Appendix B

Solutions

Ex. 1 Gram Schmidt

1. Two vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ of $\mathbb{R}^n$ are orthogonal if their inner product equals zero. If either one of the vectors is a zero vector, the vectors are trivially orthogonal to each other. Hence, assume that both $\mathbf{u}_1$ and $\mathbf{u}_2$ are nonzero. Because

$$\begin{aligned}\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_2 &= \mathbf{u}_1^{\mathrm{T}}(\mathbf{a}_2 - \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_1) \\ &= \mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2 - \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 \\ &= \mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2 - \mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2 \\ &= 0,\end{aligned}$$

the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal.

Let $\mathbf{v}$ be a linear combination of $\mathbf{a}_1$ and $\mathbf{a}_2$, i.e. $\mathbf{v} = \alpha\mathbf{a}_1 + \beta\mathbf{a}_2$ for some real numbers α and β. Since $\mathbf{u}_1$ and $\mathbf{u}_2$ were defined in terms of $\mathbf{a}_1$ and $\mathbf{a}_2$, we can write $\mathbf{v}$ as

$$\begin{aligned}\mathbf{v} &= \alpha\mathbf{a}_1 + \beta\mathbf{a}_2 \\ &= \alpha\mathbf{u}_1 + \beta(\mathbf{u}_2 + \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_1) \\ &= \alpha\mathbf{u}_1 + \beta\mathbf{u}_2 + \beta\frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_1 \\ &= (\alpha + \beta\frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1})\mathbf{u}_1 + \beta\mathbf{u}_2,\end{aligned}$$

where $\alpha + \beta((\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_2)/(\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1))$ and β are real numbers, so $\mathbf{v}$ can be written in terms of $\mathbf{u}_1$ and $\mathbf{u}_2$.

2. Recall that when showing things by induction, one has first to show that the claim holds for the first possible value, and then, that, if the claim holds for some value k, it holds also for value $k+1$. Since orthogonality is a property of two vectors, the first possible value is two, which was proved in part 1 of this exercise. So we assume that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ are orthogonal vectors and prove that, given this assumption, the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}$ are orthogonal as well:

$$\mathbf{u}_{k+1} = \mathbf{a}_{k+1} - \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_1 - \frac{\mathbf{u}_2^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_2}\mathbf{u}_2 - \ldots - \frac{\mathbf{u}_k^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k}\mathbf{u}_k, \tag{B.1}$$

and for all $i = 1, 2, \ldots, k$

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_{k+1} = \mathbf{u}_i^{\mathrm{T}}\mathbf{a}_{k+1} - \frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_1 - \ldots - \frac{\mathbf{u}_k^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_k.$$

By assumption $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 0$ if $i \neq j$, so

$$\begin{aligned}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_{k+1} &= \mathbf{u}_i^{\mathrm{T}}\mathbf{a}_{k+1} - 0 - \ldots - \frac{\mathbf{u}_i^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i \ldots - 0 \\ &= \mathbf{u}_i^{\mathrm{T}}\mathbf{a}_{k+1} - \mathbf{u}_i^{\mathrm{T}}\mathbf{a}_{k+1} \\ &= 0,\end{aligned}$$

which proves the claim. Notice that if vector $\mathbf{a}_j$ is a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{j-1}$, the vector $\mathbf{u}_j$ given by Gram Schmidt process is zero vector and the process cannot be applied for the other vectors $\mathbf{a}_{j+1}, \mathbf{a}_{j+2}, \ldots, \mathbf{a}_k$. If $k > n$, the vector $\mathbf{a}_{n+1}$ is automatically a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, since there cannot exist a set of $k > n$ linearly independent vectors. On the other hand, if $k = n$ and $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ are linearly independent, Gram Schmidt process gives an orthogonal basis of $\mathbb{R}^n$.

For the other part, notice that the case of one vector is trivial and the case of two vectors was prover in part 1 of the exercise. Using induction, we assume that the claim holds for k vectors and we will prove it for $k+1$ vectors: Let $\mathbf{v}$ be a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k+1}$, ie. $\mathbf{v} = \alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots + \alpha_k\mathbf{a}_k + \alpha_{k+1}\mathbf{a}_{k+1}$ for some real numbers $\alpha_1, \alpha_2, \ldots, \alpha_{k+1}$. Using the induction assumption, $\mathbf{v}$ can be written as

$$\mathbf{v} = \beta_1\mathbf{u}_1 + \beta_2\mathbf{u}_2 + \ldots + \beta_k\mathbf{u}_k + \alpha_{k+1}\mathbf{a}_{k+1},$$

for some real numbers $\beta_1, \beta_2, \ldots, \beta_k$ Furthermore, using equation $(B.1)$, $\mathbf{v}$ can be written as

$$\begin{aligned}\mathbf{v} = \beta_1\mathbf{u}_1 &+ \ldots + \beta_k\mathbf{u}_k + \alpha_{k+1}\mathbf{u}_{k+1} + \alpha_{k+1}\frac{\mathbf{u}_1^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1}\mathbf{u}_1 \\ &+ \ldots + \alpha_{k+1}\frac{\mathbf{u}_k^{\mathrm{T}}\mathbf{a}_{k+1}}{\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k}\mathbf{u}_k.\end{aligned}$$

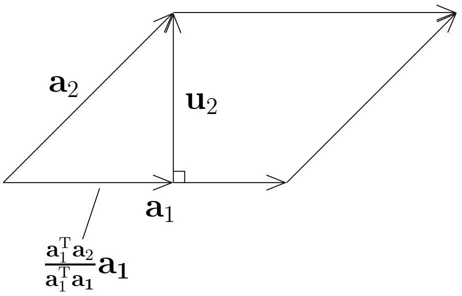
Denoting $\gamma_i = \beta_i + \alpha_{k+1}(\mathbf{u}_i^{\mathrm{T}}\mathbf{a}_{k+1})/(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i)$, $\mathbf{v}$ is then

$$\mathbf{v} = \gamma_1\mathbf{u}_1 + \gamma_2\mathbf{u}_2 + \ldots + \gamma_k\mathbf{u}_k + \alpha_{k+1}\mathbf{u}_{k+1},$$

which proves the claim.

Ex. 2 Linear Transforms

1. Let $\mathbf{a}_1$ and $\mathbf{a}_2$ be vectors that span a parallelogram. From geometry we know that the area of parallelogram is base times height, which is equivalent to length of the base vector times length of the height vector. Denote this by $S^2 = ||\mathbf{a}_1||^2||\mathbf{u}_2||^2$, where is $\mathbf{a}_1$ is the base vector and $\mathbf{u}_2$ is the height vector which is orthogonal to the base vector. Using Gram Schmidt process for vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ in that order, we get the vector $\mathbf{u}_2$ as the second output.


$\mathbf{a}_2$
$\mathbf{u}_2$
$\mathbf{a}_1$
$\frac{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}\mathbf{a}_1$

Therefore $||\mathbf{u}_2||^2$ can be written as

$$\begin{aligned}||\mathbf{u}_2||^2 &= \mathbf{u}_2^{\mathrm{T}}\mathbf{u}_2 \\ &= \left(\mathbf{a}_2 - \frac{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}\mathbf{a}_1\right)^{\mathrm{T}}\left(\mathbf{a}_2 - \frac{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}\mathbf{a}_1\right) \\ &= \mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - \frac{(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1} - \frac{(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1} + \left(\frac{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}\right)^2\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 \\ &= \mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - \frac{(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}.\end{aligned}$$

Thus, S^2 is:

$$\begin{aligned}S^2 &= ||\mathbf{a}_1||^2||\mathbf{u}_2||^2 \\ &= (\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1)(\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_2) \\ &= (\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1)\left(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - \frac{(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2}{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}\right) \\ &= (\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2)(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1) - (\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2.\end{aligned}$$

2. As requested, let

$$A = \begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22}\end{pmatrix}.$$

The determinant of A is $|A| = a_{11}a_{22} - a_{12}a_{21}$. By multiplying out $(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2)$, $(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1)$ and $(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2$, we get

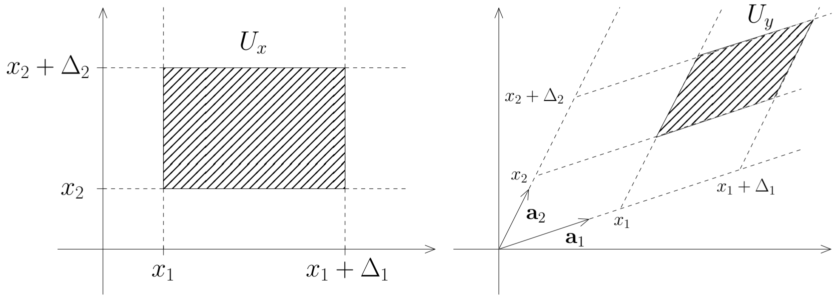
$$\begin{aligned} \mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 &= a_{12}^2 + a_{22}^2 \\ \mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 &= a_{11}^2 + a_{21}^2 \\ (\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2 &= (a_{11}a_{12} + a_{21}a_{22})^2 = a_{11}^2a_{12}^2 + a_{21}^2a_{22}^2 + 2a_{11}a_{12}a_{21}a_{22}. \end{aligned}$$

Therefore the area equals

$$\begin{aligned} S^2 &= (a_{12}^2 + a_{22}^2)(a_{11}^2 + a_{21}^2) - (\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2)^2 \\ &= a_{12}^2a_{11}^2 + a_{12}^2a_{21}^2 + a_{22}^2a_{11}^2 + a_{22}^2a_{21}^2 - \\ &\quad (a_{12}^2a_{11}^2 + a_{21}^2a_{22}^2 + 2a_{11}a_{12}a_{21}a_{22}) \\ &= a_{12}^2a_{21}^2 + a_{22}^2a_{11}^2 + 2a_{11}a_{12}a_{21}a_{22} \\ &= (a_{11}a_{22} - a_{12}a_{21})^2, \end{aligned}$$

which equals $|A|^2$.

3. Denote $A = (\mathbf{a}_1\ \mathbf{a}_2)$. A rectangle with the same area as U_x is spanned by vectors $(\Delta_1\ 0)^{\mathrm{T}}$ and $(0\ \Delta_2)^{\mathrm{T}}$. Under the linear transform A these spanning vectors become $\Delta_1\mathbf{a}_1$ and $\Delta_2\mathbf{a}_2$. Therefore a parallelogram with the same area as U_y is spanned by $\Delta_1\mathbf{a}_1$ and $\Delta_2\mathbf{a}_2$.



Using the part 2 of this exercise it is easy to calculate the area A_{U_y} of U_y as the determinant of $(\Delta_1\mathbf{a}_1\ \Delta_2\mathbf{a}_2)$:

$$\begin{aligned} A_{U_y} &= \det\begin{pmatrix}\Delta_1 a_{11} & \Delta_2 a_{12} \\ \Delta_1 a_{21} & \Delta_2 a_{22}\end{pmatrix} \\ &= \Delta_1\Delta_2 a_{11}a_{22} - \Delta_1\Delta_2 a_{12}a_{21} \\ &= \Delta_1\Delta_2(a_{11}a_{22} - a_{12}a_{21}) \\ &= \Delta_1\Delta_2|A| \end{aligned}$$

Therefore the area of U_y is the area of U_x times $|A|$.

4. This explanation is intuitive and doesn't contain any real proof. On the left hand side of the change of variables formula, $\mathbf{y}$ runs over the region U_y. On the right hand side of the formula, $\mathbf{x}$ runs over the set $U_x = A^{-1}(U_y)$ so that $A\mathbf{x}$ runs also over U_y. Hence, the functions on both sides of the equations take on the same values. In part 3 of this exercise it was shown that the area of U_y is the area of U_x times $|A|$. Hence, we need to use the compensating factor $|A|$ when we integrate over $A^{-1}(U_y)$.

Ex. 3 Eigenvalue Decomposition

This exercise is very important and will be used many times in the future exercises.

1. Let A be an $n \times n$ matrix and $\mathbf{u}_1$ and $\mathbf{u}_2$ be such vectors of $\mathbb{R}^n$ that $A\mathbf{u}_1 = \lambda\mathbf{u}_1$ and $A\mathbf{u}_2 = \lambda\mathbf{u}_2$ for some real number λ. Denote $\mathbf{u} = \alpha\mathbf{u}_1 + \beta\mathbf{u}_2$, where α and β are arbitrary real numbers. Now

$$\begin{aligned} A\mathbf{u} &= \alpha A\mathbf{u}_1 + \beta A\mathbf{u}_2 \\ &= \alpha\lambda\mathbf{u}_1 + \beta\lambda\mathbf{u}_2 \\ &= \lambda(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) \\ &= \lambda\mathbf{u}, \end{aligned}$$

so $\mathbf{u}$ is an eigenvector of A with the same eigenvalue as $\mathbf{u}_1$ and $\mathbf{u}_2$.

2. Let Λ be a diagonal $n \times n$ matrix with the eigenvalues of A on its diagonal: The elements of Λ are zero everywhere but on its diagonal where the elements are the $\lambda_1, \lambda_2, \dots, \lambda_n,$. Denote by U the matrix with has as column vectors the linearly independent eigenvectors of A: $U = (\mathbf{u}_1\ \mathbf{u}_2\ \dots\ \mathbf{u}_n)$. Using $A\mathbf{u}_i = \lambda_i \mathbf{u}_i$ for all $i = 1, 2, \dots, n$, and the properties of matrix multiplication we have

$$AU = (A\mathbf{u}_1\ A\mathbf{u}_2\ \dots\ A\mathbf{u}_n) = (\lambda_1\mathbf{u}_1\ \lambda_2\mathbf{u}_2\ \dots\ \lambda_n\mathbf{u}_n) = U\Lambda.$$

3. (i) Since the columns of U are linearly independent, U is invertible. Because $AU = U\Lambda$, multiplying from the right with the inverse of U gives $A = U\Lambda U^{-1} = U\Lambda V^{\mathrm{T}}$.

   (ii) Denote by $\mathbf{u}^{[i]}$ the ith row of U, $\mathbf{v}^{(j)}$ the jth column of V^{T} and $\mathbf{v}^{[j]}$ the jth row of V and denote $B = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$. Let $e^{[i]}$ be a row vector with 1 in the ith place and 0 elsewhere and $e^{(j)}$ be a column vector with 1 in the jth place and 0 elsewhere. Notice that because $A = U\Lambda V^{\mathrm{T}}$, the element in the ith row and jth column is

$$\begin{aligned} A_{ij} &= \mathbf{u}^{[i]}\Lambda v^{(j)} \\ &= \mathbf{u}^{[i]}\Lambda (v^{[j]})^{\mathrm{T}} \\ &= \mathbf{u}^{[i]} \begin{pmatrix} \lambda_1 V_{j1} \\ \vdots \\ \lambda_n V_{jn} \end{pmatrix} \\ &= \sum_{k=1}^{n} \lambda_k V_{jk} U_{ik}. \end{aligned}$$

   On the other hand, for matrix B the element in the ith row and jth column is

$$\begin{aligned} B_{ij} &= \sum_{k=1}^{n} \lambda_k e^{[i]} \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}} e^{(j)} \\ &= \sum_{k=1}^{n} \lambda_k U_{ik} V_{jk}, \end{aligned}$$

   which is the same as A_{ij}. Therefore $A = B$.

   (iii) Since Λ is a diagonal matrix with no zeros as diagonal elements, it is invertible (the inverse matrix of Λ is just a diagonal matrix with inverse elements of the eigenvalues of A as diagonal elements in the same order as in Λ). We have thus

$$\begin{aligned} A^{-1} &= (U\Lambda U^{-1})^{-1} = (U(\Lambda U^{-1}))^{-1} = (\Lambda U^{-1})^{-1} U^{-1} \\ &= (U^{-1})^{-1}\Lambda^{-1}U^{-1} = U\Lambda^{-1}V^{\mathrm{T}}. \end{aligned}$$

   (iv) The calculation is the same as in part 4 of this exercise when the element λ_i is substituted by $1/\lambda_i$ for every $i = 1, 2, \dots, n$.

Ex. 4 Trace, Determinants and Eigenvalues

1. Since $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and $A = U\Lambda U^{-1}$

$$\begin{aligned} \mathrm{Tr}(A) &= \mathrm{Tr}(U\Lambda U^{-1}) = \mathrm{Tr}((U)(\Lambda U^{-1})) \\ &= \mathrm{Tr}(\Lambda U^{-1}U) = \mathrm{Tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i. \end{aligned}$$

2. The determinant of a diagonal matrix is just the product of its diagonal elements and the determinant of an inverse matrix is just the inverse of the determinant of the original matrix. In addition since $\det(AB) = \det(A)\det(B)$, we get

$$\begin{aligned} \det(A) &= \det(U\Lambda U^{-1}) = \det(U)\det(\Lambda)\det(U^{-1}) \\ &= \frac{\det(U)\det(\Lambda)}{\det(U)} = \det(\Lambda) = \prod_{i=1}^{n} \lambda_i. \end{aligned}$$

Ex. 5 Eigenvalue Decomposition for Symmetric Matrices

1. Since $A\mathbf{u}_1 = \lambda_1\mathbf{u}_1$ and $A\mathbf{u}_2 = \lambda_2\mathbf{u}_2$, we also have
$$\mathbf{u}_1^{\mathrm{T}}A\mathbf{u}_2 \quad = \quad \lambda_2\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_2.$$
Taking the transpose of $\mathbf{u}_1^{\mathrm{T}}A\mathbf{u}_2$ gives us
$$(\mathbf{u}_1^{\mathrm{T}}A\mathbf{u}_2)^{\mathrm{T}} = (A\mathbf{u}_2)^{\mathrm{T}}(\mathbf{u}_1^{\mathrm{T}})^{\mathrm{T}} = \mathbf{u}_2^{\mathrm{T}}A^{\mathrm{T}}\mathbf{u}_1 = \mathbf{u}_2^{\mathrm{T}}A\mathbf{u}_1 = \lambda_1\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1$$
because A is symmetric. On the other hand, the same operation gives us
$$(\mathbf{u}_1^{\mathrm{T}}A\mathbf{u}_2)^{\mathrm{T}} = (\lambda_2\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_2)^{\mathrm{T}} = \lambda_2\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1$$
Therefore $\lambda_1\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1 = \lambda_2\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1$, which is equivalent to $\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1(\lambda_1 - \lambda_2) = 0$. Because $\lambda_1 \neq \lambda_2$, the only possibility is that $\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1 = 0$. Therefore $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal to each other.

2. Part 1 of this exercise showed that for distinct eigenvalues the corresponding eigenvectors are orthogonal, so in this case the only thing left to do is normalize each of them by multiplying the vector with the inverse of its norm. If A has some equal eigenvalues, the fact that A is symmetric tells us that there exists a set of n linearly independent eigenvectors. Applying Gram Schmidt to the set of the eigenvectors with the same eigenvalue, we can thus obtain a set of orthogonal vectors. A generalization of part one of exercise 3 (Eigenvalue Decomposition) tells us that the vectors obtained via Gram Schmidt are still eigenvectors. After normalizing the vectors, we have a set of eigenvectors that are orthogonal and of unit length.

3. Assume that $\mathbf{v}^{\mathrm{T}}A\mathbf{v} > 0$ for all $\mathbf{v} \neq 0$. Since eigenvectors are not zero vectors, the assumption holds also for eigenvector $\mathbf{u}_k$ with corresponding eigenvalue λ_k. Now
$$\mathbf{u}_k^{\mathrm{T}}A\mathbf{u}_k = \mathbf{u}_k^{\mathrm{T}}\lambda_k\mathbf{u}_k = \lambda_k(\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k) = \lambda_k||\mathbf{u}_k|| > 0$$
and because $||\mathbf{u}_k|| > 0$, we obtain $\lambda_k > 0$.

   Assume now that all the eigenvalues of A, $\lambda_1, \lambda_2, \ldots, \lambda_n$, are positive and nonzero. In part 2 of this exercise it was shown that there exists an orthogonal basis consisting of eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ and therefore every vector $\mathbf{v}$ can be written as a linear combination of those vectors. Hence for a nonzero vector $\mathbf{v}$ and for some real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ holds that
$$\begin{aligned} & \mathbf{v}^{\mathrm{T}}A\mathbf{v} \\ = \; & (\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_n\mathbf{u}_n)^{\mathrm{T}}A(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_n\mathbf{u}_n) \\ = \; & (\alpha_1\mathbf{u}_1 + \ldots + \alpha_n\mathbf{u}_n)^{\mathrm{T}}(\alpha_1A\mathbf{u}_1 + \ldots + \alpha_nA\mathbf{u}_n) \\ = \; & (\alpha_1\mathbf{u}_1 + \ldots + \alpha_n\mathbf{u}_n)^{\mathrm{T}}(\alpha_1\lambda_1\mathbf{u}_1 + \ldots + \alpha_n\lambda_n\mathbf{u}_n) \\ = \; & \sum_{i,j}\alpha_i\mathbf{u}_i^{\mathrm{T}}\alpha_j\lambda_j\mathbf{u}_j \\ = \; & \sum_i\alpha_i\alpha_i\lambda_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i \\ = \; & \sum_i(\alpha_i)^2||u_i||^2\lambda_i, \end{aligned}$$
because $\mathbf{u}_i^T\mathbf{u}_j = 0$ if $i \neq j$ due to orthogonality of the basis. Since $(\alpha_i)^2 > 0$, $||u_i||^2 > 0$ and $\lambda_i > 0$ for all i, $\mathbf{v}^{\mathrm{T}}A\mathbf{v} > 0$.

   Since every eigenvalue of A is nonzero, we can use the last part of exercise 3 (Eigenvalue Decomposition) to conclude that inverse of A exists.

Ex. 6 Gaussian Distribution

1. The best thing is to start with the cumulative distribution function (cdf) $F(y) = P(Y \leq y)$. Assume that Y has a density f and therefore equation
$$F(y) = \int_{-\infty}^{y} f(u)\,\mathrm{d}u \tag{B.2}$$
holds. Now $F(y)$ can be written as
$$\begin{aligned} F(y) &= P(Y \leq y) \\ &= P(X_1 + X_2 \leq y) \\ &= \int P(X_1 + X_2 \leq y|X_2 = x)f_2(x)\,\mathrm{d}x \\ &= \int P(X_1 \leq y - x|X_2 = x)f_2(x)\,\mathrm{d}x \\ &= \int P(X_1 \leq y - x)f_2(x)\,\mathrm{d}x, \end{aligned}$$

because X_1 and X_2 are independent. On the other hand

$$P(X_1 \leq y - x) = \int_{-\infty}^{y-x} f_1(u)\,\mathrm{d}u,$$

and therefore

$$\begin{aligned}
F(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{y-x} f_1(u) f_2(f)\,\mathrm{d}u\mathrm{d}x \\
&\overset{\tilde{u}=u+x}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{y} f_1(\tilde{u}-x) f_2(x)\,\mathrm{d}\tilde{u}\mathrm{d}x \\
&= \int_{-\infty}^{y}\left[\int_{-\infty}^{\infty} f_1(\tilde{u}-x) f_2(x)\mathrm{d}x\right]\mathrm{d}\tilde{u}.
\end{aligned}$$

Now comparison with equation $(B.2)$ shows that

$$f(y) = \int_{-\infty}^{\infty} f_1(y-u) f_2(u)\,\mathrm{d}u.$$

2. Denote again the densities of X_1 and X_2 by f_1 and f_2:

$$\begin{aligned}
f_1(x) &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{x^2}{2\sigma_1^2}\right], \\
f_2(x) &= \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{x^2}{2\sigma_2^2}\right].
\end{aligned}$$

Using the formula that was derived in the previous part of this exercise, we get

$$\begin{aligned}
f(y) &= \int f_1(y-u) f_2(u)\,\mathrm{d}u \\
&= \int \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{(y-u)^2}{2\sigma_1^2}\right]\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{u^2}{2\sigma_2^2}\right]\mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}}\frac{1}{\sqrt{2\pi\sigma_2^2}}\int \exp\left[-\frac{(y-u)^2}{2\sigma_1^2}-\frac{u^2}{2\sigma_2^2}\right]\mathrm{d}u.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\frac{(y-u)^2}{2\sigma_1^2}+\frac{u^2}{2\sigma_2^2} \\
&= \frac{y^2}{2\sigma_1^2}+u^2\left(\frac{1}{2\sigma_1^2}+\frac{1}{2\sigma_2^2}\right)-\frac{2uy}{2\sigma_1^2} \\
&= u^2\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}-2u\frac{y\sigma_2^2}{2\sigma_1^2\sigma_2^2}+\frac{y^2}{2\sigma_1^2} \\
&= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(u^2-2u\frac{y\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)+\frac{y^2}{2\sigma_1^2} \\
&= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(u-\frac{y\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2-\frac{y^2\sigma_2^2}{2(\sigma_1^2+\sigma_2^2)\sigma_1^2}+\frac{y^2(\sigma_1^2+\sigma_2^2)}{2(\sigma_1^2+\sigma_2^2)\sigma_1^2} \\
&= \frac{\left(u-\frac{y\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}+\frac{y^2}{2(\sigma_1^2+\sigma_2^2)},
\end{aligned}$$

Hence, we can write $f(y)$ as follows:

$$\begin{aligned}
& \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \int \exp\left[-\frac{(y-u)^2}{2\sigma_1^2} - \frac{u^2}{2\sigma_2^2}\right] \mathrm{d}u \\
= & \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} \cdot \\
& \int \exp\left[-\frac{\left(u - \frac{y\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} - \frac{y^2}{2(\sigma_1^2+\sigma_2^2)}\right] \\
= & \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \underbrace{\frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} \int \exp\left[-\frac{\left(u - \frac{y\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}\right] \mathrm{d}u \cdot}_{=1} \\
& \exp\left[-\frac{y^2}{2(\sigma_1^2+\sigma_2^2)}\right] \\
= & \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left[-\frac{y^2}{2(\sigma_1^2+\sigma_2^2)}\right].
\end{aligned}$$

The last lines defines a Gaussian random variable with variance $\sigma_1^2 + \sigma_2^2$, which completes the proof.

3. Recall that if X is a Gaussian random variable with mean μ and variance σ^2, then for any constant c holds that $X + c$ is a Gaussian random variable with mean $\mu + c$ and variance σ^2. Now we can write $X_1 = \tilde{X_1} + \mu_1$ and $X_2 = \tilde{X_2} + \mu_2$, where $\tilde{X_i} \sim N(0, \sigma_i^2)$ for $i = 1, 2$, and furthermore

$$Y = X_1 + X_2 = \tilde{X_1} + \tilde{X_2} + \mu_1 + \mu_2.$$

In part 2 of this exercise it was shown that $\tilde{X_1} + \tilde{X_2}$ is a Gaussian random variable with zero mean and variance $\sigma_1^2 + \sigma_2^2$. Thus $Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Ex. 7 Bivariate Gaussian

1. Recall that if A is a 2×2 matrix with $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then the inverse of A is

$$A^{-1} = 1/\det(A) \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Since the determinant of Σ is

$$|\Sigma| = \sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_2^2\rho^2 = \sigma_1^2\sigma_2^2(1-\rho^2),$$

the inverse of Σ is

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_1\sigma_2\rho \\ -\sigma_1\sigma_2\rho & \sigma_1^2 \end{pmatrix}.$$

Therefore

$$\begin{aligned}
\mathbf{x}^T\Sigma^{-1}\mathbf{x} &= (x_1\ x_2)\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_1\sigma_2\rho \\ -\sigma_1\sigma_2\rho & \sigma_1^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}(x_1\ x_2) \begin{pmatrix} \sigma_2^2 x_1 - \sigma_1\sigma_2\rho x_2 \\ -\sigma_1\sigma_2\rho x_1 + x_2\sigma_1^2 \end{pmatrix} \\
&= \frac{\sigma_2^2 x_1^2 - \sigma_1\sigma_2\rho x_1 x_2 - \sigma_1\sigma_2\rho x_1 x_2 + x_2^2\sigma_1^2}{\sigma_1^2\sigma_2^2(1-\rho^2)} \\
&= \frac{1}{1-\rho^2}\left(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right) \\
&= q(x_1, x_2)
\end{aligned}$$

and thus

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left[-\frac{1}{2}q(x_1-\mu_1, x_2-\mu_2)\right].$$

2. Write first $q(x_1, x_2)$ in a different form:

$$\begin{aligned}
& q(x_1, x_2) \\
= & \frac{1}{1-\rho^2}\left(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right) \\
= & \frac{1}{1-\rho^2}\left(\frac{x_2^2}{\sigma_2^2} - 2\rho\frac{x_1 x_2}{\sigma_1\sigma_2} + \rho^2\left(\frac{x_1}{\sigma_1}\right)^2 - \rho^2\left(\frac{x_1}{\sigma_1}\right)^2 + \frac{x_1^2}{\sigma_1^2}\right) \\
= & \frac{1}{1-\rho^2}\left(\left(\frac{x_2}{\sigma_2} - \rho\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_1}{\sigma_1}\right)^2(1-\rho^2)\right) \\
= & \frac{1}{1-\rho^2}\left(\frac{x_2}{\sigma_2} - \rho\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_1}{\sigma_1}\right)^2.
\end{aligned}$$

The marginal distribution $f(x_1)$ is

$$\begin{aligned}
f(x_1) &= \int f(x_1, x_2)\mathrm{d}x_2 \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}\int \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]\mathrm{d}x_2 \\
&\quad \exp\left[-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right] \\
&= \underbrace{\frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}}\int \exp\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)^2\right]\mathrm{d}x_2}_{=1} \\
&\quad \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right],
\end{aligned}$$

which defines a Gaussian random variable with variance $V(x_1) = \sigma_1^2$.

3. The covariance $\mathrm{cov}(x_1, x_2)$ is

$$\begin{aligned}
\mathrm{cov}(x_1, x_2) &= E((x_1-\mu_1)(x_2-\mu_2)) \\
&= \int\int (x_1-\mu_1)(x_2-\mu_2)\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \\
&\quad \exp\left[-\frac{1}{2}q(x_1-\mu_1, x_2-\mu_2)\right]\mathrm{d}x_1\mathrm{d}x_2 \\
&= \int\int x_1 x_2 \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}\exp\left[-\frac{1}{2}q(x_1, x_2)\right]\mathrm{d}x_1\mathrm{d}x_2 \\
&\overset{\text{part 2}}{=} \int x_1\left[\underbrace{\int x_2\frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}}\exp\left[-\frac{\left(x_2-\rho\frac{\sigma_2}{\sigma_1}x_1\right)^2}{2\sigma_2^2(1-\rho^2)}\right]\mathrm{d}x_2}_{=\text{ mean of a Gaussian pdf}}\right] \\
&\quad \exp\left[-\frac{x_1^2}{\sigma_1^2}\right]\mathrm{d}x_1 \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}}
\end{aligned}$$

The mean that is marked in the previous equation chain is $(\rho\sigma_2 x_1)/\sigma_1$, which can be seen by comparison with the common form of Gaussian density function. Hence

$$\begin{aligned}
\mathrm{cov}(x_1, x_2) &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\int \rho\frac{\sigma_2}{\sigma_1}x_1^2\exp\left[-\frac{x_1^2}{2\sigma_1^2}\right]\mathrm{d}x_1 \\
&= \rho\frac{\sigma_2}{\sigma_1}\underbrace{\frac{1}{\sqrt{2\pi\sigma_1^2}}\int x_1^2\exp\left[-\frac{x_1^2}{\sigma_1^2}\right]\mathrm{d}x_1}_{\text{the variance } V(x_1)=\sigma_1^2} \\
&= \rho\frac{\sigma_2}{\sigma_1}\sigma_1^2 \\
&= \rho\sigma_1\sigma_2.
\end{aligned}$$

4. Using the part 2 of this exercise it holds that

$$\begin{aligned} -q(x_1,x_2)+\frac{x_1^2}{\sigma_1^2} &= -\frac{1}{1-\rho^2}\left(\frac{x_2}{\sigma_2}-\rho\frac{x_1}{\sigma_1}\right)^2-\frac{x_1^2}{\sigma_1^2}+\frac{x_1^2}{\sigma_1^2} \\ &= -\frac{1}{\sigma_2^2(1-\rho^2)}\left(x_2-\rho\frac{x_1\sigma_2}{\sigma_1}\right)^2. \end{aligned}$$

Therefore

$$\begin{aligned} f(x_2|x_1) &= \frac{f(x_1,x_2)}{f(x_1)} \\ &= \frac{\sqrt{2\pi\sigma_1^2}}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}\cdot \\ &\quad \exp\left[-\frac{1}{2}q(x_1-\mu_1,x_2-\mu_2)+\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}}\cdot \\ &\quad \exp\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)^2\right]. \end{aligned}$$

Hence, $f(x_2|x_1)$ is a Gaussian probability density function with

$$\begin{aligned} \text{mean}: \mu_{x_2|x_1} &= \mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1) \text{ and} \\ \text{variance}: \sigma_{x_2|x_1}^2 &= \sigma_2^2(1-\rho^2). \end{aligned}$$

5. If $\rho=0$, then $\mu_{x_2|x_1}=\mu_2$ and $\sigma_{x_2|x_1}^2=\sigma_2^2$. Hence, if $\text{cov}(x_1,x_2)=\rho\sigma_1\sigma_2=0$, then x_1 and x_2 are also independent. Note that this isn't generally true for random variables.

6. Assume that f_x is the density function of X with mean vector μ_x and covariance matrix Σ_x and f_y is the density function of Y. Let U_y be a rectangle and $U_x=H^{-1}U_y$. Integrating over U_x may be hard, so making a change of variables (see exercise 2: Linear Transforms) to integrate over the rectangle U_y gives us

$$\begin{aligned} \int_{U_y} f_y(\mathbf{y})\mathrm{d}\mathbf{y} &= P(\mathbf{y}\in U_y) \\ &= P(\mathbf{x}\in H^{-1}(U_y)) \\ &= \int_{U_x=H^{-1}(U_y)} f_{\mathbf{x}}\mathrm{d}\mathbf{x} \\ &= \int_{U_y} f_x(H^{-1}\mathbf{y})|H^{-1}|\mathrm{d}\mathbf{y}. \end{aligned}$$

Hence,

$$\begin{aligned} f_y(\mathbf{y}) &= f_x(H^{-1}\mathbf{y})|H^{-1}| \\ &= f_x(H^{-1}\mathbf{y})\frac{1}{|H|} \\ &= \frac{1}{2\pi|\Sigma_x|^{1/2}|H|}\exp\left[(H^{-1}\mathbf{y}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(H^{-1}\mathbf{y}-\mu_x)\right]. \end{aligned}$$

Therefore $\mu_y=H\mu_x$, and because

$$\begin{aligned} & (H^{-1}\mathbf{y}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(H^{-1}\mathbf{y}-\mu_x) \\ =\; & (\mathbf{y}-H\mu_x)^{\mathrm{T}}(H^{\mathrm{T}})^{-1}\Sigma_x^{-1}H^{-1}(\mathbf{y}-H\mu_x), \end{aligned}$$

we have $\Sigma_y^{-1}=(H^{\mathrm{T}})^{-1}\Sigma_x^{-1}H^{-1}$, or $\Sigma_y=H\Sigma_xH^{\mathrm{T}}$.

Ex. 8 Maximum Likelihood for a Gaussian

1. The likelihood $L(\mu,\sigma)$ is the joint density of the data, treated as a function of the parameters μ and σ. Because of the iid assumption of the data, the likelihood is the product of the probability density functions $f(X_i;\mu,\sigma)$. The random variable being Gaussian, the likelihood is

$$\begin{aligned} L(\mu,\sigma) &= \prod_i^N f(X_i;\mu,\sigma) \\ &= \prod_i^N \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(X_i-\mu)^2}{2\sigma^2} \right] \right) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (X_i-\mu)^2 \right]. \end{aligned}$$

2. To make the answer to the third part of the exercise simpler, write the sum in the first part differently:

$$\begin{aligned} \sum_{i=1}^N (X_i-\mu)^2 &= \sum_{i=1}^N (X_i-\bar{X}+\bar{X}-\mu)^2 \\ &= \sum_{i=1}^N [(X_i-\bar{X})^2 + 2(X_i-\bar{X})(\bar{X}-\mu) + (\bar{X}-\mu)^2] \\ &= \sum_{i=1}^N (X_i-\bar{X})^2 + 2(\bar{X}-\mu)\underbrace{\sum_{i=1}^N (X_i-\bar{X})}_{=0} + \\ & \quad N(\bar{X}-\mu)^2 \\ &= \sum_{i=1}^N (X_i-\bar{X})^2 + N(\bar{X}-\mu)^2 \\ &= NS^2 + N(\bar{X}-\mu)^2, \end{aligned}$$

where $\bar{X}$ is the sample mean and S^2 is the sample variance. Using this form of the sum, the log-likelihood is

$$\begin{aligned} \ell(\mu,\sigma) &= \log L(\mu,\sigma) \\ &= \log\left(\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{NS^2 + N(\bar{X}-\mu)^2}{2\sigma^2} \right] \right) \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{N}{2\sigma^2}S^2 - \frac{N}{2\sigma^2}(\bar{X}-\mu)^2. \end{aligned}$$

Using the original form of the sum, the log-likelihood would be

$$\begin{aligned} \ell(\mu,\sigma) &= \log L(\mu,\sigma) \\ &= \log\left(\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\sum_{i=1}^N \frac{(X_i-\mu)^2}{2\sigma^2} \right] \right) \\ &= \log\left(\frac{1}{(2\pi\sigma^2)^{N/2}} \right) + \log\left(\exp\left[-\sum_{i=1}^N \frac{(X_i-\mu)^2}{2\sigma^2} \right] \right) \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^N \frac{(X_i-\mu)^2}{2\sigma^2}. \end{aligned}$$

3. Calculating the maximum likelihood estimate means calculating the value of the parameter that maximizes the likelihood. Because the logarithm function is strictly monotonically increasing, it doesn't change the argument where the likelihood is maximized. Since finding the maximum likelihood estimates for the log-likelihood function is simpler than finding them for the original likelihood function, it is convenient to start with the log-likelihood function. Finding the estimates is done by taking derivatives with respect to μ and σ

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{N(\bar{X}-\mu)}{\sigma^2} \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{N}{\sigma} + \frac{NS^2}{\sigma^3} + \frac{N(\bar{X}-\mu)^2}{\sigma^3}. \end{aligned}$$

Now, the only value of μ which makes the derivative with respect to μ zero is the sample mean $\bar{X}$. Setting μ to $\bar{X}$ in the derivative with respect to σ, we get

$$-\frac{N}{\sigma} + \frac{NS^2}{\sigma^3} = 0 \iff \frac{1}{\sigma} = \frac{S^2}{\sigma^3} \iff \sigma^2 = S^2.$$

This is the only value of σ^2 that makes the derivative zero. Furthermore, it is the global maximum we were looking for (this can be shown by calculating second derivatives). Therefore the maximum likelihood estimates are $\bar{X}$ and S^2.

Ex. 9 Gradient of vector-valued functions

This exercise is useful for the later exercise sessions, so remember it if you have to calculate gradients in later exercises.

1. Gradient of $J(\mathbf{w}) = \mathbf{a}^{\mathrm{T}}\mathbf{w}$ using both methods.

   1. First method

$$\begin{aligned} J(\mathbf{w}) &= \mathbf{a}^{\mathrm{T}}\mathbf{w} = \sum_{i=1}^{n} a_i w_i \\ \Longrightarrow \frac{\partial J(\mathbf{w})}{\partial w_i} &= a_i \\ \Longrightarrow \nabla J(\mathbf{w}) &= \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \mathbf{a}. \end{aligned}$$

   2. Second method

$$\begin{aligned} J(\mathbf{w}+\epsilon\mathbf{h}) &= \mathbf{a}^{\mathrm{T}}(\mathbf{w}+\epsilon\mathbf{h}) = \underbrace{\mathbf{a}^{\mathrm{T}}\mathbf{w}}_{J(\mathbf{w})} + \epsilon \underbrace{\mathbf{a}^{\mathrm{T}}}_{\nabla J(\mathbf{w})^{\mathrm{T}}} \mathbf{h} \\ \Longrightarrow \nabla J(\mathbf{w}) &= \mathbf{a}. \end{aligned}$$

2. Gradient of $J(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}A\mathbf{w}$ using both methods. Notice that for every real number x holds that x equals its own transpose.

   1. First method

$$\begin{aligned} J(\mathbf{w}) &= \mathbf{w}^{\mathrm{T}}A\mathbf{w} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i A_{ij} w_j \\ \Longrightarrow \frac{\partial J(\mathbf{w})}{\partial w_k} &= \sum_{j=1}^{n} A_{kj} w_j + \sum_{i=1}^{n} w_i A_{ik} \\ \Longrightarrow \nabla J(\mathbf{w}) &= \begin{pmatrix} \sum_{j=1}^{n} A_{1j} w_j + \sum_{i=1}^{n} w_i A_{i1} \\ \vdots \\ \sum_{j=1}^{n} A_{nj} w_j + \sum_{i=1}^{n} w_i A_{in} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^{n} A_{1j} w_j \\ \vdots \\ \sum_{j=1}^{n} A_{nj} w_j \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{n} w_i A_{i1} \\ \vdots \\ \sum_{i=1}^{n} w_i A_{in} \end{pmatrix} \\ &= \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} + \\ & \quad \begin{pmatrix} A_{11} & \dots & A_{n1} \\ \vdots & \ddots & \vdots \\ A_{1n} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \\ &= A\mathbf{w} + A^{\mathrm{T}}\mathbf{w} \end{aligned}$$

   2. Second method

$$\begin{aligned} J(\mathbf{w}+\epsilon\mathbf{h}) &= (\mathbf{w}+\epsilon\mathbf{h})^{\mathrm{T}}A(\mathbf{w}+\epsilon\mathbf{h}) \\ &= \mathbf{w}^{\mathrm{T}}A\mathbf{w} + \mathbf{w}^{\mathrm{T}}A(\epsilon\mathbf{h}) + \epsilon\mathbf{h}^{\mathrm{T}}A\mathbf{w} + \underbrace{(\epsilon\mathbf{h})^{\mathrm{T}}A\epsilon\mathbf{h}}_{=O(\epsilon^2)} \\ &= \mathbf{w}^{\mathrm{T}}A\mathbf{w} + \epsilon(\mathbf{w}^{\mathrm{T}}A\mathbf{h} + \underbrace{(\mathbf{w}^{\mathrm{T}}A^{\mathrm{T}}\mathbf{h})}_{\in\mathbb{R}}{}^{\mathrm{T}}) + O(\epsilon^2) \\ &= \mathbf{w}^{\mathrm{T}}A\mathbf{w} + \epsilon(\mathbf{w}^{\mathrm{T}}A\mathbf{h} + \mathbf{w}^{\mathrm{T}}A^{\mathrm{T}}\mathbf{h}) + O(\epsilon^2) \\ &= \underbrace{\mathbf{w}^{\mathrm{T}}A\mathbf{w}}_{=J(\mathbf{w})} + \epsilon\underbrace{(\mathbf{w}^{\mathrm{T}}A + \mathbf{w}^{\mathrm{T}}A^{\mathrm{T}})}_{=\nabla J(\mathbf{w})^{\mathrm{T}}}\mathbf{h} + O(\epsilon^2) \\ \Longrightarrow \nabla J(\mathbf{w}) &= (\mathbf{w}^{\mathrm{T}}A + \mathbf{w}^{\mathrm{T}}A^{\mathrm{T}})^{\mathrm{T}} = A^{\mathrm{T}}\mathbf{w} + A\mathbf{w}. \end{aligned}$$

3. The easiest way to calculate gradient for $J(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{w}$ is to use the previous part of this exercise by choosing $A = I$ (the identity matrix). Therefore

$$\nabla J(\mathbf{w}) = I\mathbf{w} + I^{\mathrm{T}}\mathbf{w} = \mathbf{w} + \mathbf{w} = 2\mathbf{w}.$$

4. Recall the chain rule for gradients:

$$\nabla(h \circ f)(x) = h'(f(x))\nabla f(x).$$

Using the chain rule and the part 2 of this exercise we can calculate the gradient for $J(\mathbf{w}) = ||\mathbf{w}|| = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$:

$$\nabla J(\mathbf{w}) \overset{\text{chain rule}}{=} \frac{\nabla(\mathbf{w}^{\mathrm{T}}\mathbf{w})}{2\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}} \overset{\text{part 2}}{=} \frac{2\mathbf{w}}{2||\mathbf{w}||} = \frac{\mathbf{w}}{||\mathbf{w}||}.$$

Using the chain rule is, of course, just a shortcut. Calculating the partial derivatives without the chain rule is here also possible without much more work.

5. Using the chain rule and previous part of this exercise we can calculate the gradient for $J(\mathbf{w}) = f(||\mathbf{w}||)$:

$$\nabla J(\mathbf{w}) \overset{\text{chain rule}}{=} f'(||\mathbf{w}||)\nabla||\mathbf{w}|| \overset{\text{previous part}}{=} f'(||\mathbf{w}||)\frac{\mathbf{w}}{||\mathbf{w}||}.$$

6. Using the chain rule and first part of this exercise we can calculate the gradient for $J(\mathbf{w}) = f(\mathbf{w}^{\mathrm{T}}\mathbf{a})$:

$$\begin{aligned}\nabla J(\mathbf{w}) &\overset{\text{chain rule}}{=} f'(\mathbf{w}^{\mathrm{T}}\mathbf{a})\nabla(\mathbf{w}^{\mathrm{T}}\mathbf{a}) \\ &\overset{\mathbf{w}^{\mathrm{T}}\mathbf{a} \in \mathbb{R}}{=} f'(\mathbf{w}^{\mathrm{T}}\mathbf{a})\nabla(\mathbf{a}^{\mathrm{T}}\mathbf{w}) \\ &\overset{\text{part 1}}{=} f'(\mathbf{w}^{\mathrm{T}}\mathbf{a})\mathbf{a}.\end{aligned}$$

Ex. 10 Newton's method

1. First, we write the function f in a different form:

$$\begin{aligned} f(\mathbf{w}) &= f(\mathbf{w}_0) + \mathbf{g}^{\mathrm{T}}(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T H(\mathbf{w} - \mathbf{w}_0) \\ &= \underbrace{f(\mathbf{w}_0) - \mathbf{g}^{\mathrm{T}}\mathbf{w}_0 + \frac{1}{2}\mathbf{w}_0^{\mathrm{T}} H\mathbf{w}_0}_{= c \text{ (constant)}} + \mathbf{g}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{w}^{\mathrm{T}} H\mathbf{w} \\ &\quad -\frac{1}{2}\mathbf{w}^{\mathrm{T}} H\mathbf{w}_0 - \underbrace{\frac{1}{2}\mathbf{w}_0^{\mathrm{T}} H\mathbf{w}}_{\in \mathbb{R}} \\ &= c + \mathbf{g}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{w}^{\mathrm{T}} H\mathbf{w} - \mathbf{w}^{\mathrm{T}} H\mathbf{w}_0. \end{aligned}$$

Using the exercise 9 (Gradient of vector-valued functions) we can calculate the gradient (remember that H is symmetric):

$$\begin{aligned}\nabla f(\mathbf{w}) &= \mathbf{g} + \frac{1}{2}(H^{\mathrm{T}}\mathbf{w} + H\mathbf{w}) - H\mathbf{w}_0 \\ &= \mathbf{g} + H\mathbf{w} - H\mathbf{w}_0.\end{aligned}$$

2. Set the gradient to zero and solve for $\mathbf{w}$ (assuming that H is invertible):

$$\begin{aligned}\nabla f(\mathbf{w}) = 0 \quad &\Longleftrightarrow \quad \mathbf{g} + H\mathbf{w} - H\mathbf{w}_0 = 0 \\ &\Longleftrightarrow \quad H(\mathbf{w} - \mathbf{w}_0) = -\mathbf{g} \\ &\Longleftrightarrow \quad \mathbf{w} = \mathbf{w}_0 - H^{-1}\mathbf{g}.\end{aligned}$$

Here, $\mathbf{w}_0$ stands for the starting point of the iteration or, during the iteration, for the previously obtained value of $\mathbf{w}$, and $\mathbf{w}$ stands for the updated value. As $\mathbf{g}$ is $\nabla f(\mathbf{w}_0)$ and H is the Hessian at $\mathbf{w}_0$, we obtain the Newton iteration.

Ex. 11 Gradient of matrix-valued functions

1. Gradient of $J(W) = \mathbf{u}^{\mathrm{T}} W\mathbf{v}$ using both methods.

1. First method

$$\begin{aligned} J(W) &= \sum_{i=1}^{n}\sum_{j=1}^{m} u_i W_{ij} v_j \\ \implies \frac{\partial J(W)}{W_{kl}} &= u_k v_l = (\mathbf{u}\mathbf{v}^{\mathrm{T}})_{kl} \\ \implies \nabla J(W) &= \mathbf{u}\mathbf{v}^{\mathrm{T}}. \end{aligned}$$

2. Second method

$$\begin{aligned} J(W+\epsilon\mathbf{e}^{(i)]}\mathbf{e}^{[j]}) &= \mathbf{u}^{\mathrm{T}}(W+\epsilon\mathbf{e}^{(i)]}\mathbf{e}^{[j]})\mathbf{v} \\ &= \mathbf{u}^{\mathrm{T}}W\mathbf{v} + \epsilon\,\underbrace{\mathbf{u}^{\mathrm{T}}\mathbf{e}^{(i)}}_{\in\,\mathbb{R}}\,\underbrace{\mathbf{e}^{[j]}\mathbf{v}}_{\in\,\mathbb{R}} \\ &= J(W) + \epsilon\mathbf{e}^{[i]}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{e}^{(j)} \\ \implies \nabla J(W) &= \mathbf{u}\mathbf{v}^{\mathrm{T}}. \end{aligned}$$

2. Notice that

$$J(W) = \mathbf{u}^{\mathrm{T}}(W+A)\mathbf{v} = \mathbf{u}^{\mathrm{T}}W\mathbf{v} + \mathbf{u}^{\mathrm{T}}A\mathbf{v},$$

and $\mathbf{u}^{\mathrm{T}}A\mathbf{v}$ is a constant with respect to W. Therefore

$$\nabla J(W) = \nabla\mathbf{u}^{\mathrm{T}}W\mathbf{v} + \nabla\mathbf{u}^{\mathrm{T}}A\mathbf{v} = \nabla\mathbf{u}^{\mathrm{T}}W\mathbf{v} + 0 \overset{\text{part 1}}{=} \mathbf{u}\mathbf{v}^{\mathrm{T}}.$$

3. Gradient of $J(W) = \sum_n f(\mathbf{w}_n^{\mathrm{T}}\mathbf{v})$ using both methods.

1. First method.

$$\begin{aligned} \frac{\partial J(W)}{\partial W_{ij}} &= \sum_{k=1}^{n} \frac{\partial}{\partial W_{ij}} f(\mathbf{w}_k^{\mathrm{T}}\mathbf{v}) \\ &= f'(\mathbf{w}_i^{\mathrm{T}}\mathbf{v}) \frac{\partial}{\partial W_{ij}} \underbrace{\mathbf{w}_i^{\mathrm{T}}\mathbf{v}}_{\sum_{j=1}^{m} W_{ij}v_j} \\ &= f'(\mathbf{w}_i^{\mathrm{T}}\mathbf{v})v_j \\ \implies \nabla J(W) &= f'(W\mathbf{v})\mathbf{v}^{\mathrm{T}}, \end{aligned}$$

where $f'(.)$ operates element-wise on the vector $W\mathbf{v}$.

2. Second method. Recall the theory of Taylor expansion $(*)$.

$$\begin{aligned} J(W) &= \sum_{k=1}^{n} f(\mathbf{w}_k^{\mathrm{T}}\mathbf{v}) = \sum_{k=1}^{n} f(e^{[k]}W\mathbf{v}) \\ J(W+\epsilon\mathbf{e}^{(i)}\mathbf{e}^{[j]}) &= \sum_{k=1}^{n} f(\mathbf{e}^{[k]}(W+\epsilon\mathbf{e}^{(i)}\mathbf{e}^{[j]})\mathbf{v}) \\ &= \sum_{k=1}^{n} f(\mathbf{e}^{[k]}W\mathbf{v} + \epsilon\mathbf{e}^{[k]}\mathbf{e}^{(i)}\mathbf{e}^{[j]}\mathbf{v}) \\ &\overset{(*)}{=} \sum_{k=1}^{n}(f(\mathbf{e}^{[k]}W\mathbf{v}) + \\ &\quad \epsilon f'(\mathbf{e}^{[k]}W\mathbf{v}) \quad \underbrace{\mathbf{e}^{[k]}\mathbf{e}^{(i)}}_{=\,0,\text{ unless } k\,=\,i} \quad \mathbf{e}^{[j]}\mathbf{v} + O(\epsilon^2) \\ &= \sum_{k=1}^{n} f(\mathbf{e}^{[k]}W\mathbf{v}) + \epsilon f'(\mathbf{e}^{[i]}W\mathbf{v})\underbrace{\mathbf{e}^{[i]}\mathbf{e}^{(i)}}_{=\,1}\mathbf{e}^{[j]}\mathbf{v} + \\ &\quad O(\epsilon^2) \\ &= J(W) + \epsilon\mathbf{e}^{[i]}f'(W\mathbf{v})\mathbf{v}^{\mathrm{T}}\mathbf{e}^{(j)} + O(\epsilon^2) \\ \implies \nabla J(W) &= f'(W\mathbf{v})\mathbf{v}^{\mathrm{T}}, \end{aligned}$$

where $f'(.)$ operates element-wise on the vector $W\mathbf{v}$.

4. Gradient of $J(W) = \mathbf{u}^{\mathrm{T}} W^{-1}\mathbf{v}$ using only the second method. Notice that because W is symmetric, W^{-1} is also symmetric $(*)$.

$$\begin{aligned}
& & & J(W + \epsilon \mathbf{e}^{(i)}\mathbf{e}^{[j]}) \\
& & = \quad & \mathbf{u}^{\mathrm{T}}(W + \epsilon \mathbf{e}^{(i)}\mathbf{e}^{[j]})^{-1}\mathbf{v} \\
& & \overset{\text{hint}}{=} \quad & \mathbf{u}^{\mathrm{T}}(W^{-1} - \epsilon W^{-1}\mathbf{e}^{(i)}\mathbf{e}^{[j]}W^{-1} + O(\epsilon^2))\mathbf{v} \\
& & = \quad & \mathbf{u}^{\mathrm{T}}W^{-1}\mathbf{v} - \epsilon \underbrace{\mathbf{u}^{\mathrm{T}}W^{-1}\mathbf{e}^{(i)}}_{\in\ \mathbb{R}}\underbrace{\mathbf{e}^{[j]}W^{-1}\mathbf{v}}_{\in\ \mathbb{R}} + O(\epsilon^2) \\
& & = \quad & J(W) - \epsilon(\mathbf{e}^{(i)})^{\mathrm{T}}(\mathbf{u}^{\mathrm{T}}W^{-1})^{\mathrm{T}}(W^{-1}\mathbf{v})^{\mathrm{T}}(\mathbf{e}^{[j]})^{\mathrm{T}} + \\
& & & O(\epsilon^2) \\
& & = \quad & J(W) - \epsilon \mathbf{e}^{[i]}(W^{-1})^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}(W^{-1})^{\mathrm{T}}\mathbf{e}^{(j)} + O(\epsilon^2) \\
& & \overset{(*)}{=} \quad & J(W) - \epsilon \mathbf{e}^{[i]}W^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}W^{-1}\mathbf{e}^{(j)} + O(\epsilon^2) \\
\Longrightarrow & \nabla J(W) & = \quad & -W^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}W^{-1}.
\end{aligned}$$

Ex. 12 Gradient of the Log-Determinant

1. As in exercise 3 (Eigenvalue Decomposition), let $U\Lambda V^{\mathrm{T}}$ be the eigenvalue decomposition of W (with $V^T = U^{-1}$). Then $\Lambda = V^{\mathrm{T}}WU$ and

$$\begin{aligned}
\lambda_n &= \mathbf{e}^{[n]}\Lambda\mathbf{e}^{(n)} \\
&= \mathbf{e}^{[n]}V^{\mathrm{T}}WU\mathbf{e}^{(n)} \\
&= (V\mathbf{e}^{(n)})^{\mathrm{T}}W(U\mathbf{e}^{(n)} \\
&= \mathbf{v}_n^T W\mathbf{u}_n,
\end{aligned}$$

where (as always) $\mathbf{e}^{[n]}$ is the row vector with 1 in the nth slot and 0 elsewhere and $\mathbf{e}^{(n)}$ is the corresponding column vector.

2. Using the previous part of this exercise and the exercise 11 (Gradient of matrix-valued functions), we get:

$$\nabla_W \lambda_n(W) = \nabla_W \mathbf{v}_n^{\mathrm{T}}W\mathbf{u}_n = \mathbf{v}_n\mathbf{u}_n^{\mathrm{T}}.$$

3. In exercise 4 (Trace, Determinants and Eigenvalues), we proved that $\det(W) = \prod_i \lambda_i$ and hence $|\det(W)| = \prod_i |\lambda_i|$.
   (i) If W is positive definite, its eigenvalues are positive (as we proved in exercise sheet 1) and $|\det(W)| = \prod_i \lambda_i$.
   (ii) If W is a matrix with real entries, then $W\mathbf{u} = \lambda\mathbf{u}$ implies $W\bar{\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}}$, i.e. if λ is a complex eigenvalue, then $\bar{\lambda}$ (the complex conjugate of λ) is also an eigenvalue. Since $|\lambda|^2 = \lambda\bar{\lambda}$,

$$|\det(W)| = \left(\prod_{\lambda_i \in \mathbb{C}} \lambda_i\right)\left(\prod_{\lambda_j \in \mathbb{R}} \operatorname{sign}(\lambda_j)\lambda_j\right).$$

Now we can write $J(W)$ in terms of the eigenvalues:

$$\begin{aligned}
J(W) &= \log|\det(W)| \\
&= \log\left(\prod_{\lambda_i \in \mathbb{C}} \lambda_i\right)\left(\prod_{\lambda_j \in \mathbb{R}} \operatorname{sign}(\lambda_j)\lambda_j\right) \\
&= \log\left(\prod_{\lambda_i \in \mathbb{C}} \lambda_i\right) + \log\left(\prod_{\lambda_j \in \mathbb{R}} \operatorname{sign}(\lambda_j)\lambda_j\right) \\
&= \sum_{\lambda_i \in \mathbb{C}} \log\lambda_i + \sum_{\lambda_j \in \mathbb{R}} \log(\operatorname{sign}(\lambda_j)\lambda_j),
\end{aligned}$$

thus making calculating the gradient easier:

$$\begin{aligned}
\nabla J(W) &= \nabla_W\Big(\sum_{\lambda_i \in \mathbb{C}} \log\lambda_i + \sum_{\lambda_j \in \mathbb{R}} \log(\operatorname{sign}(\lambda_j)\lambda_j)\Big) \\
&= \sum_{\lambda_i \in \mathbb{C}} \frac{1}{\lambda_i}\nabla_W\lambda_i + \sum_{\lambda_i \in \mathbb{R}} \frac{1}{\operatorname{sign}(\lambda_i)\lambda_i}\nabla_W(\operatorname{sign}(\lambda_i)\lambda_i \\
&\overset{\text{part 2}}{=} \sum_{\lambda_i \in \mathbb{C}} \frac{\mathbf{v}_i\mathbf{u}_i^{\mathrm{T}}}{\lambda_i} + \sum_{\lambda_i \in \mathbb{R}} \frac{\operatorname{sign}(\lambda_i)\mathbf{v}_i\mathbf{u}_i^{\mathrm{T}}}{\operatorname{sign}(\lambda_i)\lambda_i} \\
&= \sum_i \frac{\mathbf{v}_i\mathbf{u}_i^{\mathrm{T}}}{\lambda_i}.
\end{aligned}$$

4. Using exercise 3 (Eigenvalue Decomposition) (∗), we get

$$\nabla J(W) = \sum_i \frac{\mathbf{v}_i \mathbf{u}_i^{\mathrm{T}}}{\lambda_i} = \sum_i \frac{1}{\lambda_i} (\mathbf{u}_i \mathbf{v}_i^{\mathrm{T}})^{\mathrm{T}} \overset{(*)}{=} (W^{-1})^{\mathrm{T}}.$$

Ex. 13 Maximum Likelihood Estimation for Multivariate Gaussians

1. The likelihood function is

$$\begin{aligned} & L(\mu, \Sigma_{\mathbf{x}}) \\ = & \prod_{n=1}^{N} f(\mathbf{x}_n) \\ = & \prod_{n=1}^{N} \frac{1}{(2\pi)^{m/2} |\Sigma_{\mathbf{x}}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}_n - \mu)^{\mathrm{T}} \Sigma_{\mathbf{x}}^{-1} (\mathbf{x}_n - \mu) \right] \\ = & \frac{1}{((2\pi)^{m/2} |\Sigma|^{1/2})^N} \exp\left[-\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mu)^{\mathrm{T}} \Sigma_{\mathbf{x}}^{-1} (\mathbf{x}_n - \mu) \right]. \end{aligned}$$

The log-likelihood function is thus

$$\begin{aligned} & \ell(\mu, \Sigma_{\mathbf{x}}) \\ = & \log L(\mu, \Sigma_{\mathbf{x}}) \\ = & -N \log((2\pi)^{m/2} |\Sigma_{\mathbf{x}}|^{1/2}) - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mu)^{\mathrm{T}} \Sigma_{\mathbf{x}}^{-1} (\mathbf{x}_n - \mu) \\ = & -N \log(2\pi)^{m/2} - \frac{N}{2} \log|\Sigma_{\mathbf{x}}| - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mu)^{\mathrm{T}} \Sigma_{\mathbf{x}}^{-1} (\mathbf{x}_n - \mu). \end{aligned}$$

2. Using the given hint (∗) we get

$$\begin{aligned} J(W + \epsilon \mathbf{e}^{(i)} \mathbf{e}^{[j]}) & = \mathbf{u}^{\mathrm{T}} (W + \epsilon \mathbf{e}^{(i)} \mathbf{e}^{[j]})^{-1} \mathbf{v} \\ & \overset{(*)}{=} \mathbf{u}^{\mathrm{T}} (W^{-1} - \epsilon W^{-1} \mathbf{e}^{(i)} \mathbf{e}^{[j]} W^{-1}) + O_0(\epsilon^2)) \mathbf{v} \\ & = \mathbf{u}^{\mathrm{T}} W^{-1} \mathbf{v} - \epsilon \underbrace{\mathbf{u}^{\mathrm{T}} W^{-1} \mathbf{e}^{(i)}}_{\in \mathbb{R}} \underbrace{\mathbf{e}^{[j]} W^{-1} \mathbf{v}}_{\in \mathbb{R}} + O(\epsilon^2) \\ & = \mathbf{u}^{\mathrm{T}} W^{-1} \mathbf{v} - \epsilon \mathbf{e}^{[i]} (W^{-1})^{\mathrm{T}} \mathbf{u} \mathbf{v}^{\mathrm{T}} (W^{-}1)^{\mathrm{T}} \mathbf{e}^{(j)} + \\ & \quad O(\epsilon^2), \end{aligned}$$

so the second method for calculating gradients for matrix valued functions (introduced in exercise sheet 2) gives us

$$\nabla J(W) = -(W^{-1})^{\mathrm{T}} \mathbf{u} \mathbf{v}^{\mathrm{T}} (W^{-}1)^{\mathrm{T}}.$$

3. We can use the results from exercise 12 (Gradient of the Log-Determinant) (∗) and part 2 of this exercise to calculate the gradient with respect to $\Sigma_{\mathbf{x}}$. Remember that because $\Sigma_{\mathbf{x}}$ is symmetric, its inverse is also symmetric (∗∗).

$$\begin{aligned} \nabla_{\Sigma_{\mathbf{x}}} \ell(\mu, \Sigma_{\mathbf{x}}) & = \underbrace{\nabla_{\Sigma_{\mathbf{x}}} (-N \log(2\pi)^{m/2})}_{=0} - \underbrace{\nabla_{\Sigma_{\mathbf{x}}} (\frac{N}{2} \log|\Sigma_{\mathbf{x}}|)}_{(*)} - \\ & \quad \underbrace{\nabla_{\Sigma_{\mathbf{x}}} (\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mu)^{\mathrm{T}} \Sigma_{\mathbf{x}}^{-1} (\mathbf{x}_n - \mu))}_{\text{part 2}} \\ & = -\frac{N}{2} (\Sigma_{\mathbf{x}}^{-1})^{\mathrm{T}} + \\ & \quad \frac{1}{2} \sum_{n=1}^{N} (\Sigma_{\mathbf{x}}^{-1})^{\mathrm{T}} (\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}} (\Sigma_{\mathbf{x}}^{-1})^{\mathrm{T}} \\ & \overset{(**)}{=} -\frac{N}{2} \Sigma_{\mathbf{x}}^{-1} + \frac{1}{2} \sum_{n=1}^{N} \Sigma_{\mathbf{x}}^{-1} (\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}} \Sigma_{\mathbf{x}}^{-1}. \end{aligned}$$

Calculating the gradient with respect to μ is easier since there is no need to use any special formulas:

$$\begin{aligned}
\nabla_\mu \ell(\mu, \Sigma_{\mathbf{x}}) &= \underbrace{\nabla_\mu(-N\log(2\pi)^{m/2})}_{=0} - \underbrace{\nabla_\mu(\frac{N}{2}\log|\Sigma_{\mathbf{x}}|)}_{=0} \\
&\quad -\nabla_\mu(\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_n - \mu)) \\
&= -\frac{1}{2}\sum_{n=1}^{N}\nabla_\mu(\mathbf{x}_n^{\mathrm{T}}\Sigma_{\mathbf{x}}^{-1}\mathbf{x}_n - 2\mathbf{x}_n^{\mathrm{T}}\Sigma_{\mathbf{x}}^{-1}\mu + \mu^{\mathrm{T}}\Sigma_{\mathbf{x}}^{-1}\mu) \\
&= -\frac{1}{2}\sum_{n=1}^{N}(-2\Sigma_{\mathbf{x}}^{-1}\mathbf{x}_n + 2\Sigma_{\mathbf{x}}^{-1}\mu) \\
&= \sum_{n=1}^{N}(\Sigma_{\mathbf{x}}^{-1}\mathbf{x}_n - \Sigma_{\mathbf{x}}^{-1}\mu) \\
&= \sum_{n=1}^{N}(\Sigma_{\mathbf{x}}^{-1}\mathbf{x}_n) - N\Sigma_{\mathbf{x}}^{-1}\mu.
\end{aligned}$$

4. Setting the gradient with respect to μ to zero gives us

$$\begin{aligned}
\nabla_\mu \ell(\mu, \Sigma_{\mathbf{x}}) = 0 \quad &\Longleftrightarrow \quad \sum_{n=1}^{N}(\Sigma_{\mathbf{x}}^{-1}\mathbf{x}_n) - N\Sigma_{\mathbf{x}}^{-1}\mu = 0 \\
&\Longleftrightarrow \quad \Sigma_{\mathbf{x}}^{-1}N\mu = \Sigma_{\mathbf{x}}^{-1}\sum_{n=1}^{N}\mathbf{x}_n \\
&\Longleftrightarrow \quad N\mu = \sum_{n=1}^{N}\mathbf{x}_n \\
&\Longleftrightarrow \quad \mu = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&\Longrightarrow \quad \hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \bar{\mathbf{X}}.
\end{aligned}$$

Remember that if the Hessian is negative definite at a critical point, the function attains a local maximum in that point. Because the Hessian in this case is (up to positive a scalar) $-\Sigma_{\mathbf{x}}^{-1}$, it is negative definite, since by assumption $\Sigma_{\mathbf{x}}$ (and therefore also $\Sigma_{\mathbf{x}}^{-1}$) is positive definite. Therefore $\hat{\mu}$ really maximizes $\ell(\mu, \Sigma_{\mathbf{x}})$ and the maximum likelihood estimate $\hat{\mu}$ is the sample mean. Setting $\mu = \hat{\mu}$ gives us then

$$\begin{aligned}
&\nabla_{\Sigma_{\mathbf{x}}}\ell(\mu, \Sigma_{\mathbf{x}}) \;\Big|_{\mu=\hat{\mu}} = 0 \\
\Longleftrightarrow \quad & -\frac{N}{2}\Sigma_{\mathbf{x}}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_n - \bar{\mathbf{X}})(\mathbf{x}_n - \bar{\mathbf{X}})^{\mathrm{T}}\Sigma_{\mathbf{x}}^{-1} = 0 \\
\Longleftrightarrow \quad & \Sigma_{\mathbf{x}}^{-1} = \Sigma_{\mathbf{x}}^{-1}(\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{X}})(\mathbf{x}_n - \bar{\mathbf{X}})^{\mathrm{T}})\Sigma_{\mathbf{x}}^{-1} \\
\Longleftrightarrow \quad & \Sigma_{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{X}})(\mathbf{x}_n - \bar{\mathbf{X}})^{\mathrm{T}} \\
\Longrightarrow \quad & \hat{\Sigma}_{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{X}})(\mathbf{x}_n - \bar{\mathbf{X}})^{\mathrm{T}},
\end{aligned}$$

which is what we wanted.

Ex. 14 Derivation of the Power Method

1. Using the alternative, 2-step method:

(i) $\mathbf{v}_{k+1} \leftarrow \mathbf{w}_k + \mu\Sigma\mathbf{w}_k$

(ii) $\mathbf{w}_{k+1} \leftarrow \frac{\mathbf{v}_{k+1}}{||\mathbf{v}_{k+1}||_2}$, we see that formula (i) looks like an update step of gradient ascent $\mathbf{w} \leftarrow \mathbf{w} + \mu\nabla J(\mathbf{w})$. Thus, $\Sigma\mathbf{w}$ should be the gradient of $J(\mathbf{w})$. Using the exercise Gradient of vector-valued functions $(*)$ we get:

$$\begin{aligned} \nabla J(\mathbf{w}) &= \nabla\frac{1}{2}\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w} \\ &\overset{(*)}{=} \frac{1}{2}(\Sigma^{\mathrm{T}}\mathbf{w} + \Sigma\mathbf{w}) \\ &\overset{\Sigma \text{ symmetric}}{=} \frac{1}{2}\,2\Sigma\mathbf{w} \\ &= \Sigma\mathbf{w}. \end{aligned}$$

Hence, formula (i) is the gradient ascent method (for maximization) with

$\mathbf{w}$	as the value at previous iteration step,
μ	as the step size,
$\Sigma\mathbf{w}_k$	as the gradient of objective function J, and
$\mathbf{v}_{k+1}$	as the updated value.

In formula (ii) we normalize the obtained vector $\mathbf{v}_{k+1}$ to unit norm by dividing the vector by its 2-norm, which is the length of the vector in Euclidean space. This is a form of constraint optimization and is in this case necessary because without the constraint of unit norm for $\mathbf{w}_{k+1}$, the maximum would be obtained when $\mathbf{w}_{k+1}$ is infinitely large.

To sum up, the update rule is a constraint gradient ascent update step to optimize the objective function J.

2. Write first $\mathbf{w}_{k+1}$ in different form to make taking the limit easier:

$$\begin{aligned} &\mathbf{w}_{k+1} \\ =\ & \frac{\mathbf{w}_k + \mu\Sigma\mathbf{w}_k}{||\mathbf{w}_k + \mu\Sigma\mathbf{w}_k||_2} \\ =\ & \frac{\mathbf{w}_k + \mu\Sigma\mathbf{w}_k}{\sqrt{(\mathbf{w}_k + \mu\Sigma\mathbf{w}_k)^{\mathrm{T}}(\mathbf{w}_k + \mu\Sigma\mathbf{w}_k)}} \\ =\ & \frac{\mathbf{w}_k + \mu\Sigma\mathbf{w}_k}{\sqrt{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k + \mu(\Sigma\mathbf{w}_k)^{\mathrm{T}}\mathbf{w}_k + \mu\mathbf{w}_k^{\mathrm{T}}(\Sigma\mathbf{w}_k) + \mu^2(\Sigma\mathbf{w}_k)^{\mathrm{T}}(\Sigma\mathbf{w}_k)}} \\ =\ & \frac{\mathbf{w}_k + \mu\Sigma\mathbf{w}_k}{\sqrt{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k + 2\mu\mathbf{w}_k^{\mathrm{T}}\Sigma\mathbf{w}_k + \mu^2(\Sigma\mathbf{w}_k)^{\mathrm{T}}(\Sigma\mathbf{w}_k)}} \\ =\ & \frac{\mu\left(\frac{1}{\mu}\mathbf{w}_k + \Sigma\mathbf{w}_k\right)}{\mu\sqrt{\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{\mu^2} + \frac{2\mathbf{w}_k^{\mathrm{T}}\Sigma\mathbf{w}_k}{\mu} + ||\Sigma\mathbf{w}_k||_2^2}} \\ =\ & \frac{\frac{1}{\mu}\mathbf{w}_k + \Sigma\mathbf{w}_k}{\sqrt{\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{\mu^2} + \frac{2\mathbf{w}_k^{\mathrm{T}}\Sigma\mathbf{w}_k}{\mu} + ||\Sigma\mathbf{w}_k||_2^2}}. \end{aligned}$$

Since $1/\mu \to 0$ when $\mu \to \infty$, taking the limit is now easy.

$$\begin{aligned} \lim_{\mu\to\infty} \mathbf{w}_k &= \lim_{\mu\to\infty} \frac{\frac{1}{\mu}\mathbf{w}_k + \Sigma\mathbf{w}_k}{\sqrt{\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{\mu^2} + \frac{2\mathbf{w}_k^{\mathrm{T}}\Sigma\mathbf{w}_k}{\mu} + ||\Sigma\mathbf{w}_k||_2^2}} \\ &= \frac{\Sigma\mathbf{w}_k}{\sqrt{||\Sigma\mathbf{w}_k||_2^2}} \\ &= \frac{\Sigma\mathbf{w}_k}{||\Sigma\mathbf{w}_k||_2}. \end{aligned}$$

Ex. 15 Convergence of the Power Method

1. Since the columns of U are orthonormal (eigen)vectors, U is orthogonal, i.e. $U^{-1} = U^{\mathrm{T}}$. With Exercises 5 and 3, we get

$$\Sigma = U\Lambda U^{\mathrm{T}},$$

where Λ is the diagonal matrix with eigenvalues λ_i of Σ as diagonal elements. Assume that $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. (Remember that all eigenvalues are positive since Σ is symmetric)

2. Notice that

$$\mathbf{v}_{k+1} = \Sigma \mathbf{w}_k = U \Lambda U^{\mathrm{T}} \mathbf{w}_k \iff U^{\mathrm{T}} \mathbf{v}_{k+1} = \Lambda U^{\mathrm{T}} \mathbf{w}_k,$$

so therefore $\tilde{\mathbf{v}}_{k+1} = \Lambda \tilde{\mathbf{w}}_k$. The norm of $\tilde{\mathbf{v}}_{k+1}$ is the same as the norm of $\mathbf{v}_{k+1}$:

$$\begin{aligned}
||\tilde{\mathbf{v}}_{k+1}|| &= ||U^{\mathrm{T}} \mathbf{v}_{k+1}|| \\
&= \sqrt{(U^{\mathrm{T}} \mathbf{v}_{k+1})^{\mathrm{T}} (U^{\mathrm{T}} \mathbf{v}_{k+1})} \\
&= \sqrt{\mathbf{v}_{k+1}^{\mathrm{T}} U U^{\mathrm{T}} \mathbf{v}_{k+1}} \\
&= \sqrt{\mathbf{v}_{k+1}^{\mathrm{T}} \mathbf{v}_{k+1}} \\
&= ||\mathbf{v}_{k+1}||.
\end{aligned}$$

Hence, in terms of $\tilde{\mathbf{v}}_k$ and $\tilde{\mathbf{w}}_k$ the vector iteration is

$$\begin{aligned}
\tilde{\mathbf{v}}_{k+1} &= \Lambda \tilde{\mathbf{w}}_k \\
\tilde{\mathbf{w}}_{k+1} &= \frac{\tilde{\mathbf{v}}_{k+1}}{||\tilde{\mathbf{v}}_{k+1}||}.
\end{aligned}$$

3. Let $\tilde{\mathbf{w}}_0 = (\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_n)^{\mathrm{T}}$. Because Λ is a diagonal matrix, we get

$$\tilde{\mathbf{v}}_1 = \begin{pmatrix} \lambda_1 \alpha_1 \\ \lambda_2 \alpha_2 \\ \vdots \\ \lambda_n \alpha_n \end{pmatrix} = \lambda_1 \alpha_1 \begin{pmatrix} 1 \\ \frac{\alpha_2}{\alpha_1} \frac{\lambda_2}{\lambda_1} \\ \vdots \\ \frac{\alpha_n}{\alpha_1} \frac{\lambda_n}{\lambda_1} \end{pmatrix},$$

and therefore

$$\tilde{\mathbf{w}}_1 = \frac{\lambda_1 \alpha_1}{c_1} \begin{pmatrix} 1 \\ \frac{\alpha_2}{\alpha_1} \frac{\lambda_2}{\lambda_1} \\ \vdots \\ \frac{\alpha_n}{\alpha_1} \frac{\lambda_n}{\lambda_1} \end{pmatrix},$$

where c_1 is a normalization constant such that $\|\tilde{\mathbf{w}}_1\| = 1$ (i.e. $c_1 = \|\tilde{\mathbf{v}}_1\|$). Hence, for $\tilde{\mathbf{w}}_k$ it holds that

$$\tilde{\mathbf{w}}_k = \tilde{c}_k \begin{pmatrix} 1 \\ \frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k \\ \vdots \\ \frac{\alpha_n}{\alpha_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k \end{pmatrix},$$

where $\tilde{c}_k$ is again a normalization constant such that $||\tilde{\mathbf{w}}_k|| = 1$. As λ_1 is the dominant eigenvalue, $|\lambda_j/\lambda_1| < 1$ for $j = 2, 3, \dots, n$ and

$$\lim_{k \to \infty} \left(\frac{\lambda_j}{\lambda_1}\right)^k = 0.$$

For the normalization constant $\tilde{c}_k$, we get

$$\tilde{c}_k = \frac{1}{\sqrt{1 + \sum_{i=2}^n \left(\frac{\alpha_i}{\alpha_1}\right)^2 \left(\frac{\lambda_i}{\lambda_1}\right)^{2k}}},$$

and therefore

$$\begin{aligned}
\lim_{k \to \infty} \tilde{c}_k &= \frac{1}{\sqrt{1 + \sum_{i=2}^n \left(\frac{\alpha_i}{\alpha_1}\right)^2 \lim_{k \to \infty} \left(\frac{\lambda_i}{\lambda_1}\right)^{2k}}} \\
&= \frac{1}{\sqrt{1 + \sum_{i=2}^n \left(\frac{\alpha_i}{\alpha_1}\right)^2 \cdot 0}} \\
&= 1.
\end{aligned}$$

Because $\tilde{\mathbf{w}}_k = \tilde{c}_k \tilde{\mathbf{v}}_k$, the limit

$$\lim_{k\to\infty} \tilde{\mathbf{w}}_k$$

exists. Hence

$$\lim_{k\to\infty} \tilde{\mathbf{w}}_k = \lim_{k\to\infty} \tilde{c}_k \lim_{k\to\infty} \begin{pmatrix} 1 \\ \frac{\alpha_2}{\alpha_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k \\ \vdots \\ \frac{\alpha_n}{\alpha_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

4. Because $\mathbf{w}_k = U\tilde{\mathbf{w}}_k$, we get

$$\lim_{k\to\infty} \mathbf{w}_k = U \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{u}_1,$$

which is the eigenvector with the largest eigenvalue ("dominant eigenvector").

Ex. 16 Dimension Reduction by PCA

1. Direct calculations give

$$\begin{aligned} J(\mathbf{w}) &= \mathrm{E}(\|\mathbf{x} - \hat{\mathbf{x}}\|) \\ &= \mathrm{E}\left(\sum_{j=1}^n (x_j - w_j z)^2\right) \\ &= \mathrm{E}\left(\sum_{j=1}^n (x_j^2 - 2x_j w_j z + w_j^2 z^2)\right) \\ &= \mathrm{E}\left(\sum_{j=1}^n x_j^2\right) + \mathrm{E}\left(\sum_{j=1}^n -2x_j w_j z\right) + \mathrm{E}\left(\sum_{j=1}^n w_j^2 z^2\right) \\ &= \sum_{j=1}^n \mathrm{E}(x_j^2) - 2\sum_{j=1}^n w_j \mathrm{E}(x_j z) + \left(\sum_{j=1}^n w_j^2\right)\mathrm{E}(z^2). \end{aligned}$$

By assumption, $\mathrm{E}(x_j) = 0$, so $\mathrm{E}(x_j^2) = \mathrm{V}(x_j)$ and

$$\begin{aligned} z^2 &= \sum_{i,k} w_i w_k x_i x_k, \\ \mathrm{E}(z^2) &= \sum_{i,k} w_i w_k \mathrm{E}(x_i x_k) = \sum_{i,k} \mathrm{cov}(x_i, x_k), \\ \mathrm{E}(x_j z) &= \sum_{i=1}^n w_i \mathrm{E}(x_j x_i) = \sum_{i=1}^n w_i \mathrm{cov}(x_j, x_i). \end{aligned}$$

Hence,

$$\begin{aligned} J(\mathbf{w}) &= \sum_{j=1}^n \mathrm{V}(x_j) - 2\sum_{j=1}^n w_j \sum_{i=1}^n w_i \mathrm{cov}(x_j, x_i) + \\ &\quad \sum_{j=1}^n w_j^2 \sum_{i,k} w_i w_k \mathrm{cov}(x_i, x_k). \end{aligned}$$

2. Since $\mathrm{V}(x_j)$ is a constant, denote $\sum_{j=1}^n \mathrm{V}(x_j) = c$. Because $\|\mathbf{w}\| = 1$, we get

$$\begin{aligned} J(\mathbf{w}) &= c - 2\sum_{j=1}^n w_j \sum_{i=1}^n w_i \mathrm{cov}(x_j, x_i) + \sum_{j=1}^n w_j^2 \sum_{i,k} w_i w_k \mathrm{cov}(x_i, x_k) \\ &= c - 2\sum_{i,j} w_i w_j \mathrm{cov}(x_j, x_i) + 1^2 \sum_{i,j} w_i w_j \mathrm{cov}(x_i, x_j) \\ &= c - \sum_{i,j} w_i w_j \mathrm{cov}(x_i, x_j) \\ &= c - \mathbf{w}^{\mathrm{T}} \Sigma_{\mathbf{x}} \mathbf{w}, \end{aligned}$$

where $\Sigma_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$, i.e. $(\Sigma_{\mathbf{x}})_{ij} = \mathrm{cov}(x_i, x_j)$. This is the same as the PCA cost function up to the additive constant and the sign.

Ex. 17 Signal Approximation with PCA

1. The outer vector product $\mathbf{w}\mathbf{w}^T$ is

$$\begin{aligned}
\mathbf{w}\mathbf{w}^{\mathrm{T}} &= \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \dots & w_n \end{pmatrix} \\
&= \begin{pmatrix} w_1^2 & w_1 w_2 & \dots & w_1 w_n \\ w_2 w_1 & w_2^2 & \dots & w_2 w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n w_1 & w_n w_2 & \dots & w_n^2 \end{pmatrix} \\
\Longrightarrow \mathrm{Tr}(\mathbf{w}\mathbf{w}^{\mathrm{T}}) &= \sum_{k=1}^{n} w_k^2 = \|\mathbf{w}\|^2.
\end{aligned}$$

2.

$$\begin{aligned}
J_{\min} &= \mathrm{E}(\|\mathbf{x} - U_m U_m^{\mathrm{T}} \mathbf{x}\|^2) \\
&= \mathrm{E}(\mathrm{Tr}((\mathbf{x} - U_m U_m^{\mathrm{T}} \mathbf{x})(\mathbf{x} - U_m U_m^{\mathrm{T}} \mathbf{x})^{\mathrm{T}})) \\
&= \mathrm{E}(\mathrm{Tr}(\mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathbf{x}\mathbf{x}^T U_m U_m^{\mathrm{T}} - U_m U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} + \\
&\quad U_m U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} U_m U_m^{\mathrm{T}})) \\
&= \mathrm{E}(\mathrm{Tr}(\mathbf{x}\mathbf{x}^{\mathrm{T}})) - \mathrm{E}(\mathrm{Tr}(\mathbf{x}\mathbf{x}^{\mathrm{T}} U_m U_m^{\mathrm{T}})) - \mathrm{E}(\mathrm{Tr}(U_m U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}})) \\
&\quad + \mathrm{E}(\mathrm{Tr}(U_m U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} U_m U_m^{\mathrm{T}})) \\
&= \mathrm{E}(\mathrm{Tr}(\mathbf{x}\mathbf{x}^{\mathrm{T}})) - \mathrm{E}(\mathrm{Tr}(U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} U_m)) - \mathrm{E}(\mathrm{Tr}(U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} U_m)) \\
&\quad + \mathrm{E}(\mathrm{Tr}(\underbrace{U_m^{\mathrm{T}} U_m}_{=I} U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} U_m)) \\
&= \mathrm{Tr}(\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}})) - \mathrm{Tr}(\mathrm{E}(U_m^{\mathrm{T}} \mathbf{x}\mathbf{x}^{\mathrm{T}} U_m)) \\
&= \mathrm{Tr}(\underbrace{\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}})}_{\Sigma_{\mathbf{x}}}) - \mathrm{Tr}(U_m^{\mathrm{T}} \underbrace{\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}})}_{=\Sigma_{\mathbf{x}}} U_m) \\
&= \mathrm{Tr}(\Sigma_{\mathbf{x}}) - \mathrm{Tr}(U_m^{\mathrm{T}} \Sigma_{\mathbf{x}} U_m).
\end{aligned}$$

3. Let $U \Lambda U^{\mathrm{T}}$ be the eigenvalue decomposition of $\Sigma_{\mathbf{x}}$. Now

$$U_m^{\mathrm{T}} \Sigma_{\mathbf{x}} U_m = U_m^{\mathrm{T}} U \Lambda U^{\mathrm{T}} U_m.$$

Find out piece by piece, what $U_m^{\mathrm{T}} \Sigma_{\mathbf{x}} U_m$ looks like:

$$\begin{aligned}
U_m^{\mathrm{T}} U &= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_m & \mathbf{u}_{m+1} & \dots & \mathbf{u}_n \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix} (m \times n \text{ matrix})
\end{aligned}$$

$$U^{\mathrm{T}} U_m = (U_m^{\mathrm{T}} U)^{\mathrm{T}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} (n \times m \text{ matrix})$$

$$\begin{aligned}
\Lambda U^{\mathrm{T}} U_m &= \begin{pmatrix} \lambda_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_m & \dots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & \lambda_n \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} (n \times m \text{ matrix}).
\end{aligned}$$

Therefore

$$\begin{aligned}
U_m^{\mathrm{T}} \Sigma_{\mathbf{x}} U_m &= \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_m \end{pmatrix} \\
&= \Lambda_m.
\end{aligned}$$

4. Using the fact that $(*)$ $\mathrm{Tr}(\Sigma_{\mathbf{x}}) = \sum_{k=1}^n \lambda_k$, we get

$$\begin{aligned}
J_{\min} &\overset{\text{part 2}}{=} \mathrm{Tr}(\Sigma_{\mathbf{x}}) - \mathrm{Tr}(U_m^{\mathrm{T}} \Sigma_{\mathbf{x}} U_m) \\
&\overset{\text{part 3}}{=} \mathrm{Tr}(\Sigma_{\mathbf{x}}) - \mathrm{Tr}(\Lambda_m) \\
&\overset{(*)}{=} \sum_{k=1}^n \lambda_k - \sum_{k=1}^m \lambda_k \\
&= \sum_{k=m+1}^n \lambda_k.
\end{aligned}$$

Ex. 18 PCA and data representation

1. Since the matrix X can be written as

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{p1} & X_{p2} & \dots & X_{pn} \end{pmatrix},$$

where X_{ij} is the jth observation of the ith component of the random variable $\mathbf{x}$, we see that the kth row of X contains all n observations of the kth random variable.

2. Now $\mathrm{cov}(X_k, X_l) = E(X_k X_l)$ because the mean was assumed to be zero. The sample covariance matrix is thus

$$\frac{1}{n} \sum_{i=1}^n X_{ki} X_{li} = \frac{1}{n} \mathbf{v}_k^{\mathrm{T}} \mathbf{v}_l.$$

Since for $\mathrm{E}(\mathbf{x}) = 0$ the covariance matrix C is $C = \mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}})$,

$$\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) = \begin{pmatrix} \mathrm{E}(X_1X_1) & \mathrm{E}(X_1X_2) & \dots & \mathrm{E}(X_1X_p) \\ \mathrm{E}(X_2X_1) & \mathrm{E}(X_2X_2) & \dots & \mathrm{E}(X_2X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(X_pX_1) & \mathrm{E}(X_pX_2) & \dots & \mathrm{E}(X_pX_p) \end{pmatrix},$$

the sample covariance matrix $\hat{C}$ is thus

$$\hat{C} = \frac{1}{n} \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 & \dots & \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_p \\ \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 & \dots & \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_p^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_p^{\mathrm{T}}\mathbf{v}_2 & \dots & \mathbf{v}_p^{\mathrm{T}}\mathbf{v}_p \end{pmatrix} = \frac{1}{n} X X^{\mathrm{T}}.$$

3. We can write Z as

$$\begin{aligned} Z &= U^{\mathrm{T}} X \\ &= U^{\mathrm{T}} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{pmatrix} \\ &= \begin{pmatrix} U^{\mathrm{T}}\mathbf{x}_1 & U^{\mathrm{T}}\mathbf{x}_2 & \dots & U^{\mathrm{T}}\mathbf{x}_n \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{u}_1^{\mathrm{T}}\mathbf{x}_1 & \mathbf{u}_1^{\mathrm{T}}\mathbf{x}_2 & \dots & \mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n \\ \mathbf{u}_2^{\mathrm{T}}\mathbf{x}_1 & \mathbf{u}_2^{\mathrm{T}}\mathbf{x}_2 & \dots & \mathbf{u}_2^{\mathrm{T}}\mathbf{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_m^{\mathrm{T}}\mathbf{x}_1 & \mathbf{u}_m^{\mathrm{T}}\mathbf{x}_2 & \dots & \mathbf{u}_m^{\mathrm{T}}\mathbf{x}_n \end{pmatrix}. \end{aligned}$$

Because $\mathbf{z}_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{x}$ is the ith principal component, the ith row of Z contains thus all the realizations of the ith principal component. Notice that the row as a whole is also often called the ith principal component.

4. The ith row of Z is $\mathbf{u}_i^{\mathrm{T}} X$. Let $i \neq j$. Taking the inner product of the ith and jth row gives us

$$\mathbf{u}_i^{\mathrm{T}} X X^{\mathrm{T}} \mathbf{u}_j = n\mathbf{u}_i^{\mathrm{T}} (\frac{1}{n} X X^{\mathrm{T}}) \mathbf{u}_j \overset{\text{part 2}}{=} n\mathbf{u}_i^{\mathrm{T}} \hat{C} \mathbf{u}_j.$$

Since U has the first m principal component weights as its columns, the sample covariance matrix $\hat{C}$ can be written as UDU^{T}, where D is a diagonal matrix (see Section 4.3.3 on page 4.3.3). Thus

$$\begin{aligned} \mathbf{u}_i^{\mathrm{T}} X X^{\mathrm{T}} \mathbf{u}_j &= n\mathbf{u}_i^{\mathrm{T}} U D U^{\mathrm{T}} \mathbf{u}_j \\ &= n \begin{pmatrix} 0 & \dots & 1_i & \dots & 0 \end{pmatrix} \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_p \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1_j \\ \vdots \\ 0 \end{pmatrix} \\ &= n \begin{pmatrix} 0 & \dots & d_i & \dots & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1_j \\ \vdots \\ 0 \end{pmatrix} \\ &= 0, \end{aligned}$$

i.e. the rows of Z are orthogonal.

5. Simply put: the principal components are an orthogonal basis for the data space.

Ex. 19 Correlations, linear dependence and small eigenvalues

1. The eigenvalues can be calculated as follows (as you may have seen in your linear algebra course):

$$\begin{aligned} & \det(C - \lambda I) = \det \begin{pmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{pmatrix} = (1-\lambda)^2 - \rho^2 = 0 \\ \implies\quad & 1 - \lambda = \pm\rho \\ \implies\quad & \lambda = 1 \pm \rho. \end{aligned}$$

If $|\rho|$ is close to 1, one of the eigenvalues is close to zero, i.e. if the random variables are highly correlated, we get one small eigenvalue.

2. The variance of x_2 has to be 1, therefore

$$V(x_2) = V(ax_1 + n) = a^2 V(x_1) + V(n) = a^2 + V(n) = 1. \tag{B.3}$$

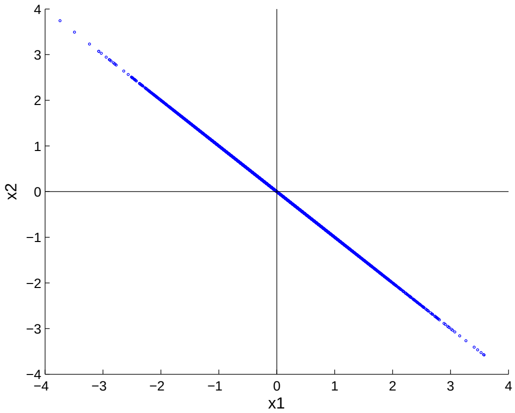
Since the mean of x_1 is zero, the variance of x_1 equals $E(x_1^2)$, which is one. Because the covariance between x_1 and x_2 has to be ρ, we get

$$\begin{aligned} \mathrm{cov}(x_1, x_2) &= E(x_1 x_2) \\ &= E(x_1(ax_1 + n)) \\ &= a \underbrace{E(x_1^2)}_{=1} + \underbrace{E(x_1)}_{=0} E(n) \\ &= \rho. \end{aligned}$$

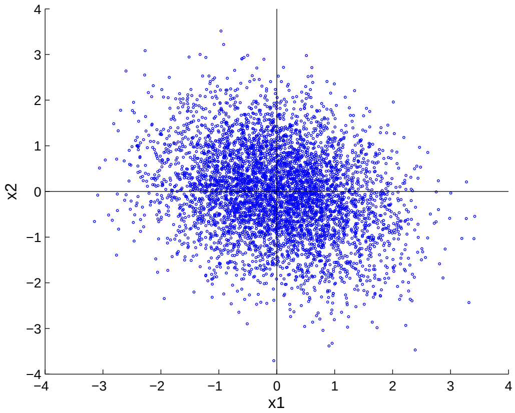
Therefore we have choose $a = \rho$. From equation $(B.3)$ we see that the noise n has to have variance $1 - \rho^2$, but that is the only criterion it has to satisfy.

3. The variances for the given ρ and the corresponding plots for 5000 samples are:

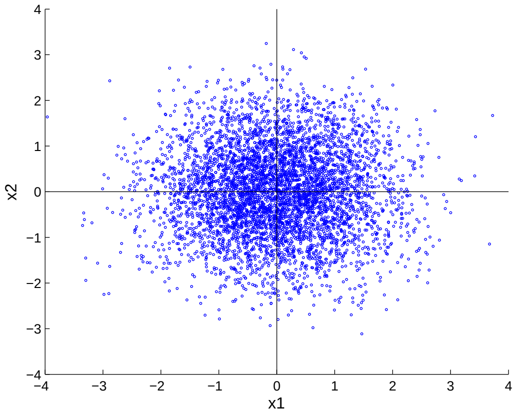
(i) $\rho = -1$: $\mathbf{V}(n) = 0$



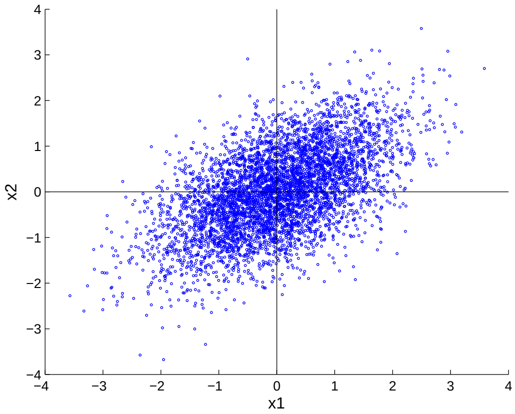
(ii) $\rho = -0.25$: $\mathbf{V}(n) = 0.9375$



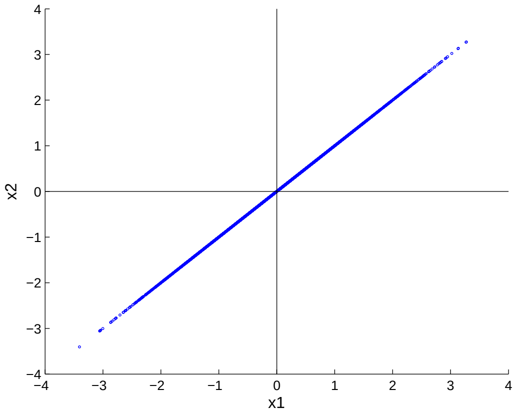
(iii) $\rho = 0$: $\mathbf{V}(n) = 1$



(iv) $\rho = 0.5$: $\mathbf{V}(n) = 0.75$



(v) $\rho = 1$: $\mathbf{V}(n) = 0$



Note that we used here a Gaussian random variable for x_1 and n. But we could also haven chosen other distributions as long as the conditions from the previous question are satisfied.

4. If $|\rho| = 1$, the variance of the noise variable is 0 and x_2 is deterministically related to x_1. Therefore $\mathbf{v}_1$ and $\mathbf{v}_2$ are linearly dependent.

   If $|\rho|$ is close to 1, the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are close to being linearly dependent. The conditioning number of C is given by

$$\frac{\lambda_{\max}}{\lambda_{\min}} = \frac{1 + |\rho|}{1 - |\rho|},$$

   which becomes arbitrary large as $|\rho| \to 1$.

   The conditioning number of $X^{\mathrm{T}} = (\mathbf{v}_1, \mathbf{v}_2)$ is a measure of the linear dependencies of $\mathbf{v}_1$ and $\mathbf{v}_2$. For any matrix M

(not necessarily square) the conditioning number is defined as

$$\operatorname{cond}(M) = \sqrt{\frac{\text{biggest eigenvalue of } M^{\mathrm{T}}M}{\text{smallest eigenvalue of } M^{\mathrm{T}}M}}.$$

In our case $M = X^{\mathrm{T}}$ and therefore $M^{\mathrm{T}}M = XX^{\mathrm{T}} = nC$, where C is the covariance matrix. The conditioning number of X^{T} is thus

$$\operatorname{cond}(X^{\mathrm{T}}) = \sqrt{\frac{n\lambda_{\max}}{n\lambda_{\min}}} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{1+|\rho|}{1-|\rho|}} = \sqrt{\operatorname{cond}(C)}.$$

If $|\rho| \to 1$, we see that the conditioning number of X^{T} and (therefore the conditioning number of C) goes up, i.e. $\mathbf{v}_1$ and $\mathbf{v}_2$ become more linearly dependent.

Ex. 20 Correlation and projection

1. Notice that for $\lambda_3 = 0$ the columns of C are linearly dependent:

$$\begin{aligned}
\cos(\alpha)\underbrace{\begin{pmatrix}1\\0\\\cos(\alpha)\end{pmatrix}}_{\text{1st column}} + \sin(\alpha)\underbrace{\begin{pmatrix}0\\1\\\sin(\alpha)\end{pmatrix}}_{\text{2nd column}} &= \begin{pmatrix}\cos(\alpha)\\\sin(\alpha)\\\cos^2(\alpha)+\sin^2(\alpha)\end{pmatrix}\\
&= \underbrace{\begin{pmatrix}\cos(\alpha)\\\sin(\alpha)\\1\end{pmatrix}}_{\text{3rd column}},
\end{aligned}$$

and therefore C is not invertible.

2. By mechanical calculation

$$\begin{aligned}
C\mathbf{u}_1 &= \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 0 & \cos(\alpha)\\0 & 1 & \sin(\alpha)\\\cos(\alpha) & \sin(\alpha) & 1\end{pmatrix}\begin{pmatrix}\cos(\alpha)\\\sin(\alpha)\\1\end{pmatrix}\\
&= \frac{1}{\sqrt{2}}\begin{pmatrix}2\cos(\alpha)\\2\sin(\alpha)\\\cos^2(\alpha)+\sin^2(\alpha)+1\end{pmatrix}\\
&= 2\frac{1}{\sqrt{2}}\begin{pmatrix}\cos(\alpha)\\\sin(\alpha)\\1\end{pmatrix}\\
&= 2\cdot\mathbf{u}_1,
\end{aligned}$$

$$\begin{aligned}
C\mathbf{u}_2 &= \begin{pmatrix}1 & 0 & \cos(\alpha)\\0 & 1 & \sin(\alpha)\\\cos(\alpha) & \sin(\alpha) & 1\end{pmatrix}\begin{pmatrix}-\sin(\alpha)\\\cos(\alpha)\\0\end{pmatrix}\\
&= \begin{pmatrix}-\sin(\alpha)\\\cos(\alpha)\\0\end{pmatrix}\\
&= 1\cdot\mathbf{u}_2,
\end{aligned}$$

$$\begin{aligned}
C\mathbf{u}_3 &= \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 0 & \cos(\alpha)\\0 & 1 & \sin(\alpha)\\\cos(\alpha) & \sin(\alpha) & 1\end{pmatrix}\begin{pmatrix}-\cos(\alpha)\\-\sin(\alpha)\\1\end{pmatrix}\\
&= \frac{1}{\sqrt{2}}\begin{pmatrix}-\cos(\alpha)+\cos(\alpha)\\-\sin(\alpha)+\sin(\alpha)\\-\cos^2(\alpha)-\sin^2(\alpha)+1\end{pmatrix}\\
&= 0\cdot\mathbf{u}_3,
\end{aligned}$$

so the eigenvalues are $\lambda_1 = 2, \lambda_2 = 1$ and $\lambda_3 = 0$.

3. Recall the exercises about eigenvalue decomposition and the formula $A = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$. Now $C = \sum_{i=1}^{3} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$:

$$\begin{aligned}
\lambda_3 \mathbf{u}_3 \mathbf{u}_3^{\mathrm{T}} &= \underbrace{\frac{1}{2}\lambda_3 \begin{pmatrix} \cos^2(\alpha) & \cos(\alpha)\sin(\alpha) & -\cos(\alpha) \\ \cos(\alpha)\sin(\alpha) & \sin^2(\alpha) & -\sin(\alpha) \\ -\cos(\alpha) & -\sin(\alpha) & 1 \end{pmatrix}}_{\text{2nd part of } C}, \\
\lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}} &= 2 \cdot \frac{1}{2} \begin{pmatrix} \cos^2(\alpha) & \cos(\alpha)\sin(\alpha) & \cos(\alpha) \\ \cos(\alpha)\sin(\alpha) & \sin^2(\alpha) & \sin(\alpha) \\ \cos(\alpha) & \sin(\alpha) & 1 \end{pmatrix}, \\
\lambda_2 \mathbf{u}_2 \mathbf{u}_2^{\mathrm{T}} &= \begin{pmatrix} \sin^2(\alpha) & -\sin(\alpha)\cos(\alpha) & 0 \\ -\sin(\alpha)\cos(\alpha) & \cos^2(\alpha) & 0 \\ 0 & 0 & 0 \end{pmatrix},
\end{aligned}$$

$$\lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}} + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^{\mathrm{T}} = \underbrace{\begin{pmatrix} 1 & 0 & \cos(\alpha) \\ 0 & 1 & \sin(\alpha) \\ \cos(\alpha) & \sin(\alpha) & 1 \end{pmatrix}}_{\text{1st part of } C}.$$

4. The principal component directions correspond to the eigenvectors of the covariance matrix C. Since we want to explain as much variance as possible, we would use the two principal components with the biggest eigenvalues, that means $s_1 = \mathbf{u}_1^{\mathrm{T}}\mathbf{x}$ and $s_2 = \mathbf{u}_2^{\mathrm{T}}\mathbf{x}$.

5. The proportion of variance explained is defined as

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i},$$

where k is the number of selected components and n the dimension of the data. Hence, for $\lambda_3 = 0.1$ we get

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{3}{3.1} \approx 0.97,$$

meaning that approximately 97% of the variance is explained by the first two principal components.

6. In this case, the projection of a point $\mathbf{x}$ is a vector defined as $\begin{pmatrix} \mathbf{u}_1^{\mathrm{T}}\mathbf{x} & \mathbf{u}_2^{\mathrm{T}}\mathbf{x} \end{pmatrix}^{\mathrm{T}}$.
For $\mathbf{y}_1 = \begin{pmatrix} x_1 & 0 & 0 \end{pmatrix}^{\mathrm{T}}$ the projection is

$$p(\mathbf{y}_1) = \begin{pmatrix} \mathbf{u}_1^{\mathrm{T}}\mathbf{y}_1 \\ \mathbf{u}_2^{\mathrm{T}}\mathbf{y}_1 \end{pmatrix} = \begin{pmatrix} x_1 \frac{1}{\sqrt{2}} \cos(\alpha) \\ -x_1 \sin(\alpha) \end{pmatrix} = x_1 \begin{pmatrix} \frac{1}{\sqrt{2}} \cos(\alpha) \\ -\sin(\alpha) \end{pmatrix}.$$

For $\mathbf{y}_2 = \begin{pmatrix} 0 & x_2 & 0 \end{pmatrix}^{\mathrm{T}}$ the projection is

$$p(\mathbf{y}_2) = \begin{pmatrix} \mathbf{u}_1^{\mathrm{T}}\mathbf{y}_2 \\ \mathbf{u}_2^{\mathrm{T}}\mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} x_2 \frac{1}{\sqrt{2}} \sin(\alpha) \\ x_2 \cos(\alpha) \end{pmatrix} = x_2 \begin{pmatrix} \frac{1}{\sqrt{2}} \sin(\alpha) \\ \cos(\alpha) \end{pmatrix}.$$

For $\mathbf{y}_3 = \begin{pmatrix} 0 & 0 & x_3 \end{pmatrix}^{\mathrm{T}}$ the projection is

$$p(\mathbf{y}_3) = \begin{pmatrix} \mathbf{u}_1^{\mathrm{T}}\mathbf{y}_3 \\ \mathbf{u}_2^{\mathrm{T}}\mathbf{y}_3 \end{pmatrix} = \begin{pmatrix} x_3 \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} = x_3 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}.$$

From these formulas we see that projecting $\mathbf{y}_i$ is the same as projecting the ith unit vector scaled by the value x_i. Thus, in the plots we only show the projection of the unit vectors $\mathbf{e}_i$. The projection of any other vector of the form $\mathbf{y}_i$ lies along the same axes.

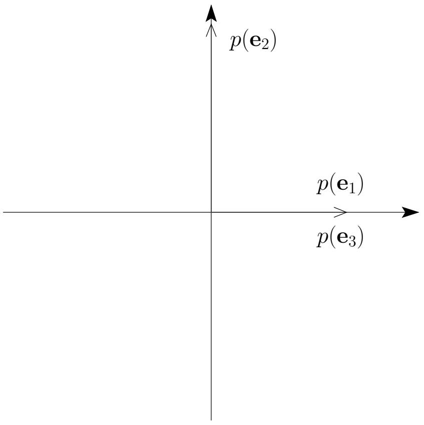
7. From the covariance matrix we see that the correlation between the first and second variable is zero, and that the correlation between the first and the third variable is given by $\cos(\alpha)$, and between the second and the third variable the correlation is given by $\sin(\alpha)$. For the values of α in the previous part of this exercise we get:

   (i) $\alpha = 0 : \rho_{13} = 1$, $\rho_{23} = 0$,

   (ii) $\alpha = \frac{\pi}{4} : \rho_{13} = \frac{1}{\sqrt{2}}$, $\rho_{23} = \frac{1}{\sqrt{2}}$,

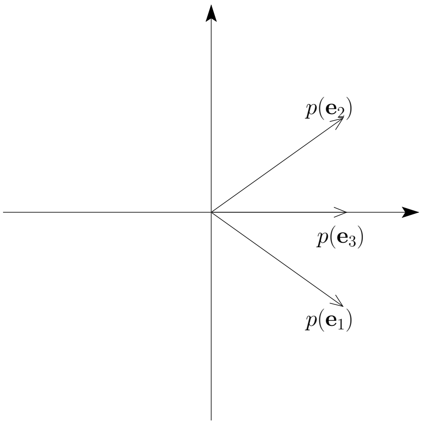
$\alpha = 0$:

$$p(\mathbf{e}_1) = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_3) = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \end{pmatrix}^{\mathrm{T}}$$



$\alpha = \frac{\pi}{4}$:

$$p(\mathbf{e}_1) = \begin{pmatrix} \frac{1}{2} & -\frac{1}{\sqrt{2}} \end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_2) = \begin{pmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} \end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_3) = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \end{pmatrix}^{\mathrm{T}}$$



(iii) $\alpha = \frac{\pi}{2} : \rho_{13} = 0$, $\rho_{23} = 0$,

(iv) $\alpha = \frac{5\pi}{6} : \rho_{13} = -\frac{\sqrt{3}}{2}$, $\rho_{23} = \frac{1}{2}$.

Relating these numbers to the projections, we see that: (1) If the axis $p(\mathbf{e}_3)$ is closer to the axis $p(\mathbf{e}_1)$ than to $p(\mathbf{e}_2)$, then the third variable is more correlated to the first variable than to the second one. (2) If the arrows, i.e. the projections, point to the same direction, the random variables are positively correlated (and otherwise negatively).

Ex. 21 PCA and linear regression

1. In matrix notation we have

$$\underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{X}^{\mathrm{T}}\beta + \underline{\epsilon}.$$

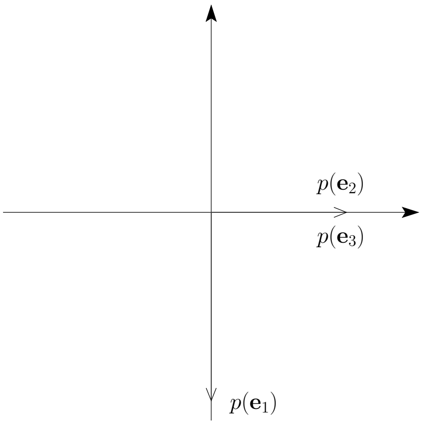
Since only $\mathbf{y}$ and $\mathbf{X}$ are observed, minimizing $J(\beta)$,

$$J(\beta) = \frac{1}{n}(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\beta),$$

gives us an estimate $\hat{\beta} = \arg\max_{\beta} J(\beta)$ of the true value of $\beta \in \mathbb{R}^p$. As usual, we calculate the gradient of J with respect to β and solve for β after setting the gradient to zero. Recall exercise 9 (Gradient of vector-valued functions)

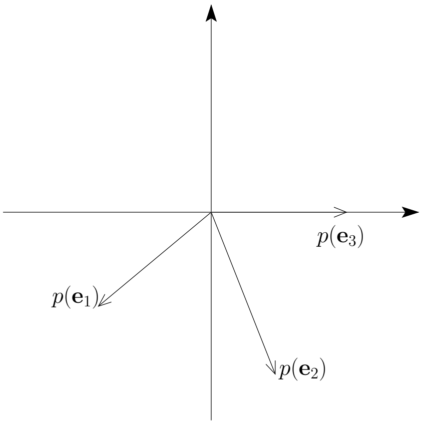
$\alpha = \frac{\pi}{2}$:

$$p(\mathbf{e}_1) = \begin{pmatrix}0 & 1\end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_2) = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_3) = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\end{pmatrix}^{\mathrm{T}}$$



$\alpha = \frac{5\pi}{6}$:

$$p(\mathbf{e}_1) \approx \begin{pmatrix}-0.61 & -0.5\end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_2) \approx \begin{pmatrix}0.35 & -0.86\end{pmatrix}^{\mathrm{T}}$$
$$p(\mathbf{e}_3) = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\end{pmatrix}^{\mathrm{T}}$$



$(*)$.

$$\begin{aligned}
& J(\beta) = \frac{1}{n}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\mathbf{X}\mathbf{y} + \beta^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\beta) \\
\Longrightarrow \quad & \nabla_\beta J(\beta) \overset{(*)}{=} \frac{1}{n}(-2\mathbf{X}\mathbf{y} + 2\mathbf{X}\mathbf{X}^{\mathrm{T}}\beta).
\end{aligned}$$

$$\begin{aligned}
& \nabla_\beta J(\beta) = 0 \\
\Longleftrightarrow \quad & \mathbf{X}\mathbf{X}^{\mathrm{T}}\beta = \mathbf{X}\mathbf{y} \\
\Longleftrightarrow \quad & \beta = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{y} = \frac{1}{n}(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{y} \overset{\text{hint}}{=} \frac{1}{n}(\hat{C}_X)^{-1}\mathbf{X}\mathbf{y} \\
\Longrightarrow \quad & \hat{\beta} = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{y} = (\hat{C}_x)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y},
\end{aligned}$$

where $\hat{C}_x$ is the sample covariance matrix. For large n, $\hat{\beta}$ converges to $\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathrm{E}(\mathbf{X}\mathbf{y})$. The first term corresponds to the covariance matrix and the second term measures the correlation between $\mathbf{X}$ and $\mathbf{y}$.

2. Since $\mathbf{y} = \mathbf{X}^{\mathrm{T}}\beta + \underline{\epsilon}$, we can write

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{y} \\
&= (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}(\mathbf{X}\mathbf{X}^{\mathrm{T}})\beta + (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\underline{\epsilon} \\
&= \beta + (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\underline{\epsilon},
\end{aligned}$$

and because every ϵ_k has zero mean, we get

$$\begin{aligned}
\mathrm{E}(\hat{\beta}|\mathbf{X}) &= \mathrm{E}(\beta|\mathbf{X}) + \mathrm{E}((\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\underline{\epsilon}|\mathbf{X}) \\
&= \beta + (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\,\underbrace{\mathrm{E}(\underline{\epsilon}|\mathbf{X})}_{=0} \\
&= \beta,
\end{aligned}$$

$$\begin{aligned}
\mathrm{V}(\hat{\beta}|\mathbf{X}) &= \mathrm{V}(\underbrace{\beta}_{\text{constant}} + \underbrace{(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}}_{\text{constant}}\underline{\epsilon}|X) \\
&= (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X})\mathrm{V}(\underline{\epsilon}|\mathbf{X})(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X})^{T} \\
&= (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\sigma^2 I((\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X})^{\mathrm{T}} \\
&= \sigma^2(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}(\mathbf{X}\mathbf{X}^{\mathrm{T}})(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1} \\
&= \sigma^2(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1} \\
&= \frac{\sigma^2}{n}(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1} \\
&= \frac{\sigma^2}{n}\hat{C}_{\mathbf{X}}^{-1}.
\end{aligned}$$

3. By writing the norm $\|\beta - \hat{\beta}\|$ differently, we get

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}(\|\beta - \hat{\beta}\||\mathbf{X}) \\
&= \mathrm{E}(\mathrm{Tr}[(\beta - \hat{\beta})(\beta - \hat{\beta})^{\mathrm{T}}]|\mathbf{X}) \\
&= \mathrm{E}(\mathrm{Tr}[(\beta - m + m - \hat{\beta})(\beta - m + m - \hat{\beta})^{\mathrm{T}}]|\mathbf{X}),
\end{aligned}$$

where $m = \mathrm{E}(\hat{\beta}|\mathbf{X})$. Notice that

$$\begin{aligned}
& (\beta - m + m - \hat{\beta})(\beta - m + m - \hat{\beta})^{\mathrm{T}} \\
= \; & (\beta - m)(\beta - m)^{\mathrm{T}} + (\beta - m)(m - \hat{\beta})^{\mathrm{T}} + \\
& (m - \hat{\beta})(\beta - m)^{\mathrm{T}} + (m - \hat{\beta})(m - \hat{\beta})^{\mathrm{T}}.
\end{aligned}$$

Because trace is a linear operation and it holds that $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$, we can take the expectation inside to get

$$\begin{aligned}
& \mathrm{MSE} \\
= \; & \mathrm{Tr}[\mathrm{E}((\beta - m)(\beta - m)^{\mathrm{T}}|\mathbf{X})] + \mathrm{Tr}[\mathrm{E}((\beta - m)(m - \hat{\beta})^{\mathrm{T}}|\mathbf{X})] + \\
& \mathrm{Tr}[\mathrm{E}((m - \hat{\beta})(\beta - m)^{\mathrm{T}})] + \mathrm{Tr}[\mathrm{E}((m - \hat{\beta})(m - \hat{\beta})^{\mathrm{T}})].
\end{aligned}$$

Now notice that

$$\begin{aligned}
\mathrm{E}((\beta - m)(m - \hat{\beta})^{\mathrm{T}}|\mathbf{X}) &= (\beta - m)(m - \mathrm{E}(\hat{\beta}|\mathbf{X}))^{\mathrm{T}} \\
&= (\beta - m)(m - m)^{\mathrm{T}} \\
&= 0,
\end{aligned}$$

which holds also for $\mathrm{E}((m - \hat{\beta})(\beta - m)^{\mathrm{T}}|\mathbf{X})$. Furthermore,

$$\mathrm{E}((\beta - m)(\beta - m)^{\mathrm{T}})|\mathbf{X}) = (\beta - m)(\beta - m)^{\mathrm{T}}),$$

because everything is deterministic here, and

$$\mathrm{E}((m - \hat{\beta})(m - \hat{\beta})^{\mathrm{T}}|\mathbf{X}) = \mathbf{V}(\hat{\beta}|X),$$

by definition of the variance and the fact that $m = \mathrm{E}(\hat{\beta}|\mathbf{X})$. Therefore

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{Tr}[(\beta - m)(\beta - m)^{\mathrm{T}}] + \mathrm{Tr}\mathbf{V}(\hat{\beta}|\mathbf{X}) \\
&= \|\beta - \mathrm{E}(\hat{\beta}|\mathbf{X})\|^2 + \mathrm{Tr}\mathbf{V}(\hat{\beta}|\mathbf{X}).
\end{aligned}$$

4. Using the previous parts of this exercise and exercise 4 (Trace, Determinants and Eigenvalues) (∗), we get

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{Tr}(\mathrm{V}(\hat{\beta}|\mathbf{X})) + \|\beta - \mathrm{E}(\hat{\beta}|\mathbf{X})\|^2 \\
&= \mathrm{Tr}(\frac{\sigma^2}{n}\hat{C}_{\mathbf{X}}^{-1}) + \|\beta - \beta\|^2 \\
&= \frac{\sigma^2}{n}\mathrm{Tr}(\hat{C}_{\mathbf{X}}^{-1}) \\
&\overset{(*)}{=} \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{1}{d_i},
\end{aligned}$$

where d_i are the eigenvalues of $\hat{C}_{\mathbf{X}}$. Hence, the small eigenvalues of $\hat{C}_{\mathbf{X}}$ cause the MSE to be large. In exercise 19 (Correlations, linear dependence and small eigenvalues) we showed that eigenvalues of $\hat{C}_{\mathbf{X}}$ are small if some random variables $\mathbf{x}_i$ are highly correlated. This means that some rows of $\mathbf{X}$ are, or are close to being, linearly dependent.

5. The vector $U_m^{\mathrm{T}}\mathbf{x}_k$ is the kth observation of the principal components. Let $z_k = U_m^{\mathrm{T}}\mathbf{x}_k$. Now

$$\begin{aligned} J(U_m\gamma) &= \frac{1}{n}\sum_{k=1}^{n}(\mathbf{y}_k - \mathbf{x}_k^{\mathrm{T}}U_m\gamma)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}(\mathbf{y}_k - (U_m^{\mathrm{T}}\mathbf{x}_k)^{\mathrm{T}}\gamma)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}(\mathbf{y}_k - \mathbf{z}_k^{\mathrm{T}}\gamma)^2 \\ &= J_{pc}(\gamma). \end{aligned}$$

Since $J(\beta) = (1/n)\sum_{k=1}^{n}(\mathbf{y}_k - \mathbf{x}_k^{\mathrm{T}}\beta)^2$, the function J_{pc} really has the same form as J, but the principal components are used instead of the original inputs.

6. Let $U = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots\mathbf{u}_m & \mathbf{u}_{m+1} & \dots\mathbf{u}_p\end{pmatrix}$ and $\mathbf{Z} = \begin{pmatrix}\mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_n\end{pmatrix}$, where $\mathbf{z}_k \in \mathbb{R}^m$ is as in the solution to the previous part of this exercise. Notice that $\mathbf{Z} = U_m^{\mathrm{T}}\mathbf{X}$. Just as in the first part (having just $\mathbf{Z}$ instead of $\mathbf{X}$), we get:

$$\begin{aligned} \hat{\gamma} &= (\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^{-1}\frac{1}{n}\mathbf{Z}\mathbf{y} \\ &= (U_m\underbrace{(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}})}_{=\hat{C}_{\mathbf{X}}}U_m)^{-1}\frac{1}{n}U_m^{\mathrm{T}}\mathbf{X}\mathbf{y} \\ &= (U_m^{\mathrm{T}}UDU^{\mathrm{T}}U_m)^{-1}U_m^{\mathrm{T}}\frac{1}{n}\mathbf{X}\mathbf{y}, \end{aligned}$$

where UDU^{T} is the eigenvalue decomposition of the sample covariance matrix $\hat{C}_{\mathbf{X}}$. On the other hand,

$$\begin{aligned} U_m^{\mathrm{T}}U &= \begin{pmatrix}\mathbf{u}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{u}_m^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_m & \mathbf{u}_{m+1} & \dots & \mathbf{u}_p\end{pmatrix} \\ &= \begin{pmatrix} 1_{11} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1_{22} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1_{mm} & 0 & \dots & 0_{mp} \end{pmatrix}, \end{aligned}$$

and therefore $U_m^{\mathrm{T}}UDU^{\mathrm{T}}U_m = D_m$, where D_m has the first m diagonal elements of D on its diagonal. Hence,

$$\hat{\gamma} = (U_m^{\mathrm{T}}UDU^{\mathrm{T}}U_m)^{-1}U_m^{\mathrm{T}}\frac{1}{n}\mathbf{X}\mathbf{y} = D_m^{-1}U_m^{\mathrm{T}}\frac{1}{n}\mathbf{X}\mathbf{y}.$$

7. The solution follows the same steps as in part 3 of this exercise: Since

$$\begin{aligned} \hat{\beta}_{pc} &= (U_mD_m^{-1}U_m^{\mathrm{T}})\frac{1}{n}\mathbf{X}\mathbf{y} \\ &= (U_mD_m^{-1}U_m^{\mathrm{T}})\frac{1}{n}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\beta + \underline{\epsilon}) \\ &= (U_mD_m^{-1}U_m^{\mathrm{T}})\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}\beta + (U_mD_m^{-1}U_m^{\mathrm{T}})\frac{1}{n}\mathbf{X}\underline{\epsilon}, \end{aligned}$$

we get

$$
\begin{aligned}
\mathrm{E}(\hat{\beta}_{pc}) &= (U_m D_m^{-1} U_m^{\mathrm{T}}) \underbrace{\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}}_{=\hat{C}_{\mathbf{X}}} \beta + (U_m D_m^{-1} U_m^{\mathrm{T}}) \frac{1}{n}\mathbf{X}\, \underbrace{\mathrm{E}(\underline{\epsilon})}_{=0} \\
&= U_m D_m^{-1} U_m^{\mathrm{T}} U D U^{\mathrm{T}} \beta \\
&= U_m D_m^{-1} \begin{pmatrix} d_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & d_m & 0 & \dots & 0 \end{pmatrix} U^{\mathrm{T}} \beta \\
&= U_m \underbrace{\begin{pmatrix} 1_{11} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1_{22} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1_{mm} & 0 & \dots & 0_{mp} \end{pmatrix} U^{\mathrm{T}}}_{=U_m^{\mathrm{T}}} \beta \\
&= U_m U_m^{\mathrm{T}} \beta,
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{V}(\hat{\beta}_{pc}|\mathbf{X}) &= \frac{1}{n^2}(U_m D_m^{-1} U_m^{\mathrm{T}}) \mathbf{X}\, \underbrace{\mathrm{E}(\underline{\epsilon\epsilon^{\mathrm{T}}})}_{=\sigma^2 I}\, \mathbf{X}^{\mathrm{T}} (U_m D_m^{-1} U_m^{\mathrm{T}})^{\mathrm{T}} \\
&= \frac{\sigma^2}{n}(U_m D_m^{-1} U_m^{\mathrm{T}}) \underbrace{\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}}_{=\hat{C}_{\mathbf{X}}} (U_m D_m^{-1} U_m^{\mathrm{T}})^{\mathrm{T}} \\
&= \frac{\sigma^2}{n} U_m D_m^{-1} \underbrace{U_m^{\mathrm{T}} U D U^{\mathrm{T}} U_m}_{=D_m} D_m^{-1} U_m^{\mathrm{T}} \\
&= \frac{\sigma^2}{n} U_m D_m^{-1} U_m^{\mathrm{T}}.
\end{aligned}
$$

8. Because $\mathrm{Tr}(cAB) = c\mathrm{Tr}(BA)$ for any constant c, the trace of $\mathrm{V}(\hat{\beta}_{pc}|\mathbf{X})$ equals

$$
\begin{aligned}
\mathrm{Tr}(\mathrm{V}(\hat{\beta}_{pc}|\mathbf{X})) &= \mathrm{Tr}(\frac{\sigma^2}{n}(U_m D_m^{-1}) U_m^{\mathrm{T}}) \\
&= \frac{\sigma^2}{n}\mathrm{Tr}(U_m^{\mathrm{T}} U_m D_m^{-1}) \\
&= \frac{\sigma^2}{n}\mathrm{Tr}(D_m^{-1}) \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{m}\frac{1}{d_i},
\end{aligned}
$$

and therefore the MSE for $\hat{\beta}$ is

$$
\begin{aligned}
\mathrm{MSE}_{pc} &= \mathrm{Tr}(\mathrm{V}(\hat{\beta}_{pc}|\mathbf{X})) + \|\beta - \mathrm{E}(\hat{(\beta)}_{pc}|\mathbf{X})\|^2 \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{m}\frac{1}{d_i} + \|\beta - U_m U_m^{\mathrm{T}}\beta\|^2 \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{m}\frac{1}{d_i} + \|\beta(I_p - U_m U_m^{\mathrm{T}})\|^2.
\end{aligned}
$$

If $m = p$, then $U_m = U$ and $UU^{\mathrm{T}} = I_p$, so MSE_{pc} becomes $(\sigma^2/n)\sum_{i=1}^{p}(1/d_i)$, i.e. it is equal to the MSE of $\hat{\beta}$.

If $m < p$, then the variance is reduced by $(\sigma^2/n)\sum_{i=m+1}^{p}(1/d_i)$ but we incur a bias since $U_m U_m^{\mathrm{T}} \neq I_p$. This is called the bias-variance trade-off: by choosing m, one can choose a certain reduction in variance, at the cost of more bias. The best m is the one which leads to the smallest MSE. The formula for the MSE_{pc} show that the best m is essentially a function of d_i and $\mathbf{u}_i$, i.e. the covariance matrix of $\mathbf{X}$.

Ex. 22 Least squares for factor analysis and PCA

1. Direct calculations give

$$\begin{aligned}
\mathrm{Tr}(A^{\mathrm{T}}B) &= \sum_i \left[\sum_j (A^{\mathrm{T}})_{ij}(B)_{ji}\right] \\
&= \sum_i \left[\sum_j (A)_{ji}(B)_{ji}\right] \\
&= \sum_{ij} (A)_{ji}(B)_{ji} \\
&= \sum_{ij} (A)_{ij}(B)_{ij}, \\
\mathrm{Tr}(AB^{\mathrm{T}}) &= \sum_i \left[\sum_j (A)_{ij}(B^{\mathrm{T}})_{ji}\right] \\
&= \sum_i \left[\sum_j (A)_{ij}(B)_{ij}\right] \\
&= \sum_{ij} (A)_{ij}(B)_{ij}.
\end{aligned}$$

2. Remember that, (a) $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$, (b) $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and (c) $\mathrm{Tr}(A) = \mathrm{Tr}(A^{\mathrm{T}})$. Remember also that the trace of a real number is just the number itself.

$$\begin{aligned}
J_{ls}(a) &= \|C - \mathbf{a}\mathbf{a}^{\mathrm{T}}\|^2 \\
&\overset{\text{part 1}}{=} \mathrm{Tr}((C - \mathbf{a}\mathbf{a}^{\mathrm{T}})(C - \mathbf{a}\mathbf{a}^{\mathrm{T}})^{\mathrm{T}}) \\
&= \mathrm{Tr}((C - \mathbf{a}\mathbf{a}^{\mathrm{T}})(C^{\mathrm{T}} - \mathbf{a}\mathbf{a}^{\mathrm{T}})) \\
&= \mathrm{Tr}(CC^{\mathrm{T}} - C\mathbf{a}\mathbf{a}^{\mathrm{T}} + \mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{a}\mathbf{a}^{\mathrm{T}} - \mathbf{a}\mathbf{a}^{\mathrm{T}}C^{\mathrm{T}}) \\
&\overset{(a)\&(c)}{=} \mathrm{Tr}(CC^{\mathrm{T}} + \mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{a}\mathbf{a}^{\mathrm{T}} - 2C\mathbf{a}\mathbf{a}^{\mathrm{T}}) \\
&\overset{(a)}{=} \mathrm{Tr}(CC^{\mathrm{T}}) + \mathrm{Tr}(\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{a}\mathbf{a}^{\mathrm{T}}) - \mathrm{Tr}(2C\mathbf{a}\mathbf{a}^{\mathrm{T}}) \\
&\overset{(b)\&(c)}{=} \mathrm{Tr}(CC^{\mathrm{T}}) + \mathrm{Tr}(\underbrace{\mathbf{a}^{\mathrm{T}}\mathbf{a}}_{=\|a\|^2}\ \mathbf{a}^{\mathrm{T}}\mathbf{a}) - 2\mathrm{Tr}(\underbrace{\mathbf{a}^{\mathrm{T}}C\mathbf{a}}_{\in\mathbb{R}}) \\
&\overset{C^{\mathrm{T}}=C}{=} \|\mathbf{a}\|^4 - 2\mathbf{a}^{\mathrm{T}}C\mathbf{a} + \mathrm{Tr}(CC).
\end{aligned}$$

3. Recall exercise 9 (Gradient of vector-valued functions):

$$\begin{aligned}
J_{ls}(a) &= (\mathbf{a}^{\mathrm{T}}\mathbf{a})^2 - 2\mathbf{a}^{\mathrm{T}}C\mathbf{a} + \mathrm{Tr}(CC) \\
\Longrightarrow \nabla J_{ls}(a) &= 2\mathbf{a}^{\mathrm{T}}\mathbf{a} \cdot 2\mathbf{a} - 2C\mathbf{a} - 2\underbrace{C^{\mathrm{T}}}_{=C}\mathbf{a} + 0 \\
&= 4\|a\|^2\mathbf{a} - 4C\mathbf{a}.
\end{aligned}$$

4. Let $\mathbf{v}$ be such vector that $\nabla J_{ls}(\mathbf{v}) = 0$. Now

$$\nabla J_{ls}(\mathbf{v}) = 0 \iff C\mathbf{v} = \|\mathbf{v}\|^2\mathbf{v}.$$

Therefore $\mathbf{v}$ is an eigenvector by definition.

Let $\mathbf{a}^* = \alpha\mathbf{e}$, where α is a scalar and $\mathbf{e}$ is an eigenvector of C with unit norm and eigenvalue λ. Since

$$\begin{aligned}
\nabla J_{ls}(\mathbf{a}^*) = 0 &\iff \|\mathbf{a}^*\|^2\mathbf{a}^* = C\mathbf{a}^* \\
&\iff \alpha\|\alpha\mathbf{e}\|^2\mathbf{e} = \alpha C\mathbf{e} \\
&\iff \alpha^2\mathbf{e} = \lambda\mathbf{e} \\
&\Longrightarrow \alpha = \pm\sqrt{\lambda},
\end{aligned}$$

the only possible scalars α are $\pm\sqrt{\lambda}$. Notice that because covariance matrices are positive-semidefinite, the eigenvalues are non-negative.

5. Notice that

$$\begin{aligned}
J_{ls}(\mathbf{a}^*) &= (\pm\sqrt{\lambda})^2 - 2(\pm\sqrt{\lambda})^2\mathbf{e}^{\mathrm{T}}C\mathbf{e} + \mathrm{Tr}(CC) \\
&= \lambda^2 - 2\lambda\lambda + \mathrm{Tr}(CC) \\
&= -\lambda^2 + \mathrm{Tr}(CC).
\end{aligned}$$

Since $J_{ls}(a^*_{\lambda_1}) < J_{ls}(a^*_{\lambda_2})$ if $\lambda_1^2 > \lambda_2^2$ and all the eigenvalues of C are non-negative, we see that $J_{ls}(a^*_{\lambda_1}) < J_{ls}(a^*_{\lambda_2})$ if $\lambda_1 > \lambda_2$. Therefore the eigenvector with the largest eigenvalue minimizes J_{ls}.

Ex. 23 Derivation of quartimax update rule

1. For $G(y) = y^4$ we have

$$J(U) = \sum_{ij} G((AU)_{ij}) = \sum_{ij}((AU)_{ij})^4 = \sum_{ij}(\sum_k a_{ik}u_{kj})^4.$$

We can use the gradient ascent to optimize the function J with the constraint that U must be orthogonal:

$$\begin{aligned} U &\leftarrow U + \mu\nabla J(U) \text{ (update step)} \\ U &\leftarrow (UU^{\mathrm{T}})^{-1/2}U \text{ (orthogonalization).} \end{aligned}$$

Thus, we have to compute the the gradient of J:

$$\nabla J(U) = \begin{pmatrix} \frac{\partial J}{\partial u_{11}} & \frac{\partial J}{\partial u_{12}} & \cdots & \frac{\partial J}{\partial u_{1n}} \\ \frac{\partial J}{\partial u_{21}} & \frac{\partial J}{\partial u_{22}} & \cdots & \frac{\partial J}{\partial u_{2n}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial J}{\partial u_{n1}} & \frac{\partial J}{\partial u_{n2}} & \cdots & \frac{\partial J}{\partial u_{nn}} \end{pmatrix}.$$

By chain rule, we get:

$$\begin{aligned} \frac{\partial J}{\partial u_{pq}} &= \frac{\partial}{\partial u_{pq}} \sum_{ij}(\sum_k a_{ik}u_{kj})^4 \\ &= \sum_i 4(\sum_k a_{ik}u_{kq})^3 \underbrace{\frac{\partial}{\partial u_{pq}} \sum_k a_{ik}u_{kq}}_{= a_{ip}} \\ &= 4\sum_i ((AU)_{iq})^3 (A)_{ip}, \end{aligned}$$

so the gradient is:

$$\nabla J(U) = 4(AU)^{(3)}A,$$

where $(AU)^{(3)}$ stands for taking the third power of matrix AU component wise.

2. Using exercise 22 (Least Squares for Factor Analysis and PCA) (∗), we get for $G(y) = y^2$

$$\begin{aligned} J(U) &= \sum_{ij} G((AU)_{ij}) \\ &= \sum_{ij} (AU)^2_{ij} \\ &= \|AU\|^2 \\ &\overset{(*)}{=} \mathrm{Tr}(AU(AU)^{\mathrm{T}}) \\ &= \mathrm{Tr}(A\underbrace{UU^{\mathrm{T}}}_{=I}A^{\mathrm{T}}) \\ &= \mathrm{Tr}(AA^{\mathrm{T}}), \end{aligned}$$

which is independent of U, i.e. $J(U)$ is constant for all orthogonal U.

Ex. 24 Kurtosis

1. Uniform distribution

$$\begin{aligned}
\mathrm{E}(x^4) &= \int_{-\sqrt{3}}^{\sqrt{3}} x^4 \frac{1}{2\sqrt{3}}\,\mathrm{d}x \\
&= \frac{1}{10\sqrt{3}}((\sqrt{3})^5 + (\sqrt{3})^5) \\
&= \frac{1}{10\sqrt{3}} 2(\sqrt{3})^5 \\
&= \frac{1}{5}(\sqrt{3})^4 \\
&= \frac{9}{5}. \\
\mathrm{E}(x^2) &= \int_{-\sqrt{3}}^{\sqrt{3}} x^2 \frac{1}{2\sqrt{3}}\,\mathrm{d}x = \frac{1}{6\sqrt{3}} 2(\sqrt{3})^3 = \frac{3}{3} = 1. \\
\Longrightarrow \mathrm{kurt}(x) &= \frac{9}{5} - 3 = \frac{9-15}{5} = -\frac{6}{5}.
\end{aligned}$$

2. Laplacian distribution

$$\begin{aligned}
& \mathrm{E}(x^2) \\
= \quad & \int_{-\infty}^{\infty} \frac{1}{\sqrt{2}} x^2 \ \exp(-\sqrt{2}|x|)\,\mathrm{d}x \\
\overset{\text{by symmetry}}{=} \quad & \sqrt{2}\int_0^{\infty} x^2 \exp(-\sqrt{2}x)\,\mathrm{d}x \\
\underset{\text{integration}}{\overset{\text{partial}}{=}} \quad & \underbrace{\left[-\frac{\sqrt{2}}{\sqrt{2}} x^2 \exp(-\sqrt{2}x)\right]_0^{\infty}}_{=0} + \int_0^{\infty} 2x \exp(-\sqrt{2}x)\,\mathrm{d}x \\
\underset{\text{integration}}{\overset{\text{partial}}{=}} \quad & \underbrace{\left[-\frac{2}{\sqrt{2}} x \exp(-2\sqrt{2}x)\right]_0^{\infty}}_{=0} + \int_0^{\infty} \frac{2}{\sqrt{2}} \exp(-\sqrt{2}x)\,\mathrm{d}x \\
= \quad & \Big[-\exp(-\sqrt{2}x)\Big]_0^{\infty} \\
= \quad & 0 + 1 \\
= \quad & 1.
\end{aligned}$$

$$\begin{aligned}
& \mathrm{E}(x^4) \\
= \quad & \int_{-\infty}^{\infty} \frac{1}{\sqrt{2}} x^4 \ \exp(-\sqrt{2}|x|)\,\mathrm{d}x \\
\overset{\text{by symmetry}}{=} \quad & \sqrt{2}\int_0^{\infty} x^4 \exp(-\sqrt{2}x)\,\mathrm{d}x \\
\underset{\text{integration}}{\overset{\text{partial}}{=}} \quad & \sqrt{2}\left(\underbrace{\left[-\frac{x^4}{\sqrt{2}} \exp(-\sqrt{2}x)\right]_0^{\infty}}_{=0} + \int_0^{\infty} \frac{4}{\sqrt{2}} x^3 \exp(-\sqrt{2}x)\,\mathrm{d}x\right) \\
\underset{\text{integration}}{\overset{\text{partial}}{=}} \quad & 4\left(\underbrace{\left[\frac{-x^3}{\sqrt{2}} \exp(-\sqrt{2}x)\right]_0^{\infty}}_{=0} + \int_0^{\infty} \frac{3}{\sqrt{2}} x^2 \exp(-\sqrt{2}x)\,\mathrm{d}x\right) \\
= \quad & 6\int_{-\infty}^{\infty} \frac{1}{\sqrt{2}} x^2 \exp(-\sqrt{2}|x|)\,\mathrm{d}x \\
= \quad & 6\mathrm{E}(x^2).
\end{aligned}$$

$$\Longrightarrow \mathrm{kurt}(x) \quad = \quad 6\mathrm{E}(x^2) - 3(\mathrm{E}(x^2))^2 = 6 - 3 = 3.$$

3. Gaussian distribution with mean zero and variance σ^2.

$$\begin{aligned}
& \mathrm{E}(x^2) \\
= \quad & \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \, \mathrm{d}x \\
\overset{\text{partial}}{\underset{\text{integration}}{=}} \quad & \frac{1}{\sqrt{2\pi\sigma^2}} \underbrace{\left[\frac{x^3}{3} \exp\left(-\frac{x^2}{2\sigma^2}\right)\right]_{-\infty}^{\infty}}_{=0} - \\
& \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{x^3}{3} \left(-\frac{x}{\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2}\right) \\
= \quad & \frac{1}{3\sigma^2} \int_{-\infty}^{\infty} x^4 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\
= \quad & \frac{1}{3\sigma^2} \mathrm{E}(x^4).
\end{aligned}$$

Since mean μ is zero, we have

$$\sigma^2 \quad = \quad \mathrm{E}(x^2)),$$

so that

$$\begin{aligned}
& \mathrm{E}(x^2) = \frac{1}{3\sigma^2}\mathrm{E}(x^4) = \frac{1}{3\mathrm{E}(x^2)}\mathrm{E}(x^4) \\
\Longrightarrow \quad & (E(x^2))^2 = \frac{1}{3}\mathrm{E}(x^4) \\
\Longrightarrow \quad & \mathrm{kurt}(x) = \mathrm{E}(x^4) - 3(\mathrm{E}(x^2))^2 = \mathrm{E}(x^4) - \mathrm{E}(x^4) = 0.
\end{aligned}$$

4. Gaussian scale mixture. Denote

$$\begin{aligned}
p(x) \quad &= \quad \frac{1}{2}\left[\frac{1}{\sqrt{2\pi\sigma_1^2}} \, \exp\left(-\frac{x^2}{2\sigma_1^2}\right) + \frac{1}{\sqrt{2\pi\sigma_2^2}} \, \exp\left(-\frac{x^2}{2\sigma_2^2}\right)\right] \\
&= \quad \frac{1}{2}(p_1(x) + p_2(x)).
\end{aligned}$$

Calculate kurtosis piece by piece:

$$\begin{aligned}
\mathrm{E}(x^2) \quad & \overset{\mu=0}{=} \quad \mathrm{V}(x) \\
&= \quad \frac{1}{2}\left[\int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x\right] \\
&= \quad \frac{1}{2}\left[\int_{-\infty}^{\infty} x^2 p_1(x)\mathrm{d}x + \int_{-\infty}^{\infty} x^2 p_2(x)\mathrm{d}x\right] \\
&= \quad \frac{1}{2}(\sigma_1^2 + \sigma_2^2).
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(x^4) \quad &= \quad \frac{1}{2}\left[\int_{-\infty}^{\infty} x^4 p(x)\mathrm{d}x\right] \\
&= \quad \frac{1}{2}\left[\int_{-\infty}^{\infty} x^4 p_1(x)\mathrm{d}x + \int_{-\infty}^{\infty} x^4 p_2(x)\mathrm{d}x\right] \\
& \overset{\text{part 1}}{=} \quad \frac{1}{2}(3(\sigma_1^2)^2 + 3(\sigma_2^2)^2) \\
&= \quad \frac{3}{2}(\sigma_1^4 + \sigma_2^4).
\end{aligned}$$

$$\begin{aligned}
\mathrm{kurt}(x) \quad &= \quad \mathrm{E}(x^4) - 3(\mathrm{E}(x^2))^2 \\
&= \quad \mathrm{E}(x^4) - 3\mathrm{V}(x)^2 \\
&= \quad \frac{3}{2}(\sigma_1^4 + \sigma_2^4) - 3 \cdot \frac{1}{4}(\sigma_1^2 + \sigma_2^2)^2 \\
&= \quad \frac{3}{2}\left[\sigma_1^4 + \sigma_2^4 - \frac{1}{2}(\sigma_1^4 + 2\sigma_1^2\sigma_2^2 + \sigma_2^4)\right] \\
&= \quad \frac{3}{2}\left[\frac{1}{2}\sigma_1^4 - \sigma_1^2\sigma_2^2 + \frac{1}{2}\sigma_2^4\right] \\
&= \quad \frac{3}{4}(\sigma_1^4 - 2\sigma_1^2\sigma_2^2 + \sigma_2^4) \\
&= \quad \frac{3}{4}(\sigma_1^2 - \sigma_2^2)^2.
\end{aligned}$$

For $\sigma_1 \neq \sigma_2$, $(\sigma_1^2 - \sigma_2^2)^2$ is always larger than zero and thus, $\mathrm{kurt}(x) > 0$. If $\sigma_1 = \sigma_2$, then $\mathrm{kurt}(x) = 0$, since $p(x)$ is then just an ordinary Gaussian distribution as in part 1 of this exercise.

5. Mixture of Gaussians for the same variance but different means. Denote

$$\begin{aligned}
p(x) &= \frac{1}{3}(p_\mu(x) + p_0(x) + p_{-\mu}(x)) \\
&= \frac{1}{3\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2}\right) \mathrm{d}x + \\
&\quad \frac{1}{3\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x + \\
&\quad \frac{1}{3\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x+\mu)^2}{2}\right) \mathrm{d}x.
\end{aligned}$$

Since $\mathrm{E}(x^2) = \mathrm{V}(x) + \mathrm{E}(x)^2$, we get

$$\begin{aligned}
\mathrm{E}(x^2) &= \frac{1}{3}\left[\underbrace{\int_{-\infty}^{\infty} x^2 p_\mu(x)\mathrm{d}x}_{=\mathrm{E}_{p_\mu}(x)} + \underbrace{\int_{-\infty}^{\infty} x^2 p_0(x)\mathrm{d}x}_{=\mathrm{E}_{p_0}(x)} + \underbrace{\int_{-\infty}^{\infty} p_{-\mu}(x)\mathrm{d}x}_{=\mathrm{E}_{p_{-\mu}}(x)}\right] \\
&= \frac{1}{3}\left((1+\mu^2) + (1+0^2) + (1+(-\mu)^2)\right) \\
&= \frac{1}{3}(3 + 2\mu^2) \\
&= 1 + \frac{2}{3}\mu^2. \\
\mathrm{E}(x^2)^2 &= 1 + \frac{4}{3}\mu^2 + \frac{4}{9}\mu^4.
\end{aligned}$$

For the calculation of the 4-th moment, notice that $h(y) = y^3 \ \exp\left(-\frac{y^2}{2\sigma_y^2}\right)$ is an odd function, i.e.

$$\begin{aligned}
h(-y) &= (-y)^3 \ \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \\
&= -y^3 \ \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \\
&= -h(y).
\end{aligned}$$

Note also that the function y^3 goes to infinity slower than the exponential function goes to zero so that the integral

$$\int_0^{\infty} h(y) \ \mathrm{d}y$$

exists. As $h(y)$ is odd symmetric, we have furthermore that

$$\int_0^{\infty} h(y) \ \mathrm{d}y = -\int_{-\infty}^{0} h(y) \ \mathrm{d}y.$$

Hence,

$$\begin{aligned}
\mathrm{E}(y^3) &= \frac{1}{\sqrt{2\pi\sigma_y^2}} \int_{-\infty}^{\infty} y^3 \ \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \ \mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi\sigma_y^2}} \left[\int_0^{\infty} h(y) \ \mathrm{d}y + \int_{-\infty}^{0} h(y) \ \mathrm{d}y\right] \\
&= 0.
\end{aligned}$$

The third moment $\mathrm{E}(y^3)$ (skewness) is zero for zero mean Gaussians.

For $\mathrm{E}(x^4)$ we need also to calculate $\mathrm{E}_{p_\mu}(x^4)$ and $\mathrm{E}_{p_{-\mu}}(x^4)$ (we can use part 1 of the exercise for $\mathrm{E}_{p_0}(x)$):

$$\begin{aligned}
\mathrm{E}_{p_\mu}(x^4) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 \ \exp\left(-\frac{(x-\mu)^2}{2}\right)\mathrm{d}x \\
&\overset{\text{change of variable}}{\underset{u=x-\mu}{=}} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (u+\mu)^4 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (u^4+4u^3\mu+6u^2\mu^2+4u\mu^3+ \\
&\quad \mu^4) \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u^4 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u + \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 4u^3\mu \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u + \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 6u^2\mu^2 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u + \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 4u\mu^3 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u + \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mu^4 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \\
&= \underbrace{\mathrm{E}(u^4)}_{\text{part 1}} + \underbrace{4\mu\mathrm{E}(u^3)}_{\text{skewness}} + \underbrace{6\mu^2\mathrm{E}(u^2)}_{\text{unit variance}} + \\
&\quad \underbrace{4\mu^3\mathrm{E}(u)}_{\text{zero mean}} + \mu^4 \\
&= 3+0+6\mu^2+0+\mu^4 \\
&= 3+6\mu^2+\mu^4.
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}_{p_{-\mu}}(x^4) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 \ \exp\left(-\frac{(x+\mu)^2}{2}\right)\mathrm{d}x \\
&\overset{\text{change of variable}}{\underset{u=x+\mu}{=}} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (u-\mu)^4 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u^4 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u - \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 4u^3\mu \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u + \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 6u^2\mu^2 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u - \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 4u\mu^3 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u + \\
&\quad \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mu^4 \ \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \\
&= \mathrm{E}(u^4) - 4\mu\mathrm{E}(u^3) + 6\mu^2\mathrm{E}(u^2) - \\
&\quad 4\mu^4\mathrm{E}(u) + \mu^4 \\
&\overset{\text{as earlier}}{=} 3+6\mu^2+\mu^4.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{E}(x^4) &= \frac{1}{3}\left[\underbrace{\int_{-\infty}^{\infty} x^4 p_\mu(x)\mathrm{d}x}_{=\mathrm{E}_{p_\mu}(x^4)} + \underbrace{\int_{-\infty}^{\infty} x^4 p_0(x)\mathrm{d}x}_{\text{part 1}} + \underbrace{\int_{-\infty}^{\infty} x^4 p_{-\mu}\mathrm{d}x}_{=\mathrm{E}_{p_{-\mu}}(x^4)}\right] \\
&= \frac{1}{3}\left((3+6\mu^2+\mu^4)+3+(3+6\mu^2+\mu^4)\right) \\
&= \frac{1}{3}(9+12\mu^2+2\mu^4) \\
&= 3+4\mu^2+\frac{2}{3}\mu^4,
\end{aligned}$$

and thus, the kurtosis is

$$\begin{aligned}
\mathrm{kurt}(x) &= \mathrm{E}(x^4) - 3\mathrm{E}(x^2)^2 \\
&= 3+4\mu^2+\frac{2}{3}\mu^4 - 3(1+\frac{4}{3}\mu^2+\frac{4}{9}\mu^4) \\
&= 3+4\mu^2+\frac{2}{3}\mu^4 - 3 - 4\mu^2 - \frac{4}{3}\mu^4 \\
&= -\frac{2}{3}\mu^4,
\end{aligned}$$

which is always negative for nonzero mean.

6. The linear properties of kurtosis were introduced in Section 7.3. Using those, we get

$$\mathrm{kurt}(x+\alpha y) = \mathrm{kurt}(x) + \alpha^4 \mathrm{kurt}(y) = \frac{3}{4}(\sigma_1^2-\sigma_2^2)^2 - \alpha^4\frac{2}{3}\mu^4.$$

Setting kurt(x) to zero gives us

$$\begin{aligned}
\alpha^4\frac{2}{3}\mu^4 = \frac{3}{4}(\sigma_1^2-\sigma_2^2)^2 \quad &\Longleftrightarrow \quad \alpha^4 = \frac{9}{8}(\sigma_1^2-\sigma_2^2)^2\frac{1}{\mu^4} \\
&\Longleftrightarrow \quad \alpha = \pm\sqrt[4]{\frac{9}{8}}\sqrt{|\sigma_1^2-\sigma_2^2|}\frac{1}{\mu}.
\end{aligned}$$

Ex. 25 Kurtosis-based ICA

1. For $g(u)=u$ we have

$$\mathrm{E}(\mathbf{z}g(\mathbf{w}^{\mathrm{T}}\mathbf{z})) = \mathrm{E}(\mathbf{z}\underbrace{\mathbf{w}^{\mathrm{T}}\mathbf{z}}_{\in\mathbb{R}}) = \mathrm{E}(\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{w}) = \mathrm{E}(\mathbf{z}\mathbf{z}^{\mathrm{T}})\mathbf{w} = \Sigma_{\mathbf{z}}\mathbf{w},$$

where $\Sigma_{\mathbf{z}}$ is the covariance matrix. The iteration is then

$$\begin{aligned}
\mathbf{w} &\leftarrow \mathbf{w}+\gamma\Sigma_{\mathbf{z}}\mathbf{w} \\
\mathbf{w} &\leftarrow \frac{\mathbf{w}}{||\mathbf{w}||},
\end{aligned}$$

which is the gradient-iteration for calculating the first principal component (see exercise 14 on the power method).

2. If $\Sigma_{\mathbf{z}} = I$ (that is if $\mathbf{z}$ is white), the direction of $\mathbf{w}$ is not changed. Nothing happens.

Ex. 26 Skewness-based ICA

1. The gradient for $J(\mathbf{w}) = \mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^3)$ is

$$\nabla J(\mathbf{w}) = \mathrm{E}(3(\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z}) = 3\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z}).$$

2. Gradient-ascent optimization becomes now

$$\begin{aligned}
\mathbf{w} &\leftarrow \mathbf{w}+\underbrace{3\tilde{\mu}}_{=\mu}\ \mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z}) \\
\mathbf{w} &\leftarrow \frac{\mathbf{w}}{||\mathbf{w}||}.
\end{aligned}$$

3. Notice that

$$\begin{aligned}\frac{\mathbf{w}}{\|\mathbf{w}\|} &= \frac{\mathbf{w}+\mu\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}{\sqrt{(\mathbf{w}+\mu\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})^{\mathrm{T}}(\mathbf{w}+\mu\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}} \\ &= \frac{\mathbf{w}+\mu\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}+2\mu\mathbf{w}^{\mathrm{T}}\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})+\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})^{\mathrm{T}}\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}} \\ &= \frac{\frac{\mathbf{w}}{\mu}+\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}{\sqrt{\frac{1}{\mu^2}+\frac{2}{\mu}\mathbf{w}^{\mathrm{T}}\,\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})+\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})^{\mathrm{T}}\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}}.\end{aligned}$$

Thus, the limit of $\mu \to \infty$ is:

$$\mathbf{w} \leftarrow \frac{0+\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}{\sqrt{0+0+\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})^{\mathrm{T}}\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}} = \frac{\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})}{\|\mathrm{E}((\mathbf{w}^{\mathrm{T}}\mathbf{z})^2\mathbf{z})\|}.$$

Ex. 27 Another reason why Gaussian variables don't work for ICA.

1. We need to know $p_{\mathbf{z}}$, the probability distribution function of $\mathbf{z}$, where $\mathbf{z} = A\mathbf{s}$, with

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{k} p_i(s_i)$$

by definition of the ICA model. Recall the theory about linear transformations of random variables. Because A is invertible, we get

$$\begin{aligned}p_{\mathbf{z}}(\mathbf{z}) &= p_{\mathbf{s}}(A^{-1}\mathbf{z}) \cdot |\det(A^{-1})| \\ &= p_{\mathbf{s}}(A^{\mathrm{T}}\mathbf{z}) \cdot \underbrace{|\det(A)|}_{=1} \\ &= \prod_{i=1}^{k} p_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{z}),\end{aligned}$$

where the $\mathbf{a}_i$ are the columns of A. Because the data $\mathbf{z}$ is iid, the likelihood (as a function of A) is therefore

$$L(A) = \prod_{j=1}^{n}\prod_{i=1}^{k} p_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{z}_i)$$

and thus, the log-likelihood (as a function of A) is

$$\begin{aligned}\log L(A) &= \sum_{j=1}^{n} \log \prod_{i=1}^{k} p_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{z}_j) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{k} \log p_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{z}_j).\end{aligned}$$

2. Assume now that the p_i are Gaussian, i.e.

$$p_i(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Now the log-likelihood becomes

$$\begin{aligned}
\log L(A) &= \sum_{j=1}^{n}\sum_{i=1}^{k}\left(\log\left[\frac{1}{\sqrt{2\pi}}\right]-\frac{(\mathbf{a}_i^{\mathrm{T}}\mathbf{z}_j)^2}{2}\right)\\
&= \underbrace{\sum_{j=1}^{n}\sum_{i=1}^{k}\log\left[\frac{1}{\sqrt{2\pi}}\right]}_{=c_1(\text{constant})}-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{k}(\underbrace{\mathbf{a}_i^{\mathrm{T}}\mathbf{z}_j}_{\in\mathbb{R}})^2\\
&= c_1-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{k}\mathbf{z}_j^{\mathrm{T}}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\mathbf{z}_j\\
&= c_1-\frac{1}{2}\sum_{j=1}^{n}\mathbf{z}_j^{\mathrm{T}}\underbrace{\left[\sum_{i=1}^{k}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}\right]}_{=AA^T=I\text{ (orthogonality of }A)}\mathbf{z}_j\\
&= c_1-\underbrace{\frac{1}{2}\sum_{j=1}^{n}\mathbf{z}_j^{\mathrm{T}}\mathbf{z}_j}_{=c_2\text{ (constant)}}\\
&= c_1-c_2.
\end{aligned}$$

Therefore the log-likelihood function is a constant and doesn't depend anymore on the matrix A. Hence, if one assumes Gaussian sources s_i all orthogonal matrices A give an equally likely fit to the data. The "true" orthogonal matrix A can thus not be found if the sources are Gaussian. In other words, the ICA model is not identifiable if the s_i are Gaussian.

Ex. 28 Maximum Likelihood Estimation of the ICA Model

1. Notice that

$$\begin{aligned}
&\int_{-\infty}^{\infty} sg(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s\\
\overset{\text{partial}}{\underset{\text{integration}}{=}}\;&\underbrace{\left[-g(s)\exp\left(-\frac{s^2}{2}\right)\right]_{-\infty}^{\infty}}_{=0}+\int_{-\infty}^{\infty}g'(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s\\
=\;&\int_{-\infty}^{\infty}g'(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s,
\end{aligned}$$

since g was assumed to grow slower than $\exp(s_i^2/2)$. Therefore we get

$$\begin{aligned}
&\mathrm{E}(sg(s)-g'(s))\\
=\;&\int_{-\infty}^{\infty}(sg(s)-g'(s))\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s\\
=\;&\frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{\infty}sg(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s-\int_{-\infty}^{\infty}g'(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s\right]\\
=\;&\frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{\infty}g'(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s-\int_{-\infty}^{\infty}g'(s)\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s\right]\\
=\;&0,
\end{aligned}$$

and hence the condition is not fulfilled.

2. For $g_i(s_i)=s_i^3$, we have

$$\begin{aligned}
\mathrm{E}(s_ig_i(s_i)-g_i'(s_i)) &= \mathrm{E}(s_i^4-3s_i^2)\\
&= \mathrm{E}(s_i^4)-3\underbrace{\mathrm{E}(s_i^2)}_{=\mathrm{V}(s_i)}\\
&= \mathrm{E}(s_i^4)-3\mathrm{E}(s_i)^2\\
&= \mathrm{kurt}(s_i),
\end{aligned}$$

so the condition corresponds to kurtosis being larger than zero.

3. For $g_i(s_i) = -s_i$, we have

$$\begin{aligned} & \frac{\tilde{p}_i'(s_i)}{\tilde{p}_i(s_i)} = -s_i \\ \Longrightarrow \quad & \frac{\partial}{\partial s} \log \tilde{p}(s_i) = -s_i \\ \Longrightarrow \quad & \log \tilde{p}_i(s_i) = -\frac{s^2}{2} + \underbrace{c}_{\text{constant}} \\ \Longrightarrow \quad & \tilde{p}_i(s_i) = c \cdot \exp\left(-\frac{s^2}{2}\right), \end{aligned}$$

which is a Gaussian distribution.

Ex. 29 On the derivation of the Natural Gradient Algorithm

1.

$$\sum_{ij} (A)_{ij}(B)_{ij} = \sum_{ij} (A)_{ij}(B^{\mathrm{T}})_{ji} = \sum_{i} (AB^{\mathrm{T}})_{ii} = \mathrm{Tr}(AB^{\mathrm{T}}).$$

2. Using the hint by setting$A = \nabla J(W)$, we obtain

$$J(W + \mu \nabla J(W)) = J(W) + \mu \sum_{ij} (\nabla J(W))_{ij} (\nabla J(W))_{ij} + O(\mu^2).$$

Here $\mu \sum_{ij} (\nabla J(W))_{ij} (\nabla J(W))_{ij} = \mu \sum_{ij} (\nabla J(W))^2_{ij}$ is positive, since $\nabla J(W) \neq 0$. Since $O(\mu^2)$ depends on μ^2 and $\mu \sum_{ij} (\nabla J(W))^2_{ij}$ depends on μ, the term $O(\mu^2)$ converges to zero quicker than the sum. Therefore ,for small enough $\mu > 0$ it holds that

$$\mu \sum_{ij} (\nabla J(W))^2_{ij} + O(\mu^2) > 0.$$

Thus, for small enough μ, we have that

$$J(W + \mu \nabla J(W)) > J(W).$$

3. Using part 1 of this exercise, the Taylor expansion in terms of the trace operator is

$$\begin{aligned} J(W + \mu A) &= J(W) + \mu \sum_{ij} (A)_{ij} (\nabla J(W))_{ij} + O(\mu^2) \\ &= J(W) + \mu \, \mathrm{Tr}(A \nabla J(W)^{\mathrm{T}}) + O(\mu^2). \end{aligned}$$

For $A = \nabla J(W) B^{\mathrm{T}} B)$, we have

$$\begin{aligned} \mathrm{Tr}(A \nabla J(W)^{\mathrm{T}}) &= \mathrm{Tr}(\nabla J(W) B^{\mathrm{T}} B \nabla J(W)^{\mathrm{T}}) \\ &= \mathrm{Tr}((\nabla J(W) B^{\mathrm{T}})(\nabla J(W) B^{\mathrm{T}})^{\mathrm{T}}) \\ &= \sum_{ij} (\nabla J(W) B^{\mathrm{T}})^2_{ij} \geq 0, \end{aligned}$$

and thus, for small enough μ the modified iteration increases the objective function (or leaves it unchanged).

Ex. 30 EM-algorithm

1. Notice that $J_t(\theta|\theta_k)$ in Equation $(A.83)$ becomes for $\theta = \theta_k$

$$J_t(\theta_k|\theta_k) = \int [\log p(x(t), \mathbf{s}(t), \theta_k)] \, p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k) \, \mathrm{d}\mathbf{s}(t),$$

which is some constant. If we subtract $J_t(\theta_k|\theta_k)$ from $J_t(\theta|\theta_k)$, we obtain

$$\begin{aligned}
& J_t(\theta|\theta_k) - J_t(\theta_k|\theta_k) \\
= & \int [\log p(\mathbf{x}(t), \mathbf{s}(t), \theta)]\, p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) - \\
& \int [\log p(\mathbf{x}(t), \mathbf{s}(t), \theta_k)]\, p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) \\
= & \int \log\left[\frac{p(\mathbf{x}(t), \mathbf{s}(t), \theta)}{p(\mathbf{x}(t), \mathbf{s}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) \\
= & \tilde{J}_t(\theta|\theta_k).
\end{aligned}$$

As additive constants don't affect the maximizing arguments, the same argument θ, which maximizes $J(\theta|\theta_k)$, maximizes also $\tilde{J}(\theta|\theta_k)$.

2. As

$$\log\left[\frac{p(\mathbf{x}(t), \mathbf{s}(t), \theta_k)}{p(\mathbf{x}(t), \mathbf{s}(t), \theta_k)}\right] = \log 1 = 0,$$

$\tilde{J}_t(\theta_k|\theta_k)$ equals zero for every t, and therefore $\tilde{J}(\theta_k|\theta_k) = 0$. In the next step, θ_{k+1} is chosen to be such that $\tilde{J}(\theta|\theta_k)$ is maximized, i.e. $\tilde{J}(\theta_{k+1}|\theta_k) \geq \tilde{J}(\theta|\theta_k)$ for every θ. Since this holds for every θ, it holds also for θ_k and hence,

$$\tilde{J}(\theta_{k+1}|\theta_k) \geq \tilde{J}(\theta_k|\theta_k) = 0.$$

3. Using the given fact, we can write

$$\begin{aligned}
\tilde{J}(\theta|\theta_k) &= \int \log\left[\frac{p(\mathbf{x}(t), \mathbf{s}(t), \theta)}{p(\mathbf{x}(t), \mathbf{s}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) \\
&= \int \log\left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t), \theta) p(\mathbf{x}(t), \theta)}{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k) p(\mathbf{x}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) \\
&= \int \log\left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t), \theta)}{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) + \\
&\quad \log\left[\frac{p(\mathbf{x}(t), \theta)}{p(\mathbf{x}(t), \theta_k)}\right] \underbrace{\int p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t)}_{=1}.
\end{aligned}$$

Now $\sum_t \log p(x(t), \theta) = \ell(\theta)$, and thus we get

$$\begin{aligned}
\tilde{J}(\theta|\theta_k) &= \sum_t \tilde{J}_t(\theta|\theta_k) \\
&= \sum_t \log\left[\frac{p(\mathbf{x}(t), \theta)}{p(\mathbf{x}(t), \theta_k)}\right] + \\
&\quad \sum_t \int \log\left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t), \theta)}{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) \\
&= \sum_t \log p(\mathbf{x}(t), \theta) - \sum_t p(\mathbf{x}(t), \theta_k) + \\
&\quad \sum_t \int \log\left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t), \theta)}{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t) \\
&= \ell(\theta) - \ell(\theta_k) + \\
&\quad \sum_t \int \log\left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t), \theta)}{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)}\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)\ \mathrm{d}\mathbf{s}(t).
\end{aligned}$$

4. From the previous part of this exercise we get

$$
\begin{aligned}
& \ell(\theta_{k+1} - \ell(\theta_k) \\
= \; & \tilde{J}(\theta_{k+1}|\theta_k) - \sum_t \int \log \left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t),\theta)}{p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k)} \right] p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k) \, \mathrm{d}\mathbf{s}(t) \\
= \; & \tilde{J}(\theta_{k+1}|\theta_k) + \sum_t \int \log \left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t),\theta)}{p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k)} \right]^{-1} p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k) \, \mathrm{d}\mathbf{s}(t) \\
= \; & \tilde{J}(\theta_{k+1}|\theta_k) + \sum_t \int \log \left[\frac{p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k)}{p(\mathbf{s}(t)|\mathbf{x}(t),\theta)} \right] p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k) \, \mathrm{d}\mathbf{s}(t) \\
= \; & \underbrace{\tilde{J}(\theta_{k+1}|\theta_k)}_{\geq 0 \text{ (part 2)}} + \sum_t \underbrace{D(p(\mathbf{s}(t)|\mathbf{x}(t),\theta_k), p(\mathbf{s}(t)|\mathbf{x}(t),\theta)}_{\geq 0} \\
\geq \; & 0,
\end{aligned}
$$

and therefore we have $\ell(\theta_{k+1}) \geq \ell(\theta_k)$. Each iteration of the EM-algorithm leads thus to an increase of the likelihood $\ell(\theta)$, which would be obtained by integrating out the latent variables $\mathbf{s}$.

Ex. 31 More on the general form of the EM-algorithm

1. If we consider $p(X, S; \theta)$ and integrate out the latent variables S, we get $p(X; \theta)$ (see Eq 11.11). Therefore

$$
\begin{aligned}
p(X;\theta) &= \int p(X,S;\theta) \, \mathrm{d}S \\
&= \int \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{s}_t;\theta) p(\mathbf{s}_t;\theta) \, \mathrm{d}S \\
&= \prod_{t=1}^{T} \int p(\mathbf{x}_t|\mathbf{s}_t;\theta) p(\mathbf{s}_t;\theta) \, \mathrm{d}\mathbf{s}_t \\
&= \prod_{t=1}^{T} p(\mathbf{x}_t;\theta).
\end{aligned}
$$

2. For continuous data we have

$$
\begin{aligned}
J(\theta) &= \int \log(p(X,S;\theta)) p(S|X;\theta_{k-1}) \, \mathrm{d}S \\
&= \int \log \left(\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{s}_t;\theta) p(\mathbf{s}_t;\theta) \right) \frac{p(X,S;\theta_{k-1})}{p(X;\theta_{k-1})} \, \mathrm{d}\mathbf{s}_1 \ldots \mathrm{d}\mathbf{s}_T \\
&= \int \left(\sum_{t=1}^{T} (\log(p(\mathbf{x}_t,\mathbf{s}_t;\theta)) \right) \cdot \\
&\quad \frac{\prod_{\tau=1}^{T} p(\mathbf{x}_\tau|\mathbf{s}_r;\theta_{k-1}) p(\mathbf{s}_\tau;\theta_{k-1})}{\prod_{\tau=1}^{T} p(\mathbf{x}_\tau;\theta_{k-1})} \, \mathrm{d}\mathbf{s}_1 \ldots \mathrm{d}\mathbf{s}_T \\
&= \int \left(\sum_{t=1}^{T} (\log(p(\mathbf{x}_t,\mathbf{s}_t;\theta)) \right) \cdot \\
&\quad \prod_{\tau=1}^{T} \frac{p(\mathbf{x}_\tau|\mathbf{s}_r;\theta_{k-1}) p(\mathbf{s}_\tau;\theta_{k-1})}{p(\mathbf{x}_\tau;\theta_{k-1})} \, \mathrm{d}\mathbf{s}_1 \ldots \mathrm{d}\mathbf{s}_T \\
&= \sum_{t=1}^{T} \left(\int \log(p(\mathbf{x}_t,\mathbf{s}_t;\theta) \prod_{\tau=1}^{T} p(\mathbf{s}_\tau|\mathbf{x}_\tau;\theta_{k-1}) \, \mathrm{d}\mathbf{s}_1 \ldots \mathrm{d}\mathbf{s}_T \right)
\end{aligned}
$$

We split now the product over τ into two parts to isolate the term involving $\tau = t$:

$$\begin{aligned}
J(\theta) &= \sum_{t=1}^{T} \int \log(p(\mathbf{x}_t, \mathbf{s}_t; \theta) p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})\, \mathrm{d}\mathbf{s}_t \cdot \\
&\quad \int \prod_{\substack{\tau=1\\ \tau\neq t}}^{T} p(\mathbf{s}_\tau|\mathbf{x}_\tau; \theta_{k-1})\, \mathrm{d}\mathbf{s}_1 \ldots \mathrm{d}\mathbf{s}_{t-1}\, \mathrm{d}\mathbf{s}_{t+1} \ldots \mathrm{d}\mathbf{s}_T \\
&= \sum_{t=1}^{T} \int \log(p(\mathbf{x}_t, \mathbf{s}_t; \theta) p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})\, \mathrm{d}\mathbf{s}_t \cdot \\
&\quad \prod_{\substack{\tau=1\\ \tau\neq t}}^{T} \underbrace{\int p(\mathbf{s}_\tau|\mathbf{x}_\tau; \theta_{k-1})\, \mathrm{d}\mathbf{s}_\tau}_{=1} \\
&= \sum_{t=1}^{T} \int \log(p(\mathbf{x}_t, \mathbf{s}_t; \theta) p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})\, \mathrm{d}\mathbf{s}_t,
\end{aligned}$$

which was what we wanted. For discrete variables the calculation is analogue. We just need to replace the integral over S with a sum over S.

This expression has a nice interpretation: $\log p(\mathbf{x}_t, \mathbf{s}_t; \theta)$, which would give the complete log-likelihood when summed-up, is replaced by an estimate, namely the conditional expectation.

3. Denote all the parameters $\mu_c, C_c, c = 1, 2, \ldots, C$, by θ. From the definition, we get:

$$\begin{aligned}
p(r(t)|\mathbf{x}(t); \theta) &= \frac{p(r(t), \mathbf{x}(t); \theta)}{p(\mathbf{x}(t); \theta)} \\
&= \frac{p(r(t), \mathbf{x}(t); \theta)}{\sum_{r(t)=1}^{C} p(r(t), \mathbf{x}(t); \theta)} \\
&= \frac{q_{t,c}}{\sum_{c=1}^{C} q_{t,c}} \\
&= q_{t,c}^{*}.
\end{aligned}$$

Ex. 32 Estimating Gaussian mixture models with EM-algorithm

1. We have

$$\begin{aligned}
p(\mathbf{x}|r = c) &= \frac{1}{(2\pi)^{n/2}|C_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x} - \mu_c)\right), \\
p(\mathbf{x}, r = c) &= p(\mathbf{x}|r = c) P(r = c) \\
&= \frac{1}{(2\pi)^{n/2}|C_c|^{1/2}} \exp\left(\frac{(\mathbf{x} - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x} - \mu_c)}{2}\right) \pi_c.
\end{aligned}$$

Since the data is iid, it holds that

$$\begin{aligned}
L_{xr}(\theta) &= \prod_{t=1}^{T} p(\mathbf{x}(t), r(t) = c_t), \\
\ell_{xr}(\theta) &= \sum_{t=1}^{T} \log(\pi_{c_t}) - \sum_{t=1}^{T} \frac{n}{2} \log(2\pi) - \sum_{t=1}^{T} \frac{1}{2}(|C_{c_t}|) - \\
&\quad \sum_{t=1}^{T} \frac{1}{2} (\mathbf{x}(t) - \mu_{c_t})^{\mathrm{T}} C_{c_t}^{-1} (\mathbf{x}(t) - \mu_{c_t}),
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\ell_{xr}(\theta) \\
= &\sum_{t=1}^{T} \sum_{c=1}^{C} \chi(c = r(t)) \cdot \\
&\left[\log(\pi_c) - \frac{n}{2} \log(2\pi) - \frac{1}{2}(|C_c|) - \frac{1}{2} (\mathbf{x}(t) - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x}(t) - \mu_c)\right],
\end{aligned}$$

where $\chi(c = r(t))$ is the indicator (or characteristic) function:

$$\chi(c = r(t)) = \begin{cases} 0 & \text{if } r(t) \neq c \\ 1 & \text{if } r(t) = c \end{cases}.$$

2. The posterior $P(r = c|\mathbf{x}, \theta)$ is calculated as follows:

$$\begin{aligned} & P(r = c|\mathbf{x}, \theta) \\ = \; & \frac{P(r = c, \mathbf{x}|\theta)}{p(\mathbf{x}|\theta)} \\ = \; & \frac{P(r = c, \mathbf{x}|\theta)}{\sum_{k=1}^{C} p(\mathbf{x}|r = k, \theta) P(r = k|\theta)} \\ = \; & \frac{\left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{|C_c|}\right)^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x} - \mu_c)\right) \pi_c}{\sum_{k=1}^{C} \left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{|C_k|}\right)^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)^{\mathrm{T}} C_k^{-1} (\mathbf{x} - \mu_k)\right) \pi_k} \\ = \; & \frac{\frac{\pi_c}{\sqrt{|C_c|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x} - \mu_c)\right)}{\sum_{k=1}^{C} \frac{\pi_k}{\sqrt{|C_k|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)^{\mathrm{T}} C_k^{-1} (\mathbf{x} - \mu_k)\right)}. \end{aligned}$$

3. The parameters θ in ℓ_{xr} are $\mu_1, \mu_2, \ldots, \mu_c$ and $C_1, C_2, \ldots, C_c$ and $\pi_1, \pi_2, \ldots, \pi_c$. As we cannot observe the value $r(t)$, we have to estimate it given all we have: the observation $\mathbf{x}(t)$ and some estimates θ_k for the parameters.

   Estimation of $r(t)$ means here taking the average value with respect to the density $P(r(t) = m|x(t), \theta_k)$:

$$\begin{aligned} & J(\theta|\theta_k) \\ = \; & \sum_{t=1}^{T} \sum_{c=1}^{C} \sum_{m=1}^{C} P(r(t) = m|x(t), \theta_k) \cdot \chi(r(t) = c) \\ & \left[\log \pi_c - \frac{1}{2} \log |C_c| - \frac{n}{2} \log(2\pi) - \frac{1}{2}(\mathbf{x}(t) - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x}(t) - \mu_c)\right] \\ = \; & \sum_{t=1}^{T} \sum_{m=1}^{C} P(r(t) = m|x(t), \theta_k) \cdot \\ & \left[\log \pi_m - \frac{1}{2} \log |C_m| - \frac{n}{2} \log(2\pi) - \frac{1}{2}(\mathbf{x}(t) - \mu_m)^{\mathrm{T}} C_m^{-1} (\mathbf{x}(t) - \mu_m)\right]. \end{aligned}$$

4. In the previous part of this exercise, in the very last equation it holds that the latter sum is constant with respect to μ_c if $m \neq c$. Therefore using exercise 13 (Maximum Likelihood Estimation for Multivariate Gaussians), we get

$$\begin{aligned} & \nabla_{\mu_c} J(\theta|\theta_k) \\ = \; & \sum_{t=1}^{T} \nabla_{\mu_c} P(r(t) = c|\mathbf{x}(t), \theta_k) \left[-\frac{1}{2}(\mathbf{x}(t) - \mu_c)^{\mathrm{T}} C_c^{-1} (\mathbf{x}(t) - \mu_c)\right] \\ = \; & \sum_{t=1}^{T} P(r(t) = c|\mathbf{x}(t), \theta_k) C_c^{-1} (\mathbf{x}(t) - \mu_c). \end{aligned}$$

   Setting the gradient to zero gives us

$$\sum_{t=1}^{T} P(r(t) = c|\mathbf{x}(t), \theta_k) C_c^{-1} \mu_c = \sum_{t=1}^{T} P(r(t) = c|\mathbf{x}(t), \theta_k) C_c^{-1} \mathbf{x}(t),$$

   from which we get

$$\mu_c = \frac{\sum_{t=1}^{T} P(r(t) = c|\mathbf{x}(t), \theta_k) \mathbf{x}(t)}{\sum_{t=1}^{T} P(r(t) = c|\mathbf{x}(t), \theta_k)},$$

   which is the value for $\mu_c(k+1)$.

   The last sum in the very last equation of the previous part of this exercise is constant with respect to C_c if $m \neq c$.

Again, we can use the results from exercise 13 to find the gradient with respect to C_c:

$$\begin{aligned}
& \nabla_{C_c} J(\theta|\theta_k) \\
= & \sum_{t=1}^{T} \nabla_{C_c} P(r(t)=c|\mathbf{x}(t),\theta_k)\left[-\frac{1}{2}\log|C_c| - \frac{1}{2}(\mathbf{x}(t)-\mu_c)^{\mathrm{T}} C_c^{-1}(\mathbf{x}(t)-\mu_c)\right] \\
= & \sum_{t=1}^{T} P(r(t)=c|\mathbf{x}(t),\theta_k)\cdot \\
& \left[-\frac{1}{2}C_c^{-1} + \frac{1}{2}C_c^{-1}(\mathbf{x}(t)-\mu_c)(\mathbf{x}(t)-\mu_c)^{\mathrm{T}} C_c^{-1}\right].
\end{aligned}$$

Setting the gradient to zero gives us

$$\begin{aligned}
& \frac{1}{2}C_c^{-1}\sum_{t=1}^{T} P(r(t)=c|\mathbf{x}(t),\theta_k)(\mathbf{x}(t)-\mu_c)(\mathbf{x}(t)-\mu_c)^{\mathrm{T}} C_c^{-1} \\
= & \frac{1}{2}C_c^{-1}\sum_{t=1}^{T} P(r(t)=c|\mathbf{x}(t),\theta_k) \\
\iff & C_c = \frac{\sum_{t=1}^{T} P(r(t)=c|\mathbf{x}(t),\theta_k)(\mathbf{x}(t)-\mu_c)(\mathbf{x}(t)-\mu_c)^{\mathrm{T}}}{\sum_{t=1}^{T} P(r(t)=c|\mathbf{x}(t),\theta_k)},
\end{aligned}$$

which gives us $C_c(k+1)$ when we use $\mu_c(k+1)$ for μ_c.

5. Because $\exp(\gamma) \geq 0$ for every $\gamma \in \mathbb{R}$, we have $\pi_c \geq 0$. Also

$$\sum_{c=1}^{C} \pi_c = \sum_{c=1}^{C} \frac{\exp(\gamma_c)}{\sum_{k=1}^{C}\exp(\gamma_k)} = \frac{\sum_{c=1}^{C}\exp(\gamma_c)}{\sum_{k=1}^{C}\exp(\gamma_k)} = 1,$$

so the trick works.

6. Using the previous part of this exercise, we get

$$\begin{aligned}
\frac{\partial J(\theta|\theta_k)}{\partial \gamma_c} &= \sum_{n=1}^{C} \frac{\partial J}{\partial \pi_n}\frac{\partial \pi_n}{\partial \gamma_c}, \\
\frac{\partial \pi_n}{\partial \gamma_c} &= -\frac{\exp(\gamma_n)}{(\sum_{k=1}^{C}\exp(\gamma_k))^2}\cdot \exp(\gamma_c) = -\pi_n\pi_c \quad (\text{if } n \neq c), \\
\frac{\partial \pi_c}{\partial \gamma_c} &= \pi_c - \pi_c\pi_c, \\
\frac{\partial J(\theta|\theta_k)}{\partial \pi_n} &= \sum_{t=1}^{T} \frac{1}{\pi_n} P(r(t)=n|x(t),\theta_k).
\end{aligned}$$

Therefore the derivative is

$$\begin{aligned}
& \frac{\partial J(\theta|\theta_k)}{\partial \gamma_c} \\
= & \sum_{\substack{k=1 \\ n\neq c}}^{C}\sum_{t=1}^{T} -\frac{\pi_n\pi_c}{\pi_n} P(r(t)=n|x(t),\theta_k) + \\
& \sum_{t=1}^{T} \frac{\pi_c - \pi_c^2}{\pi_c} P(r(t)=c|x(t),\theta_k) \\
= & -\pi_c \sum_{\substack{k=1 \\ n\neq c}}^{C}\sum_{t=1}^{T} P(r(t)=n|x(t),\theta_k) - \pi_c \sum_{t=1}^{T} P(r(t)=c|x(t),\theta_k) + \\
& \sum_{t=1}^{T} P(r(t)=c|x(t),\theta_k) \\
= & -\pi_c \sum_{t=1}^{T} \underbrace{\sum_{n=1}^{C} P(r(t)=n|x(t),\theta_k)}_{=1} + \sum_{t=1}^{T} P(r(t)=c|x(t),\theta_k) \\
= & -\pi_c \cdot T + \sum_{t=1}^{T} P(r(t)=c|x(t),\theta_k).
\end{aligned}$$

Setting the derivative to zero gives us

$$\pi_c = \frac{1}{T}\sum_{t=1}^{T} P(r(t) = c|x(t), \theta_k),$$

which gives us $\pi_c(k+1)$.

Ex. 33 K-means

1. Because $I^{-1} = I$ and $|I| = \det(I) = 1$, we have

$$\begin{aligned}
P(r = c|\mathbf{x}, \theta_k) &= \frac{\frac{1}{C}\frac{1}{\sqrt{|I|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^{\mathrm{T}} I^{-1}(\mathbf{x}-\mu_c)\right)}{\sum_{k=1}^{C}\frac{1}{C}\frac{1}{\sqrt{|I|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^{\mathrm{T}} I^{-1}(\mathbf{x}-\mu_c)\right)} \\
&= \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^{\mathrm{T}}(\mathbf{x}-\mu_c)\right)}{\sum_{k=1}^{C}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^{\mathrm{T}}(\mathbf{x}-\mu_c)\right)} \\
&= \frac{\exp\left(-\frac{1}{2}\|\mathbf{x}-\mu_c\|^2\right)}{\sum_{k=1}^{C}\exp\left(-\frac{1}{2}\|\mathbf{x}-\mu_k\|^2\right)}.
\end{aligned}$$

2. Simply put:

$$\begin{aligned}
\underset{c}{\operatorname{argmax}}\ P(r = c|\mathbf{x}, \theta_k) &= \underset{c}{\operatorname{argmax}}\ \exp\left(-\frac{1}{2}\|\mathbf{x}-\mu_c\|^2\right) \\
&= \underset{c}{\operatorname{argmin}}\ \|\mathbf{x}-\mu_c\|,
\end{aligned}$$

   i.e. the value of r that maximizes $P(r = c|\mathbf{x}, \theta_k)$ is given by the cluster whose mean is closest to $\mathbf{x}$.

3. With Exercise 32 (Estimating Gaussian Mixture Models with EM-Algorithm), the EM-update step for μ_c is

$$\begin{aligned}
\mu_c(k+1) &= \frac{\sum_{t=1}^{T}\hat{P}(r(t) = c|\mathbf{x}, \theta_k)\mathbf{x}(t)}{\sum_{t=1}^{T}\hat{P}(r(t) = c|\mathbf{x}, \theta_k)} \\
&= \frac{\sum_{t:\hat{r}_k(t)=c} 1\cdot\mathbf{x}(t) + \sum_{t:\hat{r}_k(t)\neq c} 0\cdot\mathbf{x}(t)}{\sum_{t:\hat{r}_k(t)=c} 1 + \sum_{t:\hat{r}_k(t)\neq c} 0} \\
&= \frac{\sum_{t:\hat{r}_k(t)=c}\mathbf{x}(t)}{\sum_{t:\hat{}_k(t)=c} 1} \\
&= \frac{\sum_{t:\hat{r}_k(t)=c}\mathbf{x}(t)}{\text{number of points assigned to cluster c}}.
\end{aligned}$$

   We can use the EM-update rule for μ_c (obtained in exercise 32 with $\hat{P}(r(t) = c|\mathbf{x}(t), \theta_k)$ instead of $P(r(t) = c|\mathbf{x}(t), \theta_k)$ because, formally, the cost function that must be optimized in the maximization step is the same as the one in exercise 32.

Ex. 34 Clustering for binary data
Note that many calculations are similar to those in Exercise 32.

1.

$$\begin{aligned}
\mathrm{E}(u_i) &= \sum_{u_i\in\{0,1\}} u_i p(u_i) = 0\cdot\mu_i^0(1-\mu_i)^1 + 1\cdot\mu_i^1(1-\mu_i)^0 = \mu_i, \\
\mathbf{V}(u_i) &= \mathrm{E}(u_i^2) - \mathrm{E}(u_i)^2 \\
&= \sum_{u_i\in\{0,1\}} u_i^2 p(u_i) - \mu_i^2 \\
&= 0^2\cdot\mu_i^0(1-\mu_i)^1 + 1^2\cdot\mu_i^1(1-\mu_i)^0 - \mu_i^2 \\
&= \mu_i - \mu_i^2 \\
&= \mu_i(1-\mu_i).
\end{aligned}$$

2.

$$\begin{aligned}
\mathrm{E}(\mathbf{u}) &= \mathrm{E}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} \mathrm{E}(u_1) \\ \mathrm{E}(u_2) \\ \vdots \\ \mathrm{E}(u_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}, \\
\mathrm{cov}(u_i, u_j) &= \mathrm{E}(u_i u_j) - \mathrm{E}(u_i)\mathrm{E}(u_j) \\
&\underset{\text{independent}}{\overset{u_i \text{ and } u_j}{=}} \mathrm{E}(u_i)\mathrm{E}(u_j) - \mathrm{E}(u_i)\mathrm{E}(u_j) \\
&= 0,
\end{aligned}$$

so we get

$$\begin{aligned}
\mathrm{cov}(\mathbf{u}) &= \begin{pmatrix} \mathbf{V}(u_1) & \mathrm{cov}(u_1, u_2) & \dots & \mathrm{cov}(u_1, u_n) \\ \mathrm{cov}(u_2, u_1) & \mathbf{V}(u_2) & \dots & \mathrm{cov}(u_2, u_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(u_n, u_1) & \mathrm{cov}(u_n, u_2) & \dots & \mathbf{V}(u_n) \end{pmatrix} \\
&= \begin{pmatrix} \mu_1(1-\mu_1) & 0 & \dots & 0 \\ 0 & \mu_2(1-\mu_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu_n(1-\mu_n) \end{pmatrix}.
\end{aligned}$$

The variables are thus uncorrelated.

3. To simplify notation, we denote $q(\mathbf{x}; \boldsymbol{\mu_c}, \pi_c, c = 1, 2, \dots, C)$ by $q(\mathbf{x})$ only. We can calculate the mean in a straightforward manner:

$$\begin{aligned}
\mathrm{E}(\mathbf{x}) &= \sum_{\mathbf{x} \in \{0,1\}^n} \mathbf{x} q(\mathbf{x}) \\
&= \sum_{\mathbf{x} \in \{0,1\}^n} \mathbf{x} \sum_{c=1}^{C} \pi_c p(\mathbf{x}; \boldsymbol{\mu_c}) \\
&= \sum_{c=1}^{C} \pi_c \sum_{\mathbf{x} \in \{0,1\}^n} \mathbf{x} p(\mathbf{x}; \boldsymbol{\mu_c}) \\
&\overset{\text{part 2}}{=} \sum_{c=1}^{C} \pi_c \boldsymbol{\mu_c}.
\end{aligned}$$

For the calculation of the covariance matrix of $\mathbf{x}$, we first calculate the marginal distribution of an element x_i of the vector $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}^{\mathrm{T}}$. We denote the marginal by $q(x_i)$. By definition of the marginal distribution, we get

$$q(x_i) = \sum_{\substack{x_k \\ k=1,2,\dots,n \\ k \neq i}} q(\mathbf{x}),$$

where we sum over all elements x_k of the vector $\mathbf{x}$ but element i. The marginal is thus

$$\begin{aligned}
q(x_i) &= \sum_{\substack{x_k \\ k=1,2,\dots,n \\ k \neq i}} q(\mathbf{x}) \\
&= \sum_{\substack{x_k \\ k=1,2,\dots,n \\ k \neq i}} \sum_{c=1}^{C} \pi_c p(\mathbf{x}; \mu_c) \\
&= \sum_{c=1}^{C} \pi_c \underbrace{\sum_{\substack{x_k \\ k=1,2,\dots,n \\ k \neq i}} p(\mathbf{x}; \mu_c)}_{\text{marginal of } x_i} \\
&= \sum_{c=1}^{C} \pi_c p(x_i; \mu_c).
\end{aligned}$$

The equality in the last line holds since $p(\mathbf{x};\mu_c)$, defined in Equation $(A.100)$, factorizes. We can use the marginal to calculate the covariance in few steps:

$$
\begin{aligned}
\mathrm{E}(x_i^2) &= \sum_{x_i=0}^{1} x_i^2 q(x_i) \\
&= \sum_{x_i=0}^{1} x_i^2 \sum_{c=1}^{C} \pi_c p(x_i;\mu_i) \\
&= \sum_{c=1}^{C} \pi_c \sum_{x_i=0}^{1} x_i^2 \mu_{ci}^{x_i} (1-\mu_{ci})^{(1-x_i)} \\
&= \sum_{c=1}^{C} \pi_c \mu_{ci}.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(x_i x_j) &= \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j \sum_{c=1}^{C} \pi_c \underbrace{p(x_i, x_j;\mu_c)}_{=p(x_i;\mu_c)p(x_j;\mu_c)} \\
&= \sum_{c=1}^{C} \pi_c \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i p(x_i;\mu_c) c_j p(x_j;\mu_c) \\
&= \sum_{c=1}^{C} \pi_c \left(\sum_{x_i=0}^{1} x_i p(x_i;\mu_c) \right) \left(\sum_{x_j=0}^{1} x_j p(x_j;\mu_c) \right) \\
&= \sum_{c=1}^{C} \pi_c \mu_{ci} \mu_{cj}.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{cov}(\mathbf{x}) &= \mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x})^{\mathrm{T}} \\
&= \sum_{c=1}^{C} \pi_c \begin{pmatrix} \mu_{c1} & \mu_{c1}\mu_{c2} & \cdots & \mu_{c1}\mu_{cn} \\ \mu_{c2}\mu_{c1} & \mu_{c2} & \cdots & \mu_{c2}\mu_{cn} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{cn}\mu_{c1} & \mu_{cn}\mu_{c2} & \cdots & \mu_{cn} \end{pmatrix} - \\
&\quad \sum_{c=1}^{C} \pi_c \begin{pmatrix} \mu_{c1} \\ \mu_{c2} \\ \vdots \\ \mu_{c_n} \end{pmatrix} \sum_{c=1}^{C} \pi_c \begin{pmatrix} \mu_{c1} & \mu_{c2} & \cdots & \mu_{cn} \end{pmatrix}.
\end{aligned}
$$

From this we can see that the covariance matrix is not diagonal and therefore the x_i are correlated. This shows that the mixture distribution is a richer distribution than a single multivariate Bernoulli distribution.

4. The log-likelihood $\ell(\boldsymbol{\mu_c}, \pi_c, c = 1, 2, \ldots, C))$ is

$$
\begin{aligned}
\ell &= \log \left(\prod_{t=1}^{T} q(\mathbf{x}(t); \boldsymbol{\mu_c}, \pi_c, c = 1, 2, \ldots, C) \right) \\
&= \sum_{t=1}^{T} \log q(\mathbf{x}(t); \boldsymbol{\mu_c}, \pi_c) \\
&= \sum_{t=1}^{T} \log \left(\sum_{c=1}^{C} \pi_c p(\mathbf{x}(t); \boldsymbol{\mu_c}) \right) \\
&= \sum_{t=1}^{T} \log \left(\sum_{c=1}^{C} \pi_c \prod_{i=1}^{n} p(x_i(t); \mu_{ci}) \right).
\end{aligned}
$$

5. When the class memberships are also observed, the log-likelihood $\ell(\boldsymbol{\mu}_{\mathbf{c}}, \pi_c, c = 1, 2, \dots, C))$ is

$$\begin{aligned}
\ell &= \log\left(\prod_{t=1}^{T} q(\mathbf{x}(t), r(t); \boldsymbol{\mu}_{r(t)})\right) \\
&= \sum_{t=1}^{T} \log q(\mathbf{x}(t), r(t); \boldsymbol{\mu}_{r(t)}) \\
&= \sum_{t=1}^{T} \log(\pi_{r(t)} p(\mathbf{x}(t); \boldsymbol{\mu}_{r(t)})) \\
&= \sum_{t=1}^{T} \left[\log \pi_{r(t)} + \log\left(\prod_{i=1}^{n} p(x_i(t); \mu_{r(t),i})\right)\right] \\
&= \sum_{t=1}^{T} \log(\pi_{r(t)}) + \sum_{t=1}^{T}\sum_{i=1}^{n} \log\left(\mu_{r(t),i}^{x_i(t)}(1-\mu_{r(t),i})^{(1-x_i(t))}\right) \\
&= \sum_{t=1}^{T} \log(\pi_{r(t)}) + \\
&\quad \sum_{t=1}^{T}\sum_{i=1}^{n}(x_i(t)\log \mu_{r(t),i} + (1-x_i(t))\log(1-\mu_{r(t),i})).
\end{aligned}$$

6.

$$\begin{aligned}
q(r(t) = c|\mathbf{x}(t)) &= \frac{q(r(t), \mathbf{x}(t))}{q(\mathbf{x}(t))} \\
&= \frac{\pi_{r(t)} p(\mathbf{x}(t); \boldsymbol{\mu}_{r(t)})}{\sum_{c=1}^{C} q(\mathbf{x}(t), r(t) = c)} \\
&= \frac{\pi_c p(\mathbf{x}(t); \boldsymbol{\mu}_c)}{\sum_{k=1}^{C} \pi_k p(\mathbf{x}(t); \boldsymbol{\mu}_k)}
\end{aligned}$$

This is equivalent to Eq. (11.16) for the Gaussian mixture.

7. Using exercise 31 (More on the general form of EM for Mixture of Gaussians) $(*)$, we have

$$\begin{aligned}
&\mathrm{E}(\ell(\mu_c, \pi_c)) \\
\overset{(*)}{=}\; & \sum_{t=1}^{T}\sum_{c=1}^{C} \underbrace{q(r(t) = c|\mathbf{x}(t))}_{=q^*_{t,c}\text{(notation)}} \underbrace{\log(q(\mathbf{x}(t), r(t); \mu_{r(t)}))}_{\text{part 5}} \\
=\; & \sum_{c=1}^{C}\sum_{t=1}^{T} q^*_{t,c} \log \pi_c + \\
& \sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{i=1}^{n} q^*_{t,c}(x_i(t)\log \mu_{ci} + (1-x_i(t))\log(1-\mu_{ci})).
\end{aligned}$$

8. Denote $J = \mathrm{E}(\ell(\mu_c, \pi_c))$. Now we have

$$\begin{aligned}
\frac{\partial J}{\partial \mu_{ci}} &= 0 + \sum_{t=1}^{T} q^*_{t,c}\left(x_i(t)\frac{1}{\mu_{ci}} - (1-x_i(t))\frac{1}{1-\mu_{ci}}\right) \\
&= \frac{1}{\mu_{ci}}\sum_{t=1}^{T} q^*_{t,c} x_i(t) - \frac{1}{1-\mu_{ci}}\sum_{t=1}^{T} q^*_{t,c}(1-x_i(t)).
\end{aligned}$$

Setting the derivative to zero gives us:

$$\mu_{ci}\left(\sum_{t=1}^{T} q^*_{t,c} - \sum_{t=1}^{T} q^*_{t,c} x_i(t)\right) = (1-\mu_{ci})\sum_{t=1}^{T} q^*_{t,c} x_i(t),$$

from which we get

$$\mu_{ci} = \frac{\sum_{t=1}^{T} q^*_{t,c} x_i(t)}{\sum_{t=1}^{T} q^*_{t,c}},$$

which is what we wanted, since $q^*_{t,c} = q(r(t) = c|\mathbf{x}(t))$. This is the same update rule as for Gaussian mixtures (compare with Eq. 11.17).

9. Denote $\tilde{J} = J + \lambda(1 - \sum_{k=1}^{c} \pi_k)$. Now we have

$$\frac{\partial \tilde{J}}{\partial \pi_c} = \sum_{t=1}^{T} q^*_{t,c} \frac{1}{\pi_c} - \lambda.$$

Setting the derivative to zero gives us:

$$\pi_c = \frac{\sum_{t=1}^{T} q^*_{t,c}}{\lambda}.$$

We use the constraint to calculate λ:

$$\begin{aligned}
1 &= \sum_{k=1}^{C} \pi_k = \sum_{k=1}^{C} \frac{\sum_{t=1}^{T} q^*_{t,c}}{\lambda} = \frac{1}{\lambda} \sum_{k=1}^{C} \sum_{t=1}^{T} q^*_{t,c} \\
\implies \lambda &= \sum_{k=1}^{C} \sum_{t=1}^{T} q^*_{t,c} \\
&= \sum_{t=1}^{T} \sum_{k=1}^{C} \frac{\pi_k p(\mathbf{x}(t); \mu_k)}{\sum_{j=1}^{C} \pi_j p(\mathbf{x}(t); \mu_j)} \\
&= \sum_{t=1}^{T} \frac{\sum_{k=1}^{C} \pi_k p(\mathbf{x}(t); \mu_k)}{\sum_{j=1}^{C} \pi_j p(\mathbf{x}(t); \mu_j)} \\
&= \sum_{t=1}^{T} 1 \\
&= T.
\end{aligned}$$

Therefore

$$\pi_c = \frac{\sum_{t=1}^{T} q^*_{t,c}}{T},$$

which is what we wanted. This is the same update rule as for Gaussian mixtures in Eq. (11.19).

Ex. 35 Some Verifications for Metric MDS

1. For the columns, the sum equals

$$\begin{aligned}
\sum_{i=1}^{N} \tilde{d}_{ij} &= \sum_{i=1}^{N} d_{ij} - \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{i}^{N} d_{ij} \right) - \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{j}^{N} d_{ij} \right) + \\
&\quad \sum_{i=1}^{N} \left(\frac{1}{N^2} \sum_{ij} d_{ij} \right) \\
&= \sum_{i=1}^{N} d_{ij} - N \left(\frac{1}{N} \sum_{i=1}^{N} d_{ij} \right) - \frac{1}{N} \sum_{ij} d_{ij} + \\
&\quad N \left(\frac{1}{N^2} \sum_{ij} d_{ij} \right) \\
&= \sum_{i=1}^{N} d_{ij} - \sum_{i=1}^{N} d_{ij} - \frac{1}{N} \sum_{ij} d_{ij} + \frac{1}{N} \sum_{ij} d_{ij} \\
&= 0,
\end{aligned}$$

which was what we wanted. The proof for the rows is very similar.

2. For the euclidean distance, we have

$$\begin{aligned}
d_{ij} &= \|\mathbf{x}_i - \mathbf{x}_j\|^2 \\
&= (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_j) \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - \underbrace{\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j}_{\in \mathbb{R}} - \underbrace{\mathbf{x}_j^{\mathrm{T}} \mathbf{x}_i}_{\in \mathbb{R}} + \mathbf{x}_j^{\mathrm{T}} \mathbf{x}_j \\
&= \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j,
\end{aligned}$$

and therefore

$$\begin{aligned}
\tilde{d}_{ij} &= \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^2 - \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_j\|^2 + \\
&\quad \frac{1}{N}\sum_{i=1}^{N}2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_i\|^2 - \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_j\|^2 + \\
&\quad \frac{1}{N}\sum_{j=1}^{N}2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \frac{1}{N^2}\underbrace{\sum_{ij}\|\mathbf{x}_i\|^2}_{=N\sum_{i=1}^{N}\|\mathbf{x}_i\|^2} + \frac{1}{N^2}\underbrace{\sum_{ij}\|\mathbf{x}_j\|^2}_{=N\sum_{j=1}^{N}\|\mathbf{x}_j\|^2} - \\
&\quad \frac{1}{N^2}\underbrace{\sum_{ij}2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}_{=\sum_{i=1}^{N}\sum_{j=1}^{N}2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j} \\
&= \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - N\frac{1}{N}\|\mathbf{x}_j\|^2 + 2\mathbf{x}_j^{\mathrm{T}}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i}_{=0} - \\
&\quad N\frac{1}{N}\|\mathbf{x}_i\|^2 + 2\mathbf{x}_i^{\mathrm{T}}\underbrace{\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j}_{=0} - \frac{1}{N}\sum_{i=1}^{N}2\mathbf{x}_i^{\mathrm{T}}\underbrace{\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j}_{=0} \\
&= \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - \|\mathbf{x}_i\|^2 - \|\mathbf{x}_j\|^2 \\
&= -2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j.
\end{aligned}$$