

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

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

- Given iid data $\mathcal{D} = \{x_1, \dots, x_n\}$, what is the likelihood function $L(\boldsymbol{\theta})$ for the Gaussian model?
- What is the log-likelihood function $\ell(\boldsymbol{\theta})$?
- 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

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

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

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

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

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

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

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

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
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 $p(x_i | \text{pa}_i; \boldsymbol{\theta}_i)$ 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.

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

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

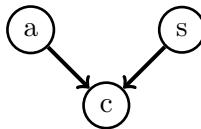
where $n_{x_i=1}^s$ is the number of times the pattern $(x_i = 1, \text{pa}_i = s)$ occurs in the data \mathcal{D} , and equivalently for $n_{x_i=0}^s$.

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

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

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.

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

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
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 $(\theta_a, \theta_s, \theta_c^1, \dots, \theta_c^4)$.

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 θ_a and θ_s
- (b) Determine the maximum-likelihood estimates of $\theta_c^1, \dots, \theta_c^4$.

Exercise 5. Bayesian inference for the Bernoulli model

Consider the Bayesian model

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

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

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

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

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

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

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

(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.

(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

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

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

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

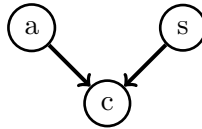
and that the parameters are also independent under the posterior.

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

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.

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

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
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, $(\theta_a, \theta_s, \theta_c^1, \dots, \theta_c^4)$, factorises and is given by beta distributions with hyperparameters $\alpha_0 = 1$ and $\beta_0 = 1$ (same for all parameters).

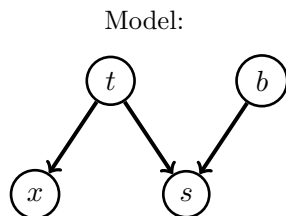
Assume we observe the following iid data (each row is a data point).

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

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
θ_s^1	0	0
θ_s^2	1	0
θ_s^3	0	1
θ_s^4	1	1

- 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^2 ?
- 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.

Exercise 9. Factor analysis

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

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

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 $\boldsymbol{\Sigma}$.

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

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

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

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_D)$ is a diagonal matrix containing the eigenvalues, and \mathbf{E} is an 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

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

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

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

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.

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.

(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.

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 $Z(\theta)$.
- (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.

