Exercise assignment 2

Each exercise assignment has two parts. The first part consists of $3-5$ elementary problems for a maximum of 10 points from each assignment. For the second part consisting of problems in applications you are to choose *one* of three problems to solve. This part can give up to 5 points from each assignment. The first part consists of elementary questions to make sure that you have understood the basic material of the course while the second part consists of larger application examples.

Solutions can either be submitted through the Blackboard page or you can submit handwritten solutions in the envelope outside of room $U3-185$ before 23.59 on Sunday 21st of December.

Each exercise assignment can give a maximum of 15 points, to pass you will need at least 20 points total from both the assignments. If you do not get enough points from the assignments you will be given the opportunity to complement your solutions to reach a passing grade. These complements need to be submitted before 23.59 on the 11th of January.

Part 1

In the first part you are to solve and hand in solutions to the questions. You are allowed to use computer software to check your results, but your calculations as well as your result should be included in the answers for full points.

1.1

Consider the three column vectors

$$\mathbf{v}_1 = (0,0,3), \quad \mathbf{v}_2 = (0,4,1), \quad \mathbf{v}_3 = (3,1,1).$$

a) (1) Find an orthogonal basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ for $\mathbb{R}^3$ with $\mathbf{b}_1 = \mathbf{v}_1$ by using the Gram–Schmidt process on $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. Show your steps.

b) (1) Using the results of part (a), let Q be the matrix with column vectors $\mathbf{b}_j$ and A be the matrix with column vectors $\mathbf{v}_j$:

$$\mathsf{Q} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \mathbf{b}_3 \end{bmatrix},$$
$$\mathsf{A} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}.$$

Find a matrix R such that

$$\mathsf{A} = \mathsf{QR}.$$

1.2

Consider the data in the table below:

x	-2	0	2	4
y	-3	1	0	2

Table 1: xy-points of some data.

a) (1) Use the least squares method to approximate a line $y = ax + b$ to the data given in table 1 in least square sense.

b) (1) Try to fit a polynomial of second degree to the data in table 1. What result do you get?

1.3

a) (1) Give a definition of the *Kroenecker product*, $\otimes$.

b) (1) Calculate

$$\begin{bmatrix} 2 & 3 & 3 \\ -1 & 6 & 2 \end{bmatrix} \otimes \begin{bmatrix} 2 & -1 \\ -1 & 2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

1.4

a) (1) Give a definition of a *Linear transformation*.

b) (1) Consider the following transformations from $\mathbb{R}^3$ to $\mathbb{R}^2$, parameterized by some scalar c:

$$T_c(\mathbf{x}) = \begin{bmatrix} c - |(x_1, x_2, x_3)|^2 - 2(x_1, x_2, x_3) \cdot (x_2, x_3, x_1) + (x_1 + x_2 + x_3 + 1)^2 \\ 2x_1 + 2x_2 - 2cx_3 \end{bmatrix}.$$

Find all possible constants c such that T_c is a linear transformation and show that such transformations T_c are linear from the *definition* of a linear transformation.

1.5

Consider 7 observations out of two classes with 2 measured variables each:

$$\begin{bmatrix} 2 & 1 & 1 & 0 & 0 & -1 & -1 \\ 1 & 1 & 2 & 3 & 4 & 2 & 3 \\ c1 & c1 & c1 & c2 & c2 & c2 & c2 \end{bmatrix}$$

where each row represent one variable (last denoting their class belonging) and each column represent one observation. Our aim is to use Linear discriminant analysis in order to be able to classify new observations for which we do no know their class already.

a) (1) For each class: calculate corresponding covariance matrix S_1, S_2 (within class scatter).

b) (1) Use LDA to find the line w^* which when projected upon best separates the two classes. Plot the 7 points and the line w^* (by hand or using a computer does not matter).

Part 2

In this second part you are to choose **ONE** example where you attempt to solve the questions presented. If you hand in answers to more than one choice you will get points corresponding to the choice which would give the least total points.

You are allowed to use computer software (and depending on which option you choose might be needed for some of the questions). If you are using a computer for some calculations you should set up the problem and present how it could theoretically be solved by hand. For example you could write 'I solved the linear system $\mathbf{Ax} = \mathbf{b}$ using Matlabs `fsolve`, another method would have been to use Gaussian elimination and solving the resulting triangular system'.

Remark: There are two different QR-factorizations of a matrix. If we QR-factorize a $n \times r$ matrix A in the way you have discussed during the lecture the Q matrix is square $n \times n$ and the triangular matrix is non-square $n \times r$. An alternative QR-factorization is taking Q non-square $n \times r$ and R square $r \times r$. These two ways are closely related and if A is square they are exactly the same. If you have calculated the first version (using Gram-Schmidt or some software) you can get the second version by taking the first r rows of R to get the new R and the first r columns of Q to get the new Q.

Estimating pressure using LSM

Consider a power plant that generates power by burning trash and using the heat to boil water and then use the resulting steam to generate electricity using turbines. There are many different parameters that need to be tuned carefully to ensure that such a power plant works efficiently.

Two important parameters are the steam pressure, P, and temperature, T, in the boiler. Changing the temperature and pressure in the boiler takes time and is expensive so in order to develop a method for tuning the boiler a few strategic measurements are made and our goal is to develop a mathematical model using these measurements.

The measurements can be found in table 2.

a) One way to the pressure would be to assume that it increases linearly with the temperature, $P(T) = a + bT$. Find the parameters a and b using the least square method.

P (MPa)	6.4	6.7	7.5	7.8	8.8
T (°C)	10	22	45	60	75

Table 2: Table of pressures, P, at different temperatures, T.

Using the QR-factorization of $\mathsf{T} = \mathsf{QR}$, where T is the matrix that describes the regression model. we can rewrite the least square method equation as

$$\mathsf{Ra} = \mathsf{Q}\mathbf{p},$$

where $\mathbf{a}$ is a column of model coefficients and $\mathbf{p}$ is a vector of pressure values.

b) Use the Gram-Schmidt process to calculate the QR-factorization of T and solve $\mathsf{Ra} = \mathsf{Q}^\top\mathbf{p}$. **Note:** make sure to read the remark at the start of part 2.

It is important to consider the way that the boilers reliability and efficiency depends on temperature. The boiler is designed such that it will be most efficient between 170°C and 230°C. To take this into account we can weight the data points in that range 4 times as strongly as the other data points below the interval and twice as high as the data points above the interval.

c) Take the weights of the data points into account fit another single linear regression model to the data. You do not need to perform QR-factorization again.

When plotting the data we notice that for the weighted version many of the points are above the line. We therefore decide to try some other models. We will then see which model gives the closest fit.

d) Assume that the relation between the pressure and temperature is a 2nd degree polynomial $P(T) = a + bx + cx^2$. Calculate the sum of squares of the residuals $S = \mathbf{e}^\top\mathbf{e}$ where $e_i = P(T_i) - P_i$.

e) Assume that the relation between the pressure and temperature is a power law $P(T) = a \cdot e^{bT}$. Calculate the sum of squares of the residuals as in d) and compare the two sums and says which is smaller.

Regression of the cumulative distribution function

An insurance company are designing a life insurance policy for a group of people who are at a higher risk of getting a certain rare disease (that is fortunately very treatable) than the average person. Since the disease is rare there is not a lot of statistical material on the chance of becoming sick so the company wants to create an approximate cumulative distribution function (CDF) to help them design the policy.

In the table below the known data about the disease are given:

We start with some simple models to describe the CDF.

Diagnosed patients, S, %	0.02	0.14	0.21	0.26	0.31
Age, T, years	22	45	50	61	75
Average time at risk, t, years	3.0	7.1	10.2	18.6	31.5

Table 3: Table of diagnosed patients in percent, S, age in years, T, and average time at risk before being diagnosed in years, t.

a) Approximate the CDF that describes the diagnosed patients, S, as a function of age, T, with a function that consists of three linear pieces. The first piece is a constant 0 and the third part is a constant 1. In between these two part the CDF increases linearly $S(T) = a + bT$. Find a and b using the least square method.

b) Approximate the CDF that describes the diagnosed patients, S, as a function of age, T, and time at risk with a function that is described by

$$S(T,t) = \begin{cases} 0 & \text{when } R(T,t) \leq 0 \\ R(T,t) = a + bT + ct & \text{when } 0 < R(T,t) < 1 \\ 1 & \text{when } R(T,t) \geq 0 \end{cases}$$

Find the parameters a, b and c using the least square method.

Using the QR-factorization of $\mathsf{T} = \mathsf{QR}$, where T is the matrix that describes the regression model in a) we can rewrite the least square method equation as

$$\mathsf{R} \begin{bmatrix} a \\ b \end{bmatrix} = \mathsf{Q}\mathbf{s},$$

where $\mathbf{s}$ is a vector of diagnosed patients.

c) Use the Gram-Schmidt process to calculate the QR-factorization of T and solve

$$\mathsf{R} \begin{bmatrix} a \\ b \end{bmatrix} = \mathsf{Q}\mathbf{s}.$$

Note: make sure to read the remark at the start of part 2.

There seems to be some strong correlation between T and t in the data so we want to see if we can improve the method using principal component regression.

The principal components for T and t is given by the matrix $W = \begin{bmatrix} 0.2953 & -0.9554 \\ -0.9554 & -0.2953 \end{bmatrix}$.

d) Use the principal components to define two new variables $r(T,t)$ and $q(T,t)$ such that $\begin{bmatrix} r \\ q \end{bmatrix} = W \begin{bmatrix} T \\ t \end{bmatrix}$. Fit the model $S = ar + bq$ to the data.

Next we want to use a more sophisticated model.

e) Find a special case of the Weibull CFD that relates diagnosed patients and their age

$$S(T) = 1 - ae^{bT^3}$$

using LSM. You can assume that it is acceptable to rewrite this as a linear regression problem.

Classification of fuel sources using SIMCA

Can you identify the species of a tree by burning the wood and measuring the volume of different components in the smoke?

To try and answer this question a lab have gathered a couple of logs from two different types of wood, burned them one at a time and measured the amount emitted of various chemical compounds. This could for example be water vapour, carbon dioxide, carbon monoxid, sulfur, etc.

Six logs each of the two tree types T_1 and T_2 have been burned and the amount emitted out of 3 types of components have been measured. This gives 12 observations in total with 3 variables (and their tree type) given in the table below:

3.5	3.9	4.2	5	5.3	6	5.5	4.8	5.1	3.9	4.1	3.5
1.3	1.7	2.1	2	3	3.3	2.5	3	3.8	4	4.4	4.5
0.5	0.7	0.8	0.7	1.1	1.2	0.9	1.1	1.4	1.5	1.6	1.7
T_1	T_1	T_1	T_1	T_1	T_1	T_2	T_2	T_2	T_2	T_2	T_2

Our aim is to use the SIMCA method in order to find a model for both types of wood such that we might classify new samples.

a) (1) Calculate the covariance matrix for each class.

b) (1) Use the covariance matrices for find the principal components (PCA) of each class. Pick the one principal component that describes the most variance within the class.

c) (1) Use the chosen principal component of each class to calculate the within class variances $(s_{T_1})^2, (s_{T_1})^2$

In order to classify new samples we start by choosing 95% as our confidence level. Since we have $n = 6$ samples in each class, $p = 3$ variables at start and use $k = 1$ principal component in each class we need.

$$F_{((p-k),(p-k)(n-k-1))\ (0.95)} = F_{(2,8)\ (0.95)} = 4.46$$

(you do not need to show this)

Now that we have everything we need we can start trying to classify new observations with unknown tree type, for this two new logs where burned:

$$\begin{bmatrix} 4.5 & 5 \\ 4 & 3 \\ 1.5 & 1.1 \end{bmatrix}$$

d) (1) Use your previous results to classify the two new logs of unknown tree-type.

d) (1) Classify all your original samples and calculate the confusion matrix. Since there is a possibility that a sample is classified as none or both types, include both of these scenarios as well in your confusion matrix (so you end up with a 4×4 matrix).