

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  $\mu$  and standard deviation  $\sigma$  is given by

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

(a) Given iid data  $\mathcal{D} = \{x_1, \dots, x_n\}$ , 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  $\mu$  and standard deviation  $\sigma$  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  $\eta$  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  $J(\mu, v)$ ,

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

with respect to  $\mu$  and  $v$ .

Taking partial derivatives gives

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

$$= \frac{1}{v} \sum_{i=1}^n x_i - \frac{n}{v} \mu \quad (\text{S.7})$$

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

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

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

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

From the first condition it follows that

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{S.11})$$

The second condition thus becomes

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

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

and hence

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

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

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

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

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

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

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

which is negative definite. Note that 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  $\mathcal{D} = \{x_1, \dots, x_n\}$ , compute  $p(\mu|\mathcal{D}, \sigma^2)$  for the Bayesian model

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

where  $\sigma^2$  is a fixed known quantity.

Hint: You may use that

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

where

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

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

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

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

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

which we can write as

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

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

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

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

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

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

The posterior is

$$p(\mu|\mathcal{D}) \propto L(\theta)p(\mu; \mu_0, \sigma_0^2) \quad (\text{S.25})$$

$$\propto \mathcal{N}(\mu; \bar{x}, \sigma^2/n)\mathcal{N}(\mu; \mu_0, \sigma_0^2) \quad (\text{S.26})$$

so that with (4), we have

$$p(\mu|\mathcal{D}) \propto \mathcal{N}(\mu; \mu_n, \sigma_n^2) \quad (\text{S.27})$$

$$\sigma_n^2 = \left( \frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2} \right)^{-1} \quad (\text{S.28})$$

$$= \frac{\sigma_0^2 \sigma^2/n}{\sigma_0^2 + \sigma^2/n} \quad (\text{S.29})$$

$$\mu_n = \sigma_n^2 \left( \frac{\bar{x}}{\sigma^2/n} + \frac{\mu_0}{\sigma_0^2} \right) \quad (\text{S.30})$$

$$= \frac{1}{\sigma_0^2 + \sigma^2/n} (\sigma_0^2 \bar{x} + (\sigma^2/n)\mu_0) \quad (\text{S.31})$$

$$= \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n} \bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \mu_0. \quad (\text{S.32})$$

As  $n$  increases,  $\sigma^2/n$  goes to zero so that  $\sigma_n^2 \rightarrow 0$  and  $\mu_n \rightarrow \bar{x}$ . 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

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

which shows more clearly that the value of  $\mu_n$  lies on a line with end-points  $\mu_0$  (for  $n = 0$ ) and  $\bar{x}$  (for  $n \rightarrow \infty$ ). As the amount of data increases,  $\mu_n$  moves from the mean under the prior,  $\mu_0$ , 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  $x_1, \dots, x_d$ ,

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

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

$p(x_3 = 1   x_1, x_2; \theta_3^1, \dots, \theta_3^4)$	$x_1$	$x_2$
$\theta_3^1$	0	0
$\theta_3^2$	1	0
$\theta_3^3$	0	1
$\theta_3^4$	1	1

with  $\boldsymbol{\theta}_3 = (\theta_3^1, \theta_3^2, \theta_3^3, \theta_3^4)$ , and where the superscripts  $j$  of  $\theta_3^j$  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  $p(x_i|\text{pa}_i; \boldsymbol{\theta}_i)$  is equivalent to writing  $p(x_i|\text{pa}_i; \boldsymbol{\theta}_i)$  as

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

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

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

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

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

$$= \theta_i^k. \quad (\text{S.36})$$

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

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

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

**Solution.** Since the data are iid, we have

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

$$(\text{S.38})$$

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

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

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

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

so that

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

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

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

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

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  $x_i^{(j)} = 0$

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

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

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

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

where

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

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

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

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

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

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

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

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

Moreover,  $\ell_{is}(\theta_i^s)$  corresponds to the log-likelihood for a Bernoulli model with success probability  $\theta_i^s$  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

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

**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

$$\hat{\theta}_i^s = \frac{n_{x_i=1}^s}{n_{x_i=1}^s + n_{x_i=0}^s} \quad (S.54)$$

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

$$n_{x_i=1}^s + n_{x_i=0}^s = \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \text{pa}_i^{(j)} = s) + \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 0, \text{pa}_i^{(j)} = s) \quad (S.55)$$

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

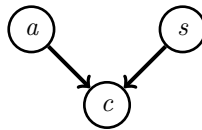
which gives

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

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)  $\theta_a$  and  $\theta_s$ , respectively, i.e.

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

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

$p(c = 1   a, s; \theta_c^1, \dots, \theta_c^4)$	$a$	$s$
$\theta_c^1$	0	0
$\theta_c^2$	1	0
$\theta_c^3$	0	1
$\theta_c^4$	1	1

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

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  $\theta_a$  and  $\theta_s$

**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  $\theta_c^1, \dots, \theta_c^4$ .

**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

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

where  $x \in \{0, 1\}$ ,  $\theta \in [0, 1]$ ,  $\alpha_0 = (\alpha_0, \beta_0)$ , 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  $\mathcal{D} = \{x_1, \dots, x_n\}$  show that the posterior of  $\theta$  given  $\mathcal{D}$  is

$$p(\theta|\mathcal{D}) = \mathcal{B}(\theta; \alpha_n, \beta_n)$$

$$\alpha_n = \alpha_0 + n_{x=1} \qquad \beta_n = \beta_0 + n_{x=0}$$

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



**Solution.** This follows from

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

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

$$L(\theta) = \prod_{i=1}^n p(x_i|\theta) \quad (\text{S.59})$$

$$= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \quad (\text{S.60})$$

$$= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{\sum_{i=1}^n (1-x_i)} \quad (\text{S.61})$$

$$= \theta^{n_{x=1}} (1 - \theta)^{n_{x=0}}, \quad (\text{S.62})$$

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

$$p(\theta|\mathcal{D}) \propto \theta^{n_{x=1}} (1 - \theta)^{n_{x=0}} \theta^{\alpha_0 - 1} (1 - \theta)^{\beta_0 - 1} \quad (\text{S.63})$$

$$\propto \theta^{\alpha_0 + n_{x=1} - 1} (1 - \theta)^{\beta_0 + n_{x=0} - 1} \quad (\text{S.64})$$

$$\propto \mathcal{B}(\theta, \alpha_0 + n_{x=1}, \beta_0 + n_{x=0}), \quad (\text{S.65})$$

which is the desired result. Since  $\alpha_0$  and  $\beta_0$  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  $f$ ,

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

using that

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

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

$$\Gamma(t) = \int_0^\infty f^{t-1} \exp(-f) df \quad (16)$$

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

$$Z(\alpha, \beta) = \int_0^1 f^{\alpha-1} (1-f)^{\beta-1} \quad (\text{S.66})$$

$$= B(\alpha, \beta). \quad (\text{S.67})$$

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

$$\mathbb{E}[f] = \int_0^1 f p(f; \alpha, \beta) df \quad (\text{S.68})$$

$$= \frac{1}{B(\alpha, \beta)} \int_0^1 f f^{\alpha-1} (1-f)^{\beta-1} df \quad (\text{S.69})$$

$$= \frac{1}{B(\alpha, \beta)} \int_0^1 f^{\alpha+1-1} (1-f)^{\beta-1} df \quad (\text{S.70})$$

$$= \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} \quad (\text{S.71})$$

$$= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad (\text{S.72})$$

$$= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad (\text{S.73})$$

$$= \frac{\alpha}{\alpha+\beta} \quad (\text{S.74})$$

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

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

**Solution.** We obtain

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1, \theta|\mathcal{D}) d\theta \quad (\text{sum rule}) \quad (\text{S.75})$$

$$= \int_0^1 p(x=1|\theta, \mathcal{D}) p(\theta|\mathcal{D}) d\theta \quad (\text{product rule}) \quad (\text{S.76})$$

$$= \int_0^1 p(x=1|\theta) p(\theta|\mathcal{D}) d\theta \quad (x \perp\!\!\!\perp \mathcal{D}|\theta) \quad (\text{S.77})$$

$$= \int_0^1 \theta p(\theta|\mathcal{D}) d\theta \quad (\text{S.78})$$

$$= \mathbb{E}[\theta|\mathcal{D}] \quad (\text{S.79})$$

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

$$p(x=1|\mathcal{D}) = \mathbb{E}[\theta|\mathcal{D}] \quad (\text{S.80})$$

$$= \frac{\alpha_n}{\alpha_n + \beta_n} \quad (\text{S.81})$$

$$= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + n_{x=1} + \beta_0 + n_{x=0}} \quad (\text{S.82})$$

$$= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + \beta_0 + n} \quad (\text{S.83})$$

where the last equation follows from the fact that  $n = n_{x=0} + n_{x=1}$ . Note that for  $n \rightarrow \infty$ , 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

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

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

where  $p(x_i|\text{pa}_i, \boldsymbol{\theta}_i)$  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  $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$  show that

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

where

$$\alpha_{i,n}^s = \alpha_{i,0}^s + n_{x_i=1}^s \quad \beta_{i,n}^s = \beta_{i,0}^s + n_{x_i=0}^s \quad (21)$$

and that the parameters are also independent under the posterior.

**Solution.** We start with

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

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

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

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

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

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

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

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  $\text{pa}_i$ , compute the posterior predictive probability  $p(x_i = 1|\text{pa}_i, \mathcal{D})$

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

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

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

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

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

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

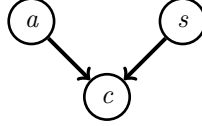
$$\stackrel{(\text{S.74})}{=} \frac{\alpha_{i,n}^s}{\alpha_{i,n}^s + \beta_{i,n}^s} \quad (\text{S.95})$$

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

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)  $\theta_a$  and  $\theta_s$ , respectively, i.e.

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

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

$p(c = 1   a, s, \theta_c^1, \dots, \theta_c^4)$	$a$	$s$
$\theta_c^1$	0	0
$\theta_c^2$	1	0
$\theta_c^3$	0	1
$\theta_c^4$	1	1

We assume that the prior over the parameters of the model,  $(\theta_a, \theta_s, \theta_c^1, \dots, \theta_c^4)$ , 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

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

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

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

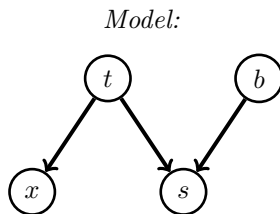
**Solution.** The parents of  $c$  are  $(a, s)$ . With Exercise 6 question (b), we have

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



$t = 1$  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:

$p(s = 1 t, b; \theta_s^1, \dots, \theta_s^4)$	$t$	$b$
$\theta_s^1$	0	0
$\theta_s^2$	1	0
$\theta_s^3$	0	1
$\theta_s^4$	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  $\theta_s^1$  and  $\theta_s^2$ ?

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

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

$$\hat{\theta}_s^2 = \frac{\sum_{i=1}^n \mathbb{1}(s_i = 1, b_i = 0, t_i = 1)}{\sum_{i=1}^n \mathbb{1}(b_i = 0, t_i = 1)} = \frac{1}{1} = 1 \quad (\text{S.100})$$

- (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

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

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

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](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

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

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

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

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

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

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  $\boldsymbol{\Sigma}$ ,

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

$$= \mathbf{c} \quad (\text{S.105})$$

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

$$= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \boldsymbol{\Psi}. \quad (\text{S.107})$$

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

$$\mathbf{C} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top \quad (24)$$

where  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_D)$  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

$$\mathbf{M}\mathbf{M} = \mathbf{C}, \quad (25)$$

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

$$\mathbf{C}^{1/2} = \mathbf{E} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \mathbf{E}^\top. \quad (26)$$

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

$$\mathbf{C}^{1/2}\mathbf{C}^{1/2} = \mathbf{E} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \mathbf{E}^\top \mathbf{E} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \mathbf{E}^\top \quad (\text{S.108})$$

$$= \mathbf{E} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \mathbf{I} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \mathbf{E}^\top \quad (\text{S.109})$$

$$= \mathbf{E} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_D}) \mathbf{E}^\top \quad (\text{S.110})$$

$$= \mathbf{E} \text{diag}(\lambda_1, \dots, \lambda_D) \mathbf{E}^\top \quad (\text{S.111})$$

$$= \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top \quad (\text{S.112})$$

$$= \mathbf{C} \quad (\text{S.113})$$

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

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

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

$$\mathbb{V}[\mathbf{v}] = \tilde{\mathbf{F}}\tilde{\mathbf{F}}^\top + \Psi \quad (\text{S.114})$$

$$= \mathbf{F}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\mathbf{F}^\top + \Psi \quad (\text{S.115})$$

$$= \mathbf{F}\mathbf{C}\mathbf{F}^\top + \Psi \quad (\text{S.116})$$

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  $\mathbf{C}^{1/2}$ ,

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

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:

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

But the term

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

can be written as

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

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.*

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

*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  $\mathbf{C} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top$ , show that*

$$\mathbf{V} = \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2})\mathbf{E}^\top \quad (28)$$

*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

$$\mathbb{V}[\mathbf{z}] = \mathbf{V}\mathbb{V}[\mathbf{x}]\mathbf{V}^\top \quad (\text{S.119})$$

$$= \mathbf{V}\mathbf{C}\mathbf{V}^\top \quad (\text{S.120})$$

$$= \mathbf{V}\mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top \quad (\text{S.121})$$

$$= \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2})\mathbf{E}^\top\mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top \quad (\text{S.122})$$

$$= \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2})\mathbf{\Lambda}\mathbf{E}^\top\mathbf{V}^\top \quad (\text{S.123})$$

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

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

we further have

$$\mathbb{V}[\mathbf{z}] = \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2})\mathbf{\Lambda}\mathbf{E}^\top\mathbf{E} \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2}) \quad (\text{S.124})$$

$$= \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2})\mathbf{\Lambda} \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2}) \quad (\text{S.125})$$

$$= \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2}) \text{diag}(\lambda_1, \dots, \lambda_d) \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2}) \quad (\text{S.126})$$

$$= \mathbf{I}, \quad (\text{S.127})$$

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

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

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

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

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

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

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

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

$$\mathbf{z} = \mathbf{V}\mathbf{v} \quad (\text{S.128})$$

$$= \mathbf{V}\mathbf{A}\mathbf{h} \quad (\text{S.129})$$

$$= \tilde{\mathbf{A}}\mathbf{h} \quad (\text{S.130})$$

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

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

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

$$\mathbf{I} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top, \quad (\text{S.132})$$

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](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

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

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

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

where  $\tilde{\mathbf{B}} = \tilde{\mathbf{A}}^{-1} = \tilde{\mathbf{A}}^\top$  since  $\tilde{\mathbf{A}}$  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

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

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. *Maximum likelihood estimation and unnormalised models*

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

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

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

**Solution.** The definition of the partition function is

$$Z(\theta) = \sum_{\{-1,1\}^2} \exp(\theta x_1 x_2 + x_1 + x_2). \quad (\text{S.136})$$

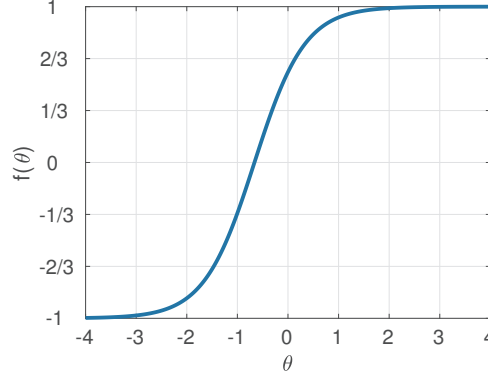
where we have to sum over  $(x_1, x_2) \in \{-1, 1\}^2 = \{(-1, 1), (1, 1), (1, -1), (-1, -1)\}$ . This gives

$$Z(\theta) = \exp(-\theta - 1 + 1) + \exp(\theta + 2) + \exp(-\theta + 1 - 1) + \exp(\theta - 2) \quad (\text{S.137})$$

$$= 2 \exp(-\theta) + \exp(\theta + 2) + \exp(\theta - 2) \quad (\text{S.138})$$

(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  $\theta$ ? Justify your answer.



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

$$\ell(\theta) = \sum_{i=1}^n \log p(x_1^i, x_2^i; \theta) \quad (\text{S.139})$$

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

$$\ell(\theta) = \sum_{i=1}^n [\theta x_1^i x_2^i + x_1^i + x_2^i] - n \log Z(\theta) \quad (\text{S.140})$$

$$= \theta \sum_{i=1}^n [x_1^i x_2^i] + \sum_{i=1}^n [x_1^i + x_2^i] - n \log Z(\theta) \quad (\text{S.141})$$

Its derivative with respect to the  $\theta$  is

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^n [x_1^i x_2^i] - n \frac{\partial \log Z(\theta)}{\partial \theta} \quad (\text{S.142})$$

$$= \sum_{i=1}^n [x_1^i x_2^i] - n f(\theta) \quad (\text{S.143})$$

Setting it to zero yields

$$\frac{1}{n} \sum_{i=1}^n [x_1^i x_2^i] = f(\theta) \quad (\text{S.144})$$

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

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(\mathbf{x}_i; \boldsymbol{\theta}) \quad (\text{S.145})$$

$$= \sum_{i=1}^n \log \phi(\mathbf{x}_i; \boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta}) \quad (\text{S.146})$$

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

$$\frac{1}{n} \sum_{i=1}^n [x_1^i x_2^i] = \frac{1}{3} (1 - 1 - 1) \quad (\text{S.147})$$

$$= -\frac{1}{3} \quad (\text{S.148})$$

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