Machine Learning 2	**Due: April 12, 2016**

Homework 2

Instructor: Ke Tran — Email: m.k.tran@uva.nl

Student: Andrea Jemmett — UvA ID: 11162929
Collaborators: N/A — Email: andreajemmett@gmail.com

You are allowed to discuss with your colleagues but you should write the answers in *your own words*. If you discuss with others, write down the name of your collaborators on top of the first page. No points will be deducted for collaborations. If we find similarities in solutions beyond the listed collaborations we will consider it as cheating.
We will not accept any late submissions under any circumstances. The solutions to the previous homework will be handed out in the class at the beginning of the next homework session. After this point, late submissions will be automatically graded zero.
$\star$ denotes bonus exercise. You earn 1 point for solving each bonus exercise. All bonus points earned will be added to your total homework points.

Problem 1. Consider three variables a, b, $c \in \{0, 1\}$ having the joint distribution $p(a, b, c)$ given in Table 1. By direct evaluation, show that

1. $p(a, b) \neq p(a)p(b)$;
2. $p(a, b|c) = p(a|c)p(b|c)$;
3. $p(a, b, c) = p(a)p(c|a)p(b|c)$, and draw the corresponding directed graph.

Table 1: The joint distribution over three binary variables.

a	b	c	$p(a,b,c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Problem 2. Consider all the Bayesian networks consisting of three vertices X, Y and Z. Group them into clusters such that all the graphs in each cluster encode the same set of independence relations. Draw those clusters and write down the set of independence relations for each cluster.

Problem 3.

1. Given distributions p and q of a continuous random variable, Kullback-Leibler divergence of q from p is defined as

$$\mathcal{KL}(p||q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx$$

Evaluate the Kullback-Leibler divergence when $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})$

2. Entropy of a distribution p is given by

$$\mathcal{H}(x) = -\int p(x) \ln p(x) dx$$

Derive the entropy of the multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Solution:

1.

$$\begin{aligned}
\mathcal{KL}(p||q) &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x} \\
&= \int \left[\ln p(\mathbf{x}) - \ln q(\mathbf{x}) \right] p(\mathbf{x}) d\mathbf{x} && (1) \\
&= \int \Big[-\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\
&\quad + \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |\mathbf{L}| + \frac{1}{2} (\mathbf{x} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\mathbf{x} - \mathbf{m}) \Big] p(\mathbf{x}) d\mathbf{x} && (2) \\
&= \int \left[\frac{1}{2} \ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2} (\mathbf{x} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\mathbf{x} - \mathbf{m}) \right] p(\mathbf{x}) d\mathbf{x} && (3) \\
&= \mathbb{E} \left[\frac{1}{2} \ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2} (\mathbf{x} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\mathbf{x} - \mathbf{m}) \right] && (4) \\
&= \mathbb{E} \left[\frac{1}{2} \ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} \right] - \mathbb{E} \left[\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] + \mathbb{E} \left[\frac{1}{2} (\mathbf{x} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\mathbf{x} - \mathbf{m}) \right] && (5) \\
&= \frac{1}{2} \ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} \mathbb{E} \left[(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] + \frac{1}{2} \mathbb{E} \left[(\mathbf{x} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\mathbf{x} - \mathbf{m}) \right] && (6) \\
&= \frac{1}{2} \ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} \left[(\boldsymbol{\mu} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}) + \mathsf{Tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}) \right] + \frac{1}{2} \left[(\boldsymbol{\mu} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}) + \mathsf{Tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) \right] \\
& && (7) \\
&= \frac{1}{2} \ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} \mathsf{Tr}(\mathbf{I}) + \frac{1}{2} (\boldsymbol{\mu} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}) + \frac{1}{2} \mathsf{Tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) && (8) \\
&= \frac{1}{2} \left[\ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - D + (\boldsymbol{\mu} - \mathbf{m})^\mathsf{T} \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}) + \mathsf{Tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) \right] && (9) \\
&= \frac{1}{2} \left[\ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - D + [(-1)(\mathbf{m} - \boldsymbol{\mu})^\mathsf{T} \mathbf{L}^{-1} (-1)(\mathbf{m} - \boldsymbol{\mu})] + \mathsf{Tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) \right] && (10) \\
&= \frac{1}{2} \left[\ln \frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} - D + (\mathbf{m} - \boldsymbol{\mu})^\mathsf{T} \mathbf{L}^{-1} (\mathbf{m} - \boldsymbol{\mu}) + \mathsf{Tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) \right] && (11)
\end{aligned}$$

2.

$$
\begin{aligned}
\mathcal{H}(\mathbf{x}) &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \\
&= -\mathbb{E}[\ln p(\mathbf{x})] && (12) \\
&= -\mathbb{E}\left[-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] && (13) \\
&= \frac{D}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] && (14) \\
&= \frac{D}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathsf{Tr}(\mathbf{I}) && (15) \\
&= \frac{1}{2}\left[D\ln 2\pi + \ln|\boldsymbol{\Sigma}| + D\right] && (16) \\
&= \frac{1}{2}(D\ln 2\pi + \ln|\boldsymbol{\Sigma}| + D\ln e) && (17) \\
&= \frac{1}{2}[\ln(2\pi)^D + \ln|\boldsymbol{\Sigma}| + \ln e^D] && (18) \\
&= \frac{1}{2}\ln[(2\pi e)^D|\boldsymbol{\Sigma}|] && (19)
\end{aligned}
$$

■

Problem 4.

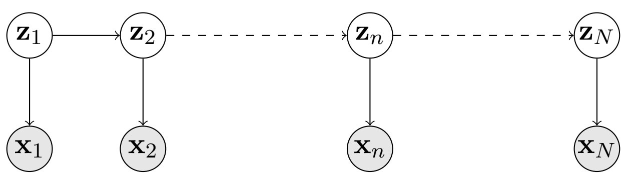

Figure 1: Markov chain of latent variables.

Given a graphical model in Figure 1. Show that:

1. $p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_n)$
2. $p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1})$
3. $p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n, \mathbf{z}_{n+1}) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_{n+1})$
4. $p(\mathbf{z}_{N+1}|\mathbf{z}_N, \mathbf{X}) = p(\mathbf{z}_{N+1}|\mathbf{z}_N)$, where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Solution:

1. If we condition on $\mathbf{z}_n$ then we have that for the d-separation principle holds

$$\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} \perp\!\!\!\perp \mathbf{x}_n \mid \mathbf{z}_n \tag{20}$$

because every path connecting $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ with $\mathbf{x}_n$ is blocked by $\mathbf{z}_n$ (they are all head-to-tail edges) and therefore we can write

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_n) \tag{21}$$

2. If we condition on $\mathbf{z}_{n-1}$, we have that all paths connecting a node in $\mathbf{x}_1, \dots, \mathbf{x}_{n-1}$ with $\mathbf{z}_n$ are blocked by $\mathbf{z}_{n-1}$ and for the d-separation principle we have

$$\mathbf{x}_1, \dots, \mathbf{x}_{n-1} \perp\!\!\!\perp \mathbf{z}_n \mid \mathbf{z}_{n-1} \tag{22}$$

and therefore we can write

$$p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) \tag{23}$$

3. If we now condition on $\mathbf{z}_{n+1}$, all paths connecting a node in $\mathbf{x}_{n+1}, \dots, \mathbf{x}_N$ with $\mathbf{z}_n$ are blocked by $\mathbf{z}_{n+1}$. For the d-separation principle we have that

$$\mathbf{x}_{n+1}, \dots, \mathbf{x}_N \perp\!\!\!\perp \mathbf{z}_n \mid \mathbf{z}_{n+1} \tag{24}$$

and therefore we can write

$$p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N | \mathbf{z}_n, \mathbf{z}_{n+1}) = p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N | \mathbf{z}_{n+1}) \tag{25}$$

4. If we condition on $\mathbf{z}_N$ then we have that all paths connecting $\mathbf{z}_{N+1}$ with a node in $\mathbf{X}$ are blocked by $\mathbf{z}_N$. For the d-separation we can then write

$$\mathbf{z}_{N+1} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{z}_N \tag{26}$$

and therefore we can write

$$p(\mathbf{z}_{N+1} | \mathbf{z}_N, \mathbf{X}) = p(\mathbf{z}_{N+1} | \mathbf{z}_N) \tag{27}$$

■

Problem 5. An edge $X \to Y$ in a graph $\mathcal{G}$ is said to be covered if $pa_Y = pa_X \cup \{X\}$.

1. Let $\mathcal{G}$ be a directed graph with a cover edge $X \to Y$, and $\mathcal{G}'$ be the graph resulted by reversing the edge $X \to Y$ to $Y \to X$, but leaving everything else unchanged. Prove that $\mathcal{G}$ and $\mathcal{G}'$ encode the same set of independent relations.
2. Provide a counterexample to this result in the case where $X \to Y$ is not a covered edge.

Solution:

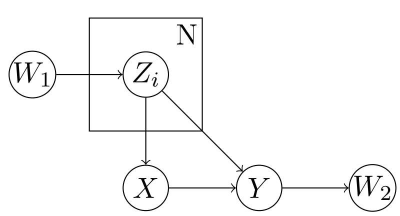

Figure 2: DAG with a cover edge $X \to Y$

1. In Figure 2 is shown a generic DAG $\mathcal{G}$ with a cover edge $X \to Y$. It is easy to see that it is the maximal generalization for which we can have a cover edge in $X \to Y$ because adding nodes and connections to X and not to Y (directed towards) would result in losing the cover edge property. Also adding nodes similar to W_1 or W_2, or changing the direction of their edges would not result

in a different set of independent relations given a change in the direction of the cover edge. We can see that the only set of independent relations are

$$W_1, Z_1, \ldots, Z_N \perp\!\!\!\perp W_2 \mid Y \tag{28}$$

If we change the direction of the cover edge only, the set of independent relations will remain the same because the blocking node Y has always head-to-tail and tail-to-tail edges.

In the case where Y has no descendant nodes, the graph $\mathcal{G}$ doesn't encode any set of independent relations and the same holds for $\mathcal{G}'$.

We can then conclude that if $X \to Y$ is a cover edge, the graph $\mathcal{G}$ will encode the same set of independent relations as $\mathcal{G}'$.

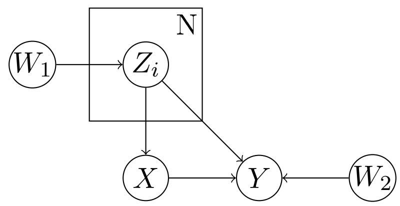

Figure 3: DAG *without* a cover edge in $X \to Y$

2. In Figure 3 is shown a graph without any cover edge. In this case, if we condition on X, then for d-separation we have

$$W_1, Z_1, \ldots, Z_N \perp\!\!\!\perp W_2 \mid X \tag{29}$$

because all paths meet at Y head-to-head but neither Y nor its descentants are in the conditioning set $\{X\}$.

Now suppose we reverse the edge $X \to Y$. In this case the set of independent relations is the empty set because now the paths meet at Y head-to-head but its descendant (X) is in the conditioning set, concluding that if $X \to Y$ is not a cover edge, reversing it will not result in the same set of independent relations being encoded in the DAG $\mathcal{G}'$.

■