Machine Learning 2 **Due: 13:00, May 21 2018**

Homework 7

Instructors: Stephan Bongers, Stephan Alaniz Kupsch
Email: s.r.bongers@uva.nl, s.a.alanizkupsch@uva.nl

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. EM for Linear Dynamical System - Bishop 13.3.2

Consider a linear dynamical system model that has linear-Gaussian conditional distributions

$$\begin{aligned} p(\mathbf{z}_n|\mathbf{z}_{n-1}) &= \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \mathbf{\Gamma}) \\ p(\mathbf{x}_n|\mathbf{z}_n) &= \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \mathbf{\Sigma}) \\ p(\mathbf{z}_1) &= \mathcal{N}(\mathbf{z}_1, |\boldsymbol{\mu}_0, \mathbf{V}_0) \end{aligned}$$

The log-likelihood of the data is given by

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=1}^{N} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \mathbf{\Gamma}) + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \mathbf{\Sigma})$$

In E step, we find $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \mathbb{E}_{\mathbf{z}|\mathbf{z}^{\text{old}}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$. Your task is to perform M step

1. *Find $\mathbf{A}^{new}$ and $\mathbf{\Gamma}^{new}$ that optimize*

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = -\frac{N-1}{2} \ln |\mathbf{\Gamma}| - \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\frac{1}{2}\sum_{n=2}^{N} (\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^\mathsf{T} \mathbf{\Gamma}^{-1} (\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right] + const$$

2. *Find $\mathbf{C}^{new}$ and $\mathbf{\Sigma}^{new}$ that optimize*

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = -\frac{N}{2} \ln |\mathbf{\Sigma}| - \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\frac{1}{2}\sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n)^\mathsf{T} \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n)\right] + const$$

Problem 2. We consider the simplest nontrivial causal Bayesian networks: those with only two variables. Call the variables X and Y.

(a) Draw all different possible structures that these networks can have.

(b) For each of the structures, write down the corresponding factorization of $p(X, Y)$.

(c) For each of the structures, write down an explicit expression for $p(Y|X)$ in terms of the factors in (b)

(d) For each of the structures, write down an explicit expression for $p(Y|do(X))$ in terms of the factors in (b).

(e) If Y means lung cancer, and X means smoking (supposing for simplicity that both are binary variables), describe in words what $p(Y|X)$ means and what $p(Y|do(X))$ means, and clearly indicate the difference in interpretation.

Problem 3. Simpson's paradox

Suppose that you are investigating the effectiveness of a drug against a deadly disease. You have gathered the following data on patients that have been admitted in the hospital in which you work; some of these patients have been treated with the drug, others haven't. Some of them recovered, others unfortunately didn't. The reasons why some patients were treated and others were not, are unknown to you.

	Recovery	No recovery	Total	Recovery rate
Drug	20	20	40	...%
No drug	16	24	40	...%
Total	36	44	80	

1a. Calculate the recovery rates (in %) for both treatment and control (i.e., untreated) group.

1b. Would you advice a new patient to take the drug, or not?

Upon closer inspection of the data, you notice something peculiar when you group patients according to gender:

Males	Recovery	No recovery	Total	Recovery rate
Drug	18	12	30	...%
No drug	7	3	10	...%
Total	25	15	40	

Females	Recovery	No recovery	Total	Recovery rate
Drug	2	8	10	...%
No drug	9	21	30	...%
Total	11	29	40	

2a. Calculate the recovery rates (in %) for both the treatment and the control groups, for both sub-populations (males and females).

2b. Would you advice a male patient to take the drug, or not? And a female patient?

3. With hindsight, what would be your advice if the gender of the patient is unknown? Is this in contradiction with your earlier advice?

This phenomenon is known as Simpson's paradox. A lot has been written about this paradox, but it dissolves once you recognize that you should not make the mistake of interpreting correlations as causations. Indeed, whether or not you should prescribe the drug depends on which causal model you believe to apply to this situation. The fact that different causal models will lead to different conclusions is not paradoxical.

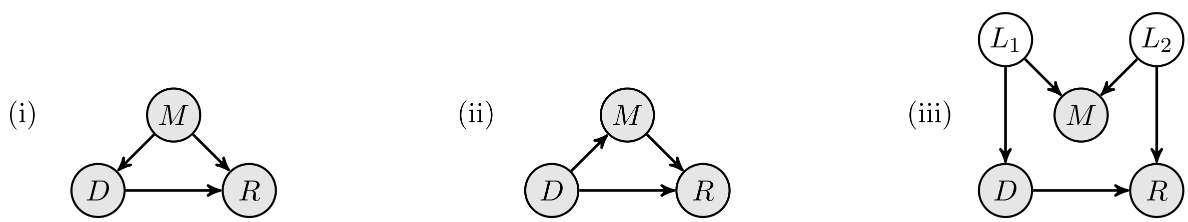


Figure 1: Different hypothetical causal models, where R stands for *Recovery*, D for taking the *Drug*, and M has different interpretations in cases (i), (ii) and (iii).

Suppose you believe that the causal model in Figure 1(i) applies, where M denotes gender of the patient (male/female).

4a. Apply Pearl's back-door criterion to obtain a formula that expresses $p(R \mid \mathrm{do}(D))$ *in terms of observable quantities (i.e., in terms of marginal or conditional distributions where the* do*-operator does not appear).*
4b. Is $p(R \mid \mathrm{do}(D)) = p(R \mid D)$ *in this case?*

4c. What would be your advice for a patient with unknown gender?

Now suppose that instead, you believe the causal model in Figure 1(ii) to apply. Intuitively, this would be quite unlikely, as we know that most drugs don't change gender, but we could have used a slightly different story where the variable M has a different interpretation (for example, "blood pressure"), and then this causal structure would also be a plausible one.

5a. Again, use Pearl's back-door criterion to express $p(R \mid \mathrm{do}(D))$ *in terms of observable quantities.*

5b. Is $p(R \mid \mathrm{do}(D)) = p(R \mid D)$ *in this case?*

5c. What would be your advice for a patient with unknown gender (or if you prefer, blood pressure) in this case?

Finally, suppose that you believe that the causal model in Figure 1(iii) applies.

6a. Invent an interpretation of M and the two latent variables L_1, L_2 yourself that could match the causal model in Figure 1(iii).

6b. Express $p(R \mid \mathrm{do}(D))$ in terms of observable *quantities.*

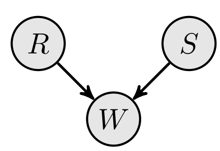
6c. Is $p(R \mid \mathrm{do}(D)) = p(R \mid D)$ in this case?

6d. Again, what would be your advice for a patient with unknown gender in this case?

Problem 4. Probabilistic and causal reasoning in SCMs

Consider three binary variables: R (cycled through the rain), S (taken a shower), W (wet hair). As R and S are *independent* causes of W, we consider the following SCM (where $\vee$ denotes logical or):

$$\begin{aligned} R &= E_R \\ S &= E_S \\ W &= R \vee S \\ p(E_R, E_S) &= p(E_R)p(E_S) \end{aligned}$$



with the following probability distribution for the exogenous variables:

$$p(E_R = 1) = 0.7, \qquad p(E_S = 1) = 0.4$$

1. Calculate the induced distribution on the endogenous variables. How does it factorize?

2. What is the probability that it rained, given that your hair is wet?

3. The probability of R increases when we observe W (in other words, R and W are correlated). Does this mean that W causes R?

We now would like to model what happens if we *intervene* on W by throwing a bucket of water over somebody.

4. Write down the intervened structural causal model for this intervention, $\mathrm{do}(W = 1)$.

5. *Calculate* $p\big(R \mid \mathrm{do}(W = w)\big)$ *and compare with* $p(R)$ *and* $p(R \mid W = w)$.

Instead of intervening on W, we now consider an intervention on S.

6. *Write down the intervened structural causal model for the intervention* $\mathrm{do}(S = s)$.

7. *Calculate* $p\big(W \mid \mathrm{do}(S = s)\big)$ *and compare with* $p(W \mid S = s)$ *and* $p(W)$.

Problem 5$\star$. Truncated factorization

Given a Markovian Structural Causal Model, or alternatively, a causal Bayesian network. Let X be an observed variable, $\boldsymbol{X}_{\mathrm{pa}(X)}$ denote all observed parents of X, and let Y be another observed variable (i.e., neither X nor one of the parents of X).

1. *By applying the truncated factorization theorem, show that*

$$p(Y \mid \mathrm{do}(X), \boldsymbol{X}_{\mathrm{pa}(X)}) = p(Y \mid X, \boldsymbol{X}_{\mathrm{pa}(X)})$$

Hint: the definition of $p(Y \mid \mathrm{do}(X), \boldsymbol{X}_{\mathrm{pa}(X)})$ is what you would expect:

$$p(Y \mid \mathrm{do}(X), \boldsymbol{X}_{\mathrm{pa}(X)}) := \frac{p(Y, \boldsymbol{X}_{\mathrm{pa}(X)} \mid \mathrm{do}(X))}{p(\boldsymbol{X}_{\mathrm{pa}(X)}) \mid \mathrm{do}(X))}$$

2. *Use that result and show by marginalizing over* $\boldsymbol{X}_{\mathrm{pa}(X)}$ *that*

$$p(Y \mid \mathrm{do}(X)) = \int p(Y \mid X, \boldsymbol{X}_{\mathrm{pa}(X)}) p(\boldsymbol{X}_{\mathrm{pa}(X)}) \, d\boldsymbol{X}_{\mathrm{pa}(X)}.$$

Hint: you also need to think about what to do with $p(\boldsymbol{X}_{\mathrm{pa}(X)} \mid \mathrm{do}(X))$.

Note that this formula is different from the following (somewhat similar) formula

$$p(Y \mid X) = \int p(Y \mid X, \boldsymbol{Z}) p(\boldsymbol{Z} \mid X) \, d\boldsymbol{Z},$$

that *always* holds for any set of variables $\boldsymbol{Z}$ (not including X and Y).