this undirected to factor graph.

Solution: When converting an undirected graph to factor graph, we just assign a factor to each maximal clique in the undirected graph. We can't just assign factors to any clique. Assigning factors to non-maximal cliques would not change conditional independences, but would imply factorization property that the undirected graph doesn't necessarily satisfy.
The resulting factor graph is as follows:

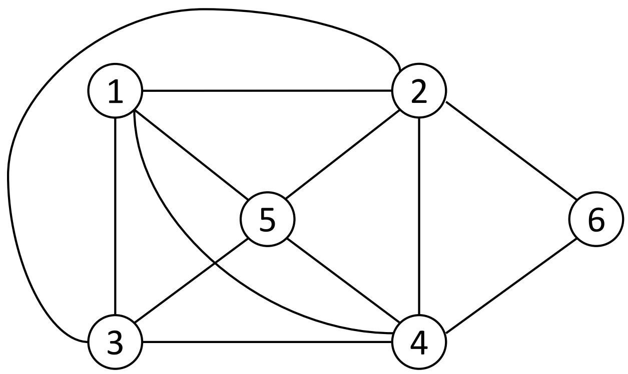


(3) Convert the factor graph to a directed acyclic graph.
Solution: When converting a factor graph to a directed one, we take an arbitrary ordering, say x_1, x_2, ..., x_n. We'll process each node in turn. For each node x_i, find a minimal set $U \subseteq \{x_1, ..., x_{i-1}\}$ s.t. $x_i \perp \{x_1, ..., x_{i-1}\} - U | U$.
The following graph is produced using ordering 1, 2, 3, 4, 5, 6. Different orderings may lead to different resulting graphs.



(4) Convert the directed graph back into undirected graph.
Solution: When converting a directed graph into an undirected one, we retain all the edges, 'marry the parents' (i.e. if the two parents in a V-structure are not directly connected, add

an edge between them), and then remove all the directions of the edges. Using the directed graph we got in part (3), we get:



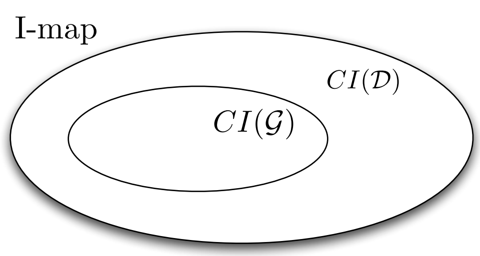
Final Comments:
1. Notice that compared to the original graph, the final undirected graph has 2 extra edges (1-4 and 2-3). We can do better by choosing a different ordering when converting to directed graph, e.g. the ordering 5, 1, 2, 3, 4, 6 will lead to a final undirected graph with only 1 extra edge (2-3).

2. For each step of the conversion, we either retain all the information, or lose some information. We can never add information. In other words, each step of the conversion finds an I-map of the previous graph.

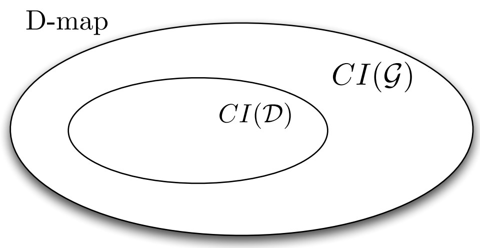
3. If the original undirected graph was chordal (e.g. if the edge 2-3 was present), we would get exactly the same undirected graph after going through these conversions.

3 I-Map, D-Map, and P-Map

Definition 1 (I-map) *We say $\mathcal{G}$ is an independence map or I-map for D if $CI(\mathcal{G}) \subseteq CI(D)$. In other words, every conditional independence implied by $\mathcal{G}$ is satisfied by D.*



Definition 2 (D-map) *We say $\mathcal{G}$ is a dependence map or D-map for D if $CI(\mathcal{G}) \supseteq CI(D)$. In other words, every conditional independence implied by D is satisfied by $\mathcal{G}$.*



Definition 3 (P-map) *We say $\mathcal{G}$ is a perfect map or P-map for D if $CI(\mathcal{G}) = CI(D)$. In other words, D and $\mathcal{G}$ have exactly the same set of conditional independences.*

$\mathcal{G}$ is a P-map of D if and only if it's both an I-map and a D-map for D.

Definition 4 (Minimal I-map) *A minimal I-map is an I-map with the property that removing any single edge would cause the graph to no longer be an I-map.*

Definition 5 (Maximal D-map) *A maximal D-map is a D-map with the property that adding any single edge would cause the graph to no longer be a D-map.*

Exercise 2: Consider a distribution D:
(1) What is a trivial example of an I-map of D?
Solution: A complete graph (directed or undirected).

(2) What is a trivial example of a D-map?
Solution: A graph with no edges.

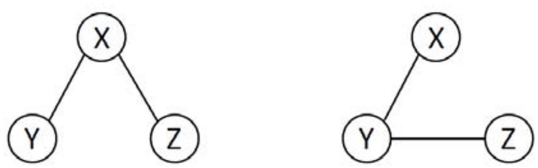
(3) Suppose we want to find the directed minimal I-map of D. Is this minimal I-map unique?
Solution: In general, the directed minimal I-map is not unique. Here's an easy counter example: if D is a distribution over variables X and Y and D has no conditional independences, then the following directed graphs are both minimal I-maps for D.

(4) Suppose we want to find the undirected minimal I-map of D, and D is a positive distribution. Is this minimal I-map unique?
Solution: Yes. We'll prove by contradiction.
Assume we have two different undirected minimal I-maps for D, denoted as $\mathcal{G}_1$ and $\mathcal{G}_2$. Since they are different, there exists at least one edge that is in one graph but not the other. Without loss of generality, assume edge $(i, j) \notin \mathcal{G}_1$ and $(i, j) \in \mathcal{G}_2$.

Since $(i, j) \notin \mathcal{G}_1$, and $\mathcal{G}_1$ is an I-map for D, we know that $x_i \perp x_j | x_{\text{rest}}$ is satisfied by D. Consider graph $\mathcal{G}'_2$, which is constructed by removing edge (i, j) from $\mathcal{G}_2$. Distribution D satisfies all the pairwise independences in $\mathcal{G}_2'$ by construction, i.e. D is pairwise Markov with respect to $\mathcal{G}_2'$. Since D is a positive distribution, pairwise Markov property is equivalent to global Markov property[2]. Thus $\mathcal{G}_2'$ is also an I-map for D. But this contradicts the fact that $\mathcal{G}_2$ is a minimal I-map!

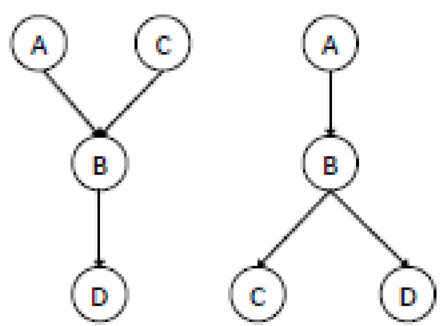
Thus the assumption is not true. In other words, the minimal I-map is unique.

(5) Again, we want to find the undirected minimal I-map of D. But D is now a general distribution (not necessarily positive). Is the minimal I-map still unique?
Solution: No. Consider the following counter example:
Let $P_{X,Y,Z}(0,0,0) = P_{X,Y,Z}(1,1,1) = \frac{1}{2}$ and all other configurations have zero probability. Then the following graph on the left is a minimal I-map for the distribution. It's an I-map because the only conditional independence it has is $Y \perp Z|X$, which is satisfied by the distribution. But if we take out the edge between X and Y, the graph implies $Y \perp (X,Z)$, which is clearly not true in the distribution. Similarly, we can't take out the edge between X and Z. Thus it is a minimal I-map. However, the graph on the right is also a minimal I-map following similar arguments. Thus for a general distribution D, its undirected minimal I-map is not unique.



Given two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, if $CI(\mathcal{G}_1) = CI(\mathcal{G}_2)$, we say $\mathcal{G}_1$ and $\mathcal{G}_2$ are 'I-equivalent'.

(6) For each of the two following graphs, determine whether there can be any other directed graph that is 'I-equivalent' to it[3].



[2] We haven't proved this fact in class. One direction of the statement (global implies pairwise) is obvious. One way to see why the other direction is true is that when we prove Hammersley-Clifford theorem, we only used pairwise independences. A more rigorous proof using induction method can be found in Koller/Friedman.

[3] Koller/Friedman Exercise 3.15

Solution: The graph on the left doesn't have an 'I-equivalent' directed graph because changing the direction of any edge would either destroy or create a V-structure where the parents are not connected. In other words, changing any edge direction would change the set of conditional independences in the graph.

The graph on the right has an 'I-equivalent' directed graph: we can reverse the arrow between A and B and the resulting directed graph has exactly the same set of conditional independences.

(7) For an undirected graph, is there any other undirected graph that is 'I-equivalent' to it?
Solution: No. In an undirected graph, the absence of edge (i,j) implies that $x_i \perp x_j | x_{\text{rest}}$. Any different undirected graphs differs by at least one edge, and the pairwise conditional independence corresponds to that edge is satisfied by one graph and not the other. Thus the two graphs cannot have exactly the same set of conditional independences.

This result shows that every undirected graph implies a different set of conditional independences, i.e. in some sense undirected graphs are not redundant, whereas many directed graphs can correspond to the same set of conditional independences.

(8) If two directed graphs have the same set of variables, the same skeleton[4] and the same set of V-structures, are they guaranteed to be 'I-equivalent'?
Solution: Yes. If you consider running Bayes Ball Algorithm on both graphs, they will obey exactly the same rules.

(9) If two directed graphs are 'I-equivalent', are they guaranteed to have the same skeleton and the same set of V-structures?
Solution: No. A simple counter example is provided by the following graphs. Both directed graphs contain no conditional independence, but clearly they have different V-structure.

[4] i.e. the underlying edges without the direction associated to them