Machine Learning 2 - Homework 7

Sindy Löwe (11594969)

May 20, 2018

Collaborators: Pascal Esser, Davide Belli, Andrii Skliar, Gautier Dagan, Linda Petrini

Problem 1

1. First, we derive $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ w.r.t. $\mathbf{A}$ and set the result to zero in order to derive $\mathbf{A}^{\text{new}}$ (in the following, we will omit the const terms, as they do not influence the final result):

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{A}} = \frac{\partial - \mathbb{E}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right]}{\partial \mathbf{A}}$$

Since we derive with respect to $\mathbf{A}$, we can interchange the derivative with the expectation:

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{A}} = \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\frac{\partial - \frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})}{\partial \mathbf{A}}\right]$$

Using Matrix Cookbook equation (88) ([1]) and setting to zero:

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \mathbf{A}} &= \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\sum_{n=2}^{N}\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\mathbf{z}_{n-1}^T\right] = 0 \\
\sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\Gamma^{-1}\mathbf{z}_n\mathbf{z}_{n-1}^T\right] &= \sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\Gamma^{-1}\mathbf{A}\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right] \\
\sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right] &= \sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{A}\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right] \\
\mathbf{A}^{\text{new}} &= \left(\sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right]\right)\left(\sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right]\right)^{-1}
\end{aligned}$$

Next, we derive $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ w.r.t. Γ and set the result to zero in order to derive Γ^{new}:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \Gamma} &= \frac{\partial - \frac{N-1}{2}\ln|\Gamma| - \mathbb{E}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right]}{\partial \Gamma} \\
&= \frac{\partial - \frac{N-1}{2}\ln|\Gamma|}{\partial \Gamma} - \frac{\partial\, \mathbb{E}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right]}{\partial \Gamma} \\
&= \frac{\partial - \frac{N-1}{2}\ln|\Gamma|}{\partial \Gamma} - \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\frac{\partial \frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})}{\partial \Gamma}\right] \\
&= -\frac{N-1}{2}\Gamma^{-1} - \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N} -\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}\right] = 0
\end{aligned}
$$

Using Matrix Cookbook equations (57) and (61).

$$
\begin{aligned}
\frac{N-1}{2}\Gamma^{-1} &= \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N}\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}\right] \\
\Gamma\frac{N-1}{2}\Gamma^{-1}\Gamma &= \Gamma \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N}\Gamma^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\Gamma^{-1}\right]\Gamma \\
\Gamma\frac{N-1}{2} &= \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\right] \\
\Gamma &= \frac{1}{N-1}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\right] \\
\Gamma &= \frac{1}{N-1}\sum_{n=2}^{N}\Bigg(\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_n^T\right] - \mathbf{A}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_{n-1}\mathbf{z}_n^T\right] \\
&\quad - \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right]\mathbf{A}^T - \mathbf{A}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right]\mathbf{A}^T\Bigg)
\end{aligned}
$$

Here, in order to determine the result for Γ^{new}, we need to evaluate $\mathbf{A}^{\text{new}}$ first, which gives us the final update rule:

$$
\begin{aligned}
\Gamma^{\text{new}} &= \frac{1}{N-1}\sum_{n=2}^{N}\Bigg(\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_n^T\right] - \mathbf{A}^{\text{new}}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_{n-1}\mathbf{z}_n^T\right] \\
&\quad - \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right](\mathbf{A}^{\text{new}})^T - \mathbf{A}^{\text{new}}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right](\mathbf{A}^{\text{new}})^T\Bigg)
\end{aligned}
$$

2. As both the transition probabilities $p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$ and the emission probabilities are modelled using Gaussian distributions, the derivation of the update rules for $\mathbf{C}^{\text{new}}$ and $\mathbf{\Sigma}^{\text{new}}$ are similar to those for $\mathbf{A}^{\text{new}}$ and Γ^{new}. This gives us:

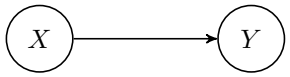
$$
\begin{aligned}
\mathbf{C}^{\text{new}} &= \left(\sum_{n=1}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{x}_n\mathbf{z}_n^T\right]\right)\left(\sum_{n=1}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_n^T\right]\right)^{-1} \\
&= \left(\sum_{n=1}^{N}\mathbf{x}_n\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n^T\right]\right)\left(\sum_{n=1}^{N}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_n^T\right]\right)^{-1}
\end{aligned}
$$

And:

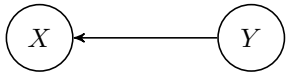
$$\begin{aligned}\boldsymbol{\Sigma}^{\text{new}} = \frac{1}{N}\sum_{n=1}^{N}\Bigg(&\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{x}_n\mathbf{x}_n^T\right] - \mathbf{C}^{\text{new}}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{x}_n^T\right] \\ &- \mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{x}_n\mathbf{z}_n^T\right](\mathbf{C}^{\text{new}})^T - \mathbf{C}^{\text{new}}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_n^T\right](\mathbf{C}^{\text{new}})^T\Bigg) \\ = \frac{1}{N}\sum_{n=1}^{N}\Bigg(&\mathbf{x}_n\mathbf{x}_n^T - \mathbf{C}^{\text{new}}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\right]\mathbf{x}_n^T \\ &-\mathbf{x}_n\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n^T\right](\mathbf{C}^{\text{new}})^T - \mathbf{C}^{\text{new}}\mathop{\mathbb{E}}_{q(\mathbf{z})}\left[\mathbf{z}_n\mathbf{z}_n^T\right](\mathbf{C}^{\text{new}})^T\Bigg)\end{aligned}$$

Problem 2

a) All possible structures are depicted in fig. 1.



(a) Structure 1



(b) Structure 2

(c) Structure 3

Figure 1: All different possible structures consisting of the two variables X and Y

b) Structure 1:

$$p(X,Y) = p(Y|X)\cdot p(X)$$

Structure 2:

$$p(X,Y) = p(Y)\cdot p(X|Y)$$

Structure 3:

$$p(X,Y) = p(Y)\cdot p(X)$$

c) Structure 1:

$$p(Y|X) = \frac{p(X,Y)}{p(X)}$$

Structure 2:

$$p(Y|X) = \frac{p(X|Y)\cdot p(Y)}{\int p(X|Y)\cdot p(Y)dY}$$

Structure 3:

$$p(Y|X) = p(Y) = \frac{p(X,Y)}{p(X)}$$

d) Structure 1:

$$p(Y|do(X)) = p(Y|X) = \frac{p(X,Y)}{p(X)}$$

Structure 2:

$$p(Y|do(X)) = p(Y)$$

Structure 3:

$$p(Y|do(X)) = p(Y|X) = p(Y)$$

e) In this setting, $p(Y|X)$ can be interpreted as the probability of a person having lung cancer, given the observation that this person smokes. On the other hand, $p(Y|do(X))$ can be interpreted as the probability of a person having lung cancer, given that we forced that person to smoke.

Problem 3

1. (a) The recovery rate for the treatment group is 50% and 40% for the control group.
   (b) Since these results are most likely not significant, it is hard to draw any conclusions from them. However, given the slightly higher recovery rate for patients in the treatment group, I would recommend taking the drug.
2. (a) In the male subpopulation, the recovery rate for the treatment group is 60% and 70% for the control group. In the female subpopulation, the recovery rate for the treatment group is 20% and 30% for the control group.
   (b) This result contradicts our previous finding that the treatment group shows a higher recovery rate and is therefore an example of Simpson's paradox. Given these results, we have to admit that we cannot conclude whether the drug is actually helpful or not. Therefore, I would not advise any of the subpopulations to take the drug, considering that any drug can have unforeseen side-effects.
3. Here, the same argument holds as given in 2.b). From the given data, we cannot conclude what the actual effect of the drug is and therefore, I would not advise to take it.
4. (a) Since $S = \{M\}$ blocks the only back-door path, it is admissible for adjustment and therefore, we can write:

$$p(R|do(D)) = \int p(R|D,M) \cdot p(M) dM$$

   (b) No, since

$$p(R|D) = \int p(R|D,M) \cdot p(M|D) dM$$

(c) For a patient with unknown gender, it would not be advisable to take the drug, since:

$$\begin{aligned} p(R=1|do(D=1)) &= p(R=1|D=1,M=1)\cdot p(M=1) \\ &\quad + p(R=1|D=1,M=0)\cdot p(M=0) \\ &= 0.6\cdot p(M=1) + 0.2\cdot p(M=0) \\ &< 0.7\cdot p(M=1) + 0.3\cdot p(M=0) \\ &= p(R=1|D=0,M=1)\cdot p(M=1) \\ &\quad + p(R=1|D=0,M=0)\cdot p(M=0) \\ &= p(R=1|do(D=0)) \end{aligned}$$

5. (a) Since there are no back-door paths, $\emptyset$ is admissible for adjustment and therefore:

$$p(R|do(D)) = \int p(R|D,M)\cdot p(M|D)dM = p(R|D)$$

(b) Yes, in this case the equation holds.

(c) Given this causal model, it would be advisable to take the drug, since

$$\begin{aligned} p(R=1|do(D=1)) &= p(R=1|D=1) = 0.5 \\ &> 0.4 = p(R=1|D=0) = p(R=1|do(D=0)) \end{aligned}$$

6. (a) Let L_1 describe the character trait of being responsible. This variable is not measurable, but can influence the intake of the drug, if an irresponsible person forgets to take his medicine. Let L_2 be the existence of a loving family, which has been shown to positively influence the recovery process ([2]), but is hard to measure as well. In this setting, M might describe the amount of stress-hormones (e.g. Cortisol) in one's blood.

(b) Since the only back-door path is blocked by $\emptyset$, we can write:

$$p(R|do(D)) = p(R|D)$$

(c) Yes, in this case the equation holds.

(d) Following the argumentation in 5.c), my advise for a patient with unknown gender would be to take the drug.

Problem 4

1.

$$\begin{aligned} p(W,R,S) &= p(R)\cdot p(S)\cdot p(W|R,S) \\ p(R) &= p(E_R) \\ p(S) &= p(E_S) \\ p(W=0) &= p(S=0)\cdot p(R=0) = 0.3\cdot 0.6 = 0.18 \\ p(W=1) &= 1 - p(W=0) = 1 - 0.18 = 0.82 \end{aligned}$$

2.

$$\begin{aligned} p(R=1|W=1) &= \frac{p(W=1,R=1)}{p(W=1)} \\ &= \frac{\sum_S p(W=1,R=1,S)}{p(W=1)} \\ &= \frac{0.28+0.42}{0.82} \\ &= \frac{0.7}{0.82} = 0.854 \end{aligned}$$

3. No. Correlation is not the same as causation.

4. The intervened structural causal model is depicted in fig. 2. Since we intervene on W, we need to remove all incoming edges to that variable. Additionally, we get the following values for the variables:

$$\begin{aligned} R &= E_R \\ S &= E_S \\ W &= 1 \end{aligned}$$

And $E_R \perp\!\!\!\perp E_S$

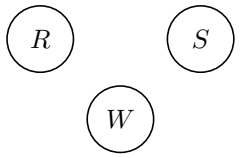


Figure 2: Intervened structural causal model

5.

$$\begin{aligned} p(R=1|do(W=w)) = p(R=1) &= 0.7 \\ p(R=1|W=1) &= 0.854 \\ p(R=1|W=0) &= 0.146 \end{aligned}$$

6. The intervened structural causal model is depicted in fig. 3. Since we intervene on S, there is no change in the model. Additionally, we get the following values for the variables:

$$\begin{aligned} R &= E_R \\ S &= s \\ W &= R \vee S \end{aligned}$$

And $E_R \perp\!\!\!\perp E_S$

7. Since $(W \perp\!\!\!\perp S)_{\mathcal{G}_{\underline{S}}}$, we can use the action/observation exchange rule from do-calculus, which

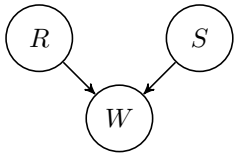


Figure 3: Intervened structural causal model

gives us $p(W|do(S=s)) = p(W|S=s)$, which is generally different from $p(W)$.

$$\begin{aligned} p(W=1|S=1) &= 1 \\ p(W=1|S=0) &= 0.7 \\ p(W=0|S=1) &= 0 \\ p(W=0|S=0) &= 0.3 \\ p(W=1) &= 0.82 \\ p(W=0) &= 0.18 \end{aligned}$$

Problem 5

1. We can show that the equation holds by:

$$\begin{aligned} p(Y|do(X), \mathbf{X}_{\text{pa}(X)}) &= \frac{p(Y, \mathbf{X}_{\text{pa}(X)}|do(X))}{p(\mathbf{X}_{\text{pa}(X)}|do(X))} \\ &\text{using the truncated factorization theorem:} \\ &= \frac{\frac{p(Y, \mathbf{X}_{\text{pa}(X)}, X)}{p(X|\mathbf{X}_{\text{pa}(X)})}}{\frac{p(\mathbf{X}_{\text{pa}(X)}, X)}{p(X|\mathbf{X}_{\text{pa}(X)})}} \\ &= \frac{p(Y, \mathbf{X}_{\text{pa}(X)}, X)}{p(\mathbf{X}_{\text{pa}(X)}, X)} \\ &= p(Y|\mathbf{X}_{\text{pa}(X)}, X) \end{aligned}$$

2. In general, we see that $p(\mathbf{X}_{\text{pa}(X)}|do(X)) = p(\mathbf{X}_{\text{pa}(X)})$. Therefore, we get:

$$\begin{aligned} p(Y|do(X)) &= \int p(Y|do(X), \mathbf{X}_{\text{pa}(X)}) \cdot p(\mathbf{X}_{\text{pa}(X)}|do(X)) d\mathbf{X}_{\text{pa}(X)} \\ &= \int p(Y|X, \mathbf{X}_{\text{pa}(X)}) \cdot p(\mathbf{X}_{\text{pa}(X)}) d\mathbf{X}_{\text{pa}(X)} \end{aligned}$$

References

[1] K. B. Petersen and M. S. Pedersen. The matrix cookbook (version: November 15, 2012), 2012.

[2] E. Tsouna-Hadjis, K. N. Vemmos, N. Zakopoulos, and S. Stamatelopoulos. First-stroke recovery process: the role of family social support. Archives of physical medicine and rehabilitation, 81(7):881–887, 2000.