Exercises

1	2	3	4	5	6	7	8	9	10

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Surname, First name

Modelling, Uncertainty and Data for Engineers (CEGM1000)
Exam 22/23 Q2

Do not open the exam until given permission by the instructor!
(You can write your name and student ID on this page)

The exam is 180 minutes. The table below gives an overview. On the following pages, some questions have a specific box for you to answer: anything written outside the boxes will not be graded. Note that we have provided a lot of space for answers. The answer space size is not an indicator of how long we expect your answers to be! (shorter is generally better). Points indicate the relative amount of time expected to be spent for each question. Scratch paper is available to use during the exam, but will not be collected or graded. You may use pen or pencil and a scientific calculator. Any required equations are generally provided in the question description.
Don't forget to write your student ID and fill in the bubbles on the top right of this page. Good luck!

No.	Question	Sub-Q	Type	Points
1	Coding	a, b, c, d	MC	4
2	Finite Difference Methods	a	MC	2
2		b, c	Open	8
3	Finite Element Methods	a, b	MC	10
3		c	Open	3
4	Optimization		Open	5
5	Optimization		Open	5
6	Optimization		Open	5
7	Signal Processing	a, b, c	Open	15
8	Time Series	a, b, c, d	Open	14
9	Machine Learning	a, b, c, d, e	MC/MS	14
10	Risk and Reliability	a	Open	5
		b, c, d	MC	6
		e	Open	4
		Total:		**100**

Part 1: Coding

1p **1a** What does the acronym FAIR stand for?

- a Flexible, Available, Interoperable, Reusable
- b Findable, Accessible, Interoperable, Repeatable
- c Flexible, Available, Interoperable, Repeatable
- d Findable, Accessible, Interoperable, Reusable

1p **1b** What does the adjective "Interoperable" stand for in FAIR data?

- a It means that data can be used by multiple operators concurrently, regardless of who they are (e.g., researchers, publishers, stakeholders, ...).
- b It means that data can be integrated with other data. To this end, data must be standardised.
- c It means that data can be used across multiple operating systems (e.g., Windows, Linux, ...).
- d It means that data can be operated regardless of its origins.

1p **1c** Is FAIR data Open Data?

- a Sometimes it can be the case. But FAIR and Open Data must be kept separate
- b Yes, every FAIR data is Open Data
- c Yes, and the converse is also true: every Open Data is FAIR
- d FAIR data is never Open data

1p **1d** Can FAIR principles prevent the fabrication of data (i.e., deliberate creation of data to propound a particular hypothesis with greater conviction)?

- a Yes
- b No

Part 2: Finite Difference Methods

Consider the linear convection equation in 2D:

$$\frac{\partial u}{\partial t} + c_x \frac{\partial u}{\partial x} + c_y \frac{\partial u}{\partial y} = 0$$

where u is the unknown and c_x and c_y are the parameters, t is time, and x and y are spatial coordinates.

2p **2a** Using the finite difference method, this PDE can be rewritten in the following discretized mathematical formulation:

$$u_{i,j}^{n+1} = u_{i,j}^{n} - c_x \frac{\Delta t}{\Delta x} \left(u_{i,j}^{n} - u_{i,j-1}^{n}\right) - c_y \frac{\Delta t}{\Delta y} \left(u_{i,j}^{n} - u_{i-1,j}^{n}\right)$$

What kind of approximation of the **time** derivative is used in the equation above?

- (a) Forward difference
- (b) Central difference
- (c) Backward difference

5p **2b** The same equation can be discretized with different choices for the finite difference approximations. Below a vectorized implementation is given:

```
u = setInitialConditions()
for n in range(nt + 1):
    un = u.copy()
    u[1:-1, 1:-1] = un[1:-1, 1:-1]
                    - cx * dt / dx / 2 * (un[1:-1, 2:] - un[1:-1, :-2])
                    - cy * dt / dy / 2 * (un[2:, 1:-1] - un[:-2, 1:-1]))
```

Give a mathematical expression for the update of $u_{i,j}^{n+1}$ from relevant u values at time step n that is implemented in the code block above. The expected answer is an equation in a similar notation as the equation given in part *(a)* above.

3p **2c** If nothing is added to the code above, what kind of boundary conditions are applied by default? Motivate your answer.

Part 3: Finite Element Methods

Recall the weak form for the Poisson equation is given as:

$\int_\Omega \nu \nabla w \cdot \nabla u d\Omega$ - $\int_\Gamma w \nu \nabla u \cdot \mathbf{n} d\Gamma$ = $\int_\Omega w f d\Omega$

Assume a Robin boundary condition is applied along the boundary Γ

$\alpha u + \nu \nabla u \cdot \mathbf{n} = \beta$

Following the conventional notation, let the $\mathbf{N}$-matrix contain shape functions and the $\mathbf{B}$-matrix contain shape function derivatives. In the discretized form with the finite element method, terms with the coefficients ν, f, α and β appear. Select the correct form of the terms for the following:

5p **3a** With ν?

- (a) $\int_\Omega \mathbf{N}^T \nu d\Omega$
- (b) $\int_\Omega \mathbf{B}^T \nu d\Omega$
- (c) $\int_\Omega \mathbf{N}^T \nu \mathbf{N} d\Omega$
- (d) $\int_\Omega \mathbf{B}^T \nu \mathbf{B} d\Omega$

5p **3b** With α?

- (a) $\int_\Gamma \mathbf{N}^T \alpha d\Gamma$
- (b) $\int_\Gamma \mathbf{B}^T \alpha d\Gamma$
- (c) $\int_\Gamma \mathbf{N}^T \alpha \mathbf{N} d\Gamma$
- (d) $\int_\Gamma \mathbf{B}^T \alpha \mathbf{B} d\Gamma$

3p **3c** For a 4-node quadrilateral (2D) element, what is the size of the $\mathbf{B}$-matrix?

Part 4: Optimization 1

Three cities (C_1, C_2, C_3) are supplied with water from three different sources (S_1, S_2, S_3). The first (S_1) is a major reservoir, the other two are local sources (S_2, S_3). Sources S_2 can supply cities C_1 and C_3 and S_1 and S_3 can supply all cities. The cities have a consumption of a minimum of R_1, R_2, R_3 respectively. The local sources can only supply a maximum of Q_2 and Q_3 of water volume. The reservoir can supply a maximum of Q_1, but there is a minimum supply of Q_{min} to be imposed.

5p **4** Establish the model that allows obtaining the optimum solution for the problem of supplying the cities in the most economical way knowing that the cost of supplying city j from source i is given by c_{ij} expressed in monetary units (m.u.) per unit of water volume.

Part 5: Optimization 2

Consider the following table of the SIMPLEX method for solving an LP maximization problem:

5p

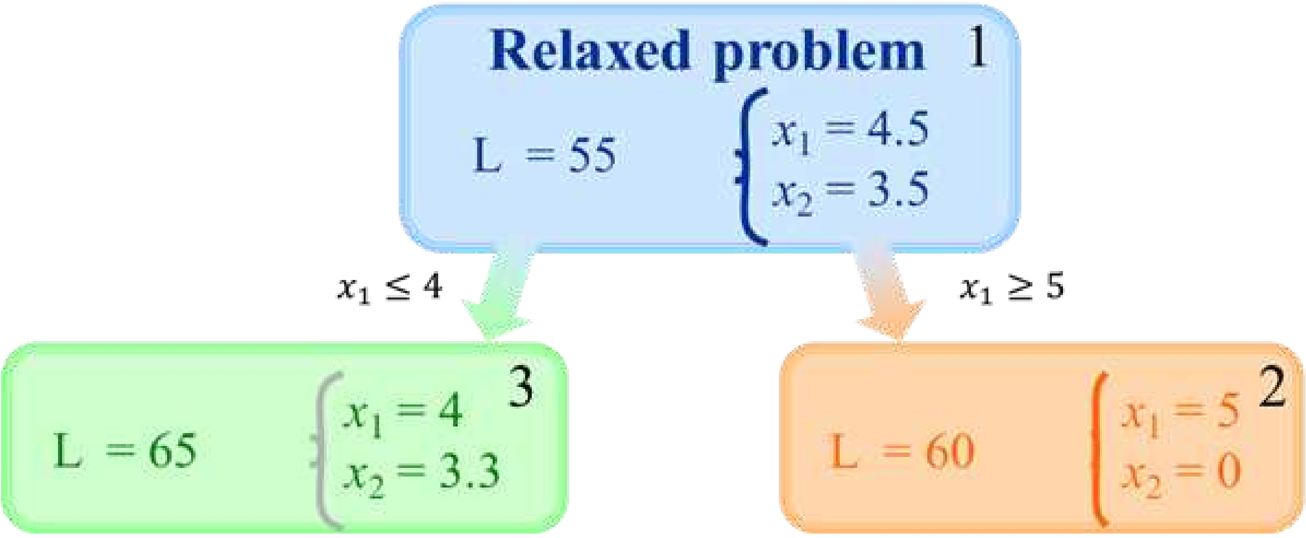
5

	Z	X1	X2	S1	S2	S3	b
Z	1	-1	-3	0	0	0	0
s1	0	0	1	1	0	0	5
s2	0	1	0	0	1	0	6
s3	0	1	3	0	0	1	21

Solve the problem.

Part 6: Optimization 3

In the next diagram you can see the solving process of the branch and bound for the minimization of an integer programming problem with two decision variables. The number in the upper right corner represents the solving order in the tree.



5p **6** Is the process finished? That is, are there more nodes to be explored? Why?

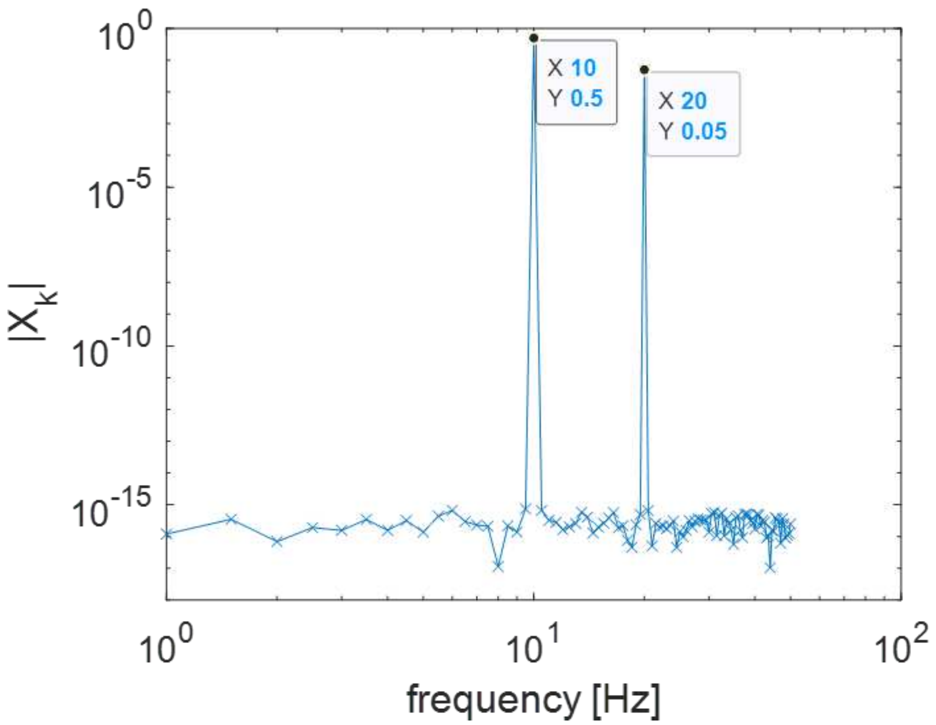
Part 7: Signal Processing
A continuous time signal $x(t)$ is given as:
$x(t) = A_1 \cos(2\pi f_1 t) + A_2 \cos(2\pi f_2 t)$
With $A_1 = 1$, $A_2 = 0.1$, $f_1 = 10$ Hz and $f_2 = 80$ Hz.

The signal has been sampled in three experiments, each time using a different sampling frequency f_s and a different measurement duration T_{meas}. The frequency domain plots (magnitude spectrum in logarithmic scale, as a result of the DFT) are shown below; the spectrum is double sided, but only shown for positive frequencies and, as commonly done in practice, up to the Nyquist frequency. The values of X_k, straight from the fft-implementation, have been divided by N, the number of samples.

Determine, for each experiment, the sampling frequency f_s, as well as the measurement duration T_{meas}. Only the final numerical answers are asked!

Some useful formulas: $f_{nyquist} = f_s/2 \qquad T_{meas} = 1/f_0$

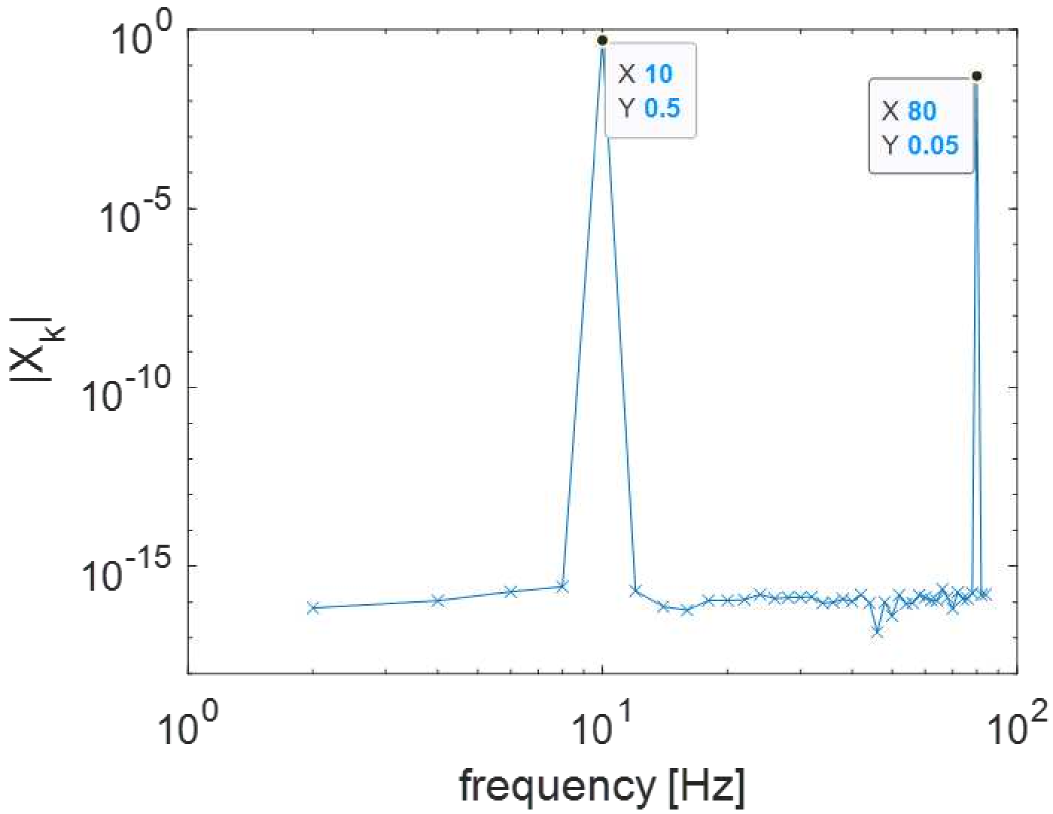
5p **7a Experiment A**



f_s =

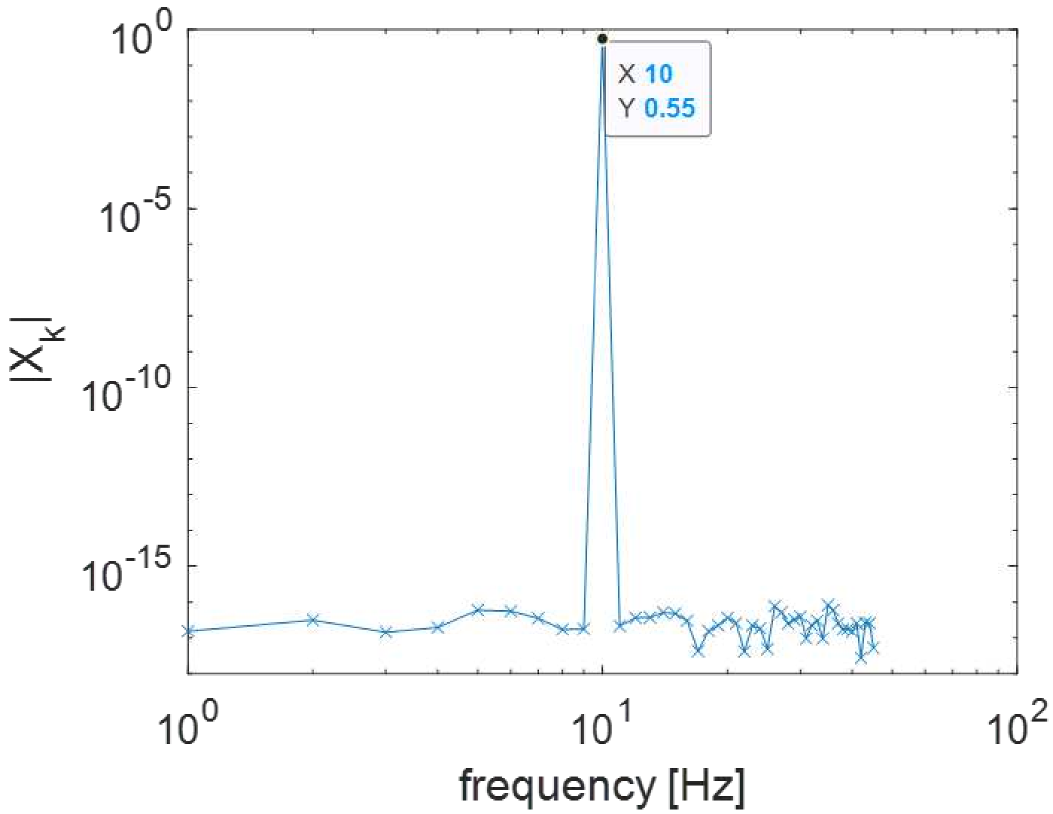
T_{meas} =

5p **7b Experiment B**



f_s = T_{meas} =

5p **7c Experiment C**



f_s =

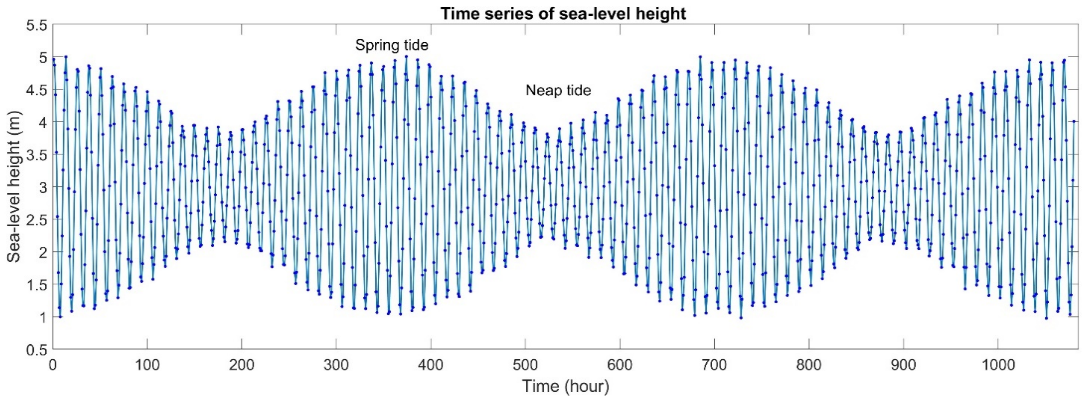
T_{meas} =

Part 8: Time series

A tide gauge station has been installed to measure the hourly sea-level variations relative to a vertical datum (reference system). These measurements are usually connected to a stable benchmark next to the tide gauge station (just to show variations with respect to a reference system). Therefore, there is a shift of approximately 3 m (the correct value should be determined) between the mean sea level (MSL) and the benchmark.

Sea level is subject to variations due to many variables like wind speed, pressure, and global warming (sea-level rise). One of such variations is caused by the forces induced by celestial bodies like the Moon and Sun (main contributors), called tide. The two major tidal constituents are the so-called M_2 (semi-lunar constituent) and S_2 (semi-solar constituent). Their periods are $T_{M_2} = 12.4206$ hour and $T_{S_2} = 12$ hour. We only use these two major constituents, and therefore ignore others.

A high-end tide gauge has been installed to measure the sea-level variations in the vertical direction. Up to now, we have collected 45 days of hourly data (so $m = 24 \cdot 45 = 1080$ observations). We assume that the measurements have been collected independently with the precision of $\sigma = 5$ cm (independent and normally distributed). The time series of the measurements $y = [y_1, \ldots, y_m]^T$ at time instances $t = [1, \ldots, m]^T$ (so t in hour) is as follows:



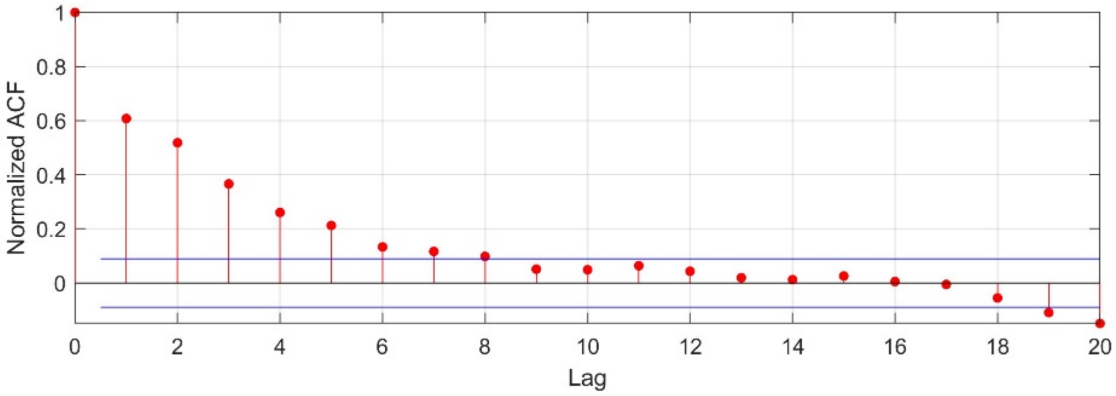
Based on the information provided above, and the fact that **sea-level rise can be neglected** here because it cannot be determined by this time series (45 days of data are too short and usually much longer time spans are required), we are interested in the functional and stochastic model $y = Ax + e$, $D(y) = Q_{yy}$.

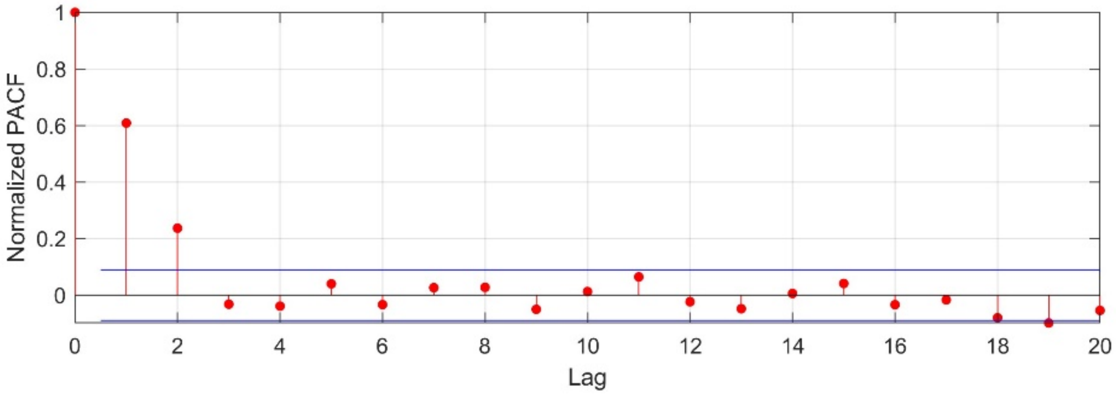
4p **8a** Specify the first and last row of the A-matrix and its dimensions. Also, specify Q_{yy}.

Assume that the two tidal constituents are not known a-priori, so we want to use spectral analysis technique to identify them. We want to compute the power spectral densities (PSD) using the least-squares harmonic estimation (LS-HE).

4p **8b** Sketch a plot of the expected PSD from the data set, where the horizontal axis is the frequency (cycle/hour). Add relevant numerical values on the horizontal axis.

Based on the observations in the past 45 days, we are now going to predict sea levels for the future. The functional part of the prediction comes from the settings of question a. For the stochastic part, we have computed the normalized Auto-Covariance Function (ACF) and partial ACF (PACF) of the least squares residuals $\hat{e}$. They are as follows:





We want to use the available data of the time series to predict the sea levels for the coming day (so 24 hours from $t = m + 1$ to $t = m + 24$). Two functional and stochastic parts can contribute to the prediction ($y_P = y_F + y_S$).

4p **8c** For the stochastic part, we need to specify the ARMA(p,q) process. How do you determine appropriate orders for the ARMA(p,q) process? So $p, q = ?$ What kinds of parameters β_i's or θ_i's of the ARMA process should we estimate?
Hint: For a zero-mean time series y_t, the ARMA process is as follows: $y_t = \sum_{i=1}^{p} \beta_i y_{t-i} + e_t + \sum_{i=1}^{q} \theta_i e_{t-i}$.

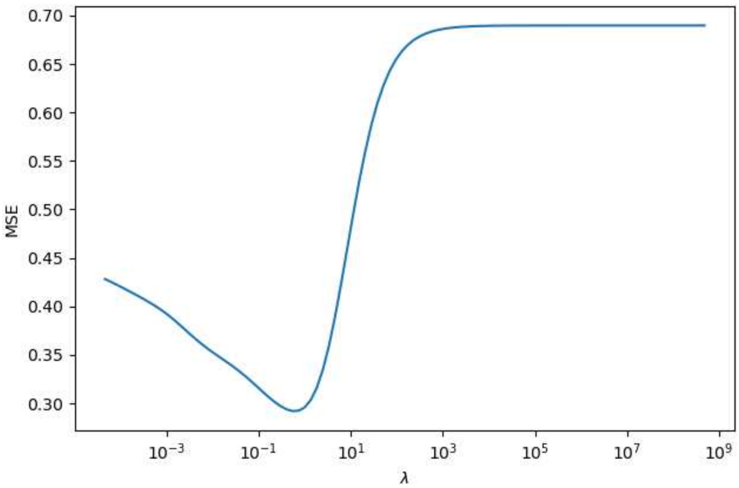
2p **8d** Based on the results for Question d, write an expression for y_S at the time instance $t = m + 1$, i.e. $y_S(m+1) = ?$

Part 9: Machine Learning

3p **9a** Assume you have a dataset with $N = 100$ data points and would like to train a linear basis function model with weights $\mathbf{w}$ which could potentially be too complex and overfit the data. You then decide to introduce an L_2 regularization term λ to the loss function and do a model selection study. Assume the number of basis functions is fixed and you cannot afford to collect more data.

- (a) You allocate 20 samples for training, 40 for validation and 40 for testing. You use the training loss to calibrate λ, the validation loss to calibrate $\mathbf{w}$ and the test set to assess the final model
- (b) You allocate all 100 samples for training and use those to obtain both λ and $\mathbf{w}$ at the same time
- (c) You allocate 70 samples for training, 20 for validation and 10 for testing. You use the validation loss to calibrate λ, the training loss to calibrate $\mathbf{w}$ and the test set to assess the final model
- (d) You allocate all 100 samples to the training set and use those to obtain $\mathbf{w}$. Then you move the samples for validation and use those to obtain λ. Finally, you move the samples to the test set and assess the final model.

4p



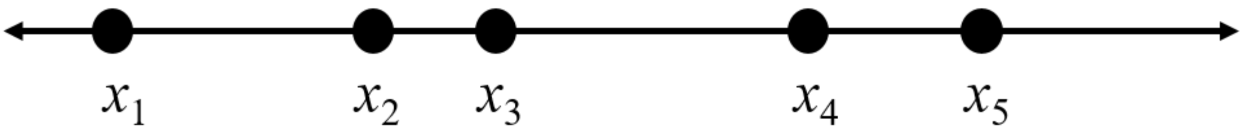
9b

A regularization term λ is added to the loss function of a neural network and a model selection study is performed by computing the mean squared error (MSE) over a validation dataset for different values of λ. The results of this study are shown above.

Regarding these results, mark **all** the options that are **TRUE**; consider that each wrong answer will result in negative points, but the lowest score for this sub-question is 0 (we will not subtract points from the rest of the exam):

- ☐ High values of λ lead to very rigid models
- ☐ Even without regularization, this specific model would already be resistant to overfitting
- ☐ The weights $\mathbf{w}$ of the neural net most likely increase as λ is decreased
- ☐ Increasing the size of the validation dataset (N) would make the "U"-shaped behavior of this curve less pronounced
- ☐ For $\lambda = 10^3$, training the model on a different dataset of the same size will lead to a very different model

2p



9c

Consider the dataset with five data samples $\{x_1, x_2, x_3, x_4, x_5\} = \{-1.6, -0.2, 0, 1.6, 2.2\}$ shown above.

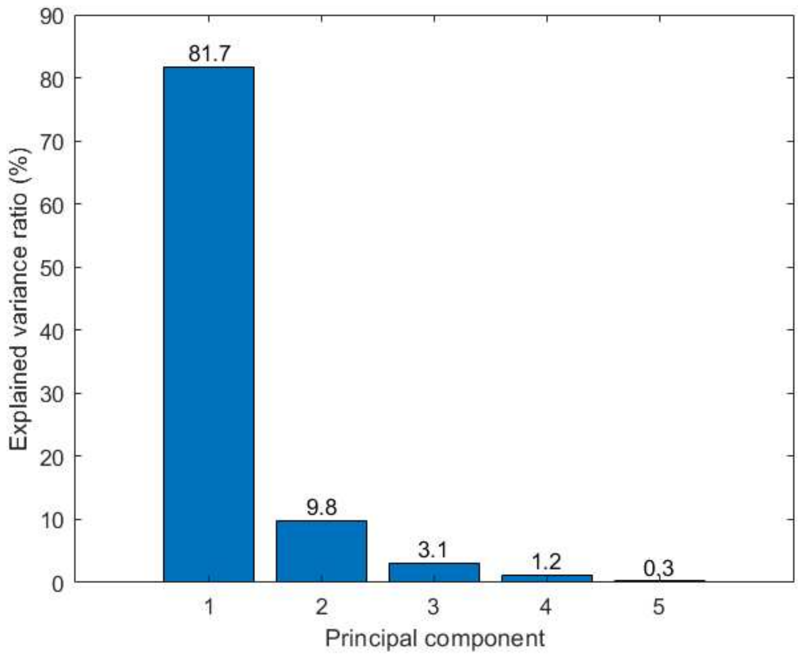
Using the Euclidean distance, whe perform K-means clustering to find the global optimal (minimum objective) when the cluster number $K = 3$. Which single data sample forms one cluster? (Euclidean distance between a and b: $d = \sqrt{(a-b)^2}$)

(a) x_1 (b) x_2 (c) x_3 (d) x_4 (e) x_5

3p **9d** Consider again the previous dataset. This time K-means clustering with Euclidean distance is used to find the global optimal (minimum objective) for $K = 2$. What are the centroids of the final clusters?

- (a) [-0.9, 1.9]
- (b) [-0.6, 1.9]
- (c) [-1.0, 2.0]
- (d) [-0.9, 1.3]
- (e) [-0.2, 1.6]

2p



9e

We perform *principal component analysis* on a given dataset. Consider the explained variance ratio with respect to the principal component number shown in the figure. What is the lowest dimension that guarantees a total explained variance ratio of 95%?

(a) 2 (b) 3 (c) 4 (d) 5

Part 10: Risk and Reliability

You are asked to evaluate the system reliability of a 2 m diameter oil pipeline that is currently operating in an earthquake region. The main objective is to evaluate the probability of failure, which in this case is defined as: the annual probability of a leak from the pipeline due to one of three different failure modes caused by an earthquake, M_i:

1. M_1: buckling of the pipe from longitudinal stress
2. M_2: high pressure failure (hoop stress)
3. M_3: failure of welded joint

Each failure mode is dependent on whether or not an earthquake occurs, which has an annual probability of occurrence of 10%. For simplicity, consider each failure mode to be mutually exclusive, and that damage can only occur once per year per failure scenario. In other words:

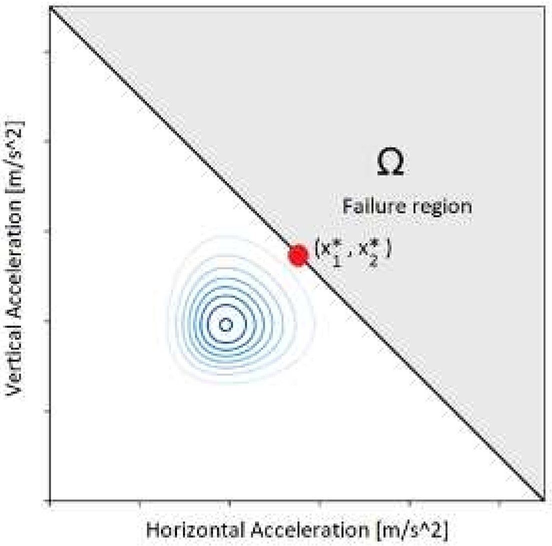
$$P(M_1 \cup M_2 \cup M_3) = P(M_1) + P(M_2) + P(M_3)$$

The probabilities of each failure mode have already been assessed and are summarized in the following table:

	$P(M_i \mid EQ)$	Damage (k€)
M_1	0.5	1000
M_2	0.3	50
M_3	0.4	10

5p **10a** Construct the FD curve (i.e., the FN curve, except with damage in place of fatalities on the x-axis) for leakage of the pipeline segment due to each of the 3 failure modes. Don't worry about the scale of your plot being precise, so long as the FD values are clearly indicated at each point.

3p **10b** Failure of one of the pipeline segments is a function of the horizontal and vertical acceleration, X_1 and X_2, respectively. The limit state can be described by a function, illustrated in the figure, where the failure region is represented by Ω. If $f_{X_1,X_2}(x_1,x_2)$ is the multivariate probability distribution of the random variables X_1 and X_2. Which of the following best defines the probability of failure:



- (a) $\int_\Omega [f_{X_1}(x_1) + f_{X_2}(x_2)]\, dX_1 f X_2$
- (b) $P(X_1 > x_1^* \cup X_2 > x_2^*)$
- (c) $\int_\Omega f_{X_1|X_2}(x_1|X_2 = x_2^*) f_{X_2}(X_2 = x_2^*) dX_1$
- (d) $\int_\Omega f_{X_1,X_2}(x_1,x_2) dX_1 dX_2$

As the pipeline is made up of many individual segments, you would like to perform a system reliability analysis to evaluate the probability of failure for the entire pipeline. Should you consider the pipeline to be a series or parallel system, and what will be the role of dependence between segments on the calculated failure probability for the entire pipeline? (use this information for the next 2 questions)

1p **10c** Should you consider the pipeline to be a series or parallel system?

(a) Series
(b) Parallel

2p **10d** What is the quantitative effect of positive dependence between segments on the calculated failure probability? (Choose only one)

(a) Increase failure probability
(b) Decrease failure probability
(c) No change (they are independent)
(d) No change (they are mutually exclusive)

The annual probability of one or more leaks is $P_1 = 0.2$ and is based on the current operating procedure of inspecting the pipeline once per year ($n = 1$). However, experience within the pipeline industry indicates the failure probability can be reduced with additional inspections, such that $P_n = P_1/n$. Environmental consequences of a leak have been estimated to be D=€100,000, and each inspection costs €1,000. Repair costs are negligible.

4p **10e** Find the optimal number of inspections per year, n, that minimizes total annual expected cost due to a pipeline leak.