Exercises for the tutorials: 5 and 9.

The other exercises are for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. *Maximum likelihood estimation for a Gaussian*

The Gaussian pdf parametrised by mean μ and standard deviation σ is given by

$$p(x; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad \boldsymbol{\theta} = (\mu, \sigma).$$

(a) Given iid data $\mathcal{D} = \{x_1, \ldots, x_n\}$, what is the likelihood function $L(\boldsymbol{\theta})$ for the Gaussian model?

Solution. For iid data, the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_i^n p(x_i; \boldsymbol{\theta}) \tag{S.1}$$

$$= \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \tag{S.2}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right]. \tag{S.3}$$

(b) What is the log-likelihood function $\ell(\boldsymbol{\theta})$?

Solution. Taking the log of the likelihood function gives

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \tag{S.4}$$

(c) Show that the maximum likelihood estimates for the mean μ and standard deviation σ are the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \tag{1}$$

and the square root of the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2. \tag{2}$$

Solution. Since the logarithm is strictly monotonically increasing, the maximiser of the log-likelihood equals the maximiser of the likelihood. It is easier to take derivatives for the log-likelihood function than for the likelihood function so that the maximum likelihood estimate is typically determined using the log-likelihood.

Given the algebraic expression of $\ell(\boldsymbol{\theta})$, it is simpler to work with the variance $v = \sigma^2$ rather than the standard deviation. (In the lecture notes, we used the variable η to denote the

transformed parameters. We could have written $\eta = \sigma^2$, but v is a more natural notation for the variance.) Since $\sigma > 0$ the function $v = g(\sigma) = \sigma^2$ is invertible, and the invariance of the MLE to re-parametrisation guarantees that

$$\hat{\sigma} = \sqrt{\hat{v}}.$$

We now thus maximise the function $J(\mu, v)$,

$$J(\mu, v) = -\frac{n}{2}\log(2\pi v) - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{S.5}$$

with respect to μ and v.

Taking partial derivatives gives

$$\frac{\partial J}{\partial \mu} = \frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) \tag{S.6}$$

$$= \frac{1}{v}\sum_{i=1}^{n} x_i - \frac{n}{v}\mu \tag{S.7}$$

$$\frac{\partial J}{\partial v} = -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{S.8}$$

A necessary condition for optimality is that the partial derivatives are zero. We thus obtain the conditions

$$\frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{S.9}$$

$$-\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \tag{S.10}$$

From the first condition it follows that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{S.11}$$

The second condition thus becomes

$$-\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 0 \qquad \text{(multiply with } v^2 \text{ and rearrange)} \tag{S.12}$$

$$\frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{n}{2}v, \tag{S.13}$$

and hence

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2, \tag{S.14}$$

We now check that this solution corresponds to a maximum by computing the Hessian matrix

$$\mathbf{H}(\mu, v) = \begin{pmatrix} \frac{\partial^2 J}{\partial \mu^2} & \frac{\partial^2 J}{\partial \mu \partial v} \\ \frac{\partial^2 J}{\partial \mu \partial v} & \frac{\partial^2 J}{\partial v^2} \end{pmatrix} \tag{S.15}$$

If the Hessian negative definite at $(\hat{\mu}, \hat{v})$, the point is a (local) maximum. Since we only have one critical point, $(\hat{\mu}, \hat{v})$, the local maximum is also a global maximum. Taking second derivatives gives

$$\mathbf{H}(\mu, v) = \begin{pmatrix} -\frac{n}{v} & -\frac{1}{v^2}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{v^2}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2}\frac{1}{v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}. \tag{S.16}$$

Substituting the values for $(\hat{\mu}, \hat{v})$ gives

$$\mathbf{H}(\hat{\mu}, \hat{v}) = \begin{pmatrix} -\frac{n}{\hat{v}} & 0 \\ 0 & -\frac{n}{2}\frac{1}{\hat{v}^2} \end{pmatrix}, \tag{S.17}$$

which is negative definite. Note that the the (negative) curvature increases with n, which means that $J(\mu, v)$, and hence the log-likelihood becomes more and more peaked as the number of data points n increases.

Exercise 2. ***Posterior of the mean of a Gaussian with known variance***

Given iid data $\mathcal{D} = \{x_1, \ldots, x_n\}$, compute $p(\mu|\mathcal{D}, \sigma^2)$ for the Bayesian model

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad p(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] \tag{3}$$

where σ^2 is a fixed known quantity.
Hint: You may use that

$$\mathcal{N}(x; m_1, \sigma_1^2)\mathcal{N}(x; m_2, \sigma_2^2) \propto \mathcal{N}(x; m_3, \sigma_3^2) \tag{4}$$

where

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{5}$$

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{6}$$

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(m_2 - m_1) \tag{7}$$

Solution. We re-use the expression for the likelihood $L(\mu)$ from Exercise 1.

$$L(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right], \tag{S.18}$$

which we can write as

$$L(\mu) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \tag{S.19}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2)\right] \tag{S.20}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right] \tag{S.21}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2n\mu\bar{x} + n\mu^2\right)\right] \tag{S.22}$$

$$\propto \exp\left[-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right] \tag{S.23}$$

$$\propto \mathcal{N}(\mu; \bar{x}, \sigma^2/n). \tag{S.24}$$

The posterior is

$$p(\mu|\mathcal{D}) \propto L(\theta)p(\mu;\mu_0,\sigma_0^2) \tag{S.25}$$
$$\propto \mathcal{N}(\mu;\bar{x},\sigma^2/n)\mathcal{N}(\mu;\mu_0,\sigma_0^2) \tag{S.26}$$

so that with (4), we have

$$p(\mu|\mathcal{D}) \propto \mathcal{N}(\mu;\mu_n,\sigma_n^2) \tag{S.27}$$
$$\sigma_n^2 = \left(\frac{1}{\sigma^2/n}+\frac{1}{\sigma_0^2}\right)^{-1} \tag{S.28}$$
$$= \frac{\sigma_0^2\sigma^2/n}{\sigma_0^2+\sigma^2/n} \tag{S.29}$$
$$\mu_n = \sigma_n^2\left(\frac{\bar{x}}{\sigma^2/n}+\frac{\mu_0}{\sigma_0^2}\right) \tag{S.30}$$
$$= \frac{1}{\sigma_0^2+\sigma^2/n}\left(\sigma_0^2\bar{x}+(\sigma^2/n)\mu_0\right) \tag{S.31}$$
$$= \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\bar{x}+\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\mu_0. \tag{S.32}$$

As n increases, σ^2/n goes to zero so that $\sigma_n^2 \to 0$ and $\mu_n \to \bar{x}$. This means that with an increasing amount of data, the posterior of the mean tends to be concentrated around the maximum likelihood estimate $\bar{x}$.

From (7), we also have that

$$\mu_n = \mu_0 + \frac{\sigma_0^2}{\sigma^2/n+\sigma_0^2}(\bar{x}-\mu_0), \tag{S.33}$$

which shows more clearly that the value of μ_n lies on a line with end-points μ_0 (for $n=0$) and $\bar{x}$ (for $n\to\infty$). As the amount of data increases, μ_n moves form the mean under the prior, μ_0, to the average of the observed sample, that is the MLE $\bar{x}$.

Exercise 3. ***Maximum likelihood estimation of probability tables in fully observed directed graphical models of binary variables***

We assume that we are given a parametrised directed graphical model for variables $x_1,\ldots,x_d$,

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i;\boldsymbol{\theta}_i) \qquad x_i \in \{0,1\} \tag{8}$$

where the conditionals are represented by parametrised probability tables, For example, if $\mathrm{pa}_3 = \{x_1,x_2\}$, $p(x_3|\mathrm{pa}_3;\boldsymbol{\theta}_3)$ is represented as

$p(x_3=1\|x_1,x_2;\theta_3^1,\ldots,\theta_3^4))$	x_1	x_2
θ_3^1	*0*	*0*
θ_3^2	*1*	*0*
θ_3^3	*0*	*1*
θ_3^4	*1*	*1*

with $\boldsymbol{\theta}_3 = (\theta_3^1,\theta_3^2,\theta_3^3,\theta_3^4)$, and where the superscripts j of θ_3^j enumerate the different states that the parents can be in.

(a) Assuming that x_i has m_i parents, verify that the table parametrisation of $p(x_i|\mathrm{pa}_i;\boldsymbol{\theta}_i)$ is equivalent to writing $p(x_i|\mathrm{pa}_i;\boldsymbol{\theta}_i)$ as

$$p(x_i|\mathrm{pa}_i;\boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i=1,\mathrm{pa}_i=s)} (1-\theta_i^s)^{\mathbb{1}(x_i=0,\mathrm{pa}_i=s)} \tag{9}$$

where $S_i = 2^{m_i}$ is the total number of states/configurations that the parents can be in, and $\mathbb{1}(x_i = 1, \mathrm{pa}_i = s)$ is one if $x_i = 1$ and $\mathrm{pa}_i = s$, and zero otherwise.

Solution. The number of configurations that m binary parents can be in is given by S_i. The questions thus boils down to showing that $p(x_i = 1|\mathrm{pa}_i = k;\boldsymbol{\theta}_i) = \theta_i^k$ for any state $k \in \{1,\dots,S_i\}$ of the parents of x_i. Since $\mathbb{1}(x_i = 1, \mathrm{pa}_i = s) = 0$ unless $s = k$, we have indeed that

$$p(x_i = 1|\mathrm{pa}_i = k;\boldsymbol{\theta}_i) = \left[\prod_{s\neq k} (\theta_i^s)^0 (1-\theta_i^s)^0\right] (\theta_i^k)^{\mathbb{1}(x_i=1,\mathrm{pa}_i=k)} (1-\theta_i^k)^{\mathbb{1}(x_i=0,\mathrm{pa}_i=k)} \tag{S.34}$$

$$= 1 \cdot (\theta_i^k)^{\mathbb{1}(x_i=1,\mathrm{pa}_i=k)} (1-\theta_i^k)^0 \tag{S.35}$$

$$= \theta_i^k. \tag{S.36}$$

(b) For iid data $\mathcal{D} = \{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}\}$ show that the likelihood can be represented as

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{n^s_{x_i=1}} (1-\theta_i^s)^{n^s_{x_i=0}} \tag{10}$$

where $n^s_{x_i=1}$ is the number of times the pattern $(x_i = 1, \mathrm{pa}_i = s)$ occurs in the data $\mathcal{D}$, and equivalently for $n^s_{x_i=0}$.

Solution. Since the data are iid, we have

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{j=1}^{n} p(\mathbf{x}^{(j)};\boldsymbol{\theta}) \tag{S.37}$$

$$\tag{S.38}$$

where each term $p(\mathbf{x}^{(j)};\boldsymbol{\theta})$ factorises as in (8),

$$p(\mathbf{x}^{(j)};\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.39}$$

with $x_i^{(j)}$ denoting the i-th element of $\mathbf{x}^{(j)}$ and $\mathrm{pa}_i^{(j)}$ the corresponding parents. The conditionals $p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i)$ factorise further according to (9),

$$p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)}, \tag{S.40}$$

so that

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{j=1}^{n}\prod_{i=1}^{d} p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.41}$$

$$= \prod_{j=1}^{n}\prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.42}$$

Swapping the order of the products so that the product over the data points comes first, we obtain

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{j=1}^{n}(\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.43}$$

We next split the product over j into two products, one for all j where $x_i^{(j)}=1$, and one for all j where $x_i^{(j)}=0$

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{\substack{j:\\ x_i^{(j)}=1}}\prod_{\substack{j:\\ x_i^{(j)}=0}}(\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.44}$$

$$= \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{\substack{j:\\ x_i^{(j)}=1}}(\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)}\prod_{\substack{j:\\ x_i^{(j)}=0}}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.45}$$

$$= \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{\sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.46}$$

$$= \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n_{x_i=1}^s}(1-\theta_i^s)^{n_{x_i=0}^s} \tag{S.47}$$

where

$$n_{x_i=1}^s = \sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s) \qquad n_{x_i=0}^s = \sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s) \tag{S.48}$$

is the number of times $x_i=1$ and $x_i=0$, respectively, with its parents being in state s.

(c) Show that the log-likelihood decomposes into sums of terms that can be independently optimised, and that each term corresponds to the log-likelihood for a Bernoulli model.

Solution. The log-likelihood $\ell(\boldsymbol{\theta})$ equals

$$\ell(\boldsymbol{\theta}) = \log p(\mathcal{D};\boldsymbol{\theta}) \tag{S.49}$$

$$= \log\prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n_{x_i=1}^s}(1-\theta_i^s)^{n_{x_i=0}^s} \tag{S.50}$$

$$= \sum_{i=1}^{d}\sum_{s=1}^{S_i}\log\left[(\theta_i^s)^{n_{x_i=1}^s}(1-\theta_i^s)^{n_{x_i=0}^s}\right] \tag{S.51}$$

$$= \sum_{i=1}^{d}\sum_{s=1}^{S_i}n_{x_i=1}^s\log(\theta_i^s) + n_{x_i=0}^s\log(1-\theta_i^s) \tag{S.52}$$

Since the parameters θ_i^s are not coupled in any way, maximising $\ell(\boldsymbol{\theta})$ can be achieved by maximising each term $\ell_{is}(\theta_i^s)$ individually,

$$\ell_{is}(\theta_i^s) = n_{x_i=1}^s\log(\theta_i^s) + n_{x_i=0}^s\log(1-\theta_i^s). \tag{S.53}$$

Moreover, $\ell_{is}(\theta_i^s)$ corresponds to the log-likelihood for a Bernoulli model with success probability θ_i^s and data with $n_{x_i=1}^s$ number of ones and $n_{x_i=0}^s$ number of zeros.

(d) Referring to the lecture material, conclude that the maximum likelihood estimates are given by

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} = \frac{\sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s)} \tag{11}$$

Solution. Given the result from the previous question, we can optimise each term $\ell_{is}(\theta_i^s)$ separately. Furthermore, each term formally corresponds to a log-likelihood for a Bernoulli model, so that we can immediately use the results derived in the lecture, which gives

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} \tag{S.54}$$

Since $n^s_{x_i=1} = \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)$ and

$$n^s_{x_i=1} + n^s_{x_i=0} = \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s) + \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 0, \mathrm{pa}_i^{(j)} = s) \tag{S.55}$$

$$= \sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s), \tag{S.56}$$

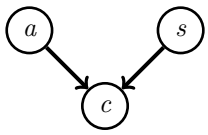
which gives

$$\hat{\theta}_i^s = \frac{\sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s)}. \tag{S.57}$$

Hence, to determine $\hat{\theta}_i^s$, we first count the number of times the parents of x_i are in state s, which gives the denominator, and then among them, count the number of times $x_i = 1$, which gives the numerator.

Exercise 4. *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.

$$p(a;\theta_a) = \theta_a^a(1-\theta_a)^{1-a} \qquad p(s;\theta_s) = \theta_s^s(1-\theta_s)^{1-s}, \tag{12}$$

and the distribution of c given the parents is parametrised 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. With (S.57), we have

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

This because, for example, we have two observations where $(a, s) = (0, 0)$, and among them, $c = 1$ never occurs, so that the MLE for $p(c = 1|a, s)$ is zero.

This 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 5. *Bayesian inference for the Bernoulli model*

Consider the Bayesian model

$$p(x|\theta) = \theta^x(1-\theta)^{1-x} \qquad\qquad p(\theta; \boldsymbol{\alpha}_0) = \mathcal{B}(\theta; \alpha_0, \beta_0)$$

where $x \in \{0, 1\}$, $\theta \in [0, 1], \boldsymbol{\alpha}_0 = (\alpha_0, \beta_0)$, *and*

$$\mathcal{B}(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \in [0, 1] \tag{13}$$

(a) Given iid data $\mathcal{D} = \{x_1, \ldots, x_n\}$ *show that the posterior of* θ *given* $\mathcal{D}$ *is*

$$p(\theta|\mathcal{D}) = \mathcal{B}(\theta; \alpha_n, \beta_n)$$
$$\alpha_n = \alpha_0 + n_{x=1} \qquad\qquad \beta_n = \beta_0 + n_{x=0}$$

where $n_{x=1}$ *denotes the number of ones and* $n_{x=0}$ *the number of zeros in the data.*

Solution. This follows from

$$p(\theta|\mathcal{D}) \propto L(\theta)p(\theta;\boldsymbol{\alpha}_0) \tag{S.58}$$

and from the expression for the likelihood function of the Bernoulli model, which is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} p(x_i|\theta) && \text{(S.59)}\\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} && \text{(S.60)}\\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)} && \text{(S.61)}\\ &= \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}, && \text{(S.62)} \end{aligned}$$

where $n_{x=1} = \sum_{i=1}^n x_i$ denotes the number of 1's in the data, and $n_{x=0} = \sum_{i=1}^n (1-x_i) = n - n_{x=1}$ the number of 0's.

Inserting the expressions for the likelihood and prior into (S.58) gives

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} && \text{(S.63)}\\ &\propto \theta^{\alpha_0+n_{x=1}-1}(1-\theta)^{\beta_0+n_{x=0}-1} && \text{(S.64)}\\ &\propto \mathcal{B}(\theta, \alpha_0+n_{x=1}, \beta_0+n_{x=0}), && \text{(S.65)} \end{aligned}$$

which is the desired result. Since α_0 and β_0 are updated by the counts of ones and zeros in the data, these hyperparameters are also referred to as "pseudo-counts". Alternatively, one can think that they are the counts that are observed in another iid data set which has been previously analysed and used to determine the prior.

(b) Compute the mean of a Beta random variable f,

$$p(f;\alpha,\beta) = \mathcal{B}(f;\alpha,\beta) \qquad f \in [0,1], \tag{14}$$

using that

$$\int_0^1 f^{\alpha-1}(1-f)^{\beta-1}\mathrm{d}f = B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{15}$$

where $B(\alpha,\beta)$ denotes the Beta function and where the Gamma function $\Gamma(t)$ is defined as

$$\Gamma(t) = \int_o^\infty f^{t-1}\exp(-f)\mathrm{d}f \tag{16}$$

and satisfies $\Gamma(t+1) = t\Gamma(t)$.

Hint: It will be useful to represent the partition function in terms of the Beta function.

Solution. We first write the partition function of $p(f;\alpha,\beta)$ in terms of the Beta function

$$\begin{aligned} Z(\alpha,\beta) &= \int_0^1 f^{\alpha-1}(1-f)^{\beta-1} && \text{(S.66)}\\ &= B(\alpha,\beta). && \text{(S.67)} \end{aligned}$$

We then have that the mean $\mathbb{E}[f]$ is given by

$$\mathbb{E}[f] = \int_0^1 f p(f;\alpha,\beta)\mathrm{d}f \tag{S.68}$$
$$= \frac{1}{B(\alpha,\beta)}\int_0^1 f f^{\alpha-1}(1-f)^{\beta-1}\mathrm{d}f \tag{S.69}$$
$$= \frac{1}{B(\alpha,\beta)}\int_0^1 f^{\alpha+1-1}(1-f)^{\beta-1}\mathrm{d}f \tag{S.70}$$
$$= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \tag{S.71}$$
$$= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{S.72}$$
$$= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{S.73}$$
$$= \frac{\alpha}{\alpha+\beta} \tag{S.74}$$

where we have used the definition of the Beta function in terms of the Gamma function and the property $\Gamma(t+1) = t\Gamma(t)$.

(c) Show that the predictive posterior probability $p(x=1|\mathcal{D})$ for a new independently observed data point x equals the posterior mean of $p(\theta|\mathcal{D})$, which in turn is given by

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\alpha_0 + n_{x=1}}{\alpha_0+\beta_0+n}. \tag{17}$$

Solution. We obtain

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1,\theta|\mathcal{D})\mathrm{d}\theta \qquad \text{(sum rule)} \tag{S.75}$$
$$= \int_0^1 p(x=1|\theta,\mathcal{D})p(\theta|\mathcal{D})\mathrm{d}\theta \qquad \text{(product rule)} \tag{S.76}$$
$$= \int_0^1 p(x=1|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta \qquad (x \perp\!\!\!\perp \mathcal{D}|\theta) \tag{S.77}$$
$$= \int_0^1 \theta p(\theta|\mathcal{D})\mathrm{d}\theta \tag{S.78}$$
$$= \mathbb{E}[\theta|\mathcal{D}] \tag{S.79}$$

From the previous question we know the mean of a Beta random variable. Since $\theta \sim \mathcal{B}(\theta;\alpha_n,\beta_n)$, we obtain

$$p(x=1|\mathcal{D}) = \mathbb{E}[\theta|\mathcal{D}] \tag{S.80}$$
$$= \frac{\alpha_n}{\alpha_n+\beta_n} \tag{S.81}$$
$$= \frac{\alpha_0+n_{x=1}}{\alpha_0+n_{x=1}+\beta_0+n_{x=0}} \tag{S.82}$$
$$= \frac{\alpha_0+n_{x=1}}{\alpha_0+\beta_0+n} \tag{S.83}$$

where the last equation follows from the fact that $n = n_{x=0}+n_{x=1}$. Note that for $n\to\infty$, the posterior mean tends to the MLE $n_{x=1}/n$.

Exercise 6. ***Bayesian inference of probability tables in fully observed directed graphical models of binary variables***

This is the Bayesian analogue of Exercise 3 and the notation follows that exercise. We consider the Bayesian model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i) \qquad x_i \in \{0,1\} \tag{18}$$

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) \tag{19}$$

where $p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i)$ is defined via (9), $\boldsymbol{\alpha}_0$ is a vector of hyperparameters containing all $\alpha_{i,0}^s$, $\boldsymbol{\beta}_0$ the vector containing all $\beta_{i,0}^s$, and as before $\mathcal{B}$ denotes the Beta distribution. Under the prior, all parameters are independent.

(a) For iid data $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ show that

$$p(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s, \alpha_{i,n}^s, \beta_{i,n}^s) \tag{20}$$

where

$$\alpha_{i,n}^s = \alpha_{i,0}^s + n_{x_i=1}^s \qquad\qquad \beta_{i,n}^s = \beta_{i,0}^s + n_{x_i=0}^s \tag{21}$$

and that the parameters are also independent under the posterior.

Solution. We start with

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0). \tag{S.84}$$

Inserting the expression for $p(\mathcal{D}|\boldsymbol{\theta})$ given in (10) and the assumed form of the prior gives

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) \tag{S.85}$$

$$\propto \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) \tag{S.86}$$

$$\propto \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} (\theta_i^s)^{\alpha_{i,0}^s - 1} (1-\theta_i^s)^{\beta_{i,0}^s - 1} \tag{S.87}$$

$$\propto \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{\alpha_{i,0}^s + n_{x_i=1}^s - 1} (1-\theta_i^s)^{\beta_{i,0}^s + n_{x_i=0}^s - 1} \tag{S.88}$$

$$\propto \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s + n_{x_i=1}^s, \beta_{i,0}^s + n_{x_i=0}^s) \tag{S.89}$$

It can be immediately verified that $\mathcal{B}(\theta_i^s; \alpha_{i,0}^s + n_{x_i=1}^s, \beta_{i,0}^s + n_{x_i=0}^s)$ is proportional to the marginal $p(\theta_i^s|\mathcal{D})$ so that the parameters are independent under the posterior too.

(b) For a variable x_i with parents pa_i, compute the posterior predictive probability $p(x_i = 1|\mathrm{pa}_i, \mathcal{D})$

Solution. The solution is analogue to the solution for question (c), using the sum rule, independencies, and properties of beta random variables:

$$p(x_i = 1|\text{pa}_i = s, \mathcal{D}) = \int p(x_i = 1, \theta_i^s|\text{pa}_i = s, \mathcal{D})\text{d}\theta_i^s \tag{S.90}$$

$$= \int p(x_i = 1|\theta_i^s, \text{pa}_i = s, \mathcal{D})p(\theta_i^s|\text{pa}_i = s, \mathcal{D}) \tag{S.91}$$

$$= \int p(x_i = 1|\theta_i^s, \text{pa}_i = s)p(\theta_i^s|\mathcal{D}) \tag{S.92}$$

$$= \int \theta_i^s p(\theta_i^s|\mathcal{D}) \tag{S.93}$$

$$= \mathbb{E}[\theta_i^s|\mathcal{D})] \tag{S.94}$$

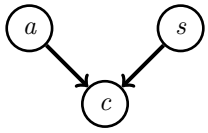
$$\overset{\text{(S.74)}}{=} \frac{\alpha_{i,n}^s}{\alpha_{i,n}^s + \beta_{i,n}^s} \tag{S.95}$$

$$= \frac{\alpha_{i,0}^s + n_{x_i=1}^s}{\alpha_{i,0}^s + \beta_{i,0}^s + n^s} \tag{S.96}$$

where $n^s = n_{x_i=0}^s + n_{x_i=1}^s$ denotes the number of times the parent configuration s occurs in the observed data $\mathcal{D}$.

Exercise 7. ***Cancer-asbestos-smoking example: Bayesian inference***

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.

$$p(a|\theta_a) = \theta_a^a(1-\theta_a)^{1-a} \qquad p(s|\theta_s) = \theta_s^s(1-\theta_s)^{1-s}, \tag{22}$$

and the distribution of c given the parents is parametrised 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*

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).

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 posterior predictive probabilities $p(a = 1|\mathcal{D})$ and $p(s = 1|\mathcal{D})$.

Solution. With Exercise 5 question (c), we have

$$p(a = 1|\mathcal{D}) = \mathbb{E}(\theta^a|\mathcal{D}) = \frac{1+1}{1+1+5} = \frac{2}{7} \tag{S.97}$$

$$p(s = 1|\mathcal{D}) = \mathbb{E}(\theta^s|\mathcal{D}) = \frac{1+2}{1+1+5} = \frac{3}{7} \tag{S.98}$$

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

Solution. The parents of c are (a, s). With Exercise 6 question (b), we have

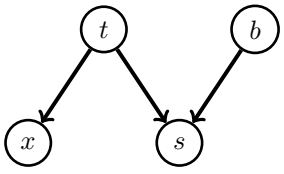
$p(c = 1\|a, s, \mathcal{D})$	a	s
$(1+0)/(1+1+2) = 1/4$	0	0
$(1+1)/(1+1+1) = 2/3$	1	0
$(1+1)/(1+1+2) = 1/2$	0	1
$(1+0)/(1+1) = 1/2$	1	1

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 equals the prior, unlike for the mle where the estimate is not defined.

Exercise 8. *Learning parameters of a directed graphical model*

We consider the directed graphical model shown below on the left for the four binary variables t, b, s, x, each being either zero or one. Assume that we have observed the data shown in the table on the right.

Model:



$t = 1$ *has tuberculosis*
$b = 1$ *has bronchitis*
$s = 1$ *has shortness of breath*
$x = 1$ *has positive x-ray*

Observed data:

x	s	t	b
0	*1*	*0*	*1*
0	*0*	*0*	*0*
0	*1*	*0*	*1*
0	*1*	*0*	*1*
0	*0*	*0*	*0*
0	*0*	*0*	*0*
0	*1*	*0*	*1*
0	*1*	*0*	*1*
0	*0*	*0*	*1*
1	*1*	*1*	*0*

We assume the (conditional) pmf of $s|t,b$ is specified by the following parametrised probability table:

$p(s=1\|t,b;\theta_s^1,\dots,\theta_s^4))$	t	b
θ_s^1	*0*	*0*
θ_s^2	*1*	*0*
θ_s^3	*0*	*1*
θ_s^4	*1*	*1*

(a) What are the maximum likelihood estimates for $p(s=1|b=0,t=0)$ and $p(s=1|b=0,t=1)$, i.e. the parameters θ_s^1 and θ_s^3?

Solution. The maximum likelihood estimates (MLEs) are equal to the fraction of occurrences of the relevant events.

$$\hat{\theta}_s^1 = \frac{\sum_{i=1}^n \mathbb{1}(s_i=1,b_i=0,t_i=0)}{\sum_{i=1}^n \mathbb{1}(b_i=0,t_i=0)} = \frac{0}{3} = 0 \tag{S.99}$$

$$\hat{\theta}_s^3 = \frac{\sum_{i=1}^n \mathbb{1}(s_i=1,b_i=0,t_i=1)}{\sum_{i=1}^n \mathbb{1}(b_i=0,t_i=1)} = \frac{1}{1} = 1 \tag{S.100}$$

(b) Assume each parameter in the table for $p(s|t,b)$ has a uniform prior on $(0,1)$. Compute the posterior mean of the parameters of $p(s=1|b=0,t=0)$ and $p(s=1|b=0,t=1)$ and explain the difference to the maximum likelihood estimates.

Solution. A uniform prior corresponds to a Beta distribution with hyperparameters $\alpha_0=\beta_0=1$. With Exercise 6 question (b), we have

$$\mathbb{E}(\theta_s^1|\mathcal{D}) = \frac{\alpha_0+0}{\alpha_0+\beta_0+3} = \frac{1}{5} \tag{S.101}$$

$$\mathbb{E}(\theta_s^3|\mathcal{D}) = \frac{\alpha_0+1}{\alpha_0+\beta_0+1} = \frac{2}{3} \tag{S.102}$$

Compared to the MLE, the posterior mean is less extreme. It can be considered a "smoothed out" or regularised estimate, where $\alpha_0>0$ and $\beta_0>0$ provides regularisation (see https://en.wikipedia.org/wiki/Additive_smoothing). We can see a pull of the parameters towards the prior predictive mean, which equals 1/2.

Exercise 9. *Factor analysis*

A friend proposes to improve the factor analysis model by working with correlated latent variables. The proposed model is

$$p(\mathbf{h};\mathbf{C}) = \mathcal{N}(\mathbf{h};\mathbf{0},\mathbf{C}) \qquad p(\mathbf{v}|\mathbf{h};\mathbf{F},\mathbf{\Psi},\mathbf{c}) = \mathcal{N}(\mathbf{v};\mathbf{F}\mathbf{h}+\mathbf{c},\mathbf{\Psi}) \tag{23}$$

where $\mathbf{C}$ is some covariance matrix, and the other variables are defined as in the lecture slides. $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{\Sigma})$ denotes the pdf of a Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.

(a) What is marginal distribution of the visibles $p(\mathbf{v};\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ stands for the parameters $\mathbf{C},\mathbf{F},\mathbf{c},\mathbf{\Psi}$?

Solution. The model specifications are equivalent to the following data generating process:

$$\mathbf{h} \sim \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{C}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi}) \qquad \mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon} \tag{S.103}$$

Recall the basic result on the distribution of linear transformations of Gaussians: if $\mathbf{x}$ has density $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbf{C}_x)$, $\mathbf{z}$ density $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_z, \mathbf{C}_z)$, and $\mathbf{x} \perp\!\!\!\perp \mathbf{z}$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$ has density

$$\mathcal{N}(\mathbf{y}; \mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{\mu}_z, \mathbf{A}\mathbf{C}_x\mathbf{A}^\top + \mathbf{C}_z).$$

It thus follows that $\mathbf{v}$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$,

$$\boldsymbol{\mu} = \mathbf{F}\underbrace{\mathbb{E}[\mathbf{h}]}_{\mathbf{0}} + \mathbf{c} + \underbrace{\mathbb{E}[\boldsymbol{\epsilon}]}_{\mathbf{0}} \tag{S.104}$$

$$= \mathbf{c} \tag{S.105}$$

$$\boldsymbol{\Sigma} = \mathbf{F}\mathbb{V}[\mathbf{h}]\mathbf{F}^\top + \mathbb{V}[\boldsymbol{\epsilon}] \tag{S.106}$$

$$= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \boldsymbol{\Psi}. \tag{S.107}$$

(b) Assume that the singular value decomposition of $\mathbf{C}$ *is given by*

$$\mathbf{C} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top \tag{24}$$

where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_H)$ *is a diagonal matrix containing the eigenvalues, and* $\mathbf{E}$ *is a orthonormal matrix containing the corresponding eigenvectors. The matrix square root of* $\mathbf{C}$ *is the matrix* $\mathbf{M}$ *such that*

$$\mathbf{M}\mathbf{M} = \mathbf{C}, \tag{25}$$

and we denote it by $\mathbf{C}^{1/2}$*. Show that the matrix square root of* $\mathbf{C}$ *equals*

$$\mathbf{C}^{1/2} = \mathbf{E}\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\mathbf{E}^\top. \tag{26}$$

Solution. We verify that $\mathbf{C}^{1/2}\mathbf{C}^{1/2} = \mathbf{C}$:

$$\mathbf{C}^{1/2}\mathbf{C}^{1/2} = \mathbf{E}\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\mathbf{E}^\top\mathbf{E}\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\mathbf{E}^\top \tag{S.108}$$

$$= \mathbf{E}\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\ \mathbf{I}\ \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\mathbf{E}^\top \tag{S.109}$$

$$= \mathbf{E}\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_D})\mathbf{E}^\top \tag{S.110}$$

$$= \mathbf{E}\operatorname{diag}(\lambda_1, \ldots, \lambda_D)\mathbf{E}^\top \tag{S.111}$$

$$= \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top \tag{S.112}$$

$$= \mathbf{C} \tag{S.113}$$

(c) Show that the proposed factor analysis model is equivalent to the original factor analysis model

$$p(\mathbf{h}; \mathbf{I}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I}) \qquad p(\mathbf{v}|\mathbf{h}; \tilde{\mathbf{F}}, \boldsymbol{\Psi}, \mathbf{c}) = \mathcal{N}(\mathbf{v}; \tilde{\mathbf{F}}\mathbf{h} + \mathbf{c}, \boldsymbol{\Psi}) \tag{27}$$

with $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{C}^{1/2}$*, so that the extra parameters given by the covariance matrix* $\mathbf{C}$ *are actually redundant and nothing is gained with the richer parametrisation.*

Solution. We verify that the model has the same distribution for the visibles. As before $\mathbb{E}[\mathbf{v}] = \mathbf{c}$, and the covariance matrix is

$$\mathbb{V}[\mathbf{v}] = \tilde{\mathbf{F}}\mathbf{I}\tilde{\mathbf{F}}^\top + \boldsymbol{\Psi} \tag{S.114}$$

$$= \mathbf{F}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\mathbf{F}^\top + \boldsymbol{\Psi} \tag{S.115}$$

$$= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \boldsymbol{\Psi} \tag{S.116}$$

where we have used that $\mathbf{C}^{1/2}$ is a symmetric matrix. This means that the correlation between the $\mathbf{h}$ can be absorbed into the factor matrix $\mathbf{F}$ and the set of pdfs defined by the proposed model equals the set of pdfs of the original factor analysis model.

Another way to see the result is to consider the data generating process and noting that we can sample $\mathbf{h}$ from $\mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{C})$ by first sampling $\mathbf{h}'$ from $\mathcal{N}(\mathbf{h}'; \mathbf{0}, \mathbf{I})$ and then transforming the sample by $\mathbf{C}^{1/2}$,

$$\mathbf{h} \sim \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{C}) \qquad \Longleftrightarrow \qquad \mathbf{h} = \mathbf{C}^{1/2}\mathbf{h}' \qquad \mathbf{h}' \sim \mathcal{N}(\mathbf{h}'; \mathbf{0}, \mathbf{I}). \tag{S.117}$$

This follows again from the basic properties of linear transformations of Gaussians, i.e.

$$\mathbb{V}(\mathbf{C}^{1/2}\mathbf{h}') = \mathbf{C}^{1/2}\mathbb{V}(\mathbf{h}')(\mathbf{C}^{1/2})^\top = \mathbf{C}^{1/2}\mathbf{I}\mathbf{C}^{1/2} = \mathbf{C}$$

and $\mathbb{E}(\mathbf{C}^{1/2}\mathbf{h}') = \mathbf{C}^{1/2}\mathbb{E}(\mathbf{h}') = \mathbf{0}$.

To generate samples from the proposed factor analysis model, we would thus proceed as follows:

$$\mathbf{h}' \sim \mathcal{N}(\mathbf{h}'; \mathbf{0}, \mathbf{I}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi}) \qquad \mathbf{v} = \mathbf{F}(\mathbf{C}^{1/2}\mathbf{h}') + \mathbf{c} + \boldsymbol{\epsilon} \tag{S.118}$$

But the term

$$\mathbf{v} = \mathbf{F}(\mathbf{C}^{1/2}\mathbf{h}') + \mathbf{c} + \boldsymbol{\epsilon}$$

can be written as

$$\mathbf{v} = (\mathbf{F}\mathbf{C}^{1/2})\mathbf{h}' + \mathbf{c} + \boldsymbol{\epsilon} = \tilde{\mathbf{F}}\mathbf{h}' + \mathbf{c} + \boldsymbol{\epsilon}$$

and since $\mathbf{h}'$ follows $\mathcal{N}(\mathbf{h}'; \mathbf{0}, \mathbf{I})$, we are back at the original factor analysis model.

Exercise 10. *Independent component analysis*

(a) Whitening corresponds to linearly transforming a random variable $\mathbf{x}$ *(or the corresponding data) so that the resulting random variable* $\mathbf{z}$ *has an identity covariance matrix, i.e.*

$$\mathbf{z} = \mathbf{V}\mathbf{x} \quad \textit{with} \quad \mathbb{V}[\mathbf{x}] = \mathbf{C} \quad \textit{and} \quad \mathbb{V}[\mathbf{z}] = \mathbf{I}.$$

The matrix $\mathbf{V}$ *is called the whitening matrix. We do not make a distributional assumption on* $\mathbf{x}$*, in particular* $\mathbf{x}$ *may or may not be Gaussian.*

Given the eigenvalue decomposition $\mathbf{C} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top$*, show that*

$$\mathbf{V} = \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top \tag{28}$$

is a whitening matrix.

Solution. From $\mathbb{V}[\mathbf{z}] = \mathbb{V}[\mathbf{V}\mathbf{x}] = \mathbf{V}\mathbb{V}[\mathbf{x}]\mathbf{V}^\top$, it follows that

$$\begin{aligned}
\mathbb{V}[\mathbf{z}] &= \mathbf{V}\mathbb{V}[\mathbf{x}]\mathbf{V}^\top && \text{(S.119)}\\
&= \mathbf{V}\mathbf{C}\mathbf{V}^\top && \text{(S.120)}\\
&= \mathbf{V}\mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top && \text{(S.121)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top\mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top && \text{(S.122)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top && \text{(S.123)}
\end{aligned}$$

where we have used that $\mathbf{E}^\top\mathbf{E} = \mathbf{I}$. Since

$$\mathbf{V}^\top = \left[\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top\right]^\top = \mathbf{E}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})$$

we further have

$$\begin{aligned}
\mathbb{V}[\mathbf{z}] &= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{\Lambda}\mathbf{E}^\top\mathbf{E}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}) && \text{(S.124)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{\Lambda}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}) && \text{(S.125)}\\
&= \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\operatorname{diag}(\lambda_1, \ldots, \lambda_d)\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}) && \text{(S.126)}\\
&= \mathbf{I}, && \text{(S.127)}
\end{aligned}$$

so that $\mathbf{V}$ is indeed a valid whitening matrix. Note that whitening matrices are not unique. For example,

$$\tilde{\mathbf{V}} = \mathbf{E}\operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\mathbf{E}^\top$$

is also a valid whitening matrix. More generally, if $\mathbf{V}$ is a whitening matrix, then $\mathbf{R}\mathbf{V}$ is also a whitening matrix when $\mathbf{R}$ is an orthonormal matrix. This is because

$$\mathbb{V}[\mathbf{R}\mathbf{V}\mathbf{x}] = \mathbf{R}\mathbb{V}[\mathbf{V}\mathbf{x}]\mathbf{R}^\top = \mathbf{R}\mathbf{I}\mathbf{R}^\top = \mathbf{I}$$

where we have used that $\mathbf{V}$ is a whitening matrix so that $\mathbf{V}\mathbf{x}$ has identity covariance matrix.

(b) Consider the ICA model

$$\mathbf{v} = \mathbf{A}\mathbf{h}, \qquad \mathbf{h} \sim p_{\mathbf{h}}(\mathbf{h}), \qquad p_{\mathbf{h}}(\mathbf{h}) = \prod_{i=1}^{D} p_h(h_i), \tag{29}$$

where the matrix $\mathbf{A}$ is invertible and the h_i are independent random variables of mean zero and variance one. Let $\mathbf{V}$ be a whitening matrix for $\mathbf{v}$. Show that $\mathbf{z} = \mathbf{V}\mathbf{v}$ follows the ICA model

$$\mathbf{z} = \tilde{\mathbf{A}}\mathbf{h}, \qquad \mathbf{h} \sim p_{\mathbf{h}}(\mathbf{h}), \qquad p_{\mathbf{h}}(\mathbf{h}) = \prod_{i=1}^{D} p_h(h_i), \tag{30}$$

where $\tilde{\mathbf{A}}$ is an orthonormal matrix.

Solution. If $\mathbf{v}$ follows the ICA model, we have

$$\begin{aligned}
\mathbf{z} &= \mathbf{V}\mathbf{v} && \text{(S.128)}\\
&= \mathbf{V}\mathbf{A}\mathbf{h} && \text{(S.129)}\\
&= \tilde{\mathbf{A}}\mathbf{h} && \text{(S.130)}
\end{aligned}$$

with $\tilde{\mathbf{A}} = \mathbf{VA}$. By the whitening operation, the covariance matrix of $\mathbf{z}$ is identity, so that

$$\mathbf{I} = \mathbb{V}(\mathbf{z}) = \tilde{\mathbf{A}}\mathbb{V}(\mathbf{h})\tilde{\mathbf{A}}^\top. \tag{S.131}$$

By the ICA model, $\mathbb{V}(\mathbf{h}) = \mathbf{I}$, so that $\tilde{\mathbf{A}}$ must satisfy

$$\mathbf{I} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top, \tag{S.132}$$

which means that $\tilde{\mathbf{A}}$ is orthonormal.

In the original ICA model, the number of parameters is given by the number of elements of the matrix $\mathbf{A}$, which is D^2 if $\mathbf{v}$ is D-dimensional. An orthogonal matrix contains $D(D-1)/2$ degrees of freedom (see e.g. https://en.wikipedia.org/wiki/Orthogonal_matrix), so that we can think that whitening "solves half of the ICA problem". Since whitening is a relatively simple standard operation, many algorithms, e.g. "fastICA", first reduce the complexity of the estimation problem by whitening the data. Moreover, due to the properties of the orthogonal matrix, the log-likelihood for the ICA model also simplifies for whitened data: The log-likelihood for ICA model without whitening is

$$\ell(\mathbf{B}) = \sum_{i=1}^{n}\sum_{j=1}^{D} \log p_h(\mathbf{b}_j\mathbf{v}_i) + n\log|\det\mathbf{B}| \tag{S.133}$$

where $\mathbf{B} = \mathbf{A}^{-1}$. If we first whiten the data, the log-likelihood becomes

$$\ell(\tilde{\mathbf{B}}) = \sum_{i=1}^{n}\sum_{j=1}^{D} \log p_h(\tilde{\mathbf{b}}_j\mathbf{z}_i) + n\log|\det\tilde{\mathbf{B}}| \tag{S.134}$$

where $\tilde{\mathbf{B}} = \tilde{\mathbf{A}}^{-1} = \tilde{\mathbf{A}}^\top$ since $\mathbf{A}$ is an orthogonal matrix. This means $\tilde{\mathbf{B}}^{-1} = \tilde{\mathbf{A}} = \tilde{\mathbf{B}}^\top$ and $\tilde{\mathbf{B}}$ is an orthogonal matrix. Hence $\det\tilde{\mathbf{B}} = 1$, and the $\log\det$ term is zero. Hence, the log-likelihood on whitened data simplifies to

$$\ell(\tilde{\mathbf{B}}) = \sum_{i=1}^{n}\sum_{j=1}^{D} \log p_h(\tilde{\mathbf{b}}_j\mathbf{z}_i). \tag{S.135}$$

While the log-likelihood takes a simpler form, the optimisation problem is now a constrained optimisation problem: $\tilde{\mathbf{B}}$ is constrained to be orthonormal. For further information, see e.g. Chapter 9 of *Independent Component Analysis* by Hyvärinen, Karhunen, and Oja.

Exercise 11. *Score matching for the exponential family*

The objective function $J(\boldsymbol{\theta})$ that is minimised in score matching is

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2\right], \tag{31}$$

where ψ_j is the partial derivative of the log model-pdf $\log p(\mathbf{x};\boldsymbol{\theta})$ with respect to the j-th coordinate (slope) and $\partial_j\psi_j$ its second partial derivative (curvature). The observed data are denoted by $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and $\mathbf{x}\in\mathbb{R}^m$.

The goal of this exercise is to show that for statistical models of the form

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K}\theta_k F_k(\mathbf{x}) - \log Z(\boldsymbol{\theta}), \qquad \mathbf{x}\in\mathbb{R}^m, \tag{32}$$

the score matching objective function becomes a quadratic form, which can be optimised efficiently (see e.g. Barber Appendix A.5.3).

The set of models above are called the (continuous) exponential family, or also log-linear models because the models are linear in the parameters θ_k. Since the exponential family generally includes probability mass functions as well, the qualifier "continuous" may be used to highlight that we are here considering continuous random variables only. The functions $F_k(\mathbf{x})$ are assumed to be known (they are called the sufficient statistics).

(a) Denote by $\mathbf{K}(\mathbf{x})$ the matrix with elements $K_{kj}(\mathbf{x})$,

$$K_{kj}(\mathbf{x}) = \frac{\partial F_k(\mathbf{x})}{\partial x_j}, \qquad k = 1 \dots K, \quad j = 1 \dots m, \tag{33}$$

and by $\mathbf{H}(\mathbf{x})$ the matrix with elements $H_{kj}(\mathbf{x})$,

$$H_{kj}(\mathbf{x}) = \frac{\partial^2 F_k(\mathbf{x})}{\partial x_j^2}, \qquad k = 1 \dots K, \quad j = 1 \dots m. \tag{34}$$

Furthermore, let $\mathbf{h}_j(\mathbf{x}) = (H_{1j}(\mathbf{x}), \dots, H_{Kj}(\mathbf{x}))^\top$ be the j–th column vector of $\mathbf{H}(\mathbf{x})$.
Show that for the continuous exponential family, the score matching objective in Equation (31) becomes

$$J(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{r} + \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{M}\boldsymbol{\theta}, \tag{35}$$

where

$$\mathbf{r} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{h}_j(\mathbf{x}_i), \qquad \mathbf{M} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top. \tag{36}$$

Solution. For

$$\log p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k F_k(\mathbf{x}) - \log Z(\boldsymbol{\theta}) \tag{S.136}$$

the first derivative with respect to x_j, the j-th element of $\mathbf{x}$, is

$$\psi_j(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial x_j} \tag{S.137}$$

$$= \sum_{k=1}^{K} \theta_k \frac{\partial F_k(\mathbf{x})}{\partial x_j} \tag{S.138}$$

$$= \sum_{k=1}^{K} \theta_k K_{kj}(\mathbf{x}). \tag{S.139}$$

The second derivative is

$$\partial_j \psi_j(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial^2 \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial x_j^2} \tag{S.140}$$

$$= \sum_{k=1}^{K} \theta_k \frac{\partial^2 F_k(\mathbf{x})}{\partial x_j^2} \tag{S.141}$$

$$= \sum_{k=1}^{K} \theta_k H_{kj}(\mathbf{x}), \tag{S.142}$$

which we can write more compactly as

$$\partial_j \psi_j(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{h}_j(\mathbf{x}). \tag{S.143}$$

The score matching objective in Equation (31) features the sum $\sum_j \psi_j(\mathbf{x};\boldsymbol{\theta})^2$. The term $\psi_j(\mathbf{x};\boldsymbol{\theta})^2$ equals

$$\psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \left[\sum_{k=1}^{K} \theta_k K_{kj}(\mathbf{x})\right]^2 \tag{S.144}$$

$$= \sum_{k=1}^{K}\sum_{k'=1}^{K} K_{kj}(\mathbf{x}) K_{k'j}(\mathbf{x}) \theta_k \theta_{k'}, \tag{S.145}$$

so that

$$\sum_{j=1}^{m} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \sum_{j=1}^{m}\sum_{k=1}^{K}\sum_{k'=1}^{K} K_{kj}(\mathbf{x}) K_{k'j}(\mathbf{x}) \theta_k \theta_{k'} \tag{S.146}$$

$$= \sum_{k=1}^{K}\sum_{k'=1}^{K} \theta_k \theta_{k'} \left[\sum_{j=1}^{m} K_{kj}(\mathbf{x}) K_{k'j}(\mathbf{x})\right], \tag{S.147}$$

which can be more compactly expressed using matrix notation. Noting that

$$\sum_{j=1}^{m} K_{kj}(\mathbf{x}_i) K_{k'j}(\mathbf{x}_i)$$

equals the (k, k') element of the matrix-matrix product $\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top$,

$$\sum_{j=1}^{m} K_{kj}(\mathbf{x}_i) K_{k'j}(\mathbf{x}_i) = \left[\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]_{k,k'}, \tag{S.148}$$

we can write

$$\sum_{j=1}^{m} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \sum_{k=1}^{K}\sum_{k'=1}^{K} \theta_k \theta_{k'} \left[\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]_{k,k'} \tag{S.149}$$

$$= \boldsymbol{\theta}^\top \mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top \boldsymbol{\theta} \tag{S.150}$$

where we have used that for some matrix $\mathbf{A}$

$$\boldsymbol{\theta}^\top \mathbf{A} \boldsymbol{\theta} = \sum_{k,k'} \theta_k \theta_{k'} [\mathbf{A}]_{k,k'} \tag{S.151}$$

where $[\mathbf{A}]_{k,k'}$ is the (k, k') element of the matrix $\mathbf{A}$.

Inserting the expressions into Equation (31) gives

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2\right] \tag{S.152}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2 \tag{S.153}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{\theta}^\top\mathbf{h}_j(\mathbf{x}_i) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}^\top\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\boldsymbol{\theta} \tag{S.154}$$

$$= \boldsymbol{\theta}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{h}_j(\mathbf{x}_i)\right] + \frac{1}{2}\boldsymbol{\theta}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]\boldsymbol{\theta} \tag{S.155}$$

$$= \boldsymbol{\theta}^\top\mathbf{r} + \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta}, \tag{S.156}$$

which is the desired result.

(b) The pdf of a zero mean Gaussian parametrised by the variance σ^2 is

$$p(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{37}$$

The (multivariate) Gaussian is a member of the exponential family. By comparison with Equation (32), we can re-parametrise the statistical model $\{p(x;\sigma^2)\}_{\sigma^2}$ and work with

$$p(x;\theta) = \frac{1}{Z(\theta)}\exp\left(\theta x^2\right), \qquad \theta < 0, \qquad x \in \mathbb{R}, \tag{38}$$

instead. The two parametrisations are related by $\theta = -1/(2\sigma^2)$. Using the previous result on the (continuous) exponential family, determine the score matching estimate $\hat{\theta}$, and show that the corresponding $\hat{\sigma}^2$ is the same as the maximum likelihood estimate. This result is noteworthy because unlike in maximum likelihood estimation, score matching does not need the partition function $Z(\theta)$ for the estimation.

Solution. By comparison with Equation (32), the sufficient statistics $F(x)$ is x^2.

We first determine the score matching objective function. For that, we need to determine the quantities $\mathbf{r}$ and $\mathbf{M}$ in Equation (36). Here, both $\mathbf{r}$ and $\mathbf{M}$ are scalars, and so are the matrices $\mathbf{K}$ and $\mathbf{H}$ that define $\mathbf{r}$ and $\mathbf{M}$. By their definitions, we obtain

$$K(x) = \frac{\partial F(x)}{\partial x} = 2x \tag{S.157}$$

$$H(x) = \frac{\partial^2 F(x)}{\partial x^2} = 2 \tag{S.158}$$

$$r = 2 \tag{S.159}$$

$$M = \frac{1}{n}\sum_{i=1}^{n}K(x_i)^2 \tag{S.160}$$

$$= 4m_2 \tag{S.161}$$

where m_2 denotes the second empirical moment,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2. \tag{S.162}$$

With Equation (31), the score matching objective thus is

$$J(\theta) = 2\theta + \frac{1}{2}4m_2\theta^2 \tag{S.163}$$

$$= 2\theta + 2m_2\theta^2 \tag{S.164}$$

A necessary condition for the minimiser to satisfy is

$$\frac{\partial J(\theta)}{\partial \theta} = 2 + 4\theta m_2 \tag{S.165}$$

$$= 0 \tag{S.166}$$

The only parameter value that satisfies the condition is

$$\hat{\theta} = -\frac{1}{2m_2}. \tag{S.167}$$

The second derivative of $J(\theta)$ is

$$\frac{\partial^2 J(\theta)}{\theta^2} = m_2, \tag{S.168}$$

which is positive (as long as all data points are non-zero). Hence $\hat{\theta}$ is a minimiser.
From the relation $\theta = -1/(2\sigma^2)$, we obtain that the score matching estimate of the variance σ^2 is

$$\hat{\sigma}^2 = -\frac{1}{2\hat{\theta}} = m_2. \tag{S.169}$$

We can obtain the score matching estimate $\hat{\sigma}^2$ from $\hat{\theta}$ in this manner for the same reason that we were able to work with transformed parameters in maximum likelihood estimation.
For zero mean Gaussians, the second moment m_2 is the maximum likelihood estimate of the variance, which shows that the score matching and maximum likelihood estimate are here the same. While the two methods generally yield different estimates, the result also holds for multivariate Gaussians where the score matching estimates also equal the maximum likelihood estimates (see the original article on score matching http://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf).

Exercise 12. *Maximum likelihood estimation and unnormalised models*

Consider the Ising model for two binary random variables (x_1, x_2),

$$p(x_1, x_2; \theta) \propto \exp\left(\theta x_1 x_2 + x_1 + x_2\right), \qquad x_i \in \{-1, 1\},$$

(a) Compute the partition function $Z(\theta)$.

Solution. The definition of the partition function is

$$Z(\theta) = \sum_{\{-1,1\}^2} \exp\left(\theta x_1 x_2 + x_1 + x_2\right). \tag{S.170}$$

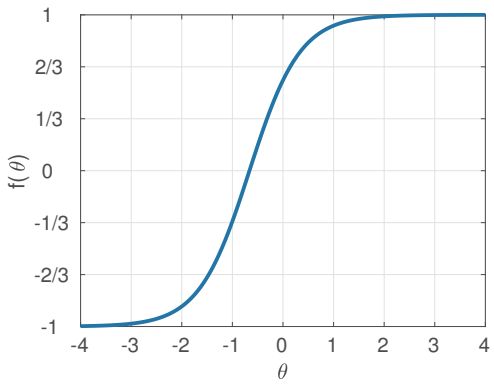
where have have to sum over $(x_1, x_2) \in \{-1, 1\}^2 = \{(-1, 1), (1, 1), (1, -1)\,(-1 - 1)\}$. This gives

$$Z(\theta) = \exp(-\theta - 1 + 1) + \exp(\theta + 2) + \exp(-\theta + 1 - 1) + \exp(\theta - 2) \tag{S.171}$$

$$= 2\exp(-\theta) + \exp(\theta + 2) + \exp(\theta - 2) \tag{S.172}$$

(b) The figure below shows the graph of $f(\theta) = \frac{\partial \log Z(\theta)}{\partial \theta}$.

Assume you observe three data points (x_1, x_2) *equal to* $(-1, -1)$, $(-1, 1)$, *and* $(1, -1)$. *Using the figure, what is the maximum likelihood estimate of* θ*? Justify your answer.*



Solution. Denoting the i-th observed data point by (x_1^i, x_2^i), the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_1^i, x_2^i; \theta) \tag{S.173}$$

Inserting the definition of the $p(x_1, x_2; \theta)$ yields

$$\ell(\theta) = \sum_{i=1}^{n} \left[\theta x_1^i x_2^i + x_1^i + x_2^i\right] - n \log Z(\theta) \tag{S.174}$$

$$= \theta \sum_{i=1}^{n} \left[x_1^i x_2^i\right] + \sum_{i=1}^{n} \left[x_1^i + x_2^i\right] - n \log Z(\theta) \tag{S.175}$$

Its derivative with respect to the θ is

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left[x_1^i x_2^i\right] - n \frac{\partial \log Z(\theta)}{\partial \theta} \tag{S.176}$$

$$= \sum_{i=1}^{n} \left[x_1^i x_2^i\right] - n f(\theta) \tag{S.177}$$

Setting it to zero yields

$$\frac{1}{n} \sum_{i=1}^{n} \left[x_1^i x_2^i\right] = f(\theta) \tag{S.178}$$

An alternative approach is to start with the more general relationship that relates the gradient of the partition function to the gradient of the log unnormalised model. For example, if

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{\phi(\mathbf{x}; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

we have

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta}) \tag{S.179}$$

$$= \sum_{i=1}^{n} \log \phi(\mathbf{x}_i; \boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta}) \tag{S.180}$$

Setting the derivative to zero gives,

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\log\phi(\mathbf{x}_i;\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\log Z(\boldsymbol{\theta})$$

In either case, numerical evaluation of $1/n\sum_{i=1}^{n}x_1^i x_2^i$ gives

$$\frac{1}{n}\sum_{i=1}^{n}\left[x_1^i x_2^i\right] = \frac{1}{3}\left(1-1-1\right) \tag{S.181}$$

$$= -\frac{1}{3} \tag{S.182}$$

From the graph, we see that $f(\theta)$ takes on the value $-1/3$ for $\theta = -1$, which is the desired MLE.

Exercise 13. ***Parameter estimation for unnormalised models***

Let $p(\mathbf{x};\mathbf{A}) \propto \exp(-\mathbf{x}^\top\mathbf{A}\mathbf{x})$ be a parametric statistical model for $\mathbf{x} = (x_1,\ldots,x_{100})$, where the parameters are the elements of the matrix $\mathbf{A}$. Assume that $\mathbf{A}$ is symmetric and positive semi-definite, i.e. $\mathbf{A}$ satisfies $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$ for all values of $\mathbf{x}$.

(a) For n iid data points $\mathbf{x}_1,\ldots,\mathbf{x}_n$, a friend proposes to estimate $\mathbf{A}$ by maximising $J(\mathbf{A})$,

$$J(\mathbf{A}) = \prod_{k=1}^{n}\exp\left(-\mathbf{x}_k^\top\mathbf{A}\mathbf{x}_k\right). \tag{39}$$

Explain why this procedure cannot give reasonable parameter estimates.

Solution. We have that $\mathbf{x}_k^\top\mathbf{A}\mathbf{x}_k \geq 0$ so that $\exp\left(-\mathbf{x}_k^\top\mathbf{A}\mathbf{x}_k\right) \leq 1$. Hence $\exp\left(-\mathbf{x}_k^\top\mathbf{A}\mathbf{x}_k\right)$ is maximal if the elements of $\mathbf{A}$ are zero. This means that $J(\mathbf{A})$ is maximal if $\mathbf{A} = 0$ whatever the observed data, which does not correspond to a meaningful estimation procedure (estimator).

(b) Explain why maximum likelihood estimation is easy when the x_i are real numbers, i.e. $x_i \in \mathbb{R}$, while typically very difficult when the x_i are binary, i.e. $x_i \in \{0,1\}$.

Solution. For maximum likelihood estimation, we needed to normalise the model by computing the partition function $Z(\boldsymbol{\theta})$, which is defined as the sum/integral of $\exp(-\mathbf{x}^\top\mathbf{A}\mathbf{x})$ over the domain of $\mathbf{x}$.

When the x_i are numbers, we can here obtain an analytical expression for $Z(\boldsymbol{\theta})$. However, if the x_i are binary, no such analytical expression is available and computing $Z(\boldsymbol{\theta})$ is then very costly.

(c) Can we use score matching instead of maximum likelihood estimation to learn $\mathbf{A}$ if the x_i are binary?

Solution. No, score matching cannot be used for binary data.