1. (a) Given disjoint subsets A, B, S of V, we say that A is *separated* from B by S if every path from any $a \in A$ to any $b \in B$ passes through a vertex in S.

Now, we say that p satisfies the global Markov property with respect to $\mathcal{G}$ if whenever A and B are separated by S in $\mathcal{G}$ we have $X_A \perp\!\!\!\perp X_B \mid X_S$ in p.

(b) A decomposition (A, S, B) is a triple of disjoint sets such that (i) $V = A \cup B \cup S$; (ii) A and B are separated by S; and (iii) S is a complete set in $\mathcal{G}$ (i.e. every pair of vertices in S are joined by an edge). The decomposition is proper if A and B are non-empty.

(c) Let π be a path from $v \in A \cup S$ to $w \in A \cup S$. If there are no vertices in B on the path then we are done; otherwise, there must be at least two vertices in S on the path, since the path needs to traverse into B and out again, and no vertices in A are adjacent to any in B by the existence of the decomposition. Take the first and last elements of S on the path, say s, t; these are adjacent since S is complete, so take the subpath given by joining them directly together (i.e. missing out any vertices in between). Then this subpath contains no vertices in B as required.

(d) Suppose that C is separated from D by E in $\mathcal{G}_{A\cup S}$; we claim that this separation is also present in $\mathcal{G}$. To see this, we work with the contrapositive: suppose that C is not separated from D by E in $\mathcal{G}$, so that by definition there is a path in $\mathcal{G}$ from $c \in C$ to $d \in D$ that does not pass through any element of E. Then by the previous part there is also a subpath in $\mathcal{G}_{A\cup S}$ that does not pass through any element of E.

(e) A graph is decomposable if either (i) it is complete (all pairs of vertices are joined by edges) or (ii) there is a proper decomposition (A, S, B), and each of the subgraphs $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are themselves decomposable.

(f) A graph is decomposable if and only if its cliques $C_1, \ldots, C_k$ can be ordered so as to satisfy the running intersection property. We proceed by induction on the number of cliques: if $k = 1$ then the graph is complete and the result is trivially true. Otherwise, there is a decomposition $(H, S_k, C_k \setminus S_k)$, where $H = \left(\bigcup_{i<k} C_i\right) \setminus S_k$. Then by the previous part,

$$p(x_V) = \frac{p(x_{C_k})p(x_{H\cup S_k})}{p(x_{S_k})}$$

where, by (d), $p(x_{H\cup S_k})$ obeys the global Markov property with respect to $\mathcal{G}_{H\cup S_k}$. But it is easy to see that this graph has cliques $C_1, \ldots, C_{k-1}$, so by the induction hypothesis the result follows.

2. (a) The *parents* of $v \in V$, denoted $\text{pa}(v)$, are those vertices w such that $w \to v$. We say that p factorizes with respect to $\mathcal{G}$ if

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\text{pa}(v)}), \qquad \forall x_V$$

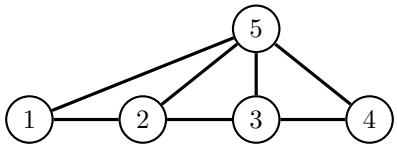
where $\text{pa}(v)$ is the set of parents of v in $\mathcal{G}$.

(b) The factorization is

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1) \cdot p(x_5) \cdot p(x_2 \mid x_1, x_5) \cdot p(x_3 \mid x_2) \cdot p(x_4 \mid x_3, x_5).$$

(i) One could note, for example, that the margin over x_1, x_5 above is just $p(x_1) \cdot p(x_5)$, so the independence does indeed hold.

(ii) To check this independence using the global Markov property we would first consider the ancestral subgraph over ancestors of 1, 3 and 4 (which is just $\mathcal{G}$), and then look at its moral graph $\mathcal{G}^m$:



Now we see that 1 is not separated from 4 by 3, since there is a path $1 - 5 - 4$, and hence the independence is not implied by the global Markov property. By the completeness of the global Markov property, this independence does not generally hold.
[We could also use d-separation to see that the path $1 \to 2 \leftarrow 5 \to 4$ is open given 3, because 2 is an ancestor of 3.]

(iii) Using, for example, the local Markov property, we see that X_3 is independent of the non-descendants X_1, X_5 given its parents X_2, which is precisely this result. It can also be seen directly from the factorization

(c) This interventional distribution is defined as

$$\begin{aligned} p(x_2, x_4, x_5 \mid do(x_1, x_3)) &= \prod_{v \in \{2,4,5\}} p(x_v \mid x_{\text{pa}(v)}) \\ &= p(x_5) \cdot p(x_2 \mid x_1, x_5) \cdot p(x_4 \mid x_3, x_5). \end{aligned}$$

From the factorization in (b) we see that this is the same as dividing the joint distribution by $p(x_1) \cdot p(x_3 \mid x_2)$.

(d) We have

$$\begin{aligned} p(x_2, x_4 \mid do(x_1, x_3)) &= \sum_{x_5} p(x_2, x_4, x_5 \mid do(x_1, x_3)) \\ &= \sum_{x_5} \frac{p(x_1, x_2, x_3, x_4, x_5)}{p(x_1) \cdot p(x_3 \mid x_2)} \\ &= \frac{\sum_{x_5} p(x_1, x_2, x_3, x_4, x_5)}{p(x_1) \cdot p(x_3 \mid x_2)} \\ &= \frac{p(x_1, x_2, x_3, x_4)}{p(x_1) \cdot p(x_3 \mid x_2)} \end{aligned}$$

Now, since $X_3 \perp\!\!\!\perp X_1 \mid X_2$ (see, for example, (b)(iii)), then we can rewrite this as

$$\begin{aligned} p(x_2, x_4 \mid do(x_1, x_3)) &= \frac{p(x_1, x_2, x_3, x_4)}{p(x_1) \cdot p(x_3 \mid x_1, x_2)} \\ &= \frac{p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \cdot p(x_4 \mid x_1, x_2, x_3)}{p(x_1) \cdot p(x_3 \mid x_1, x_2)} \\ &= p(x_2 \mid x_1) \cdot p(x_4 \mid x_1, x_2, x_3), \end{aligned}$$

and summing over x_2 gives the result.
Using the original expression for the causal distribution above,

$$\begin{aligned} p(x_4 \mid do(x_1, x_3)) &= \sum_{x_2, x_5} p(x_2, x_4, x_5 \mid do(x_1, x_3)) \\ &= \sum_{x_2, x_5} p(x_5) \cdot p(x_2 \mid x_1, x_5) \cdot p(x_4 \mid x_3, x_5) \\ &= \sum_{x_5} p(x_5) \cdot p(x_4 \mid x_3, x_5) \cdot \sum_{x_2} p(x_2 \mid x_1, x_5) \\ &= \sum_{x_5} p(x_5) \cdot p(x_4 \mid x_3, x_5) \end{aligned}$$

which does not depend upon x_1. Hence

$$p(x_4 \mid do(x_1, x_3)) = \sum_{x_2} p(x_2 \mid x_1) \cdot p(x_4 \mid x_1, x_2, x_3)$$

does not depend upon x_1.

3. (a) The concentration matrix is $K = \Sigma^{-1}$. The density of a multivariate Gaussian with mean 0 and concentration matrix K is

$$f(\boldsymbol{x}; K) = \frac{|K|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T K \boldsymbol{x}\right\}, \qquad \boldsymbol{x} \in \mathbb{R}^p.$$

(b) The conditional distribution is just obtained by dropping terms not containing x_1, so (looking at the log-density for convenience) we get

$$\begin{aligned}\log f(x_1; x_2, \ldots, x_k, K) &= -\frac{1}{2}k_{11}x_1^2 - \sum_{i=2}^{k} x_1 x_i k_{1i} + \text{const} \\ &= -\frac{1}{2}k_{11}\left(x_1 + \sum_{i=2}^{k} x_i k_{1i}/k_{11}\right)^2 + \text{const},\end{aligned}$$

which is the log-density of a normal distribution with mean $\mu = -\sum_{i=2}^{k} x_i k_{1i}/k_{11}$ and variance k_{11}^{-1}.

(c) The conditional distribution of X_1 from does not depend on X_j if and only if $k_{1j} = 0$, so therefore $X_1 \perp\!\!\!\perp X_j \mid X_{V\setminus\{1,j\}}$ if and only if $k_{1j} = 0$. Since there is nothing special about 1, this clearly also holds if we replace it by $i \neq j$.

(d) Gibbs sampling proceeds by sequentially sampling univariate distributions from a vector, conditional on all the other components of the vector being fixed.

For example,

- for each $i = 1, \ldots, k$, sample $X_i^{(t)}$ from $p(x_i \mid x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_k^{(t-1)})$

In our case, we could start with, for example, each $x_i^{(0)} = 0$, and then simulate

- for each $i = 1, \ldots, k$, sample $x_i^{(t)}$ from $N(\mu_i^{(t)}, k_{ii}^{-1})$ where

$$\mu_i^{(t)} = \frac{\sum_{j=1}^{i-1} k_{ij} x_j^{(t)} + \sum_{j=i+1}^{k} k_{ij} x_j^{(t-1)}}{k_{ii}}$$

(e) One advantage is that it does not require us to invert the matrix K or perform any other sort of numerical decomposition. One disadvantage is that we only have a sequence that converges in distribution to the correct stationary distribution, so sampling is not from exactly the target.

4. (a) ψ_C and ψ_D are consistent if $\sum_{x_{C\setminus D}} \psi_C(x_C) = \sum_{x_{D\setminus C}} \psi_D(x_D)$, i.e. they agree on their common margin S. Passing a message from ψ_C to ψ_D with separator potential ψ_S involves replacing ψ_S and ψ_D respectively by:

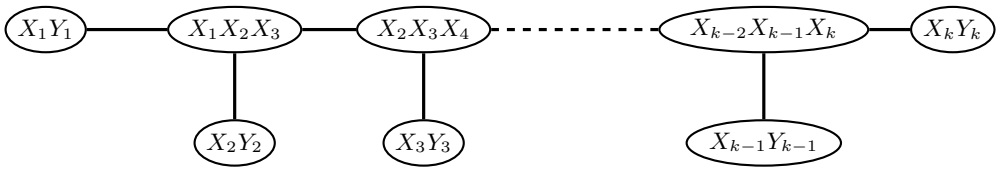
$$\psi'_S(x_S) \equiv \sum_{x_{C\setminus S}} \psi_C(x_C), \qquad \psi'_D(x_D) \equiv \psi_D(x_D)\frac{\psi'_S(x_S)}{\psi_S(x_S)}.$$

(b) Clearly ψ'_S and ψ_C are consistent by definition. On the other hand,

$$\begin{aligned}\sum_{x_{D\setminus S}} \psi'_D(x_D) &= \sum_{x_{D\setminus S}} \psi_D(x_D)\frac{\psi'_S(x_S)}{\psi_S(x_S)} \\ &= \frac{\psi'_S(x_S)}{\psi_S(x_S)} \sum_{x_{D\setminus S}} \psi_D(x_D)\end{aligned}$$

so if ψ_D and ψ_S were consistent before then we are just left with $\psi'_S(x_S)$.

(c) One possible junction tree in this case is:



[Some variation is possible by joining each X_iY_i node to any of the (up to three) other nodes containing X_i.]

(d) The *junction tree algorithm* works by 'collecting' and then 'distributing' messages in the tree. Pick an arbitrary root node, R; for collection, starting from the leaves of the tree and working inwards towards R, pass a message from each node towards R. For distribution, starting from R pass messages back towards the leaves.

This algorithm gives consistency because after collection, each separator node will be consistent with the node adjacent to it that is further away from R. After distribution this consistency is maintained, but the separator will also be consistent with the node nearer to R. Hence all nodes are consistent.

(e) Consistency in a junction tree means that each potential represents a margin of the distribution obtained by their product. This is useful for probabilistic inference.

(f) The junction tree is consistent and therefore each potential represents the relevant marginal distribution. Starting with (say) Y_1, replace $\psi_{X_1Y_1} = p(x_1, y_1)$ with $p(x_1 \mid Y_1 = y_1)$. Then pass messages to the rest of the tree to regain consistency (in fact, passing messages up to X_2Y_2 is sufficient). The potential $\psi_{X_2Y_2}$ is now $p(x_2, y_2 \mid Y_1 = y_1)$, so use it to calculate $p(x_2 \mid Y_2 = y_2, Y_1 = y_1)$. Repeating this process, we will have each potential ψ_C consistent and equal to $p(x_C \mid Y_1 = y_1, \ldots, Y_k = y_k)$. Taking the usual product:

$$\frac{\prod_C \psi_C(x_C)}{\prod_S \psi_S(x_S)}$$

gives the result.

5. (a) Let $\mathrm{pa}_{\mathcal{G}}(v) = \{w \in V : w \to v \text{ in } \mathcal{G}\}$ be the *parents* of v in $\mathcal{G}$. We say that w is a *non-descendant* of v if $w \neq v$ and there is no sequence of edges $v \to \cdots \to w$ from v to w. Denote the set of non-descendants of v by $\mathrm{nd}_{\mathcal{G}}(v)$.

We say that $p(x_V)$ satisfies the local Markov property if $X_v \perp\!\!\!\perp X_{\mathrm{nd}(v)\setminus\mathrm{pa}(v)} \mid X_{\mathrm{pa}(v)}\,[p]$ holds for each $v \in V$.

(b) Since the variables are jointly Gaussian, we can write $X = \alpha Y + \beta Z + \varepsilon$, where ε is a normal random variable independent of Y and Z. The conditional independence shows that $\beta = 0$. Similarly, $Z = \gamma Y + \varepsilon'$.

Then we see directly that $\mathrm{Cov}(X, Y) = \alpha$ and $\mathrm{Cov}(Z, Y) = \gamma$. Also $\mathrm{Cov}(X, Z) = \mathrm{Cov}(\alpha Y, \gamma Y) = \alpha\gamma \operatorname{Var} Y$. This gives the result.

(c) Now, consider a distribution in which $X_1 \sim N(0, 1)$, and $X_i = X_{i-1} + \varepsilon_i$ for $i = 2, \ldots, k$ with ε_i independent Gaussians; any variables not in the path are set to be jointly independent of all other variables. Without loss of generality we may assume that $V = \{1, \ldots, k\}$, since all independences between variables outside the given chain automatically hold, and $X_1 \perp\!\!\!\perp X_k \mid X_C$ (for C not containing any of $1, \ldots, k$) if and only if $X_1 \perp\!\!\!\perp X_k$.

By repeated substitution, $X_k = X_1 + \sum_{i=2}^{k} \varepsilon_i$, so certainly $X_1 \not\perp\!\!\!\perp X_k$.

We claim that this distribution satisfies the local Markov property. [As already noted, for variables i not in the chain this is immediate, since $X_i \perp\!\!\!\perp X_{V\setminus\{i\}}$ will certainly imply $X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)\setminus\mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)}$.]

For $i = 1, 2$ the local Markov property gives only trivial independence statements. For each $i \in \{3, \ldots, k\}$ we have $X_i \perp\!\!\!\perp X_1, \ldots, X_{i-2} \mid X_{i-1}$. Now, the vertices $i, i+1, \ldots, k$ are descendants of i, so $\mathrm{nd}_{\mathcal{G}}(i) = \{1, \ldots, i-1\}$. Applying the graphoid axioms (and noting that $i - 1 \in \mathrm{pa}_{\mathcal{G}}(i)$, this gives us $X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)\setminus\mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)}$.

(d) If j is a non-descendant of i, then $X_i \perp\!\!\!\perp X_j \mid X_{\mathrm{pa}(i)}$ by application of the local Markov property (and the graphoid axioms). If j is a descendant of i, then there exists a directed path $i \to \cdots \to j$ in the graph. Hence, by considering the directed path from i to j and applying (c), there exists a distribution that satisfies the local Markov property but for which $X_i \not\perp\!\!\!\perp X_j \mid X_{\mathrm{pa}(i)}$ (note that $\mathrm{pa}(i)$ cannot contain any vertices on the directed path, else there would be a cycle).

(e) The PC algorithm works by testing conditional independences $X_i \perp\!\!\!\perp X_j \mid X_C$, and removing the i, j edge whenever such an independence is found to hold.

1. Start with a complete undirected graph, say $\mathcal{H}$.
2. For each $c = 0, 1, \ldots, |V| - 2$:
   i. For each pair (i, j) still adjacent in $\mathcal{H}$, and $C \subseteq \mathrm{ne}_{\mathcal{H}}(i) \setminus \{j\}$ with $|C| = c$, test $X_i \perp\!\!\!\perp X_j \mid X_C$.
   ii. If $X_i \perp\!\!\!\perp X_j \mid X_C$ then remove the i, j edge.

In words, we first test all marginal independences, and then for remaining pairs we test conditional on one other neighbour, then on two other neighbours, and so on.

(f) Suppose we have a distribution is faithful to $\mathcal{G}$ and an oracle independence test. Then if i, j are really adjacent there is no conditional independence between them, and we will certainly never remove the i, j edge. If i, j are not adjacent, then by the discussion above we have $X_i \perp\!\!\!\perp X_j \mid X_{\mathrm{pa}(i)}$ (without loss of generality). Since we never remove the edges between i and the parents of i, we will eventually test this independence, and hence we will remove the edge between i and j.