Introduction to Machine Learning
Fall 2022
University of Science and Technology of China

Lecturer: Jie Wang — Homework 2
Name: Yunqin Zhu — ID: PB20061372

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Linear regression

Consider a data set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R}$.

1. If we want to fit the data by a linear model

$$y = w_0 + w_1 x, \tag{1}$$

please find $\hat{w}_0$ and $\hat{w}_1$ by the least squares approach (you need to find expressions of $\hat{w}_0$ and $\hat{w}_1$ by $\{(x_i, y_i)\}_{i=1}^n$, respectively).

Solution:

Denote $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$. The average fitting error is defined as $L_S(\mathbf{w}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$. The least squares solution is given by $\hat{\mathbf{w}}_{LS} \in \mathbf{argmin}_{\mathbf{w}} L_S(\mathbf{w})$. Since

$$\nabla_{\mathbf{w}} L_S(\mathbf{w}) = -\frac{2}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) = \mathbf{0} \implies \mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y},$$
$$\nabla^2_{\mathbf{ww}} L_S(\mathbf{w}) = \frac{2}{n}\mathbf{X}^\top\mathbf{X} \geq 0,$$

we know $\hat{\mathbf{w}}_{LS}$ can be any solution of the normal equation, i.e.

$$\hat{\mathbf{w}}_{LS} = \left(\mathbf{X}^\top\mathbf{X}\right)^+ \mathbf{X}^\top\mathbf{y} + \left(\mathbf{I} - \left(\mathbf{X}^\top\mathbf{X}\right)^+ \mathbf{X}^\top\mathbf{X}\right)\mathbf{u},\ \forall \mathbf{u} \in \mathbb{R}^n,$$

where $\left(\mathbf{X}^\top\mathbf{X}\right)^+$ denotes the Moore-Penrose pseudoinverse of $\mathbf{X}^\top\mathbf{X}$. Suppose $\mathbf{X}$ has full column rank. Then $\left(\mathbf{X}^\top\mathbf{X}\right)^+ = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$, and thus

$$\hat{\mathbf{w}}_{LS} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Specifically,

$$\hat{w}_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, \quad \hat{w}_1 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}. \tag{2}$$

■

2. **Programming Exercise:** We provide you a data set $\{(x_i, y_i)\}_{i=1}^{30}$. Consider the model in (1) and the one as follows:

$$y = w_0 + w_1 x + w_2 x^2. \tag{3}$$

Which model do you think fits better the data? Please detail your approach first and then implement it by your favorite programming language. The required output includes

(a) your detailed approach step by step;

(b) your code with detailed comments according to your planned approach;

(c) a plot showing the data and the fitting models;

(d) the model you finally choose ($\hat{w}_0$ and $\hat{w}_1$ if you choose the model in (1), or $\hat{w}_0$, $\hat{w}_1$, and $\hat{w}_2$ if you choose the model in (3)).

Solution:

(a) First, we fit the linear model by computing (2). The results are $\hat{w}_0 = 1.000387$ and $\hat{w}_1 = 0.430838$.

Next, to fit the quadratic model, we define the design matrix as $\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$ and let $\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$. Analogously, by solving the normal equation $\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{X}^\top \mathbf{y}$ which has the unique solution $\hat{\mathbf{w}}_{LS} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$, we get the coefficients $\hat{w}_0 = 1.029568$, $\hat{w}_1 = 0.386143$ and $\hat{w}_2 = -0.142151$.

After that, we compute the predicted values $\hat{y}_i, i = 1, \ldots, 30$ for the linear and the quadratic model respectively, and compare $L_S(\hat{\mathbf{w}}_{LS})$ for the two models.

The average fitting error of the linear model is $L_S(\hat{\mathbf{w}}_{LS}) = 0.009405$, while the quadratic model has smaller $L_S(\hat{\mathbf{w}}_{LS}) = 0.008083$, implying that it fits better than the linear one.

Accordingly, we choose the quadratic model, i.e. the model in (3), as the final model.

(b) The python code is as follows:

```python
# %%
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import pandas as pd
mpl.rcParams['mathtext.fontset'] = 'cm'

# %%
df = pd.read_csv('HW2_DataSet/Ex1 data', sep='\t', header=None, names=['x',
↪  'y']).sort_values('x')
df['x^2'] = df['x'] ** 2
df['1'] = 1
df

# %%
# solve normal equation
def linear_reg(X, y):
    return np.linalg.inv(X.T @ X) @ X.T @ y


# linear model
X = df[['1', 'x']].values
y = df['y'].values
w = linear_reg(X, y)
df['linear'] = X @ w
mse = np.mean((df['linear'] - df['y'])**2)
print(f'Linear Model: \n w0 = {w[0]} \n w1 = {w[1]} \n MSE = {mse} \n')

# quadratic model
X = df[['1', 'x', 'x^2']].values
y = df['y'].values
w = linear_reg(X, y)
df['quadratic'] = X @ w
mse = np.mean((df['quadratic'] - df['y'])**2)
print(
    f'Quadratic Model: \n w0 = {w[0]} \n w1 = {w[1]} \n w2 = {w[2]} \n MSE =
    ↪  {mse} \n'
)

# %%
# plot
plt.plot(df['x'], df['y'], 'o', mfc='none', label='Data')
plt.plot(df['x'], df['linear'], label='Linear Reg.')
plt.plot(df['x'], df['quadratic'], label='Quad. Reg.')
plt.legend()
plt.xlabel(r'$x$')
plt.ylabel(r'$y$')

plt.savefig('HW2_Ex1.pdf')
plt.show()
```

(c) A plot of the data and the fitting models is shown in Figure 1.

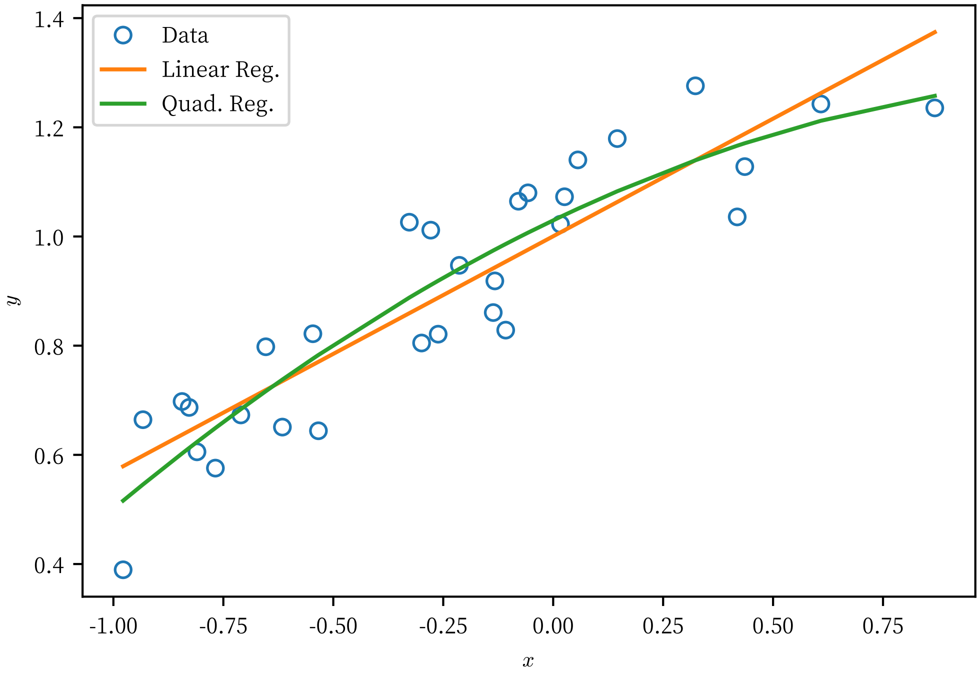


Figure 1: Linear and quadratic fitting models in Exercise 1.

(d) We choose the quadratic model in (3), i.e.

$$y = 1.029568 + 0.386143x - 0.142151x^2.$$

■

Exercise 2: Projection

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{x} \in \mathbb{R}^m$. Define

$$\Pi_{\mathbf{A}}\left(\mathbf{x}\right) = \underset{\mathbf{z}\in\mathbb{R}^m}{\mathbf{argmin}} \left\{\|\mathbf{x}-\mathbf{z}\|_2 : \mathbf{z} \in \mathcal{C}(\mathbf{A})\right\}.$$

We call $\Pi_{\mathbf{A}}\left(\mathbf{x}\right)$ the projection of the point $\mathbf{x}$ onto the column space of $\mathbf{A}$.

1. Please prove that $\Pi_{\mathbf{A}}(\mathbf{x})$ is unique for any $\mathbf{x} \in \mathbb{R}^m$.

   Solution:
   Denote the objective function by $f(\mathbf{z}) = \|\mathbf{x}-\mathbf{z}\|_2^2$, which is clearly a continuous function. Let $\mathbf{z}_0 \in \mathcal{C}(A)$ and $r = \|\mathbf{x}-\mathbf{z}_0\|_2$. If we denote $C = \mathcal{C}(A) \cap B(\mathbf{x}, r)$, we can conclude that C is nonempty, as it at least contains $\mathbf{z}_0$. Since both $B(\mathbf{x}, r)$ and $\mathcal{C}(A)$ are closed, the set C is closed as well. Moreover, the boundedness of $B(\mathbf{x}, r)$ implies that C must be bounded. All together, we conclude that the set C is compact. By the Extreme Value Theorem, there exists $\mathbf{z}_1 \in C$ such that $f(\mathbf{z}_1) \leq f(\mathbf{z})$ for all $\mathbf{z} \in C$, and also for all $\mathbf{z} \in \mathcal{C}(A)$, i.e. $\mathbf{z}_1 \in \mathbf{argmin}\left\{\|\mathbf{x}-\mathbf{z}\|_2 : \mathbf{z} \in \mathcal{C}(\mathbf{A})\right\}$.

   Suppose there exists another $\mathbf{z}_2 \in \mathcal{C}(A)$ such that $f(\mathbf{z}_2) = f(\mathbf{z}_1)$. Then we have

   $$f\left(\frac{\mathbf{z}_1+\mathbf{z}_2}{2}\right) = \left\|\mathbf{x} - \frac{\mathbf{z}_1+\mathbf{z}_2}{2}\right\|_2^2 = \frac{1}{2}\|\mathbf{x}-\mathbf{z}_1\|_2^2 + \frac{1}{2}\|\mathbf{x}-\mathbf{z}_2\|_2^2 - \frac{1}{4}\|\mathbf{z}_1-\mathbf{z}_2\|_2^2 < f(\mathbf{z}_1),$$

   which is a contradiction. This shows that $\mathbf{z}_1 = \Pi_{\mathbf{A}}\left(\mathbf{x}\right)$ is unique. ■

2. Let $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1, \ldots, d$ with $d \leq n$, which are linearly independent.

   (a) For any $\mathbf{w} \in \mathbb{R}^n$, please find $\Pi_{\mathbf{v}_1}\left(\mathbf{w}\right)$, which is the projection of $\mathbf{w}$ onto the subspace spanned by $\mathbf{v}_1$.

   (b) Please show $\Pi_{\mathbf{v}_1}\left(\cdot\right)$ is a linear map, i.e.,

   $$\Pi_{\mathbf{v}_1}\left(\alpha\mathbf{u} + \beta\mathbf{w}\right) = \alpha\Pi_{\mathbf{v}_1}\left(\mathbf{u}\right) + \beta\Pi_{\mathbf{v}_1}\left(\mathbf{w}\right),$$

   where $\alpha, \beta \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^n$.

   (c) Please find the projection matrix corresponding to the linear map $\Pi_{\mathbf{v}_1}\left(\cdot\right)$, i.e., find the matrix $\mathbf{H}_1 \in \mathbb{R}^{n\times n}$ such that

   $$\Pi_{\mathbf{v}_1}\left(\mathbf{w}\right) = \mathbf{H}_1\mathbf{w}.$$

   (d) Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$.

      i. For any $\mathbf{w} \in \mathbb{R}^n$, please find $\Pi_{\mathbf{V}}\left(\mathbf{w}\right)$ and the corresponding projection matrix $\mathbf{H}$.

      ii. Please find $\mathbf{H}$ if we further assume that $\mathbf{v}_i^\top\mathbf{v}_j = 0$, $\forall\, i \neq j$.

   Solution:

   (a) Define $f(x) = \|\mathbf{w} - \mathbf{v}_1 x\|_2^2$, $x \in \mathbb{R}$. To minimize $f(x)$, let $f'(x) = -2\mathbf{v}_1^\top(\mathbf{w} - \mathbf{v}_1 x) = 0$ $\implies x = \left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1^\top\mathbf{w}$. Therefore, $\Pi_{\mathbf{v}_1}\left(\mathbf{w}\right) = \mathbf{v}_1\left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1^\top\mathbf{w}$.

(b) $\Pi_{\mathbf{v}_1}\left(\alpha\mathbf{u}+\beta\mathbf{w}\right)=\mathbf{v}_1\left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1^\top(\alpha\mathbf{u}+\beta\mathbf{w})=\alpha\mathbf{v}_1\left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1^\top\mathbf{u}+\beta\mathbf{v}_1\left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1^\top\mathbf{w}$ $=\alpha\Pi_{\mathbf{v}_1}\left(\mathbf{u}\right)+\beta\Pi_{\mathbf{v}_1}\left(\mathbf{w}\right).$

(c) $\mathbf{H}_1=\mathbf{v}_1\left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1.$

(d) i. Define $f(\mathbf{x})=\|\mathbf{w}-\mathbf{V}\mathbf{x}\|_2^2,\ \mathbf{x}\in\mathbb{R}^d$. To minimize $f(\mathbf{x})$, let

$$f'(\mathbf{x})=-2\mathbf{V}^\top(\mathbf{w}-\mathbf{V}\mathbf{x})=0\implies\mathbf{x}=\left(\mathbf{V}^\top\mathbf{V}\right)^{-1}\mathbf{V}^\top\mathbf{w}.$$

Therefore, $\Pi_{\mathbf{V}}\left(\mathbf{w}\right)=\mathbf{V}\left(\mathbf{V}^\top\mathbf{V}\right)^{-1}\mathbf{V}^\top\mathbf{w}$, and thus $\mathbf{H}=\mathbf{V}\left(\mathbf{V}^\top\mathbf{V}\right)^{-1}\mathbf{V}^\top$.

ii.
$$\mathbf{V}^\top\mathbf{V}=\begin{pmatrix}\mathbf{v}_1^\top\\ \vdots\\ \mathbf{v}_d^\top\end{pmatrix}\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_d\end{pmatrix}=\begin{pmatrix}\mathbf{v}_1^\top\mathbf{v}_1 & & \\ & \ddots & \\ & & \mathbf{v}_d^\top\mathbf{v}_d\end{pmatrix},$$

$$\begin{aligned}\mathbf{H}=\mathbf{V}\left(\mathbf{V}^\top\mathbf{V}\right)^{-1}\mathbf{V}^\top&=\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_d\end{pmatrix}\begin{pmatrix}\mathbf{v}_1^\top\mathbf{v}_1 & & \\ & \ddots & \\ & & \mathbf{v}_d^\top\mathbf{v}_d\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{v}_1^\top\\ \vdots\\ \mathbf{v}_d^\top\end{pmatrix}\\&=\mathbf{v}_1\left(\mathbf{v}_1^\top\mathbf{v}_1\right)^{-1}\mathbf{v}_1+\cdots+\mathbf{v}_d\left(\mathbf{v}_d^\top\mathbf{v}_d\right)^{-1}\mathbf{v}_d\\&=\mathbf{H}_1+\cdots+\mathbf{H}_d.\end{aligned}$$
■

3. (a) Suppose that

$$\mathbf{A}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

What are the coordinates of $\Pi_{\mathbf{A}}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x}\in\mathbb{R}^2$? Are the coordinates unique?

(b) Suppose that

$$\mathbf{A}=\begin{pmatrix}1 & 2\\ 1 & 2\end{pmatrix}.$$

What are the coordinates of $\Pi_{\mathbf{A}}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x}\in\mathbb{R}^2$? Are the coordinates unique?

Solution:

(a) $\Pi_{\mathbf{A}}\left(\mathbf{x}\right)=\mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{x}=\mathbf{x}$. The coordinate $\mathbf{x}$ is unique.

(b) $\Pi_{\mathbf{A}}\left(\mathbf{x}\right)=\Pi_{\mathbf{a}_1}\left(\mathbf{x}\right)=\mathbf{a}_1\left(\mathbf{a}_1^\top\mathbf{a}_1\right)^{-1}\mathbf{a}_1^\top\mathbf{x}=\frac{1}{2}\begin{pmatrix}x_1+x_2\\ x_1+x_2\end{pmatrix}$. The coordinates $\mathbf{w}$ with respect to $\{\mathbf{a}_1,\mathbf{a}_2\}$ can be found by solving $\mathbf{A}\mathbf{w}=\Pi_{\mathbf{A}}\left(\mathbf{x}\right)$, which has infinitely many solutions. Hence the coordinates are not unique. ■

4. A matrix $\mathbf{P}$ is called a projection matrix if $\mathbf{Px}$ is the projection of $\mathbf{x}$ onto $\mathcal{C}(\mathbf{P})$ for any $\mathbf{x}$.

   (a) Let λ be the eigenvalue of $\mathbf{P}$. Show that λ is either 1 or 0. (*Hint: you may want to figure out what the eigenspaces corresponding to $\lambda = 1$ and $\lambda = 0$ are, respectively.*)

   (b) Show that $\mathbf{P}$ is a projection matrix if and only if $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}$ is symmetric.

   Solution:
   First, we show that $\Pi_{\mathbf{A}}(\mathbf{x}) = \mathbf{z}_1$ if and only if $\mathbf{x} - \mathbf{z}_1$ is orthogonal to $\mathcal{C}(A)$. Note that any $\mathbf{z} \in \mathcal{C}(A)$ can be represented by $\mathbf{z}_1 + \epsilon\mathbf{y}$, where $\epsilon > 0$ and $\mathbf{y} \in \mathcal{C}(A)$ are arbitrary. We expand $f(\mathbf{z})$ as

$$\begin{aligned} f(\mathbf{z}) &= \|\mathbf{x} - \mathbf{z}_1 + \mathbf{z}_1 - \mathbf{z}\|_2^2 \\ &= \|\mathbf{x} - \mathbf{z}_1\|_2^2 + \|\mathbf{z}_1 - \mathbf{z}\|_2^2 + 2\langle \mathbf{x} - \mathbf{z}_1, \mathbf{z}_1 - \mathbf{z}\rangle \\ &= f(\mathbf{z}_1) + \epsilon^2\|\mathbf{y}\|_2^2 + 2\epsilon\langle \mathbf{x} - \mathbf{z}_1, \mathbf{y}\rangle. \end{aligned}$$

   Therefore,

$$\begin{aligned} \forall\, \mathbf{z} \in \mathcal{C}(A),\ f(\mathbf{z}) - f(\mathbf{z}_1) \geq 0 &\iff \forall\, \mathbf{y} \in \mathcal{C}(A),\ \forall\, \epsilon > 0,\ \epsilon\|\mathbf{y}\|_2^2 + 2\langle \mathbf{x} - \mathbf{z}_1, \mathbf{y}\rangle \geq 0 \\ &\iff \forall\, \mathbf{y} \in \mathcal{C}(A),\ \langle \mathbf{x} - \mathbf{z}_1, \mathbf{y}\rangle \geq \sup_{\epsilon > 0} -\frac{\epsilon}{2}\|\mathbf{y}\|_2^2 = 0, \end{aligned}$$

   which leads to the lemma. Next, we show the statements in the problem.

   (a) For any $\mathbf{x} \in \mathcal{C}(\mathbf{P})$, we have $\Pi_{\mathbf{P}}(\mathbf{x}) = \mathbf{x}$, while for any $\mathbf{x} \in \mathcal{C}(\mathbf{P})^{\perp}$, $\Pi_{\mathbf{P}}(\mathbf{x}) = \mathbf{0}$. Therefore, λ is either 1 with geometric multiplicity $\mathbf{rank}(\mathbf{P})$, or 0 with geometric multiplicity $n - \mathbf{rank}(\mathbf{P})$.

   (b) (⇒) If $\mathbf{P}$ is a projection matrix, then $\mathbf{P}^2 = \mathbf{P}$, because each column of $\mathbf{P}$ is also its own projection onto $\mathcal{C}(\mathbf{P})$. Moreover, given arbitrary $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\langle \mathbf{x}, \mathbf{Py}\rangle = \langle \mathbf{Px}, \mathbf{Py}\rangle = \langle \mathbf{Px}, \mathbf{y}\rangle$, which implies $\mathbf{P}$ is symmetric.

   (⇐) Given arbitrary $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{Py} \in \mathcal{C}(\mathbf{P})$, we have $\langle \mathbf{x}, \mathbf{Py}\rangle = \langle \mathbf{x}, \mathbf{P}^2\mathbf{y}\rangle = \langle \mathbf{Px}, \mathbf{Py}\rangle$, i.e. $\langle \mathbf{x} - \mathbf{Px}, \mathbf{Py}\rangle = 0$, and hence $\Pi_{\mathbf{P}}(\mathbf{x}) = \mathbf{Px}$. Therefore, $\mathbf{P}$ must be a projection matrix. ■

5. Let $\mathbf{B} \in \mathbb{R}^{m \times s}$ and $\mathcal{C}(\mathbf{B})$ be its column space. Suppose that $\mathcal{C}(\mathbf{B})$ is a proper subspace of $\mathcal{C}(\mathbf{A})$. Is $\Pi_{\mathbf{B}}(\mathbf{x})$ the same as $\Pi_{\mathbf{B}}(\Pi_{\mathbf{A}}(\mathbf{x}))$? Please show your claim rigorously.

   Solution:
   $\Pi_{\mathbf{B}}(\mathbf{x})$ is the same as $\Pi_{\mathbf{B}}(\Pi_{\mathbf{A}}(\mathbf{x}))$.

   As both $\mathbf{x} - \Pi_{\mathbf{A}}(\mathbf{x})$ and $\Pi_{\mathbf{A}}(\mathbf{x}) - \Pi_{\mathbf{B}}(\Pi_{\mathbf{A}}(\mathbf{x}))$ is orthogonal to $\mathcal{C}(\mathbf{B})$, it follows that their sum $\mathbf{x} - \Pi_{\mathbf{B}}(\Pi_{\mathbf{A}}(\mathbf{x}))$ is also orthogonal to $\mathcal{C}(\mathbf{B})$. By the lemma shown in Exercise 2.4, we have $\Pi_{\mathbf{B}}(\mathbf{x}) = \Pi_{\mathbf{B}}(\Pi_{\mathbf{A}}(\mathbf{x}))$. ■

Exercise 3: Linear regression by maximum likelihood (optional)

Suppose that the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d., where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. For any $i \in \{1, \ldots, n\}$, we assume that

$$y_i = w_0 + w_1 x_{i,1} + \cdots + w_d x_{i,d} + \epsilon_i,$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_d)^\top \in \mathbb{R}^{d+1}$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. For simplicity, we define $\bar{\mathbf{x}}_i = (1, x_{i,1}, \ldots, x_{i,d})^\top$, $\mathbf{X} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_n)^\top$, and $\mathbf{y} = (y_1, \ldots, y_n)^\top$, where $\mathbf{X}$ has full rank.

1. Please find the maximum likelihood estimation (MLE) $\hat{\mathbf{w}}$ of the weights $\mathbf{w}$. Specifically, please give the expression of $\hat{w}_0$.

   Solution:
   The probability density function of y_i conditioned on the model parameters and the input variables is

   $$p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma) = \mathcal{N}(\mathbf{w}^\top \bar{\mathbf{x}}_i, \sigma^2).$$

   Suppose that y_i are mutually independent given the model parameters and the input variables. Then the likelihood function is

   $$L(\mathbf{w}, \sigma) = p(\mathbf{y}|\mathbf{x}_i, \mathbf{w}, \sigma) = \prod_{i=1}^n p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma).$$

   The log-likelihood is

   $$\begin{aligned}\ell(\mathbf{w}, \sigma) = \log L(\mathbf{w}) &= \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{w}^\top \bar{\mathbf{x}}_i)^2}{2\sigma^2} \right) \right) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{w}^\top \bar{\mathbf{x}}_i)^2 - n \log \sigma + C \\ &= -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - n \log \sigma + C,\end{aligned}$$

   where C is a constant independent of $\mathbf{w}$ and σ. Then $\hat{\mathbf{w}} = \mathbf{argmax}_{\mathbf{w}}\, \ell(\mathbf{w}, \sigma)$. From

   $$\begin{aligned}\frac{\partial \ell}{\partial \mathbf{w}} = -\frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{X}\mathbf{w} - \mathbf{y}) = 0 &\implies \mathbf{w} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \\ \frac{\partial^2 \ell}{\partial \mathbf{w}^2} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} &= -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} < 0,\end{aligned}$$

   we conclude that $\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$. ■

2. Please find the MLE of σ.

   Solution:
   The MLE of σ is given by

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma} &= \frac{1}{\sigma^3}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 - \frac{n}{\sigma} = 0 \implies \sigma = \frac{\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2}{\sqrt{n}}, \\ \frac{\partial^2 \ell}{\partial \sigma^2} &= -\frac{3}{\sigma^4}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \frac{n}{\sigma^2} < 0, \implies \sigma < \frac{\sqrt{3}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2}{\sqrt{n}}.\end{aligned}$$

   Hence $\hat{\sigma} = \frac{\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2}{\sqrt{n}}$. ■

Exercise 4: Multicollinearity

Consider the linear regression problem formulated as below:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e},\ \mathbb{E}\left(\mathbf{e}\right) = \mathbf{0},\ \operatorname{Cov}\left(\mathbf{e}\right) = \sigma^2 \mathbf{I_n},$$

where $\mathbf{y} = \left(y_1, \ldots, y_n\right)^\top$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Suppose that $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ is the least squares estimator of $\mathbf{w}$.

1. Recall that the covariance matrix of p-dimensional random vectors is defined as

$$\operatorname{Cov}\left(\hat{\mathbf{w}}\right) = \mathbb{E}\Big[(\hat{\mathbf{w}} - \mathbb{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbb{E}(\hat{\mathbf{w}}))^\top\Big].$$

Please show that

(a) $\mathbb{E}\left(\hat{\mathbf{w}}\right) = \mathbf{w}$;

(b) $\operatorname{Cov}\left(\hat{\mathbf{w}}\right) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

Solution:

(a)
$$\begin{aligned}\mathbb{E}\left(\hat{\mathbf{w}}\right) &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right] = \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}+\mathbf{e})\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}\left(\mathbf{e}\right) = \mathbf{w}\end{aligned}$$

(b)
$$\begin{aligned}\operatorname{Cov}\left(\hat{\mathbf{w}}\right) &= \operatorname{Cov}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right) = \operatorname{Cov}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left(\mathbf{X}\mathbf{w}+\mathbf{e}\right)\right) \\ &= \operatorname{Cov}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{e}\right) = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\operatorname{Cov}\left(\mathbf{e}\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)^\top \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.\end{aligned}$$

■

2. We usually measure the quality of an estimator by mean squared error (MSE). The mean squared error (MSE) of estimator $\hat{\mathbf{w}}$ is defined as

$$\operatorname{MSE}\left(\hat{\mathbf{w}}\right) = \mathbb{E}\left[\|\hat{\mathbf{w}} - \mathbf{w}\|^2\right].$$

Please derive that MSE can be decomposed into the variance of the estimator and the squared bias of the estimator, i.e.,

$$\begin{aligned}\operatorname{MSE}\left(\hat{\mathbf{w}}\right) &= \operatorname{tr}\operatorname{Cov}\left(\hat{\mathbf{w}}\right) + \|\mathbb{E}\left(\hat{\mathbf{w}}\right) - \mathbf{w}\|^2 \\ &= \sum_{i=1}^{p}\operatorname{Var}\left(\hat{w}_i\right) + \sum_{i=1}^{p}(\mathbb{E}\left(\hat{w}_i\right) - w_i)^2.\end{aligned}$$

Solution:

$$\begin{aligned}\operatorname{MSE}(\hat{\mathbf{w}}) &= \mathbb{E}\left[\|\hat{\mathbf{w}}-\mathbf{w}\|^2\right] = \mathbb{E}\left[\|\hat{\mathbf{w}}-\mathbb{E}(\hat{\mathbf{w}})+\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w}\|^2\right] \\ &= \mathbb{E}\left[\|\hat{\mathbf{w}}-\mathbb{E}(\hat{\mathbf{w}})\|^2\right] + \mathbb{E}\left[\|\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w}\|^2\right] + 2\mathbb{E}\left[(\hat{\mathbf{w}}-\mathbb{E}(\hat{\mathbf{w}}))^\top(\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w})\right] \\ &= \operatorname{Var}(\hat{\mathbf{w}}) + \|\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w}\|^2 + 2\mathbb{E}\left[\hat{\mathbf{w}}-\mathbb{E}(\hat{\mathbf{w}})\right]^\top(\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w}) \\ &= \operatorname{Var}(\hat{\mathbf{w}}) + \|\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w}\|^2 + 0 \\ &= \operatorname{tr}\operatorname{Cov}(\hat{\mathbf{w}}) + \|\mathbb{E}(\hat{\mathbf{w}})-\mathbf{w}\|^2 = \sum_{i=1}^p \operatorname{Var}(\hat{w}_i) + \sum_{i=1}^p (\mathbb{E}(\hat{w}_i)-w_i)^2. \end{aligned} \quad \blacksquare$$

3. Please show that

$$\operatorname{MSE}(\hat{\mathbf{w}}) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{X}^\top\mathbf{X}$.

Solution:
Because $\mathbb{E}(\hat{\mathbf{w}}) - \mathbf{w} = \mathbf{0}$, we have $\operatorname{MSE}(\hat{\mathbf{w}}) = \operatorname{tr}\operatorname{Cov}(\hat{\mathbf{w}}) = \sigma^2 \operatorname{tr}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]$. If $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{X}^\top\mathbf{X}$, then $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_p}$ are the eigenvalues of $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$, and thus $\operatorname{tr}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right] = \sum_{i=1}^p \frac{1}{\lambda_i}$. Therefore, $\operatorname{MSE}(\hat{\mathbf{w}}) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i}$. ■

4. What would happen if there exists an eigenvalue $\lambda_k \approx 0$?

Solution:
If there exists an eigenvalue $\lambda_k = 0$, it implies that $\mathbf{X}$ is not full rank, or there exists some explanatory variables that are linearly dependent. In this case, the least squares estimator $\hat{\mathbf{w}}$ is not unique, and the MSE of $\hat{\mathbf{w}}$ is not well defined.

If there exists an eigenvalue $\lambda_k \approx 0$, it means that there exists high multicollinearity among the explanatory variables. In this case, the least squares estimator $\hat{\mathbf{w}}$ is still unique and unbiased, but as $\lambda_k \to 0$, it follows that $\operatorname{MSE}(\hat{\mathbf{w}}) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \to \infty$, implying that $\hat{\mathbf{w}}$ may not be a good estimator. ■

Exercise 5: Regularized least squares

Suppose that $\mathbf{X} \in \mathbb{R}^{n \times d}$.

1. Please show that $\mathbf{X}^\top \mathbf{X}$ is always positive semi-definite. Moreover, $\mathbf{X}^\top \mathbf{X}$ is positive definite if and only if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly independent.

   Solution:
   For any $\mathbf{u} \in \mathbf{R}^d$, $\mathbf{u}^\top \mathbf{X}^\top \mathbf{X} \mathbf{u} = \|\mathbf{X}\mathbf{u}\|_2^2 \geq 0$, so $\mathbf{X}^\top \mathbf{X}$ is positive semi-definite.

   If $\mathbf{X}^\top \mathbf{X}$ is positive definite, then $\mathbf{X}^\top \mathbf{X}$ as well as $\mathbf{X}$ must have full rank. Hence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly independent.

   If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly independent, then all eigenvalues of $\mathbf{X}$ are nonzero. The same is true for $\mathbf{X}^\top \mathbf{X}$. Since $\mathbf{X}^\top \mathbf{X}$ is positive semi-definite, all eigenvalues of $\mathbf{X}^\top \mathbf{X}$ must be positive. Therefore, $\mathbf{X}^\top \mathbf{X}$ is positive definite. ■

2. Please show that $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is always invertible, where $\lambda > 0$ and $\mathbf{I} \in \mathbb{R}^{d \times d}$ is an identity matrix.

   Solution:
   For any nonzero $\mathbf{u} \in \mathbf{R}^d$, $\mathbf{u}^\top \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right) \mathbf{u} = \|\mathbf{X}\mathbf{u}\|_2^2 + \lambda \|\mathbf{u}\|_2^2 > 0$, so $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is positive definite and thus invertible. ■

3. Consider the regularized least squares linear regression and denote

$$\mathbf{w}^*(\lambda) = \underset{\mathbf{w}}{\mathbf{argmin}}\ L(\mathbf{w}) + \lambda \Omega(\mathbf{w}),$$

   where $L(\mathbf{w}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ and $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$. For regular parameters $0 < \lambda_1 < \lambda_2$, please show that $L(\mathbf{w}^*(\lambda_1)) < L(\mathbf{w}^*(\lambda_2))$ and $\Omega(\mathbf{w}^*(\lambda_1)) > \Omega(\mathbf{w}^*(\lambda_2))$. Explain intuitively why this holds.

   Solution:
   $\mathbf{w}^*(\lambda)$ is given by

$$\nabla L(\mathbf{w}) + \lambda \nabla \Omega(\mathbf{w}) = \frac{2}{n}\left(\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{y}\right) + 2\lambda \mathbf{w} = 0,$$
$$\nabla^2 L(\mathbf{w}) + \lambda \nabla^2 \Omega(\mathbf{w}) = \frac{2}{n}\mathbf{X}^\top \mathbf{X} + 2\lambda \mathbf{I} > 0.$$

   So $\nabla L(\mathbf{w}^*) + \lambda \nabla \Omega(\mathbf{w}^*) = 0$, from which we derive that

$$\frac{\mathrm{d}\mathbf{w}^*}{\mathrm{d}\lambda} = -\frac{\nabla \Omega(\mathbf{w}^*)}{\nabla^2 L(\mathbf{w}^*) + \lambda \nabla^2 \Omega(\mathbf{w}^*)},$$

   and hence

$$\frac{\mathrm{d}L(\mathbf{w}^*)}{\mathrm{d}\lambda} = -\frac{\nabla L(\mathbf{w}^*) \nabla \Omega(\mathbf{w}^*)}{\nabla^2 L(\mathbf{w}^*) + \lambda \nabla^2 \Omega(\mathbf{w}^*)} > 0,$$
$$\frac{\mathrm{d}\Omega(\mathbf{w}^*)}{\mathrm{d}\lambda} = -\frac{(\nabla \Omega(\mathbf{w}^*))^2}{\nabla^2 L(\mathbf{w}^*) + \lambda \nabla^2 \Omega(\mathbf{w}^*)} < 0.$$

   The conclusion follows.

Intuitively speaking, the regular parameter λ controls the trade-off between the two terms in the objective function. Note that $L(\mathbf{w})$ and $\Omega(\mathbf{w})$ are minimized at different points, i.e. $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ and $\mathbf{0}$, respectively. As λ increases, $\Omega(\mathbf{w})$ becomes more important in the objective function, and $L(\mathbf{w})$ becomes less important. In consequence, $\mathbf{w}^*$ moves closer to $\mathbf{0}$ (in the sense that $\Omega(\mathbf{w}^*)$ decreases), and farther away from $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ (in the sense that $L(\mathbf{w}^*)$ increases). ■

Exercise 6: Conditional Expectations (optional)

Recall that, for supervised learning problems, each data instance consists of a D-dimensional input feature vector $X \in \mathbb{R}^D$ and the corresponding output $Y \in \mathbb{R}$. We would like to find a mapping $f(X)$ to estimate the value of Y given a sample of X. Let

$$\ell(y, f(\mathbf{x})) = (f(\mathbf{x}) - y)^2$$

be the square loss. We choose the function $f(X)$ by minimizing the expectation of the square loss:

$$J[f] := \mathbb{E}[\ell(Y, f(X)] = \iint (y - f(\mathbf{x}))^2 p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y,$$

where $p(\mathbf{x}, y)$ is the joint PDF.

1. Let h be a function of X and $\epsilon > 0$. Please calculate $J[f + \epsilon h] - J[f]$.

   Solution:

$$\begin{aligned} J[f + \epsilon h] - J[f] &= \iint \left\{(y - f(\mathbf{x}) - \epsilon h(\mathbf{x}))^2 - (y - f(\mathbf{x}))^2\right\} p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y \\ &= \iint \left\{\epsilon^2 \left(h(\mathbf{x})\right)^2 - 2\epsilon h(\mathbf{x})(y - f(\mathbf{x}))\right\} p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y \\ &= \epsilon \int h(\mathbf{x}) \left\{\int -2\left(y - f(\mathbf{x})\right) p(\mathbf{x}, y)\mathrm{d}y\right\} \mathrm{d}\mathbf{x} + \epsilon^2 \int \left(h(\mathbf{x})\right)^2 p(\mathbf{x}, y)\mathrm{d}y. \qquad \blacksquare \end{aligned}$$

2. Prove that $J[f + \epsilon h] - J[f] \geq 0$ for any $\epsilon > 0$ if and only if

$$\int h(\mathbf{x}) \left\{\int -2\left(y - f(\mathbf{x})\right) p(\mathbf{x}, y)\mathrm{d}y\right\} \mathrm{d}\mathbf{x} \geq 0.$$

   Solution:

$$\begin{aligned} & \forall\, \epsilon > 0,\ J[f + \epsilon h] - J[f] \geq 0 \\ \Longleftrightarrow\ & \forall\, \epsilon > 0,\ \int h(\mathbf{x}) \left\{\int -2\left(y - f(\mathbf{x})\right) p(\mathbf{x}, y)\mathrm{d}y\right\} \mathrm{d}\mathbf{x} + \epsilon \int \left(h(\mathbf{x})\right)^2 p(\mathbf{x}, y)\mathrm{d}y \geq 0 \\ \Longleftrightarrow\ & \int h(\mathbf{x}) \left\{\int -2\left(y - f(\mathbf{x})\right) p(\mathbf{x}, y)\mathrm{d}y\right\} \mathrm{d}\mathbf{x} \geq \sup_{\epsilon > 0} \left\{-\epsilon \int \left(h(\mathbf{x})\right)^2 p(\mathbf{x}, y)\mathrm{d}y\right\} = 0 \qquad \blacksquare \end{aligned}$$

3. Please show that $f^*(X) = \mathbb{E}[Y|X]$ is a solution to

$$J[f^*] = \min_f\{J[f]\}.$$

   Solution:
   If $J[f^*] = \min_f\{J[f]\}$, then $J[f^* + \epsilon h] - J[f^*] \geq 0$ for any $\epsilon > 0$ and any $h(X)$, i.e.

$$\int h(\mathbf{x}) \left\{\int -2\left(y - f^*(\mathbf{x})\right) p(\mathbf{x}, y)\mathrm{d}y\right\} \mathrm{d}\mathbf{x} \geq 0.$$

for any $h(X)$. So we must have

$$\int -2\left(y - f^*(\mathbf{x})\right) p(\mathbf{x}, y) \mathrm{d}y = 0,$$

which leads to

$$f^*(\mathbf{x}) = \frac{\int y p(\mathbf{x}, y) \mathrm{d}y}{\int p(\mathbf{x}, y) \mathrm{d}y} = \mathbb{E}[Y|X].$$

■

4. Please deduce that

$$\mathbb{E}[\ell(Y, f(X))] = \int \{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} + \iint \{\mathbb{E}[y|\mathbf{x}] - y\}^2 p(\mathbf{x}, y) \mathrm{d}\mathbf{x} \mathrm{d}y.$$

Solution:
We expand the square loss as

$$\begin{aligned} \ell(Y, f(X)) &= \left(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y\right)^2 \\ &= \left\{f(X) - \mathbb{E}[Y|X]\right\}^2 + \left\{\mathbb{E}[Y|X] - Y\right\}^2 + 2\left\{f(X) - \mathbb{E}[Y|X]\right\}\left\{\mathbb{E}[Y|X] - Y\right\}. \end{aligned}$$

Then we have

$$\begin{aligned} \mathbb{E}[\ell(Y, f(X))] = &\int \left\{f(\mathbf{x}) - \mathbb{E}[Y|\mathbf{x}]\right\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} + \iint \left\{\mathbb{E}[Y|\mathbf{x}] - y\right\}^2 p(\mathbf{x}, y) \mathrm{d}\mathbf{x} \mathrm{d}y \\ &- \int \left\{f(\mathbf{x}) - \mathbb{E}[Y|\mathbf{x}]\right\} \left\{\int -2\left\{y - \mathbb{E}[Y|\mathbf{x}]\right\} p(\mathbf{x}, y) \mathrm{d}y\right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

According to Exercise 6.3, the last term on the RHS equals zero. The conclusion follows. ■

Exercise 7: Bias-Variance Trade-off (Programming Exercise. You are required to finish at least one of Exercises 7 and 8.)

We provide you with $L = 100$ data sets, each having $N = 25$ points:

$$\mathcal{D}^{(l)} = \{(x_n, y_n^{(l)})\}_{n=1}^N, \quad l = 1, 2, \cdots, L,$$

where x_n are uniformly taken from $[-1, 1]$, and all points $(x_n, y_n^{(l)})$ are independently from the sinusoidal curve $h(x) = \sin(\pi x)$ with an additional disturbance.

1. For each data set $\mathcal{D}^{(l)}$, consider fitting a model with 24 Gaussian basis functions

$$\phi_j(x) = e^{-(x-\mu_j)^2}, \quad \mu_j = 0.2 \cdot (j - 12.5), \quad j = 1, \cdots 24$$

by minimizing the regularized error function

$$L^{(l)}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^N (y_n^{(l)} - \mathbf{w}^\top \boldsymbol{\phi}(x_n))^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w},$$

where $\mathbf{w} \in \mathbb{R}^{25}$ is the parameter, $\boldsymbol{\phi}(x) = (1, \phi_1(x), \cdots, \phi_{24}(x))^\top$ and λ is the regular coefficient. What's the closed form of the parameter estimator $\hat{\mathbf{w}}^{(l)}$ for the data set $\mathcal{D}^{(l)}$?

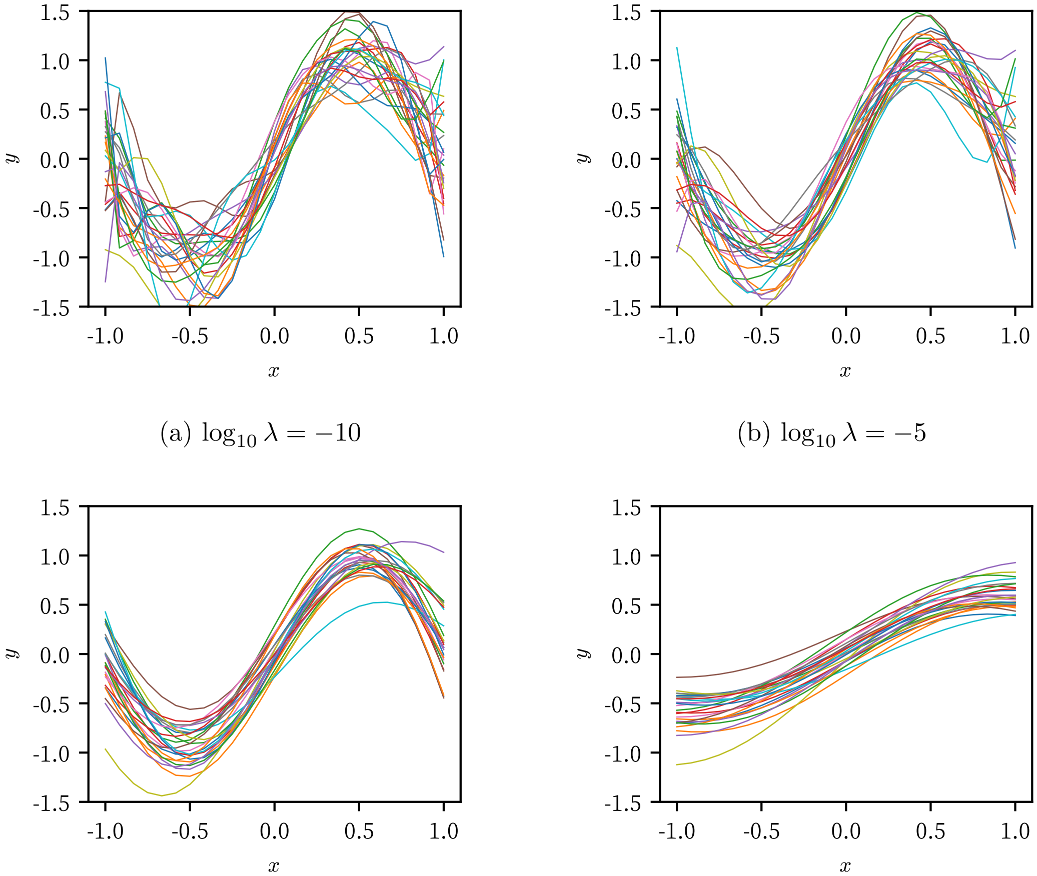
Solution:

$$\nabla L^{(l)}(\mathbf{w}) = -\sum_{n=1}^N (y_n^{(l)} - \mathbf{w}^\top \boldsymbol{\phi}(x_n))\boldsymbol{\phi}(x_n) + \lambda \mathbf{w} = 0$$

$$\implies \hat{\mathbf{w}}^{(l)} = \left(\sum_{n=1}^N \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^\top + \lambda \mathbf{I}\right)^{-1} \sum_{n=1}^N y_n^{(l)} \boldsymbol{\phi}(x_n).$$ ■

2. For $\log_{10} \lambda = -10, -5, -1, 1$, plot the prediction functions $y^{(l)}(x) = f_{\mathcal{D}^{(l)}}(x)$ on $[-1, 1]$ respectively. For clarity, show only the first 25 fits in the figure for each λ.

Solution:
The plots are shown in Figure 2. ■



(a) $\log_{10}\lambda = -10$

(b) $\log_{10}\lambda = -5$

(c) $\log_{10}\lambda = -1$

(d) $\log_{10}\lambda = 1$

Figure 2: Prediction functions $y^{(l)}(x) = f_{\mathcal{D}^{(l)}}(x)$ for $\log_{10}\lambda = -10, -5, -1, 1$.

3. For $\log_{10}\lambda \in [-3, 1]$, calculate the followings:

$$\bar{y}(x) = \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)] = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

$$(\text{bias})^2 = \mathbb{E}_X[(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(X)] - h(X))^2] = \frac{1}{N}\sum_{n=1}^{N}(\bar{y}(x_n) - h(x_n))^2$$

$$\text{variance} = \mathbb{E}_X[\mathbb{E}_{\mathcal{D}}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2]] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}(y^{(l)}(x_n) - \bar{y}(x_n))^2$$

Plot the three quantities, $(\text{bias})^2$, variance and $(\text{bias})^2$+variance in one figure, as the functions of $\log_{10}\lambda$. (**Hint:** see [1] for an example.)

Solution:
The plot is shown in Figure 3. ■

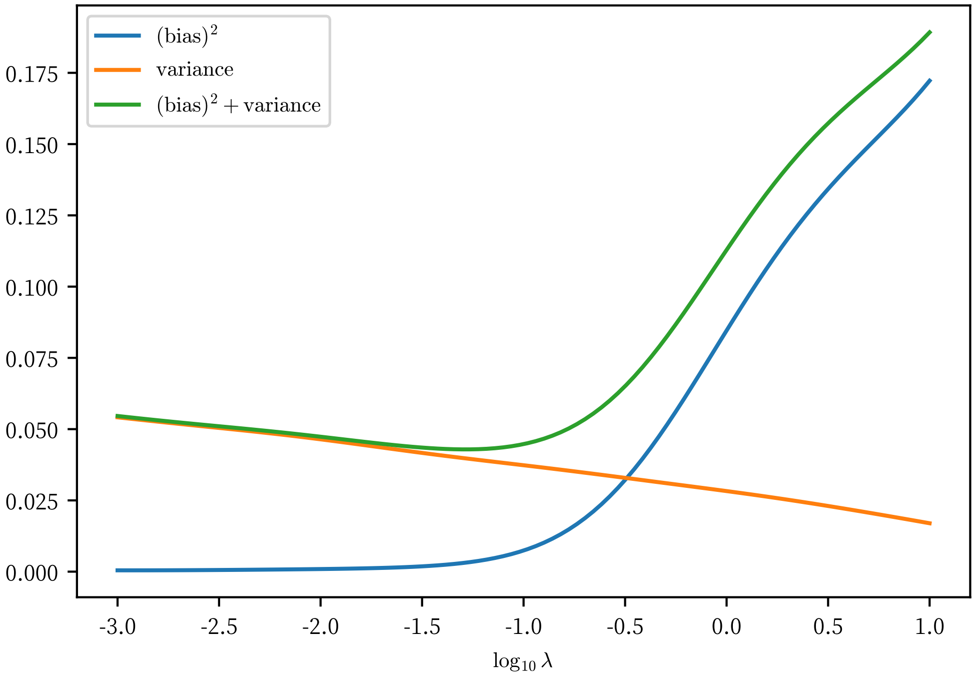


Figure 3: Bias and variance as functions of $\log_{10} \lambda$.

Exercise 8: Bayesian Linear Regression (Programming Exercise. You are required to finish at least one of Exercises 7 and 8.)

Consider a single input variable $\mathbf{x}$, a single output variable $\mathbf{y}$ and a linear model of the form $\mathbf{y} = w_0 + w_1\mathbf{x} + \epsilon$, where ϵ is Gaussian distributed with mean of 0 and standard deviation of 0.25.

1. Suppose that, the model parameter $\mathbf{w} = (w_0, w_1)^T \in \mathbb{R}^2$ has a Gaussian prior of the form

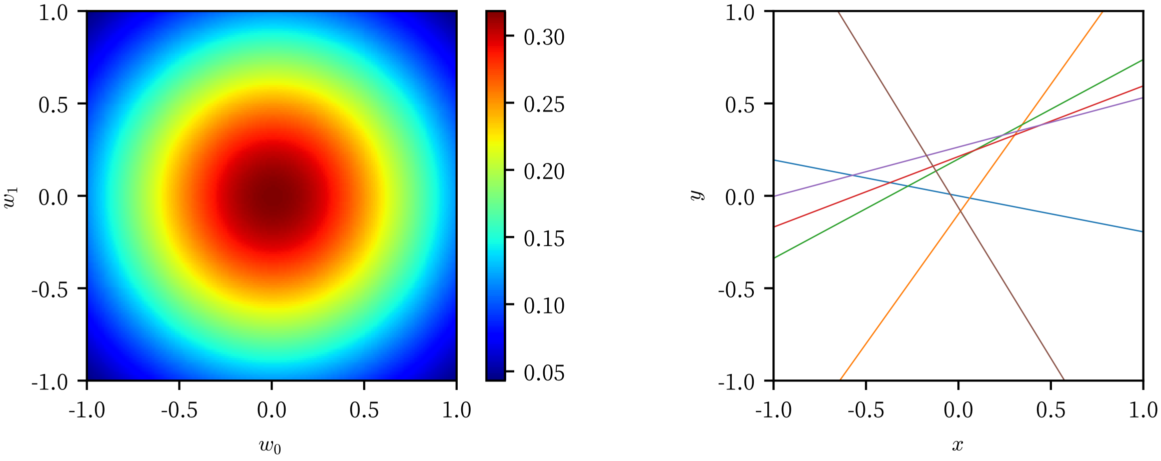
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \frac{1}{2\pi}\frac{1}{|\boldsymbol{\Sigma}_0|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{w} - \boldsymbol{\mu}_0)\}$$

where $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \frac{1}{2}\mathbf{I}$. Please plot this Gaussian distribution in the form of heat map.

Solution:
The heat map is shown in Figure 4a. ■

2. Sample six times independently from the prior Gaussian distribution defined above. Please plot the six straight lines $y = w_0 + w_1 x$ using these samples.

Solution:
The plot of samples is shown in Figure 4b. ■



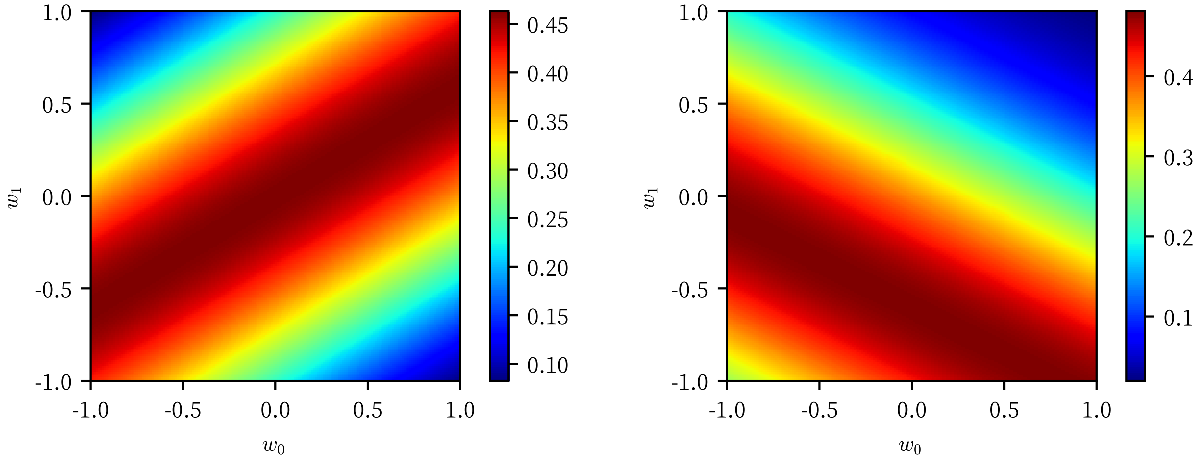
(a) $p(\mathbf{w})$

(b) $y = w_0 + w_1 x$

Figure 4: Gaussian prior and samples.

3. Now, suppose that we have observed a single data point $(x_1, y_1) = (0.6, 0)$. Please plot the likelihood function $p(y_1|x_1, \mathbf{w})$ for this data point as the function of $\mathbf{w}$, still in the form of heat map.

Solution:
The plot of likelihood is shown in Figure 5a. ■



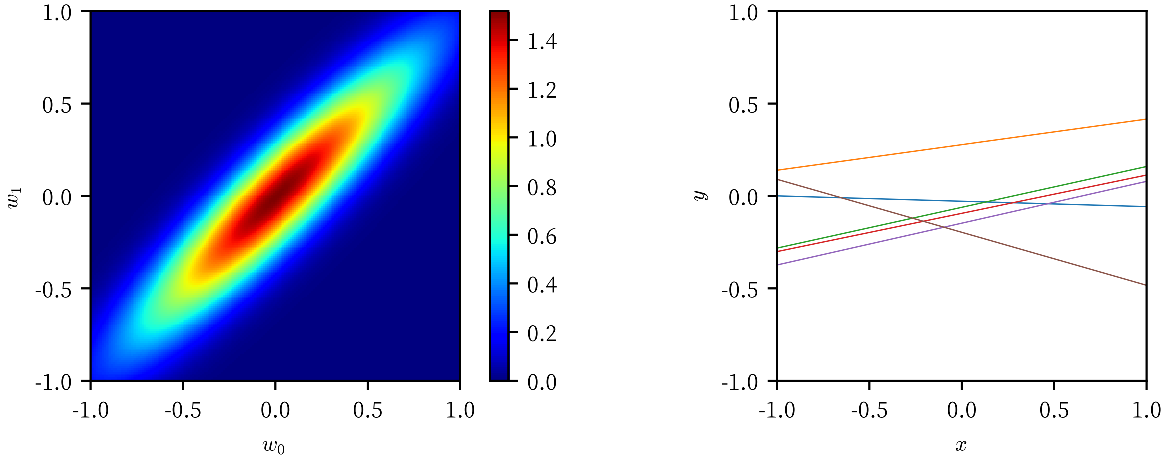
(a) $p(y_1|x_1, \mathbf{w})$

(b) $p(y_2|x_2, \mathbf{w})$

Figure 5: Likelihood functions for (x_1, y_1) and (x_2, y_2).

4. Calculate the posterior distribution of $\mathbf{w}$, denoted by $p(\mathbf{w}|y_1, x_1)$. Please plot the posterior distribution.

   Solution:
   The heat map of posterior is shown in Figure 6a. ■



(a) $p(\mathbf{w}|y_1, x_1)$

(b) $y = w_0 + w_1 x$

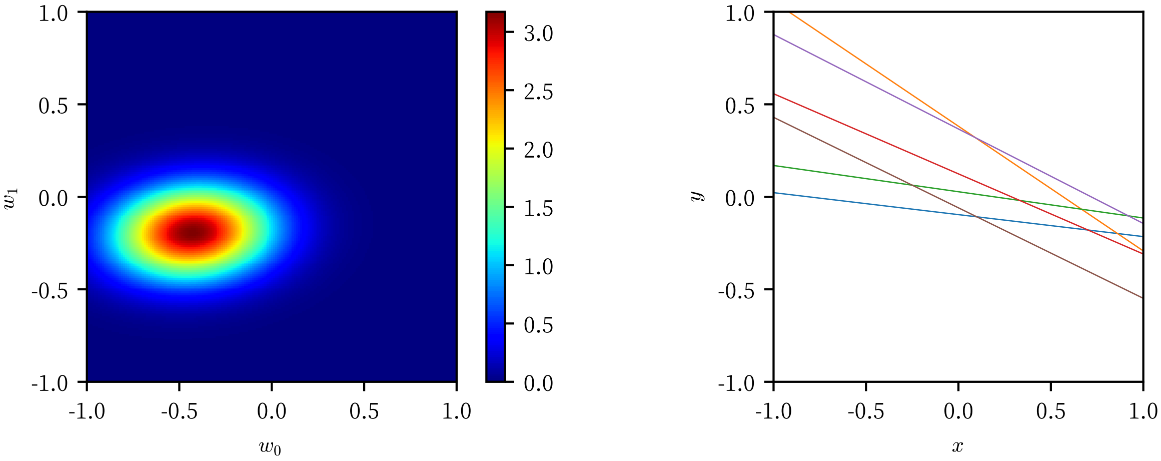
Figure 6: Posterior and samples for (x_1, y_1).

5. Sample six times independently from this posterior distribution of $\mathbf{w}$ and plot the six straight lines $y = w_0 + w_1 x$.

   Solution:
   The plot of samples is shown in Figure 6b. ■

6. Then, suppose we observe a new single data point $(x_2, y_2) = (-0.5, 0.6)$. Please plot the corresponding likelihood function $p(y_2|x_2, \mathbf{w})$ of this second point alone, the posterior distribution of $\mathbf{w}$, denote by $p(\mathbf{w}|y_1, y_2, x_1, x_2)$, and six samples drawn from the current posterior function.

   Solution:
   The plot of likelihood is shown in Figure 5b. The plots of posterior and samples are shown in Figure 7a and 7b. ■



(a) $p(\mathbf{w}|y_1, y_2, x_1, x_2)$ (b) $y = w_0 + w_1 x$

Figure 7: Posterior and samples for (x_1, y_1).

7. If we can observe new data points continuously, and then observe the posterior distributions and their sampled linear regression models sequentially, what will you infer from them? Please write down your conclusions. (**Hint:** see [1] for an example.)

   Solution:
   As more and more new data points are observed, the posterior distribution would become sharper and sharper, and the sampled linear regression models would become more and more accurate. In the limit of an infinite number of data points, the posterior distribution would become a delta function centred on the true parameter values, and the sampled models would become the true linear regression model. ■

Exercise 9: Covariance Matrix and Gaussian Distribution

Let $\mathbf{X} = (X_1, X_2, \cdots, X_D)^\top \in \mathbb{R}^D$ be a D-dimensional random vector. The covariance matrix of $\mathbf{X}$, denoted by $\mathbf{\Sigma_X}$, is defined as

$$\operatorname{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\right].$$

1. Please show that $\mathbf{\Sigma_X}$ is positive semi-definite.

   Solution:
   For any $\mathbf{u} \in \mathbb{R}^D$, we have $\mathbf{u}^\top \mathbf{\Sigma_X} \mathbf{u} = \mathbb{E}\left[\left\|\mathbf{u}^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right\|_2^2\right] \geq 0$. Thus $\mathbf{\Sigma_X}$ is positive semi-definite. ■

2. Please show that $\mathbf{\Sigma_X}$ doesn't have full rank if and only if $\{X_i - \mathbb{E}[X_i]\}_{i=1}^D$ are linearly dependent.

   Solution:
   Since $\mathbf{\Sigma_X}$ is positive semi-definite, $\mathbf{\Sigma_X}$ doesn't have full rank if and only if for some $\mathbf{u} \in \mathbb{R}^D$, $\mathbf{u}^\top \mathbf{\Sigma_X} \mathbf{u} = \mathbb{E}\left[\left\|\mathbf{u}^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right\|_2^2\right] = 0$. Note that $\left\|\mathbf{u}^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right\|_2^2 \geq 0$, so the expectation equals zero if and only if $\mathbf{u}^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}]) = \mathbf{0}$, i.e. $\{X_i - \mathbb{E}[X_i]\}_{i=1}^D$ are linearly dependent. ■

3. Suppose that, the random vector $\mathbf{X}$ has a multivariate Gaussian distribution with the mean vector being $\boldsymbol{\mu}$ and the covariance matrix being $\mathbf{\Sigma}$, respectively. The probability density function of $\mathbf{X}$ is

   $$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\},$$

   where $\mathbf{x}$ is a realization of the random vector $\mathbf{X}$. For notational simplicity, let

   $$c = (2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}. \tag{4}$$

   Clearly, we must have

   $$\int_{\mathbb{R}^D} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \mathrm{d}\mathbf{x} = c. \tag{5}$$

   Now, let us denote the first M components of $\mathbf{X}$ by $\mathbf{X}_a$, and the remaining $D - M$ ones by $\mathbf{X}_b$, so that

   $$\mathbf{X} = \begin{pmatrix} \mathbf{X}_a \\ \mathbf{X}_b \end{pmatrix}.$$

   We denote the corresponding partitions of the mean vector by

   $$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

and the covariance matrix by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

Please show that $\mathbf{X}_a$ has a Gaussian distribution with its mean vector being $\boldsymbol{\mu}_a$ and the covariance matrix being $\boldsymbol{\Sigma}_{aa}$. In other words, please show that

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\mathrm{d}\mathbf{x}_b = \frac{1}{(2\pi)^{M/2}}\frac{1}{|\boldsymbol{\Sigma}_{aa}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right\}.$$

(**Hint:**

(a) you can make use of identities similar to Eq. (4) and Eq. (5) to integrate out $\mathbf{x}_b$.

(b) you may find the following identity useful:

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{aa}||\boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}| = |\boldsymbol{\Sigma}_{bb}||\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}|.$$

Solution:
Denote $\boldsymbol{\Sigma}^{-1}$ by

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa}^{-1} + \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Lambda}_{bb}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1} & -\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Lambda}_{bb} \\ -\boldsymbol{\Lambda}_{bb}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}.$$

where $\boldsymbol{\Lambda}_{bb} = (\boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab})^{-1}$ is the inverse of the Schur complement of $\boldsymbol{\Sigma}_{aa}$ in $\boldsymbol{\Sigma}$. We have

$$\begin{aligned}
&(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \\
=& \begin{pmatrix} \mathbf{X}_a - \boldsymbol{\mu}_a \\ \mathbf{X}_b - \boldsymbol{\mu}_b \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{X}_a - \boldsymbol{\mu}_a \\ \mathbf{X}_b - \boldsymbol{\mu}_b \end{pmatrix} \\
=& (\mathbf{X}_a - \boldsymbol{\mu}_a)^\top (\boldsymbol{\Sigma}_{aa}^{-1} + \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Lambda}_{bb}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1})(\mathbf{X}_a - \boldsymbol{\mu}_a) \\
&+2(\mathbf{X}_a - \boldsymbol{\mu}_a)^\top(-\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Lambda}_{bb})(\mathbf{X}_b - \boldsymbol{\mu}_b) + (\mathbf{X}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{bb}(\mathbf{X}_b - \boldsymbol{\mu}_b) \\
=& (\mathbf{X}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{X}_a - \boldsymbol{\mu}_a) \\
&+(\mathbf{X}_b - \boldsymbol{\mu}_b - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{X}_a - \boldsymbol{\mu}_a))^\top \boldsymbol{\Lambda}_{bb}(\mathbf{X}_b - \boldsymbol{\mu}_b - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{X}_a - \boldsymbol{\mu}_a))
\end{aligned}$$

Let $\mathbf{X}_{b'} = \mathbf{X}_b - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\mathbf{X}_a$ and $\boldsymbol{\mu}_{b'} = \boldsymbol{\mu}_b - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\mu}_a$. Then the PDF of $\mathbf{X}$ becomes

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{M/2}}\frac{1}{|\boldsymbol{\Sigma}_{aa}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right\} \\
&\quad \cdot \frac{1}{(2\pi)^{(D-M)/2}}\frac{1}{|\boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_{b'} - \boldsymbol{\mu}_{b'})^\top \boldsymbol{\Lambda}_{bb}(\mathbf{x}_{b'} - \boldsymbol{\mu}_{b'})\right\}.
\end{aligned}$$

Hence, by changing variables, we integrate out $\mathbf{X}_b$ and obtain the marginal PDF of $\mathbf{X}_a$ as

$$\begin{aligned} p(\mathbf{x}_a) = &\int_{\mathbb{R}^{D-M}} p(\mathbf{x}_a, \mathbf{x}_b)\mathrm{d}\mathbf{x}_b \\ = &\frac{1}{(2\pi)^{M/2}} \frac{1}{|\mathbf{\Sigma}_{aa}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \mathbf{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right\} \\ &\cdot \frac{1}{(2\pi)^{(D-M)/2}} \frac{1}{|\mathbf{\Lambda}_{bb}^{-1}|^{1/2}} \int_{\mathbb{R}^{D-M}} \exp\left\{-\frac{1}{2}(\mathbf{x}_{b'} - \boldsymbol{\mu}_{b'})^\top \mathbf{\Lambda}_{bb}(\mathbf{x}_{b'} - \boldsymbol{\mu}_{b'})\right\} \mathrm{d}\mathbf{x}_{b'} \\ = &\frac{1}{(2\pi)^{M/2}} \frac{1}{|\mathbf{\Sigma}_{aa}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \mathbf{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right\}, \end{aligned}$$

where we have used the fact that $\mathbf{\Sigma}_{bb} - \mathbf{\Sigma}_{ba}\mathbf{\Sigma}_{aa}^{-1}\mathbf{\Sigma}_{ab} = \mathbf{\Lambda}_{bb}^{-1}$ and $|\mathbf{\Sigma}| = |\mathbf{\Sigma}_{aa}||\mathbf{\Lambda}_{bb}^{-1}|$. ■

Exercise 10: Determinant and geometric (optional)

The determinant of a square matrix can be viewed as the signed volume spanned by the columns or rows of the matrix.

1. Consider three vectors $\mathbf{a} = (1, 2, 3, 4)^\top$, $\mathbf{b} = (5, 6, 7, 8)^\top$ and $\mathbf{c} = (7, -11, 1, 3)^\top$ in $\mathbb{R}^4$. Please find the volume of the parallelepipedon spanned by $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. You may first find a unit vector $\mathbf{n}$ such that $\mathbf{n} \perp \operatorname{Span}(\{\mathbf{a}, \mathbf{b}, \mathbf{c}\})$ and then calculate the volume of the parallelogram spanned by $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{n}$. Explain why you can do so.

   Solution:
   By solving the system

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 7 & -11 & 1 & 3 \\ x & y & z & w \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

   we obtain $\mathbf{n} = \pm \left(\frac{1}{6}, \frac{1}{6}, -\frac{5}{6}, \frac{1}{2}\right)^\top$, which is a unit vector orthogonal to $\operatorname{Span}(\{\mathbf{a}, \mathbf{b}, \mathbf{c}\})$. The volume of the parallelogram spanned by $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{n}$ is

$$\left\| \begin{matrix} 1 & 5 & 7 & \frac{1}{6} \\ 2 & 6 & -11 & \frac{1}{6} \\ 3 & 7 & 1 & -\frac{5}{6} \\ 4 & 8 & 3 & \frac{1}{2} \end{matrix} \right\| = 240,$$

   which can also be viewed as the volume of the parallelepipedon spanned by $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. This is because the volume of a parallelogram is the product of its base and altitude. Specifically, the volume of the parallelepipedon spanned by $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ can be viewed as a base of the parallelogram spanned by $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{n}$, with the corresponding altitude $|\mathbf{n}| = 1$. ■

2. Consider two vectors $\mathbf{a} = (1, 2, 3, 4)^\top$ and $\mathbf{b} = (5, 6, 7, 8)^\top$ in $\mathbb{R}^4$. Please find the area of the parallelogram spanned by $\mathbf{a}$ and $\mathbf{b}$.

   Solution:
   By applying Gram-Schmidt orthogonalization, we obtain $\mathbf{c} = \left(-\frac{\sqrt{70}}{70}, -\frac{\sqrt{70}}{35}, \frac{\sqrt{70}}{10}, -\frac{2\sqrt{70}}{35}\right)^\top$ and $\mathbf{d} = \left(\frac{\sqrt{14}}{7}, -\frac{3\sqrt{14}}{14}, 0, \frac{\sqrt{14}}{14}\right)^\top$, which form an orthonormal basis of $\operatorname{Span}(\{\mathbf{a}, \mathbf{b}\})^\top$. The area of the parallelogram spanned by $\mathbf{a}$ and $\mathbf{b}$ is

$$\left\| \begin{matrix} 1 & 5 & -\frac{\sqrt{70}}{70} & \frac{\sqrt{14}}{7} \\ 2 & 6 & -\frac{\sqrt{70}}{35} & -\frac{3\sqrt{14}}{14} \\ 3 & 7 & \frac{\sqrt{70}}{10} & 0 \\ 4 & 8 & -\frac{2\sqrt{70}}{35} & \frac{\sqrt{14}}{14} \end{matrix} \right\| = 8\sqrt{5},$$

   ■

3. Now we want to calculate the n-dimension volume of an n-dimension parallelepiped in $\mathbb{R}^m$ $(n \leq m)$. The parallelepiped P has the form

$$P = \left\{ \mathbf{a} + \sum_{i=1}^{n} \lambda_i \mathbf{b}_i ; 0 \leq \lambda_i \leq 1, 1 \leq i \leq n \right\},$$

where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b}_i \in \mathbb{R}^m$, $1 \leq i \leq n$. Show that the volume is given by

$$\mathrm{V}(P) = \sqrt{\det\left(\mathbf{B}^\top \mathbf{B}\right)},$$

where $\mathbf{B}$ denotes the $m \times n$ matrix whose i^{th} column is the vector $\mathbf{b}_i$.

[Hint: you may follow the idea in question 1.]

Solution:
Prove by induction on n. The base case $n = 1$ has $\mathbf{B}^\top \mathbf{B}$ equal to the square of the length of the vector $\mathbf{b}_1$, and hence is trivially $(\mathrm{V}(P))^2$. Assume that the result is true for $n - 1$. We denote the $m \times (n-1)$ matrix whose i^{th} column is $\mathbf{b}_i$ by $\mathbf{B}_{n-1}$ and the corresponding parallelepiped by P_{n-1}. We can write $\mathbf{b}_n$ as $\mathbf{B}_{n-1}\mathbf{u} + \mathbf{b}$, where $\mathbf{B}_{n-1}\mathbf{u}$ is its projection onto $\mathcal{C}(\mathbf{B}_{n-1})$ and $\mathbf{b} \in \mathcal{C}(\mathbf{B}_{n-1})^\top$. Then

$$\mathbf{B}^\top \mathbf{B} = \begin{pmatrix} \mathbf{B}_{n-1}^\top \\ \mathbf{b}_n^\top \end{pmatrix} \begin{pmatrix} \mathbf{B}_{n-1} & \mathbf{b}_n \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{n-1}^\top \mathbf{B}_{n-1} & \mathbf{B}_{n-1}^\top \mathbf{B}_{n-1}\mathbf{u} \\ \mathbf{u}^\top \mathbf{B}_{n-1}^\top \mathbf{B}_{n-1} & \mathbf{u}^\top \mathbf{B}_{n-1}^\top \mathbf{B}_{n-1}\mathbf{u} + \mathbf{b}^\top \mathbf{b} \end{pmatrix}$$

$$\implies \det\left(\mathbf{B}^\top \mathbf{B}\right) = \det\left(\mathbf{B}_{n-1}^\top \mathbf{B}_{n-1}\right) \det\left(\mathbf{b}^\top \mathbf{b}\right),$$

where $\sqrt{\det\left(\mathbf{B}_{n-1}^\top \mathbf{B}_{n-1}\right)} = \mathrm{V}_{n-1}(P)$ can be viewed as a base of P and $\sqrt{\det\left(\mathbf{b}^\top \mathbf{b}\right)} = |\mathbf{b}|$ as the corresponding altitude. Therefore, the volume of P is

$$\mathrm{V}(P) = \mathrm{V}_{n-1}(P)|\mathbf{b}| = \sqrt{\det\left(\mathbf{B}^\top \mathbf{B}\right)},$$

as desired. ■

4. Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$, please show that

$$|\det(\mathbf{A})| \leq \prod_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_2,$$

where $\boldsymbol{\alpha}_i$ is the i^{th} row of $\mathbf{A}$. Then explain the geometrical meaning of this inequality.

Solution:
Denote the $n \times k$ matrix whose i^{th} row is $\boldsymbol{\alpha}_i$ by $\mathbf{A}_k$. We prove $\sqrt{\det(\mathbf{A}_k^\top \mathbf{A}_k)} \leq \prod_{i=1}^{k} \|\boldsymbol{\alpha}_i\|_2$ by induction on k. The base case $k = 1$ is trivial. Assume that the result is true for $k - 1$. We

can write $\boldsymbol{\alpha}_k$ as $\mathbf{A}_{k-1}\mathbf{u}+\boldsymbol{\alpha}$, where $\mathbf{A}_{k-1}\mathbf{u}$ is its projection onto $\mathcal{C}(\mathbf{A}_{k-1})$ and $\boldsymbol{\alpha} \in \mathcal{C}(\mathbf{A}_{k-1})^\top$. Then

$$\mathbf{A}_k^\top \mathbf{A}_k = \begin{pmatrix} \mathbf{A}_{k-1}^\top \\ \boldsymbol{\alpha}_k^\top \end{pmatrix} \begin{pmatrix} \mathbf{A}_{k-1} & \boldsymbol{\alpha}_k \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{k-1}^\top \mathbf{A}_{k-1} & \mathbf{A}_{k-1}^\top \mathbf{A}_{k-1}\mathbf{u} \\ \mathbf{u}^\top \mathbf{A}_{k-1}^\top \mathbf{A}_{k-1} & \mathbf{u}^\top \mathbf{A}_{k-1}^\top \mathbf{A}_{k-1}\mathbf{u} + \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \end{pmatrix}$$

$$\Longrightarrow \det\left(\mathbf{A}_k^\top \mathbf{A}_k\right) = \det\left(\mathbf{A}_{k-1}^\top \mathbf{A}_{k-1}\right)\det\left(\boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right) \le \det\left(\mathbf{A}_{k-1}^\top \mathbf{A}_{k-1}\right)\|\boldsymbol{\alpha}_k\|_2^2 .$$

Hence $\sqrt{\det\left(\mathbf{A}_k^\top \mathbf{A}_k\right)} \le \sqrt{\det\left(\mathbf{A}_{k-1}^\top \mathbf{A}_{k-1}\right)}\,\|\boldsymbol{\alpha}_k\|_2 \le \prod_{i=1}^k \|\boldsymbol{\alpha}_i\|_2$. When $k = n$, we have $|\det(\mathbf{A})| \le \prod_{i=1}^n \|\boldsymbol{\alpha}_i\|_2$, as desired.

The geometric meaning of this inequality is that the volume of the parallelepiped spanned by the columns of $\mathbf{A}$ is bounded by the product of the lengths of the columns. ■

Exercise 11: Calculus in Bayesian linear regression

1. In Bayesian linear regression lecture, we suppose that the model parameter $\mathbf{w}$ has a Gaussian prior of the form $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$. Then, given a set of input data instances $\{\mathbf{x}_i\}_{i=1}^n$, the joint distribution of the corresponding target variables is a Gaussian $p(\mathbf{y}|\mathbf{w}, \mathbf{X}) = \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)$. We first find the joint distribution over $\mathbf{w}$ and $\mathbf{y}$. Let

$$\mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix}.$$

The log of the joint distribution is

$$\ln p(\mathbf{z}) = \ln p(\mathbf{w}) + \ln p(\mathbf{y}|\mathbf{w}).$$

The calculation of Gaussian density $p(\mathbf{z})$ is a tedious work. We are going to find a simpler way to calculate $p(\mathbf{z})$.

(a) Let $\boldsymbol{\xi}$ is an n-dimension Gaussian random vector, satisfying $\mathbb{E}(\boldsymbol{\xi}) = \boldsymbol{\mu}$, $\operatorname{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Sigma}$. Show that for any $n \times n$ matrix $\mathbf{A}$, $\mathbf{A}\boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$.

(b) We can rewrite $\mathbf{y}$ as $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and is independent of $\mathbf{w}$. Please find the covariance matrix of $\mathbf{y}$ and the matrix $\mathbb{E}\left[(\mathbf{w} - \mathbb{E}[\mathbf{w}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^\top\right]$. Then write down the mean and covariance matrix of $\mathbf{z}$.

Solution:

(a) Let $\delta > 0$ be small enough such that $\mathbf{A}' = \mathbf{A} + \delta\mathbf{I}$ is invertible. The PDF of $\mathbf{A}'\xi$ is

$$p_{\mathbf{A}'\boldsymbol{\xi}}(\mathbf{x}) = \frac{p_{\boldsymbol{\xi}}(\mathbf{A}'^{-1}\mathbf{x})}{|\mathbf{A}'|} = \frac{\mathcal{N}(\mathbf{A}'^{-1}\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{|\mathbf{A}'|}$$

Suppose that $\boldsymbol{\Sigma}$ is invertible, then

$$\begin{aligned} p_{\mathbf{A}'\boldsymbol{\xi}}(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{|\mathbf{A}'||\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{A}'^{-1}\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{A}'^{-1}\mathbf{x} - \boldsymbol{\mu})\right\} \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{|\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}'^\top|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{A}'\boldsymbol{\mu})^\top (\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}'^\top)^{-1}(\mathbf{x} - \mathbf{A}'\boldsymbol{\mu})\right\} \\ &= \mathcal{N}\left(\mathbf{x}\middle|\mathbf{A}'\boldsymbol{\mu}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}'^\top\right) \end{aligned}$$

Let $\delta \to 0$. Then $\mathbf{A}'\boldsymbol{\mu} \to \mathbf{A}\boldsymbol{\mu}$ and $\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}'^\top \to \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$. We assert $\mathbf{A}\boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$ without rigorous proof, which is out of the scope. When $\boldsymbol{\Sigma}$ is not invertible, the PDF of $\boldsymbol{\xi}$ can be written with rank, pseudoinverse and pseudo-determinant of $\boldsymbol{\Sigma}$ and the same conclusion holds.

(b) By (a), we have

$$\begin{aligned} \begin{pmatrix} \mathbf{w} \\ \boldsymbol{\epsilon} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\right) \\ \implies \mathbf{y} = \begin{pmatrix} \mathbf{X} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \boldsymbol{\epsilon} \end{pmatrix} &\sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\mu}_0, \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top\right) \end{aligned}$$

Hence $\boldsymbol{\Sigma}_{\mathbf{y}} = \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top$. Analogously, we have

$$\mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{X} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_0 \\ \mathbf{X}\boldsymbol{\mu}_0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0\mathbf{X}^\top \\ \mathbf{X}\boldsymbol{\Sigma}_0 & \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top \end{pmatrix} \right).$$

Hence $\boldsymbol{\Sigma}_{\mathbf{wy}} = \mathbb{E}\left[(\mathbf{w} - \mathbb{E}[\mathbf{w}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^\top\right] = \boldsymbol{\Sigma}_0\mathbf{X}^\top$. Moreover, $\boldsymbol{\mu}_{\mathbf{z}} = \begin{pmatrix} \boldsymbol{\mu}_0 \\ \mathbf{X}\boldsymbol{\mu}_0 \end{pmatrix}$, $\boldsymbol{\Sigma}_{\mathbf{z}} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0\mathbf{X}^\top \\ \mathbf{X}\boldsymbol{\Sigma}_0 & \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top \end{pmatrix}$. ■

2. (Optional) We know that a multivariate Gaussian random vector $\mathbf{X}$ with uncorrelated components has mean vector $\boldsymbol{\mu}$ and the invertible covariance matrix $\boldsymbol{\Sigma}$. The probability density function of $\mathbf{X}$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

But now we concentrate on the multivariate Gaussian random vector $\mathbf{X}$ with correlated components, which means that $|\boldsymbol{\Sigma}| = 0$ and $\boldsymbol{\Sigma}$ is not invertible.

Specifically, Suppose X is a Gaussian random variable with mean μ and variance σ^2. Another random variable $Y = aX + b$, where a, b are non-zero real numbers. Please show that X and Y are correlated, then find the joint density function of X and Y.

Hint: you may use the Dirac Delta function. This function, typically written $\delta(x)$, is defined as:

$$\delta(x) = \begin{cases} \infty & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

with the properties that

(a) $\int_{-\infty}^{\infty} \delta(x) dx = 1$;

(b) $\int_{-\infty}^{\infty} f(x)\delta(x - x_0)\, dx = f(x_0)$ for any function $f(x)$ that is continuous around $x = x_0$.

Solution:

$$\begin{aligned} \begin{pmatrix} X \\ 1 \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} \right) \\ \implies \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & b \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mu \\ a\mu + b \end{pmatrix}, \begin{pmatrix} \sigma^2 & a\sigma^2 \\ a\sigma^2 & a^2\sigma^2 \end{pmatrix} \right). \end{aligned}$$

The covariance matrix of $\begin{pmatrix} X \\ Y \end{pmatrix}$ has rank 1, which means that X and Y are correlated.

The joint PDF can be obtained by

$$\begin{aligned} p_{X,Y}(x, y) &= p_{Y|X}(y|x)\, p_X(x) \\ &= \delta(y - ax - b) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}. \end{aligned}$$

■

Exercise 12: Inverse of block matrix

Please prove Lemma 1 in Bayesian linear regression lecture.

Lemma 1.

Suppose that the involved matrices are invertible. Then,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}$$

where

$$\mathbf{M} = \left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}.$$

Please imitate the process of finding an inverse of a matrix $\mathbf{X}$, i.e., we first write $(\mathbf{X}, \mathbf{I})$ and then executing elementary row operations to get $(\mathbf{I}, \mathbf{X}^{-1})$.

Solution:

$$\left(\begin{array}{cc|cc} \mathbf{A} & \mathbf{B} & \mathbf{I} & \mathbf{O} \\ \mathbf{C} & \mathbf{D} & \mathbf{O} & \mathbf{I} \end{array}\right) \to \left(\begin{array}{cc|cc} \mathbf{A} & \mathbf{B} & \mathbf{I} & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} & \mathbf{O} & \mathbf{D}^{-1} \end{array}\right) \to \left(\begin{array}{cc|cc} \mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{O} & \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} & \mathbf{O} & \mathbf{D}^{-1} \end{array}\right)$$

$$\to \left(\begin{array}{cc|cc} \mathbf{I} & \mathbf{O} & \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} & \mathbf{O} & \mathbf{D}^{-1} \end{array}\right) \to \left(\begin{array}{cc|cc} \mathbf{I} & \mathbf{O} & \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} & -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{array}\right) \quad \blacksquare$$

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.