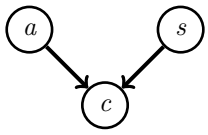
Exercise 1. ***Cancer-asbestos-smoking example: MLE***

Consider the model specified by the DAG



The distribution of a and s are Bernoulli distributions with parameter (success probability) θ_a and θ_s, respectively, i.e.

$$p(a;\theta_a) = \theta_a^a(1-\theta_a)^{1-a} \qquad p(s;\theta_s) = \theta_s^s(1-\theta_s)^{1-s}, \tag{1}$$

and the distribution of c given the parents is parameterised as specified in the following table

$p(c=1\|a,s;\theta_c^1,\ldots,\theta_c^4))$	a	s
θ_c^1	*0*	*0*
θ_c^2	*1*	*0*
θ_c^3	*0*	*1*
θ_c^4	*1*	*1*

The free parameters of the model are $(\theta_a, \theta_s, \theta_c^1, \ldots, \theta_c^4)$.

Assume we observe the following iid data (each row is a data point).

a	s	c
0	*1*	*1*
0	*0*	*0*
1	*0*	*1*
0	*0*	*0*
0	*1*	*0*

(a) Determine the maximum-likelihood estimates of θ_a and θ_s.

Solution. The maximum likelihood estimate (MLE) $\hat{\theta}_a$ is given by the fraction of times that a is 1 in the data set. Hence $\hat{\theta}_a = 1/5$. Similarly, the MLE $\hat{\theta}_s$ is $2/5$.

(b) Determine the maximum-likelihood estimates of $\theta_c^1, \ldots, \theta_c^4$.

Solution. The maximum likelihood estimate of the conditional is the fraction of times $c = 1$ among the data points that satisfy the constraints given by the conditioning set.

- For θ_c^1, we have two observations where $(a, s) = (0, 0)$, and among them, $c = 1$ never occurs. Hence, the MLE $\hat{\theta}_c^1 = \hat{p}(c = 1|a = 0, s = 0)$ is zero.
- For θ_c^2, we have one observation where $(a, s) = (1, 0)$, and for that data point $c = 1$. Hence, the MLE $\hat{\theta}_c^2 = \hat{p}(c = 1|a = 1, s = 0)$ is one.

- For $\hat{\theta}_c^3$, we have two data points with $(a, s) = (0, 1)$. Among them $c = 1$ occurs once, hence $\hat{\theta}_c^3 = \hat{p}(c = 1|a = 0, s = 1) = 1/2$.
- For $\hat{\theta}_c^4$, there are no data points where $(a, s) = (1, 1)$, which means that the MLE is not defined.

In summary, we thus obtain the following maximum likelihood estimates:

$\hat{p}(c = 1\|a, s)$	a	s
$\hat{\theta}_c^1 = 0$	0	0
$\hat{\theta}_c^2 = 1$	1	0
$\hat{\theta}_c^3 = 1/2$	0	1
$\hat{\theta}_c^4$ not defined	1	1

The example illustrates some issues with maximum likelihood estimates: We may get extreme probabilities, zero or one, or if the parent configuration does not occur in the observed data, the estimate is undefined.

Exercise 2. ***Cancer-asbestos-smoking example: Bayesian inference***

We here perform Bayesian inference for the model from Question 1.

We assume that the prior over the parameters of the model, $(\theta_a, \theta_s, \theta_c^1, \ldots, \theta_c^4)$, factorises and is given by Beta distributions with hyperparameters $\alpha_0 = 1$ and $\beta_0 = 1$ (same for all parameters). The posterior then factorises too, with each parameter i following a Beta distribution with hyperparameters equal to

$$\alpha_{i,n}^k = \alpha_{i,0}^k + n_{x_i=1}^k, \qquad \beta_{i,n}^k = \beta_{i,0}^k + n_{x_i=0}^k. \tag{2}$$

Here x_i is the random variable associated with the parameter, e.g. a for θ_a or c for θ_c^k, and k enumerates the possible configurations of its parents. $n_{x_i=1}^k$ denotes the number of times variable x_i equals 1 when its parents are in configuration k, and $n_{x_i=0}^k$ is defined analogously.

(a) Determine the posteriors for θ_a and θ_s.

Solution. We count the number of times a variable equals 1 or 0 in the data set. This gives us the counts $n_{x_i=1}$ and $n_{x_i=0}$. To obtain the posterior hyperparameters, we then add one to them since the prior hyperparameters are $\alpha_0 = 1$ and $\beta_0 = 1$.

- We have $n_{a=1} = 1$ and $n_{a=0} = 4$. Hence $\theta_a \sim \mathcal{B}(2, 5)$.
- We have $n_{s=1} = 2$ and $n_{s=0} = 3$. Hence $\theta_s \sim \mathcal{B}(3, 4)$.

(b) Determine the posteriors for θ_c^k, $k = 1, \ldots, 4$.

Solution. We proceed as in the previous question, just restricting the counts to the different parent configurations. This gives the following results table:

Configuration k	a	s	$n_{c=1}^k$	$n_{c=0}^k$	Posterior
1	0	0	0	2	$\theta_c^1 \sim \mathcal{B}(1, 3)$
2	1	0	1	0	$\theta_c^2 \sim \mathcal{B}(2, 1)$
3	0	1	1	1	$\theta_c^3 \sim \mathcal{B}(2, 2)$
4	1	1	0	0	$\theta_c^4 \sim \mathcal{B}(1, 1)$

Since the configuration $(a, s) = (1, 1)$ does not occur in the data, the posterior for θ^4_c is the same as the prior.

(c) Determine the posterior predictive probabilities $p(a = 1|\mathcal{D}), p(s = 1|\mathcal{D})$, and $p(c = 1|\text{pa}, \mathcal{D})$ for all possible parent configurations.

Solution. We compute the posterior predictive probability for a generic Bernoulli-distributed random variable x. Given data $\mathcal{D}$, let $\mathcal{B}(\alpha, \beta)$ be the posterior distribution of success parameter θ. We then have

$$
\begin{aligned}
p(x = 1|\mathcal{D}) &= \int_0^1 p(x = 1, \theta|\mathcal{D})\mathrm{d}\theta && \text{(sum rule)} && \text{(S.1)}\\
&= \int_0^1 p(x = 1|\theta, \mathcal{D})p(\theta|\mathcal{D})\mathrm{d}\theta && \text{(product rule)} && \text{(S.2)}\\
&= \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta && (x \perp\!\!\!\perp \mathcal{D}|\theta) && \text{(S.3)}\\
&= \int_0^1 \theta p(\theta|\mathcal{D})\mathrm{d}\theta && \text{(Bernoulli model)} && \text{(S.4)}\\
&= \mathbb{E}[\theta|\mathcal{D}] && && \text{(S.5)}
\end{aligned}
$$

Hence the posterior predictive probability for $x = 1$ equals the posterior mean of θ. As the mean of $\mathcal{B}(\alpha, \beta) = \alpha/(\alpha + \beta)$, we have

$$p(x = 1|\mathcal{D}) = \frac{\alpha}{\alpha + \beta} \tag{S.6}$$

Plugging-in the values for α and β computed in the previous question gives:

$$p(a = 1|\mathcal{D}) = \mathbb{E}(\theta^a|\mathcal{D}) = \frac{2}{2 + 5} = \frac{2}{7} \tag{S.7}$$

$$p(s = 1|\mathcal{D}) = \mathbb{E}(\theta^s|\mathcal{D}) = \frac{3}{3 + 4} = \frac{3}{7} \tag{S.8}$$

and

$p(c = 1\|a, s, \mathcal{D})$	a	s
$1/(1 + 3) = 1/4$	0	0
$2/(2 + 1) = 2/3$	1	0
$2/(2 + 2) = 1/2$	0	1
$1/(1 + 1) = 1/2$	1	1

Compared to the MLE solution in Exercise (b) question (b), we see that the estimates are less extreme. This is because they are a combination of the prior knowledge and the observed data. Moreover, when we do not have any data, the posterior predictive equals the prior predictive probability, unlike for the MLE where the estimate is not defined.

Exercise 3. *Independent component analysis*

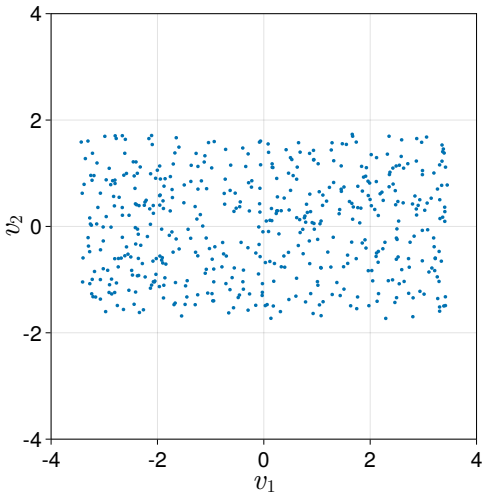
The two scatter plots below show two-dimensional data $\mathbf{v} = (v_1, v_2)^\top$ that were generated by sampling from the noise-free square ICA model

$$\mathbf{v} = \mathbf{A}\mathbf{h}, \tag{3}$$

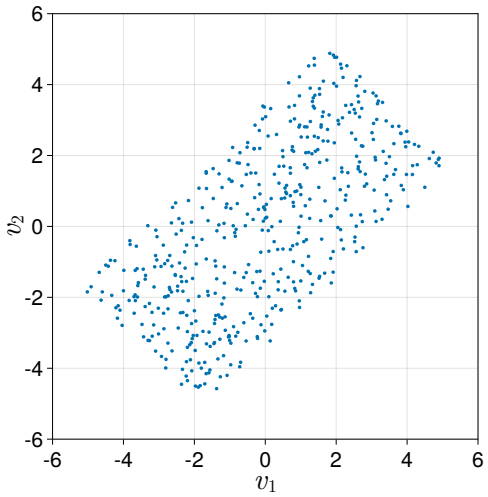
where $\mathbf{A}$ *is a* 2×2 *matrix and* $\mathbf{h} = (h_1, h_2)^\top$ *contains the independent sources. Both* h_1 *and* h_2 *were sampled from a uniform distribution of mean zero and variance one.*

For each scatter plot, select among the following 4 mixing matrices the one that has most likely generated the data. Justify your answer.

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \quad \mathbf{A}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} \quad \mathbf{A}_3 = \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix} \quad \mathbf{A}_4 = \begin{pmatrix} -1 & 2 \\ 1 & 2 \end{pmatrix} \tag{4}$$

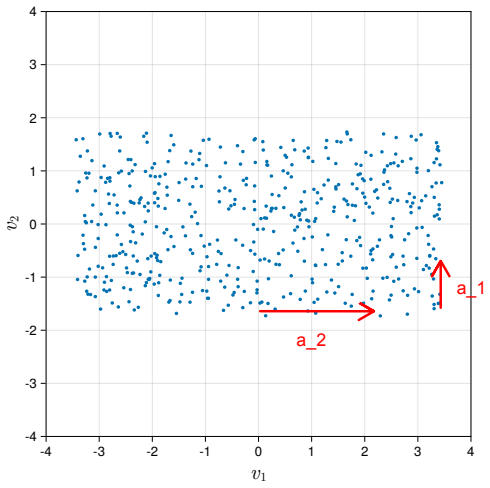


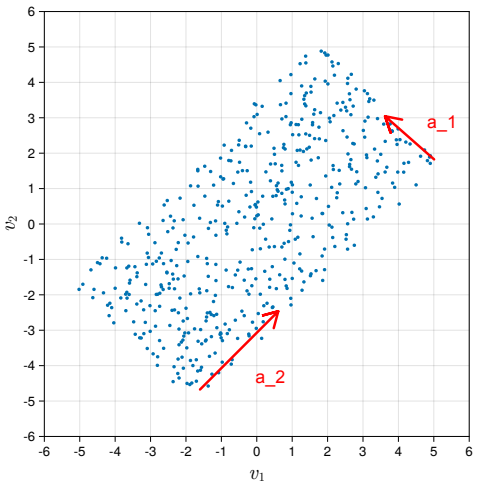
(a)



(b)

Solution. The generative model $\mathbf{v} = \mathbf{A}\mathbf{h}$ can be written as $\mathbf{v} = \mathbf{a}_1 h_1 + \mathbf{a}_2 h_2$ where $\mathbf{a}_1$ and $\mathbf{a}_2$ are the first and second column of $\mathbf{A}$, respectively. Since h_1 and h_2 are independent and uniformly distributed, the edges of the data-parallelogram correspond to $\mathbf{a}_1$ and $\mathbf{a}_2$. This is shown in the following figures.





The data-parallelogram shown in Figure (a) is aligned with the axes, with the data being spread out twice as much along the v_1 axis than along the v_2 axis. This means that we are looking for vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ that are multiples of the unit vectors, that are orthogonal to each other, and where one vector has an element in the first slot that is roughly twice as large as the element in the second slot. The correct mixing matrix for Figure (a) thus is $\mathbf{A}_1$. Matrix $\mathbf{A}_2$ is not correct

because it would assert that the spread along the v_2 axis is three times as large as the spread along the v_1. Matrices $\mathbf{A}_3$ and $\mathbf{A}_4$ are not correct because they rotate the data parallelogram.

The data-parallelogram shown in Figure (b) is not aligned with the axes, which excludes matrices $\mathbf{A}_1$ and $\mathbf{A}_2$. We also see that top-right and bottom-left edge of the data-parallelogram corresponds to a column vector with a negative element in the first slot. The correct mixing matrix for figure (b) thus is $\mathbf{A}_4$.