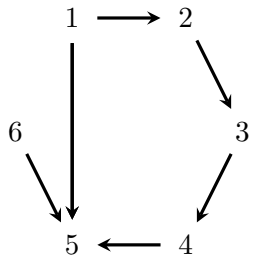
Graphical Models: Worksheet 3 MT 2024

A: Warm Up

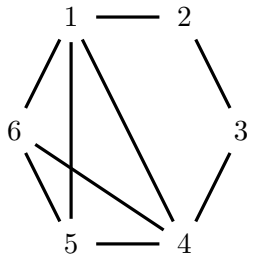
A1. Directed Acyclic Graphs



(a) List all the conditional independences implied by applying the local Markov property to the DAG $\mathcal{G}$ shown above.
These are

$$X_1 \perp\!\!\!\perp X_6, \qquad X_2 \perp\!\!\!\perp X_6 \mid X_1$$
$$X_3 \perp\!\!\!\perp X_1, X_6 \mid X_2 \qquad X_4 \perp\!\!\!\perp X_1, X_2, X_6 \mid X_3$$
$$X_5 \perp\!\!\!\perp X_2, X_3 \mid X_1, X_4, X_6 \qquad X_6 \perp\!\!\!\perp X_1, X_2 X_3, X_4.$$

(b) Find the moral graph. Is it decomposable?
The moral graph is:



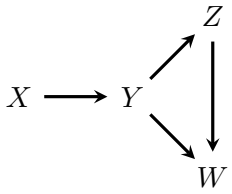
This graph is not decomposable, since there is a chordless cycle $1, 2, 3, 4$.

(c) Find all the sets of vertices $C \subseteq \{2, 3, 5, 6\}$ such that $X_1 \perp\!\!\!\perp X_4 \mid X_C$ according to the global Markov property.
The ancestors of 1 and 4 include 2 and 3, so any subset C must include one of these. If we include 5 then they will not be separated because of the v-structure $1 \to 5 \leftarrow 4$. *On the other hand, we can always include 6 without problems.*

$$X_1 \perp\!\!\!\perp X_4 \mid X_2, \qquad X_1 \perp\!\!\!\perp X_4 \mid X_3, \qquad X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$$
$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_6, \qquad X_1 \perp\!\!\!\perp X_4 \mid X_3, X_6, \qquad X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3, X_6.$$

A2. Markov Equivalence

List all the graphs (either undirected or directed acyclic) that are Markov equivalent to the one shown.

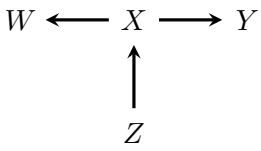


There are no v-structures, so the undirected graph with the same skeleton is Markov equivalent. For directed graphs we just need any graph with the same skeleton and no v-structures. If $X \to Y$ then we need $Y \to Z$ and $Y \to W$ to avoid a v-structure, so there are two possibilities depending on the orientation of the remaining edge. If $Y \to X$ then we can have the other three edges in any orientation that accounts for a topological ordering of Y, W, Z: there are six such orderings.

In total there are 8 directed graphs (including the one shown) and one undirected.

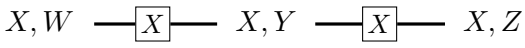
A3. Junction Trees

Consider the graphical model shown.



(a) Draw a junction tree suitable for performing probability inference on a distribution that is Markov with respect to this graph.

The graph has no v-structures, so is Markov equivalent with respect to the (decomposable) undirected graph with the same skeleton. One possible junction tree for the cliques is



Though any of the nodes can be in the centre.

(b) Suppose that the distribution is given by

$$p(z) = \begin{array}{c|cc} z = 0 & 1 \\ \hline 0.4 & 0.6 \end{array} \qquad p(x \mid z) = \begin{array}{c|cc} z & x = 0 & 1 \\ \hline 0 & 0.9 & 0.1 \\ 1 & 0.4 & 0.6 \end{array}$$

$$p(w \mid x) = \begin{array}{c|cc} x & w = 0 & 1 \\ \hline 0 & 0.1 & 0.9 \\ 1 & 0.2 & 0.8 \end{array} \qquad p(y \mid x) = \begin{array}{c|cc} x & y = 0 & 1 \\ \hline 0 & 0.7 & 0.3 \\ 1 & 0.4 & 0.6 \end{array}$$

Give an initialization of potentials in your junction tree consistent with this joint distribution. *[Hint: you shouldn't need to do any calculations.]*

The obvious way to do this is to set $\psi_{XW} = p(w \mid x)$, $\psi_{XY} = p(y \mid x)$ and $\psi_{XZ} = p(z) \cdot p(x \mid z)$; the two separators can just be $\psi_X = \tilde{\psi}_X = 1$.

(c) Using the junction tree algorithm, calculate the consistent potentials for this junction tree.

Choosing XW as the root node, we will collect the evidence first. We have

$$\psi_{XZ} = p(x,z) = \begin{array}{c|cc} z\backslash x & 0 & 1 \\ \hline 0 & 0.36 & 0.04 \\ 1 & 0.24 & 0.36 \end{array}$$

so the marginal distribution of X is $(0.6, 0.4)$. This becomes the value of the potential $\tilde{\psi}_X$, replacing 1, so then

$$\psi'_{XY} = \frac{\tilde{\psi}'_X}{1}\psi_{XY} = p(x)p(y\,|\,x) = \begin{array}{c|cc} x & y=0 & 1 \\ \hline 0 & 0.42 & 0.18 \\ 1 & 0.16 & 0.24 \end{array}$$

Repeating this with a propagation to the final table gives

$$\psi'_{XW} = \frac{\psi'_X}{1}\psi_{XW} = p(x)p(w\,|\,x) = \begin{array}{c|cc} x & w=0 & 1 \\ \hline 0 & 0.06 & 0.54 \\ 1 & 0.08 & 0.32 \end{array}$$

One can verify that all the potentials are now consistent, and therefore the distribute step will not change anything.

(d) Use the junction tree to compute $p(w\,|\,y=1)$.

Introducing the evidence $\{Y=1\}$ into the relevant clique changes that table to

$$\psi_{XY} = \begin{array}{c|cc} x & y=0 & 1 \\ \hline 0 & 0 & 0.429 \\ 1 & 0 & 0.571 \end{array}$$

so the marginal of X is $(0.429, 0.571)$ (as opposed to $0.6, 0.4$ before). Propagating to the XW table gives

$$\psi'_{XW} = \psi_{XW}\frac{\psi'_X}{\psi_X} = \begin{array}{c|cc} x & w=0 & 1 \\ \hline 0 & 0.043 & 0.386 \\ 1 & 0.114 & 0.457 \end{array}$$

giving a marginal distribution of $(0.157, 0.843)$ for W. Hence $p(w=1\,|\,y=1) = 0.843$ (note it was $p(w=1) = 0.86$ before this evidence was introduced, so there is no dramatic change!)

B: Core Questions

B1. Markov Blanket

Let $\mathcal{G}$ be a DAG. The *Markov blanket* of a vertex v is

$$\mathrm{mb}_{\mathcal{G}}(v) \equiv \mathrm{ch}_{\mathcal{G}}(v) \cup \mathrm{pa}_{\mathcal{G}}(\{v\} \cup \mathrm{ch}_{\mathcal{G}}(v)) \setminus \{v\}.$$

(i.e., the parents of v, children of v, and the other parents of children of v, but not v itself).

(a) Show that, in the moral graph $\mathcal{G}^m$, the boundary of v is precisely $\mathrm{bd}_{\mathcal{G}^m}(v) = \mathrm{mb}_{\mathcal{G}}(v)$.

The neighbours of v in the original graph are its parents and children. The only edges that are added in $\mathcal{G}^m$ are between the parents of a common child, so k (say) will become a neighbour of v if and only if $v \to i \leftarrow k$ for some i. Hence, the neighbours are precisely the parents, children, and other parents of children.

(b) Deduce that if p is Markov with respect to $\mathcal{G}$ then

$$X_v \perp\!\!\!\perp X_{V\setminus(\mathrm{mb}(v)\cup\{v\})} \mid X_{\mathrm{mb}(v)} \,[p] \qquad\qquad v \in V. \tag{1}$$

We know that if p is Markov with respect to $\mathcal{G}$ then it is also Markov with respect to $\mathcal{G}^m$. Hence, it satisfies the local Markov property for $\mathcal{G}^m$, which means $X_v \perp\!\!\!\perp X_{V\setminus(\mathrm{bd}(v)\cup\{v\})} \mid X_{\mathrm{bd}(v)}\,[p]$. But we have shown that $\mathrm{bd}_{\mathcal{G}^m}(v) = \mathrm{mb}_{\mathcal{G}}(v)$, so this gives the result.

(c) Suppose p satisfies (1). Does this imply that p is Markov with respect to $\mathcal{G}$? Justify your answer.

No, as can be seen by considering the graph $a \to c \leftarrow b$, or indeed any graph with a v-structure.

B2. Structural Equation Models

Let $\mathcal{G}$ be a DAG and let X_V be a multivariate normal vector with zero mean and positive definite covariance matrix Σ.

(a) Let v be a vertex that has no children in $\mathcal{G}$, and denote $W = V \setminus \{v\}$. Show that

$$X_v \mid X_W = x_W \sim N(b_{vW}x_W,\ \Sigma_{vv\cdot W}).$$

where $\Sigma_{vv\cdot W} = \Sigma_{vv} - \Sigma_{vW}(\Sigma_{WW})^{-1}\Sigma_{Wv}$ is the Schur complement (see Worksheet 0, question 5) and $b_{vW} = (b_{vw})_{w\in W}$ is a vector which you should find.

The log density function of X_V (ignoring constant terms) is

$$\begin{aligned}&\log f(x_V)\\ &= -\frac{1}{2}\begin{pmatrix}x_v\\ x_W\end{pmatrix}^T\begin{pmatrix}\Sigma_{vv} & \Sigma_{vW}\\ \Sigma_{Wv} & \Sigma_{WW}\end{pmatrix}^{-1}\begin{pmatrix}x_v\\ x_W\end{pmatrix} + \textit{const.}\end{aligned}$$

where (using Sheet 0 q5)

$$\begin{pmatrix}\Sigma_{vv} & \Sigma_{vW}\\ \Sigma_{Wv} & \Sigma_{WW}\end{pmatrix}^{-1} = \begin{pmatrix}\Sigma_{vv\cdot W}^{-1} & -\Sigma_{vv\cdot W}^{-1}\Sigma_{vW}(\Sigma_{WW})^{-1}\\ -\Sigma_{vv\cdot W}^{-1}\Sigma_{vW}(\Sigma_{WW})^{-1} & \Sigma_{WW\cdot v}^{-1}\end{pmatrix}.$$

Expanding this out and ignoring terms not depending on x_v we get

$$= -\frac{1}{2}\Sigma_{vv\cdot W}^{-1}x_v^2 + x_v\Sigma_{vv\cdot W}^{-1}\Sigma_{vW}(\Sigma_{WW})^{-1}x_W + \textit{const.}$$

so completing the square:

$$= \frac{1}{2}\Sigma_{vv\cdot W}^{-1}(x_v - \Sigma_{vW}(\Sigma_{WW})^{-1}x_W)^2 + \textit{const.};$$

hence the result holds with $b_{vW} = \Sigma_{vW}(\Sigma_{WW})^{-1}$.

(b) Show that Σ is Markov with respect to $\mathcal{G}$ if and only if both (i) Σ_{WW} is Markov with respect to $\mathcal{G}_W$ and (ii) $b_{vw} = 0$ for each $w \notin \mathrm{pa}_{\mathcal{G}}(v)$.

Using the local Markov property from lectures under an ordering in which v comes last, we have that Σ is Markov with respect to $\mathcal{G}$ if and only if X_W is Markov with respect to $\mathcal{G}_W$ and $X_v \perp\!\!\!\perp X_{W\setminus \mathrm{pa}(v)} \mid X_{\mathrm{pa}(v)}$. The distribution of X_W is Gaussian with covariance Σ_{WW} which gives (i). The latter condition is clearly equivalent to saying that the entries in the vector b_{vW} corresponding to non-parents are zero, since otherwise the conditional distribution will depend upon their value; this gives (ii).

(c) Deduce that Σ is Markov with respect to $\mathcal{G}$ if and only if we can write

$$X_v = \sum_{w \in \mathrm{pa}_{\mathcal{G}}(v)} b_{vw} X_w + \varepsilon_v,$$

for all $v \in V$, where $\varepsilon_V = (\varepsilon_v)_{v \in V}$ is a Gaussian random vector with independent components. (Here an empty sum is zero by convention.)

We have shown that the expression above holds for the final vertex in a topological ordering, and it follows from the form of the conditional distribution above that $\varepsilon_v \perp\!\!\!\perp X_W$. Hence, invoking an inductive argument, we have the equation for all v, and since ε_W is a function of X_W, we also have the necessary independences.

(d) By writing the previous result in matrix form, show that

$$\Sigma = (I - B)^{-1} D (I - B)^{-T},$$

where I is the identity matrix, D is diagonal, and B is a lower triangular matrix with (i, j)th entry b_{ij}.

We have

$$X_V = B X_V + \boldsymbol{\varepsilon},$$

where B is the specified lower triangular matrix, and $\boldsymbol{\varepsilon}$ has diagonal covariance matrix, say D. It follows that

$$X_V = (I - B)^{-1} \boldsymbol{\varepsilon}$$

(the invertibility of $I - B$ follows from its form), and hence $\Sigma = \mathrm{Cov}\, X_V = (I - B)^{-1} (\mathrm{Cov}\, \varepsilon)(I - B)^{-T}$, giving the result.

(e) Let $K = \Sigma^{-1}$ be the concentration matrix for X_V. Show that if $i \neq j$ then

$$k_{ij} = \sum_{\ell \in C_{ij}} d_{\ell\ell}^{-1} b_{\ell i} b_{\ell j} - d_{jj}^{-1} b_{ji} - d_{ii}^{-1} b_{ij}$$

where $C_{ij} = \mathrm{ch}_{\mathcal{G}}(i) \cap \mathrm{ch}_{\mathcal{G}}(j)$. Deduce a graphical condition (i.e. a condition on $\mathcal{G}$) that will ensure $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$.

We have $K = (I - B)^T D^{-1} (I - B)$, so this is easily reduced to

$$k_{ij} = \sum_{\ell} (I - B)_{\ell i} (I - B)_{\ell j} d_{\ell\ell}^{-1}.$$

If $\ell \neq i, j$ the only terms are $B_{\ell i} B_{\ell j} d_{\ell\ell}^{-1}$ which is only non-zero if ℓ is a child of i and j; this gives the first term. If $i = \ell$ we obtain $-B_{ij} d_{ii}^{-1} = -d_{ii}^{-1} b_{ij}$, and similarly for $\ell = j$.

We deduce that $k_{ij} = 0$ if i and j are not adjacent (so that $b_{ij} = b_{ji} = 0$) and do not share any common children (so then $b_{\ell i} b_{\ell j} = 0$).

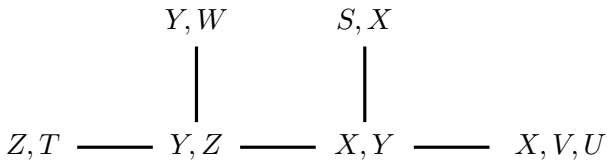
B3. Evidence Propagation

Let $\mathcal{T}$ be a junction tree with cliques $C_1, \ldots, C_k$, and suppose that all potentials are consistent.

(a) Let $e \in C_i$ and $f \in C_j$. Explain why the calculation of $p(x_f \,|\, \{X_e = y_e\})$ only requires messages to be passed from ψ_{C_i} along the (unique) path in $\mathcal{T}$ to ψ_{C_j}.

We can imagine a sub-junction tree that only consists of these cliques: then these sets will also be consistent, and after introducing evidence we only need to make these sets consistent to know that each of our potentials contains the relevant marginal. Further (as covered in lectures), we only need to propagate in one direction, because each clique is already consistent with the separator set 'away' from ψ_{C_i}, and will remain so after the update is passed through it.

Suppose we have random variables S, T, U, V, W, X, Y, Z all taking values in $\{0, 1\}$, arranged in the junction tree below. Initially, the potentials are all consistent.



(b) How would you calculate $p(z = 0 \,|\, s = 1)$ in the most efficient way possible using the tree?

We can ignore all but the nodes S, X, X, Y and Y, Z, since they are sufficient to answer the query. Then replacing $\psi_{SX} = p(x, s)$ with $p(x \,|\, s = 1)$ and passing messages from S, X to X, Y and from X, Y to Y, Z will leave $\psi_{YZ}(y, z) = p(y, z \,|\, s = 1)$. The solution can be computed just by summing over y.

(c) How many additions and multiplications do you need to perform in order to calculate $p(z = 0 \,|\, s = 1)$ using (i) the method above; (ii) from the joint distribution directly?

Calculating $p(x \,|\, s = 1)$ from $p(s, x)$ requires one addition for $p(s = 1) = p(x = 0, s = 1) + p(x = 1, s = 1)$, and then two multiplications to get the conditionals.

Each message pass from ψ_C to ψ_D involves $2^{|C \setminus S|}$ additions to compute the new separator ψ_S, and $2^{|S|}$ multiplications to compute the ratio ψ'_S/ψ_S. Then we need $2^{|D|}$ multiplications to compute ψ'_D. In our case, this amounts to 2 additions and 6 multiplications per message.

Finally, we need one addition to compute $p(z = 0 \,|\, s = 1) = p(x = 0, z = 0 \,|\, s = 1) + p(x = 1, z = 0 \,|\, s = 1)$. This gives a total of 6 additions and 14 multiplications.

The naïve method certainly involves computing $p(s = 1)$ which means at least $2^7 - 1 = 127$ additions (note $p(z = 0, s = 1)$ can be computed as part of this calculation). Then one multiplication is needed to get the final answer.

Some variation is possible depending on the exact approach taken.

(d) How would you calculate $p(t = 0 \,|\, s = 1, y = 1)$?

Again we can ignore all but the nodes on the path from S, X to Z, T. Proceed as in (a), until we have reached the point where $\psi_{YZ}(y, z) = p(y, z \,|\, s = 1)$. Now we can replace this with $\psi_{YZ}(y, z) = p(z \,|\, s = 1, y = 1)$, and pass a final message

from Y, Z to Z, T to give the solution. since they are sufficient to answer the query. Then replacing $\psi_{SX} = p(x, s)$ with $p(x \mid s = 1)$ and passing messages from S, X to X, Y and from X, Y to Y, Z will give the answer.

*The important point here is that we **must** update in a clique that already has the information $S = 1$, otherwise the distribution we divide by will be $p(y = 1)$ rather than $p(y = 1 \mid s = 1)$.*

C: Optional

C1. Junction Tree Efficiency

Let $X_1, \ldots, X_k$ be binary random variables arranged in a junction tree with maximum clique size c and diameter d (the diameter is the length of the longest path in the tree). What is the maximum complexity (in terms of the number of additions, multiplications, divisions) required to calculate $p(x_i \mid x_j = 0)$? What about $p(x_i \mid x_j = 0, x_k = 0)$?

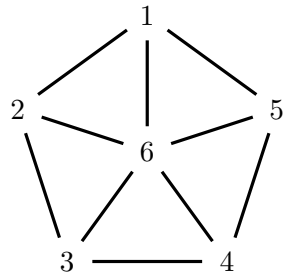
C2. Triangulation

The *Tarjan elimination algorithm* is a method for taking an undirected graph $\mathcal{G}$ and an ordering of the vertices of $\mathcal{G}$, and returning a triangulated graph $\mathcal{G}' \supseteq \mathcal{G}$.

1. Pick the largest element v of V under the ordering;
2. Join together all neighbours of v, and remove v from $\mathcal{G}$;
3. Repeat 1–2 until all vertices have been eliminated;
4. Construct a new graph which contains all the additional edges.

The ordering of the vertices used above is called an *elimination order*. An elimination order is said to be *perfect* if $\mathcal{G}' = \mathcal{G}$.

(a) Apply the algorithm to the graph below using the elimination orderings (i) $1, 2, 3, 4, 5, 6$; (ii) $6, 1, 2, 3, 4, 5$. (NB: the *last* vertex is eliminated first.) What are the resulting cliques?



(i) Eliminating 6 first immediately results in a complete graph. (ii) Eliminating 5 adds an edge between 1 and 4, and then eliminating 4 adds an edge between 1 and 3. The resulting graph is triangulated and has cliques $\{1, 4, 5, 6\}, \{1, 3, 4, 6\}, \{1, 2, 3, 6\}$, so this is a significantly smaller graph.

(b) Show that the graph returned by the Tarjan Elimination algorithm is triangulated.

We proceed by induction on the number of vertices p. All graphs of size $p \leq 3$ are triangulated, so the result holds. Otherwise if v is the largest vertex, the algorithm constructs a graph that has a decomposition $(v, \mathrm{bd}_{\mathcal{G}}(v), V \setminus (\{v\} \cup \mathrm{bd}_{\mathcal{G}}(v)))$ and such that $\mathrm{bd}_{\mathcal{G}}(v)$ is complete. By the induction hypothesis, the graph $\mathcal{G}'_{V\setminus\{v\}}$ is decomposable, so therefore $\mathcal{G}'$ is also decomposable (hence triangulated).

(c) Show that there exists a perfect elimination order if and only if $\mathcal{G}$ is triangulated.
By definition, if there is a perfect elimination order the resulting graph is the same as the original one, and therefore both are triangulated by the previous part. For the converse, we use induction: if the graph is decomposable then it has cliques $C_1, \ldots, C_k$ satsifying the RIP. Then let $v \in C_k \setminus S_k$, and note that its boundary is $C_k \setminus \{v\}$ and that $(v_k, C_k \setminus \{v\}, V \setminus C_k)$ is a proper decomposition. Hence, by the induction hypothesis there is a perfect elimination ordering on $\mathcal{G}_{V\setminus\{v\}}$, and so adding v to the end of it gives one for $\mathcal{G}$.