Tips for task 3

Exercise 1

Important: Suppose that the samples are independent and identically distributed. Our likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has then the following form:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})} = (2\pi)^{-\frac{dN}{2}} |\Sigma|^{-\frac{N}{2}} e^{-\frac{1}{2}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})}$$

Do not forget that $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix. As usual, because of the exponential function and the positivity of the function it is easier to derive

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{dN}{2} \log 2\pi - \frac{N}{2} \log|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}).$$

1. Take the derivative of $\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the appropriate variable
2. Solve $\nabla \ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0$
3. Check if solution is indeed extreme value

Again, do these steps for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ separately i.e. work with partial derivatives.

Let us do it for $\boldsymbol{\mu}$. In contrast to the last exercise, $\boldsymbol{\mu}$ is now a vector and not a scalar. Because of each component μ_i obeys symmetry properties concerning the other μ_j, we can just take the partial derivative of one component and stack the final solution into a vector.

It is necessary to conduct some auxiliary calculations:

$$(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x}_n - \boldsymbol{\mu}) = \sum_{k=1}^{d} (x_{nk} - \mu_k) \cdot \sum_{p=1}^{d} \Lambda_{kp} (x_{np} - \mu_p)$$

Which leads to

$$\begin{aligned}
\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x}_n - \boldsymbol{\mu}) &= \sum_{n=1}^{N} \sum_{k=1}^{d} (x_{nk} - \mu_k) \cdot \sum_{p=1}^{d} \Lambda_{kp} (x_{np} - \mu_p) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \sum_{p=1}^{d} \Lambda_{kp} (x_{np} - \mu_p) - \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \sum_{p=1}^{d} \Lambda_{kp} (x_{np} - \mu_p) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \sum_{p=1}^{d} \Lambda_{kp} x_{np} - \sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \sum_{p=1}^{d} \Lambda_{kp} \mu_p \\
&\quad - \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \sum_{p=1}^{d} \Lambda_{kp} x_{np} + \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \sum_{p=1}^{d} \Lambda_{kp} \mu_p \,.
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \mu_i} \ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2} \frac{\partial}{\partial \mu_i} \sum_{n=1}^{N} \sum_{k=1}^{d} (x_{nk} - \mu_k) \cdot \sum_{p=1}^{d} \Lambda_{kp} (x_{np} - \mu_p) \\
&= -\frac{1}{2} \frac{\partial}{\partial \mu_i} \left[\sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \sum_{p=1}^{d} \Lambda_{kp} \mu_p + \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \sum_{p=1}^{d} \Lambda_{kp} \mu_p \right] \\
&= -\frac{1}{2} \frac{\partial}{\partial \mu_i} \left[-\sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \sum_{p=1}^{d} \Lambda_{kp} \mu_p - \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \sum_{p=1}^{d} \Lambda_{kp} x_{np} + \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \sum_{p=1}^{d} \Lambda_{kp} \mu_p \right] \\
&= -\frac{1}{2} \left[-\sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \Lambda_{ki} - \sum_{n=1}^{N} \sum_{k=1}^{d} \Lambda_{ik} x_{nk} + \underbrace{\sum_{n=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{d} \mu_k \Lambda_{ik} + \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \Lambda_{ki}}_{\Lambda_{ik} = \Lambda_{ki}} + 2 \sum_{n=1}^{N} \mu_i \Lambda_{ii} \right]. \\
&= -\frac{1}{2} \left[-2 \sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \Lambda_{ki} + 2 \sum_{n=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{d} \mu_k \Lambda_{ik} + 2 \sum_{n=1}^{N} \mu_i \Lambda_{ii} \right] \\
&= -\frac{1}{2} \left[-2 \sum_{n=1}^{N} \sum_{k=1}^{d} x_{nk} \Lambda_{ki} + 2 \sum_{n=1}^{N} \sum_{k=1}^{d} \mu_k \Lambda_{ik} \right] \\
&= \left[\sum_{n=1}^{N} \left(\sum_{k=1}^{d} x_{nk} \Lambda_{ki} - \sum_{k=1}^{d} \mu_k \Lambda_{ik} \right) \right] \\
&= \left[\sum_{n=1}^{N} \left(\sum_{k=1}^{d} \Lambda_{ik} \left(x_{nk} - \mu_k \right) \right) \right] \\
&= \left[\sum_{n=1}^{N} \langle \mathrm{row_i} \left(\boldsymbol{\Lambda} \right), (\mathbf{x_n} - \boldsymbol{\mu}) \rangle \right]
\end{aligned}
$$

Where $\langle \mathrm{row_i} \left(\Lambda \right), (\mathbf{x_n} - \boldsymbol{\mu}) \rangle$ denotes the dot product of the ith row of Λ with $(\mathbf{x_n} - \boldsymbol{\mu})$. Also, do not forget that $\Lambda_{ik} = \Lambda_{ki}$ as $\boldsymbol{\Lambda}$ is symmetric!

This allows us to determine the whole gradient:

$\frac{\partial}{\partial \boldsymbol{\mu}} \ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} [\boldsymbol{\Lambda}(\mathbf{x_n} - \boldsymbol{\mu})] = 0.$

Since $\boldsymbol{\Lambda}$ has inverse $\boldsymbol{\Sigma}$ we just multiply the equation from the left side with $\boldsymbol{\Sigma}$**:**

$$
\sum_{n=1}^{N} [\boldsymbol{\Lambda}(\mathbf{x_n} - \boldsymbol{\mu})] = 0
$$

$$
\begin{aligned}
& \sum_{n=1}^{N}[\boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})] && = 0 \quad | \cdot \boldsymbol{\Sigma}\ from\ the\ left \\
\Leftrightarrow \quad & \sum_{n=1}^{N}[(\mathbf{x}_n - \boldsymbol{\mu})] && = 0 \\
\Leftrightarrow \quad & \sum_{n=1}^{N}\mathbf{x}_n && = \sum_{n=1}^{N}\boldsymbol{\mu} \\
\Leftrightarrow \quad & \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n && = \boldsymbol{\mu}.
\end{aligned}
$$

Now do the same for the covariance matrix! This is easier done with the Matrix differentiation rules from "The Matrix Cookbook" which can be found online.

Exercise 3

First, calculate the product:

$$N(x;\mu_1,\sigma_1^2)N(x;\mu_2,\sigma_2^2) = (2\pi)^{-\frac{1}{4}}\sigma_1\sigma_2 e^{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(x-\mu_2)^2}{\sigma_2^2}\right]}.$$

Now calculate the exponent and solve for x^2. The procedure is called "completing the square "(google it). Do not forget: We are only looking for proportionality, so you can drop all terms that do not depend on x!

$$
\begin{aligned}
\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(x-\mu_2)^2}{\sigma_2^2} &= \frac{x^2-2\mu_1 x+\mu_1^2}{\sigma_1^2}+\frac{x^2-2\mu_2 x+\mu_2^2}{\sigma_2^2} \\
&= \frac{(\sigma_2^2+\sigma_1^2)x^2-2x(\sigma_2^2\mu_1+\sigma_1^2\mu_2)+\sigma_2^2\mu_1^2+\sigma_1^2\mu_2^2}{\sigma_1^2\sigma_2^2} \\
&= \frac{x^2-2x\frac{(\sigma_2^2\mu_1+\sigma_1^2\mu_2)}{(\sigma_2^2+\sigma_1^2)}+\frac{\sigma_2^2\mu_1^2}{(\sigma_2^2+\sigma_1^2)}+\frac{\sigma_1^2\mu_2^2}{(\sigma_2^2+\sigma_1^2)}}{\frac{\sigma_1^2\sigma_2^2}{(\sigma_2^2+\sigma_1^2)}}
\end{aligned}
$$

Now complete the square of the term in the numerator. If you look closely, you can already recognize the wanted expectation and variance.

Exercise 4

What a wonder! Maximum likelihood! I wonder what we have to do now? Of course the same as always! This time we do not have to use the logarithm, as now it would actually make our job harder!

Hint: Instead of solving $\nabla SE(\mathbf{w})$ directly, again solve the partial derivatives $\frac{\partial SE(\mathbf{w})}{\partial w_i}$ and combine final solution into a vector (or matrix) if possible.

$$\frac{\partial}{\partial w_i} 0.5 \sum_{n=1}^{N} r_n (w^T \phi(x_n) - t_n)^2 = \sum_{n=1}^{N} \phi_i(x_n) r_n (w^T \phi(x_n) - t_n)$$
$$= \sum_{n=1}^{N} \phi_i(x_n) r_n \left(w_{-i}^T \phi_{-i}(x_n) - t_n\right) + \sum_{n=1}^{N} \phi_i(x_n)^2 \cdot r_n \cdot w_i$$
$$= 0$$

Now solve for w_i:

$$\sum_{n=1}^{N} \phi_i(x_n) r_n \left(w_{-i}^T \phi_{-i}(x_n) - t_n\right) + \sum_{n=1}^{N} \phi_i(x_n)^2 \cdot r_n \cdot w_i = 0$$
$$\Leftrightarrow \quad w_i \sum_{n=1}^{N} \phi_i(x_n)^2 \cdot r_n = \sum_{n=1}^{N} \phi_i(x_n) r_n \left(t_n - w_{-i}^T \phi_{-i}(x_n)\right)$$
$$\Leftrightarrow \quad w_i = \frac{\sum_{n=1}^{N} \phi_i(x_n) r_n \left(t_n - w_{-i}^T \phi_{-i}(x_n)\right)}{\sum_{n=1}^{N} \phi_i(x_n)^2 \cdot r_n}$$

It appears that w_i is dependent on the other w_{-i}! Are we now stuck? Look at the mean calculation in exercise one, where we converted our equation into a matrix. Is it possible here?

To answer that what do you think about the following decomposition:

$$\sum_{n=1}^{N} \phi_i(x_n) r_n (w^T \phi(x_n) - t_n) = \left(r_1 \phi_i(x_1), \ldots., r_n \phi_i(x_n)\right)[(\mathbf{w}^T \mathbf{\Phi}^T)^T - \mathbf{t}].$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(x_1) & \ldots & \phi_n(x_1) \\ \ldots & \ldots & \ldots \\ \phi_1(x_n) & \ldots & \phi_n(x_n) \end{pmatrix}$$

Is the design matrix.

If this decomposition is correct, this would lead to:

$$\frac{\partial SE(\mathbf{w})}{\partial \mathbf{w}} = \begin{pmatrix} r_1 \phi_1(x_1) & \ldots & r_n \phi_1(x_n) \\ \ldots & \ldots & \ldots \\ r_1 \phi_d(x_1) & \ldots & \ldots r_n \phi_d(x_n) \end{pmatrix} [(\mathbf{w}^T \mathbf{\Phi}^T)^T - \mathbf{t}] = \mathbf{0}.$$

Can we solve this for **w?** Hint: It might be possible to use the Penrose Pseudoinverse if a matrix is originally not invertible.