IE30301-Datamining Assignment 3 (70 Points)

Eldor Fozilov

April 27, 2021

Exercise 1

Write a detailed description of the following concepts and their differences. (Explain at least 2 lines about each concepts and write differences in 1 sentence. If not, there are 3 points deduction per problem. [20 pts, 5 pts for each.]

1. **Likelihood** & **Probability**
   Probability, in simple words, can be defined as the chance that a given event will occur. It is calculated as the ratio of the number of outcomes in a sample space representing a given event of interest to the total number of possible outcomes in that sample space.
   The word **likelihood** is often interchangeably used used with the word **probability** in daily conversations. However, it has a different meaning in statistics and machine learning. In those areas, **likelihood** generally concerned with finding the best distribution of the data given a sample data set. It is represented as a function (called **likelihood function**), which calculates the joint probability distribution of given data as a function of model parameters. That function can be thought as the probability of obtaining the given data given the parameters.

2. **Feature Selection** & **Feature Extraction**
   Feature selection is referred to the selection of a subset of features out of the original features with the aim of simplifying a model, reducing computational cost and the influence of noise, caused by redundant features in the model. So, the result of successful feature selection can be an improvement in the model accuracy and speed.
   Feature extraction serves similar purposes mentioned above, but it tries to achieve it in a different way compared to **feature selection**: instead of choosing which variables to include in a model, feature extraction methods create new variables from all given feature variables by combining them in various ways so that those new variables retain most of the useful information. For example PCA creates new variables by taking a **linear** combination of given feature variables.

3. **PC Score** & **PC loading**
   We know the fact that **PCA** produces a low-dimensional representation of a data set by finding a sequence of linear combinations of the variables that retain maximum variance, and are uncorrelated with each other. The particular values in those derived variable vectors are called **PC scores**.
   To find those **PC scores**, we need to take a dot product between each normalized sample

record, which can be thought as being a vector, and a parameter vector. The values in that parameter vector are called **PC loadings**, which represent the contribution of each feature variable to the computation of **PC score**. Those **PC loadings** can also be interpreted as correlations between feature variables and the loading vector (principal axis) they belong to. So we can see that PC score and PC loading are closely related, but not the same thing.

4. **Newton-Raphson Method** & **Gradient Descent**
   Gradient descent is an iterative optimization algorithm for finding a local minimum of a differentiable function using its first-order derivative. It is grounded in the idea that if a derivative of a multivariate function f exists at a point $\mathbf{x}$, then the direction of fastest decrease/descent in the function value will happen when if one goes from the point $\mathbf{x}$ in the direction of the negative gradient of f at $\mathbf{x}$, $-\nabla f(\mathbf{x})$. That negative gradient can be multiplied by a some value to increase or decrease the step size (change in value of $\mathbf{x}$).

   Newton-Raphson method is an iterative method used for finding the roots of a twice differentiable function f. The main difference between those two methods is that gradient descent algorithm is parametric in nature (it usually requires a learning rate μ), while the other method does not need such a parameter, and thus we can apply it without worrying for changing any hyperparameter.

Exercise 2

In multiple linear regression, we can estimate $\hat{\boldsymbol{\beta}}$ as follows by least square method.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \in \mathbb{R}^{(p+1)\times 1}, \quad \mathbf{X} \in \mathbb{R}^{N\times(p+1)}, \quad \mathbf{y} \in \mathbb{R}^{N\times 1} \tag{2.1}$$

The regression model can be derived as follows:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y} \tag{2.2}$$

And in the above equation, $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is specifically referred to as $\mathbf{H}$, hat matrix.

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \tag{2.3}$$

2.1

Show that $\mathbf{H}$ is symmetric ($\mathbf{H}^{\mathrm{T}} = \mathbf{H}$) and idempotent ($\mathbf{H}^2 = \mathbf{H}$). [3 pts]

Answer:
First we will show that the matrix $\mathbf{H}$ is symmetric.

$$\mathbf{H}^{\mathrm{T}} = \left[\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\right]^{T}$$

We will use the following property of the transpose applied when multiplying two matrices:

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} \tag{2.4}$$

Let's label the matrix $(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}$ as M. Then we write $\mathbf{H}^\mathrm{T}$ as

$$\mathbf{H}^\mathrm{T} = [\mathbf{XM}]^T = \mathbf{M}^\mathrm{T}\mathbf{X}^\mathrm{T} = \left[(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\right]^\mathrm{T}\mathbf{X}^\mathrm{T} = \left[(\mathbf{X}^\mathrm{T})^\mathrm{T}\left((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\right)^\mathrm{T}\right]\mathbf{X}^\mathrm{T}$$

We know that $(\mathbf{X}^\mathrm{T})^\mathrm{T} = \mathbf{X}$. We can also show that $\left((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\right)^\mathrm{T} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$.

Let $\mathbf{A} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$. Then $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. If we take the transpose of both sides in those equations and apply (2.4), then we get the following:

$$\mathbf{A}^\mathrm{T}(\mathbf{A}^{-1})^\mathrm{T} = \mathbf{I}^\mathrm{T} = \mathbf{I}, \quad (\mathbf{A}^{-1})^\mathrm{T}\mathbf{A}^\mathrm{T} = \mathbf{I}^\mathrm{T} = \mathbf{I}$$

From the above equations, we can see that $(\mathbf{A}^\mathrm{T})^{-1} = (\mathbf{A}^{-1})^\mathrm{T}$.

So, $\left((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\right)^\mathrm{T} = \left((\mathbf{X}^\mathrm{T}\mathbf{X})^\mathrm{T}\right)^{-1} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$

Thus, $\mathbf{H}^\mathrm{T}$ can be expresses as

$$\mathbf{H}^\mathrm{T} = \left[(\mathbf{X}^\mathrm{T})^\mathrm{T}\left((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\right)^\mathrm{T}\right]\mathbf{X}^\mathrm{T} = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T} = \mathbf{H}$$

Now, we will show that the matrix $\mathbf{H}$ is idempotent.

$$\begin{gathered}\mathbf{H}^2 = \mathbf{HH} = \left(\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\right)\left(\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\right) = \\ = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}(\mathbf{X}^\mathrm{T}\mathbf{X})(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T} = \mathbf{XI}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T} = \\ = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T} = \mathbf{H}\end{gathered}$$

2.2

An estimator of a given parameter is said to be unbiased if its expected value is equal to the true value of the parameter. ($\mathrm{E}[\hat{\theta}] = \theta$)

Show that $MSE = \hat{\sigma}^2 = \dfrac{SSE}{N-p-1} = \dfrac{(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}})}{N-p-1}$ **is unbiased estimator** through using following properties (2.5) - (2.9) & results of above problem **2.1** [10 pts]

If c is a scalar, and $\mathbf{A}$ is a n × n square matrix, ($c \in \mathbb{R}, \mathbf{A} \in \mathbb{R}^{n\times n}$)

$$c = \mathrm{Tr}(c), \quad \mathrm{Tr}(c\mathbf{A}) = c\mathrm{Tr}(\mathbf{A}) \tag{2.5}$$

If **A, B** are n × n square matrix that have same dimension, ($\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$)

$$\mathrm{Tr}(\mathbf{A}+\mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}), \quad \mathrm{Tr}(\mathbf{A}-\mathbf{B}) = \mathrm{Tr}(\mathbf{A}) - \mathrm{Tr}(\mathbf{B}) \tag{2.6}$$

If $\mathbf{A}$ is a n × n square matrix, $\mathbf{A} \in \mathbb{R}^{n\times n}$

$$\mathrm{E}[\mathrm{Tr}(\mathbf{A})] = \mathrm{Tr}(\mathrm{E}[\mathbf{A}]) \tag{2.7}$$

If $\mathbf{A}$ is a n × m matrix and $\mathbf{B}$ is a m × n matrix ($\mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{B} \in \mathbb{R}^{m\times n}$)

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}) \tag{2.8}$$

If $\mathbf{x}$ is random vector,

$$\mathrm{Var}[\mathbf{x}] = \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}]^T, \quad \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathrm{Var}[\mathbf{x}] + \mathrm{E}[\mathbf{x}]\mathrm{E}[\mathbf{x}]^T \tag{2.9}$$

Answer:
Finding the expectation of **SSE** will give us clue to the expectation of **MSE** since **MSE** is a just a SSE divided by $N - p - 1$. So we will tackle it first.

$$\mathrm{E}[SSE] = \mathrm{E}[(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})] = \mathrm{E}[(\mathbf{y} - \mathbf{H}\mathbf{y})^T(\mathbf{y} - \mathbf{H}\mathbf{y})], \quad \text{where } \mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

$$\mathrm{E}[(\mathbf{y} - \mathbf{H}\mathbf{y})^T(\mathbf{y} - \mathbf{H}\mathbf{y})] = \mathrm{E}[((\mathbf{I} - \mathbf{H})\mathbf{y})^{\mathrm{T}}((\mathbf{I} - \mathbf{H})\mathbf{y})] =$$

$$= \mathrm{E}[\mathrm{Tr}\big(((\mathbf{I} - \mathbf{H})\mathbf{y})^{\mathrm{T}}((\mathbf{I} - \mathbf{H})\mathbf{y})\big)] \quad (\because \text{property (2.5)})$$

$$\mathrm{E}[\mathrm{Tr}\big(((\mathbf{I} - \mathbf{H})\mathbf{y})^{\mathrm{T}}((\mathbf{I} - \mathbf{H})\mathbf{y})\big)] = \mathrm{E}[\mathrm{Tr}\big(\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{y})\big)] =$$

$$= \mathrm{E}[\mathrm{Tr}\big(\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\mathbf{y}\big)] \quad (\because \mathbf{H} \text{ and } \mathbf{I} \text{ are symmetric and thus } (\mathbf{I} - \mathbf{H}) \text{ is also symmetric})$$

$$\mathrm{E}[\mathrm{Tr}\big(\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\mathbf{y}\big)] = \mathrm{E}[\mathrm{Tr}\big(\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})^2\mathbf{y}\big)] =$$

$$= \mathrm{E}[\mathrm{Tr}\big(\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{y}\big)] \quad (\because \mathbf{I} \text{ and } \mathbf{H} \text{ are idempotent and thus } (\mathbf{I} - \mathbf{H}) \text{ is idempotent})$$

$$\mathrm{E}[\mathrm{Tr}\big(\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{y}\big)] = \mathrm{E}[\mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathbf{y}\mathbf{y}^{\mathrm{T}}\big)] \quad (\because \text{property (2.8)})$$

$$\mathrm{E}[\mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathbf{y}\mathbf{y}^{\mathrm{T}}\big)] = \mathrm{Tr}\big(\mathrm{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}\mathbf{y}^{\mathrm{T}}]\big) \quad (\because \text{property (2.7)})$$

$$\mathrm{Tr}\big(\mathrm{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}\mathbf{y}^{\mathrm{T}}]\big) = \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}]\big) \quad (\because \mathbf{X} \text{ is considered to be given, thus } \mathbf{H} \text{ is a constant})$$

$$\mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}]\big) = \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\big(\mathrm{Var}[\mathbf{y}] + \mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big)\big) \quad (\because \text{property (2.9)})$$

$$\mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\big(\mathrm{Var}[\mathbf{y}] + \mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big)\big) = \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\big(\sigma^2\mathbf{I} + \mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big)\big) =$$

$$= \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I} + (\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) = \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}\big) + \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) \quad (\because \text{property (2.6)})$$

$$\mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}\big) + \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) = \mathrm{Tr}\big(\sigma^2\mathbf{I} - \sigma^2\mathbf{H}\big) + \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) =$$

$$= \mathrm{Tr}\big(\sigma^2\mathbf{I}\big) - \mathrm{Tr}\big(\sigma^2\mathbf{H}\big) + \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) \quad (\because \text{property (2.6)})$$

$$\mathrm{Tr}\big(\sigma^2\mathbf{I}\big) - \mathrm{Tr}\big(\sigma^2\mathbf{H}\big) + \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) =$$

$$= \sigma^2\mathrm{Tr}\big(\mathbf{I}\big) - \sigma^2\mathrm{Tr}\big(\mathbf{H}\big) + \mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) \quad (\because \text{property (2.5)})$$

Since $\mathbf{I} \in \mathbb{R}^{N\times 1}$, $\mathrm{Tr}(\mathbf{I}) = N$. And $\mathrm{Tr}(\mathbf{H}) = \mathrm{Tr}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}) = \mathrm{Tr}((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X}))$ because of property (2.8). So, $\mathrm{Tr}(\mathbf{H}) = \mathrm{Tr}(\mathbf{I}) = \mathrm{p} + 1$ ($\because I \in \mathbb{R}^{(p+1)\times(p+1)}$)

$$\mathrm{Tr}\big((\mathbf{I} - \mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) = \mathrm{Tr}\big((\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})(\mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta})^T\big) =$$

$$= \mathrm{Tr}\big(\big(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\big)(\mathbf{X}\boldsymbol{\beta})^T\big) = \mathrm{Tr}\big(\big(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\mathbf{I}\boldsymbol{\beta}\big)(\mathbf{X}\boldsymbol{\beta})^T\big) =$$

$$= \mathrm{Tr}\big(\big(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}\big)(\mathbf{X}\boldsymbol{\beta})^T\big) = \mathrm{Tr}\big(\mathbf{0}(\mathbf{X}\boldsymbol{\beta})^T\big), \quad \text{where } \mathbf{0} \text{ is a zero vector} \in \mathbb{R}^{N\times 1}$$

$$\mathrm{Tr}\big(\mathbf{0}(\mathbf{X}\boldsymbol{\beta})^T\big) = \mathrm{Tr}\big(\mathbf{0}\big), \quad \mathbf{0} \text{ now represents a zero matrix} \in \mathbb{R}^{N\times N}$$

$$\mathrm{Tr}\big(\mathbf{0}\big) = 0 \quad (\text{a constant})$$

Therefore,

$$\mathrm{E}[SSE] = \sigma^2\mathrm{Tr}\big(\mathbf{I}\big) - \sigma^2\mathrm{Tr}\big(\mathbf{H}\big) + \mathrm{Tr}\big((\mathbf{I}-\mathbf{H})\mathrm{E}[\mathbf{y}]\mathrm{E}[\mathbf{y}]^T\big) =$$
$$= \sigma^2 N - \sigma^2(p+1) + 0 = \sigma^2(N-p-1)$$

So, $\mathrm{E}[MSE] = \mathrm{E}\left[\frac{SSE}{N-p-1}\right] = \frac{1}{N-p-1}\mathrm{E}[SSE] = \frac{1}{N-p-1}\sigma^2(N-p-1) = \sigma^2$

Exercise 3

For matrix $\mathbf{A}$, solve the following problems

$$\mathbf{A} = \begin{pmatrix} 6 & -3 \\ 5 & -2 \end{pmatrix}$$

3.1

Compute the eigenvalues λ_1, λ_2 ($\lambda_1 < \lambda_2$) and its corresponding eigenvectors v_1, v_2 of matrix $\mathbf{A}$ [2 pts]

Answer:
Let's first compute eigenvalues.

$$\det\big(\mathbf{A} - \lambda\mathbf{I}\big) = \det\Big(\begin{bmatrix} 6-\lambda & -3 \\ 5 & -2-\lambda \end{bmatrix}\Big) = (6-\lambda)(-2-\lambda) - (-3)\times 5 =$$
$$= \lambda^2 - 4\lambda + 3 = 0 \Longrightarrow \boxed{\lambda_1 = 3, \lambda_2 = 1}$$

Now we will find eigenvectors.

$$\mathbf{A}v_1 = \lambda_1 v_1 \quad \Longrightarrow \quad \mathbf{A}v_1 - \lambda_1 v_1 = \overline{0}, \quad \overline{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$(\mathbf{A} - \lambda_1\mathbf{I})v_1 = \overline{0}$$
$$\begin{pmatrix} 6-\lambda_1 & -3 \\ 5 & -2-\lambda_1 \end{pmatrix} v_1 = \begin{pmatrix} 6-3 & -3 \\ 5 & -2-3 \end{pmatrix} v_1 = \begin{pmatrix} 3 & -3 \\ 5 & -5 \end{pmatrix} v_1 = \overline{0}$$

We will perform Gauss-Jordan elimination (but we will not follow the exact algorithm since the matrix is too simple) to find the vector $v_1 = \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix}$ and later the vector $v_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$:

$$\begin{pmatrix} 3 & -3 \\ 5 & -5 \end{pmatrix} v_1 = \overline{0} \Longrightarrow \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} v_1 = \overline{0} \Longrightarrow \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} v_1 = \overline{0} \Longrightarrow v_{11} = v_{12}$$

We can freely choose a value for either v_{11} or v_{12}. Let's choose a value of 1 for v_{11}, which will mean that v_{12} will be equal to 1. So, $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Now, we will find v_2.

$$\mathbf{A}v_2 = \lambda_2 v_2 \quad \Longrightarrow \quad \mathbf{A}v_2 - \lambda_2 v_2 = \bar{0}, \quad \bar{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$(\mathbf{A} - \lambda_2 \mathbf{I})v_2 = \bar{0}$$

$$\begin{pmatrix} 6-\lambda_2 & -3 \\ 5 & -2-\lambda_2 \end{pmatrix} v_2 = \begin{pmatrix} 6-1 & -3 \\ 5 & -2-1 \end{pmatrix} v_2 = \begin{pmatrix} 5 & -3 \\ 5 & -3 \end{pmatrix} v_2 = \bar{0}$$

$$\begin{pmatrix} 5 & -3 \\ 5 & -3 \end{pmatrix} v_2 = \bar{0} \Longrightarrow \begin{pmatrix} 1 & -0.6 \\ 1 & -0.6 \end{pmatrix} v_2 = \bar{0} \Longrightarrow \begin{pmatrix} 1 & -0.6 \\ 0 & 0 \end{pmatrix} v_2 = \bar{0} \Longrightarrow v_{21} = 0.6 v_{22}$$

We can freely choose a value for either v_{21} or v_{22}. Let's choose a value of 1 for v_{22}, which will mean that v_{21} will be equal to 0.6. So, $v_2 = \begin{pmatrix} 0.6 \\ 1 \end{pmatrix}$

3.2

Find matrix $\mathbf{P}$ to diagonalize $\mathbf{A}$. Here $\mathbf{D}$ is a diagonal matrix of size 2×2 [3 pts]

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$$

Answer:

Let $\mathbf{V} = [v_1 \; v_2] = \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix}$, and $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$.

Then the following equality holds:

$$\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda} \tag{3.1}$$

Since the eigenvectors v_1 and v_2 that we found in problem 3.1 are independent, $\mathbf{V}^{-1}$ exists, and thus we can multiply the both sides of the equation (3.1) by $\mathbf{V}^{-1}$ from the left side:

$$\begin{aligned} &\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda} \Longrightarrow \\ &\Longrightarrow \mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{I}\mathbf{\Lambda} = \mathbf{\Lambda} \end{aligned} \tag{3.2}$$

Thus, matrix $\boxed{\mathbf{P} = \mathbf{V}}$

3.3

Compute the determinant of $\mathbf{A}^{2521}$. The calculation should be trivial if you use the properties of determinant. [3 pts]

Answer:

We are aware of the following property of the determinant when we have two square matrices $\mathbf{A}$ and $\mathbf{B}$, both $\in \mathbb{R}^{n \times n}$:

$$\det(\mathbf{A}\mathbf{B}) = \det(\mathbf{A})\det(\mathbf{B}) \tag{3.3}$$

The $\mathbf{B}$ matrix can be $\mathbf{A}$ also, so we can write the property (3.3) as

$$\begin{aligned} \det(\mathbf{A}^2) &= \det(\mathbf{A})\det(\mathbf{A}) \Longrightarrow \\ \det(\mathbf{A}^3) &= \det(\mathbf{A}^2)\det(\mathbf{A}) = \det(\mathbf{A})\det(\mathbf{A})\det(\mathbf{A}) \end{aligned}$$

So, we can see the general trend as the power of $\mathbf{A}$ grows, and thus we can write the following:

$$\det\left(\mathbf{A}^{2521}\right) = [\det(\mathbf{A})]^{2521}$$

Thus, if we can find the determinant of the matrix $\mathbf{A}$ alone, we can find the determinant of $\mathbf{A}^{2521}$ easily.

$$\det(\mathbf{A}) = \det\left(\begin{bmatrix} 6 & -3 \\ 5 & -2 \end{bmatrix}\right) = 6 \times (-2) - (-3) \times 5 = 3 \Longrightarrow$$

$$\boxed{\det\left(\mathbf{A}^{2521}\right) = 3^{2521}}$$

Exercise 4

Given data $\mathbf{X}$, solve following problems

$$\mathbf{X} = \begin{pmatrix} 1 & 3 & 9 \\ 2 & 5 & 7 \\ 4 & 4 & 6 \\ 9 & 8 & 2 \end{pmatrix}$$

i.e., data with four samples and three features (predictors).

4.1

Find the mean value of each column. [1 pts]

Answer:

$$\mu_1 = \frac{1 + 2 + 4 + 9}{4} = 4$$
$$\mu_2 = \frac{3 + 5 + 4 + 8}{4} = 5$$
$$\mu_3 = \frac{9 + 7 + 6 + 2}{4} = 6$$

4.2

Subtract each mean from each element of the corresponding column.
Now, let us set the derived matrix as $\mathbf{X}'$. [1 pts]

Answer:

$$\mathbf{X}' = \begin{pmatrix} 1-4 & 3-5 & 9-6 \\ 2-4 & 5-5 & 7-6 \\ 4-4 & 4-5 & 6-6 \\ 9-4 & 8-5 & 2-6 \end{pmatrix} = \begin{pmatrix} -3 & -2 & 3 \\ -2 & 0 & 1 \\ 0 & -1 & 0 \\ 5 & 3 & -4 \end{pmatrix}$$

4.3

Calculate $\frac{1}{4-1}\mathbf{X'}^T\mathbf{X'}$. [1 pts]

Answer:

$$\frac{1}{4-1}\mathbf{X'}^T\mathbf{X'} = \frac{1}{3}\begin{pmatrix}-3 & -2 & 3\\ -2 & 0 & 1\\ 0 & -1 & 0\\ 5 & 3 & -4\end{pmatrix}^T\begin{pmatrix}-3 & -2 & 3\\ -2 & 0 & 1\\ 0 & -1 & 0\\ 5 & 3 & -4\end{pmatrix} =$$

$$= \frac{1}{3}\begin{pmatrix}-3 & -2 & 0 & 5\\ -2 & 0 & -1 & 3\\ 3 & 1 & 0 & -4\end{pmatrix}\begin{pmatrix}-3 & -2 & 3\\ -2 & 0 & 1\\ 0 & -1 & 0\\ 5 & 3 & -4\end{pmatrix} = \frac{1}{3}\begin{pmatrix}38 & 21 & -31\\ 21 & 14 & -18\\ -31 & -18 & 26\end{pmatrix} = \begin{pmatrix}\frac{38}{3} & 7 & \frac{-31}{3}\\ 7 & \frac{14}{3} & -6\\ \frac{-31}{3} & -6 & \frac{26}{3}\end{pmatrix}$$

4.4

For the calculated matrix in **4.3**, find eigenvalues $\lambda_1, \lambda_2, \lambda_3$ in descending order ($\lambda_1 > \lambda_2 > \lambda_3$) up to four decimal places. [2 pts]
(https://www.symbolab.com/solver/matrix-eigenvalues-calculator/eigenvalues)

Answer:

$\lambda_1 = 25.3084$
$\lambda_2 = 0.6044$
$\lambda_3 = 0.0872$

4.5

Calculate $\frac{\lambda_1}{(\lambda_1+\lambda_2+\lambda_3)}$. (up to four decimal places) [2 pts]

Answer:

$$\frac{\lambda_1}{(\lambda_1+\lambda_2+\lambda_3)} = \frac{25.3084}{25.3084+0.6044+0.0872} = 0.9734$$

4.6

Find eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ corresponding to $\lambda_1, \lambda_2, \lambda_3$. [2 pts]
(https://www.symbolab.com/solver/matrix-eigenvalues-calculator/eigenvectors)

Answer:

$$\mathbf{a}_1 = \begin{pmatrix}-1.2045\\ -0.6992\\ 1\end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix}138.5863\\ -237.3327\\ 1\end{pmatrix}, \quad \mathbf{a}_3 = \begin{pmatrix}0.6182\\ 0.3652\\ 1\end{pmatrix}$$

Exercise 5

Consider three random variables X, Y, Z. The three variables have the covariance matrix in the form of:

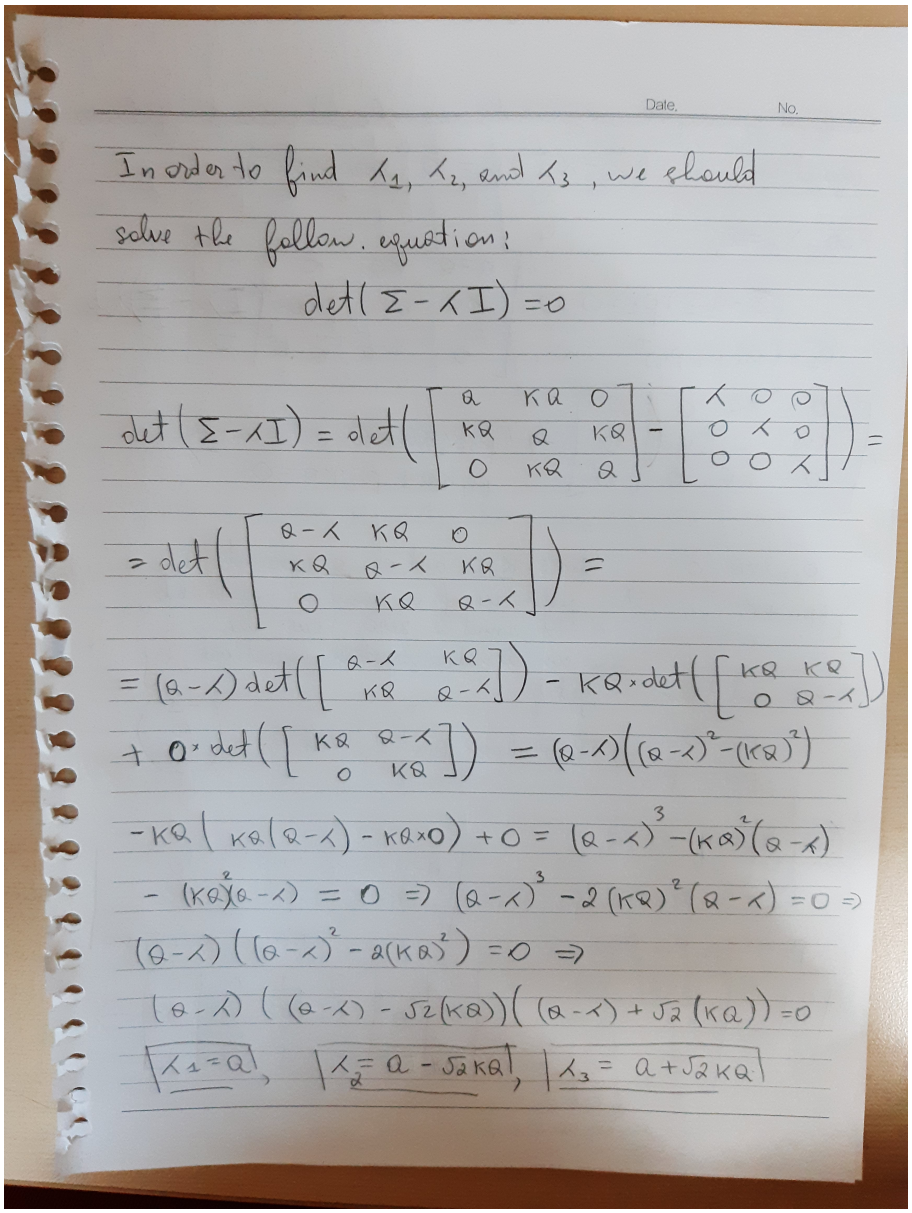
$$\Sigma = \begin{pmatrix} a & ka & 0 \\ ka & a & ka \\ 0 & ka & a \end{pmatrix}$$

, where $0 < k < \frac{1}{\sqrt{2}}$

5.1

Calculate the eigenvalues $\lambda_1, \lambda_2, \lambda_3$. (Show process of calculation neatly) [4 pts]

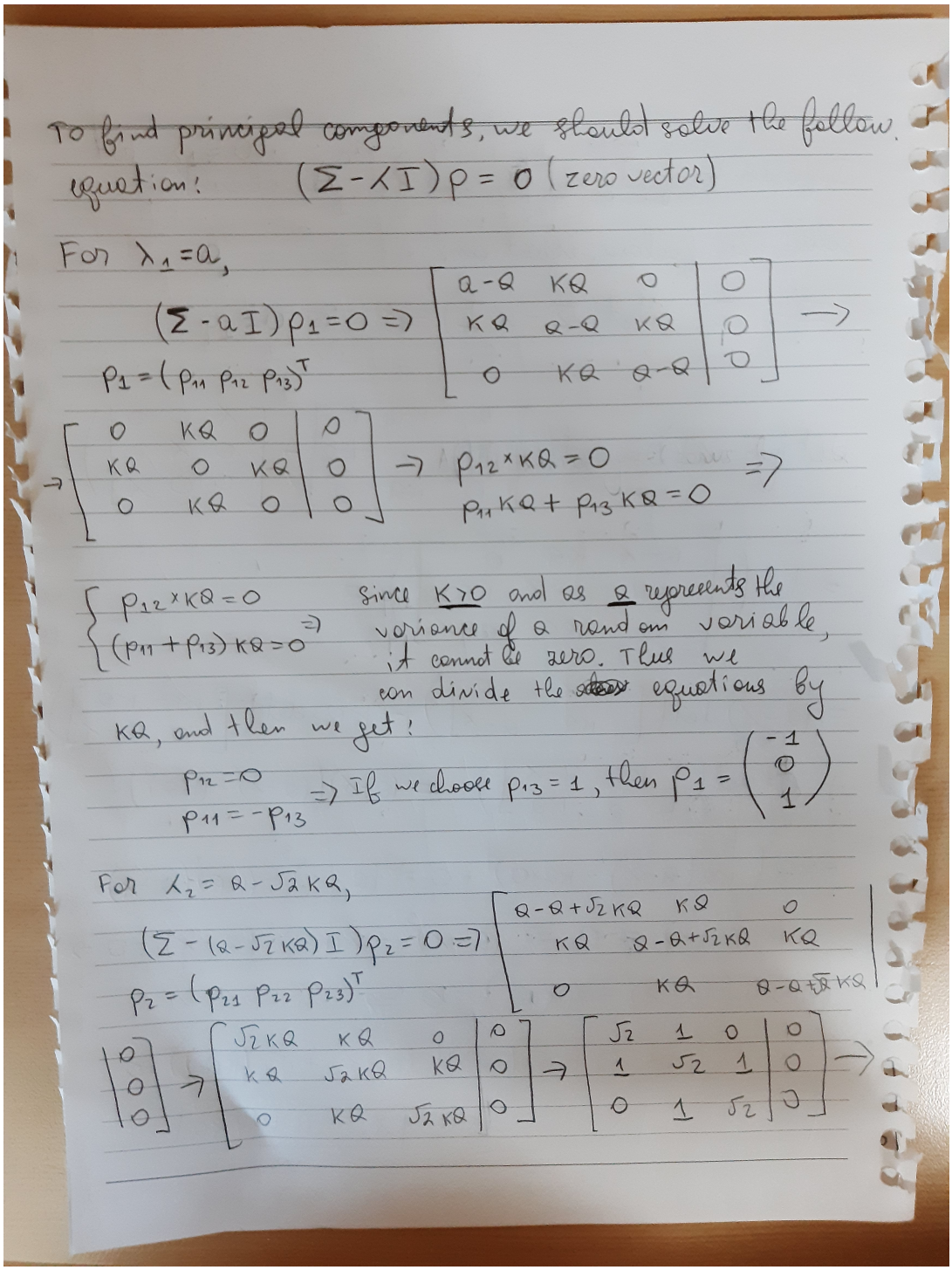
Answer:

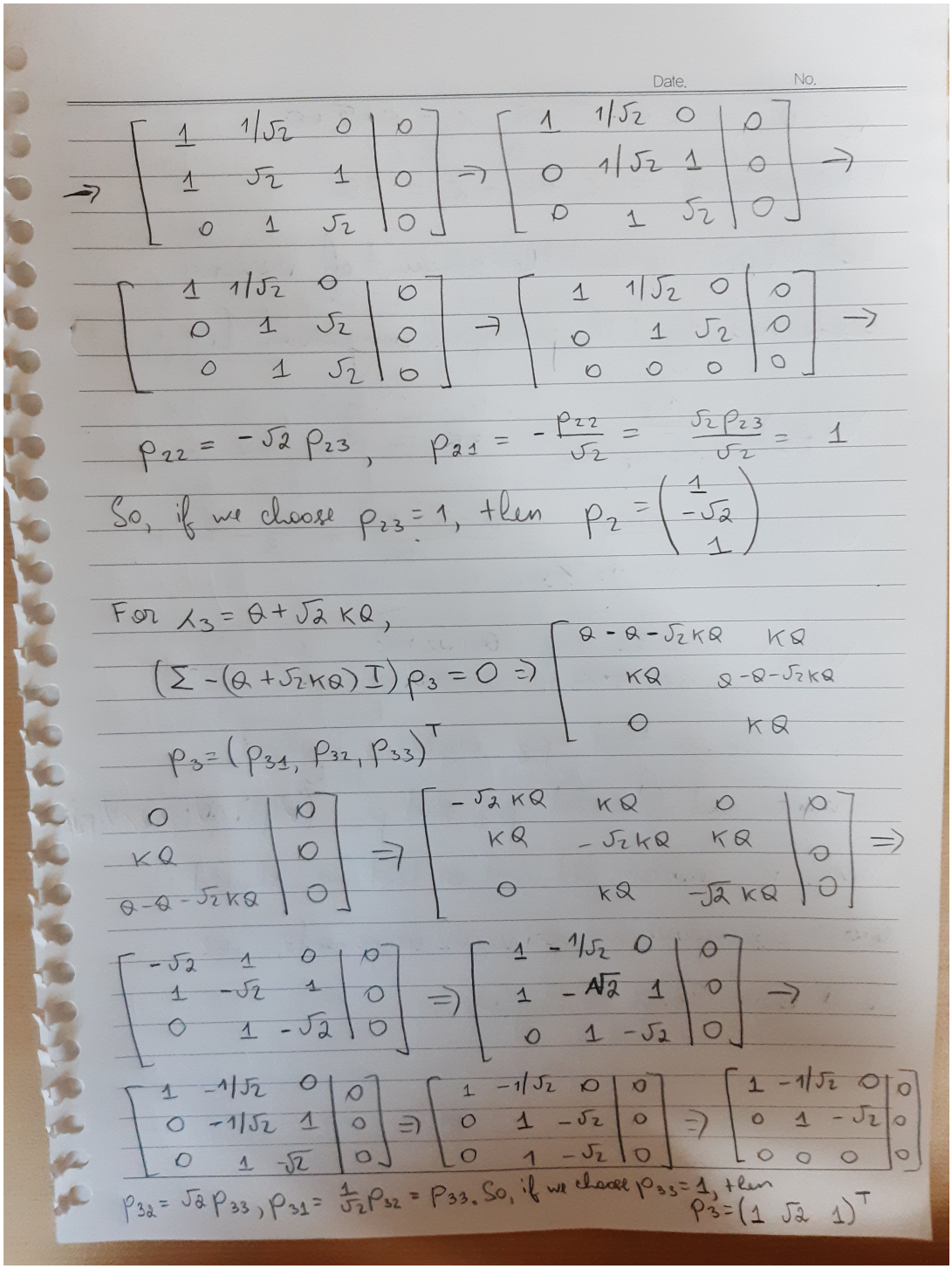


5.2

Find PC(Principal Component) $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ of each random variable X, Y, Z. [3 pts]

Answer:

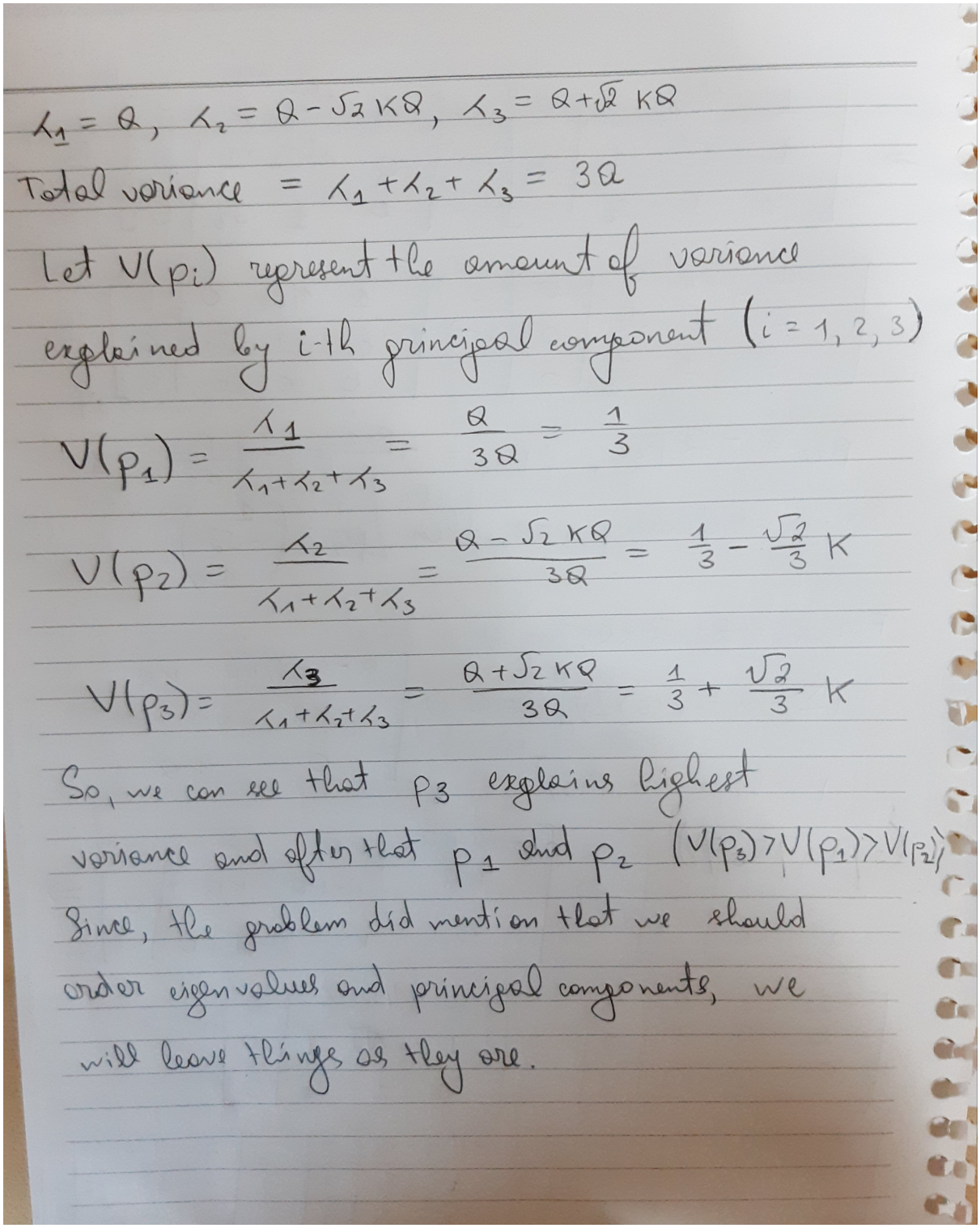




5.3

Calculate how much total variance is explained by each principal component. [3 pts]

Answer:



Exercise 6

The following table is the outcome of the logistic regression model for an iris flower being species Versicolor versus species Virginica. ($Y = 1$ for Versicolor, $Y = 0$ for Virginica) [9 pts]

Variables	Intercept	Length of Sepals	Width of Sepals	Length of petals	Width of Petals
Estimated Coefficient	25.21	-3	-0.7	2.4	-10.3

6.1

Interpret the effect of length of sepals on the relative risks of an iris flower being species Versicolor versus species Virginica. Fill up the **A,B,C** in the below interpretation of the outcome and select the appropriate word for **D**. (Tips from TA : Think about the meaning of odds(X)!) [4 pts]

Answer:

Interpretation: Species **Virginica** is $\frac{1}{e^{-3}}$ = 20 times more probable than Species **Versicolor** when the length of sepals D **increases** by 1 unit.

6.2

What is the predicted class for the following new data $\mathbf{x}_1$ and $\mathbf{x}_2$? You should provide the probability of $P(Y_1 = 1|\mathbf{x}_1), P(Y_1 = 0|\mathbf{x}_1), P(Y_2 = 1|\mathbf{x}_2), P(Y_2 = 0|\mathbf{x}_2)$. [6 pts]

$\mathbf{x}_1$	6.4	3.1	4.3	1.3
$\mathbf{x}_2$	6.9	3.0	3.9	1.4

Answer:

$$P(Y_1 = 1|\mathbf{x}_1) = \frac{1}{1 + e^{-(25.21+(-3)\times 6.4+(-0.7)\times 3.1+2.4\times 4.3+(-10.3)\times 1.3)}} = \frac{1}{1 + e^{-0.77}} = 0.6835$$

$$P(Y_1 = 0|\mathbf{x}_1) = 1 - P(Y_1 = 1|\mathbf{x}_1) = 1 - 0.6835 = 0.3165$$

So, the predicted class for $\mathbf{x}_1$ will be **Versicolor**.

$$P(Y_2 = 1|\mathbf{x}_2) = \frac{1}{1 + e^{-(25.21+(-3)\times 6.9+(-0.7)\times 3.0+2.4\times 3.9+(-10.3)\times 1.4)}} = \frac{1}{1 + e^{2.65}} = 0.066$$

$$P(Y_2 = 0|\mathbf{x}_2) = 1 - P(Y_2 = 1|\mathbf{x}_2) = 1 - 0.066 = 0.934$$

So, the predicted class for $\mathbf{x}_2$ will be **Virginica**.