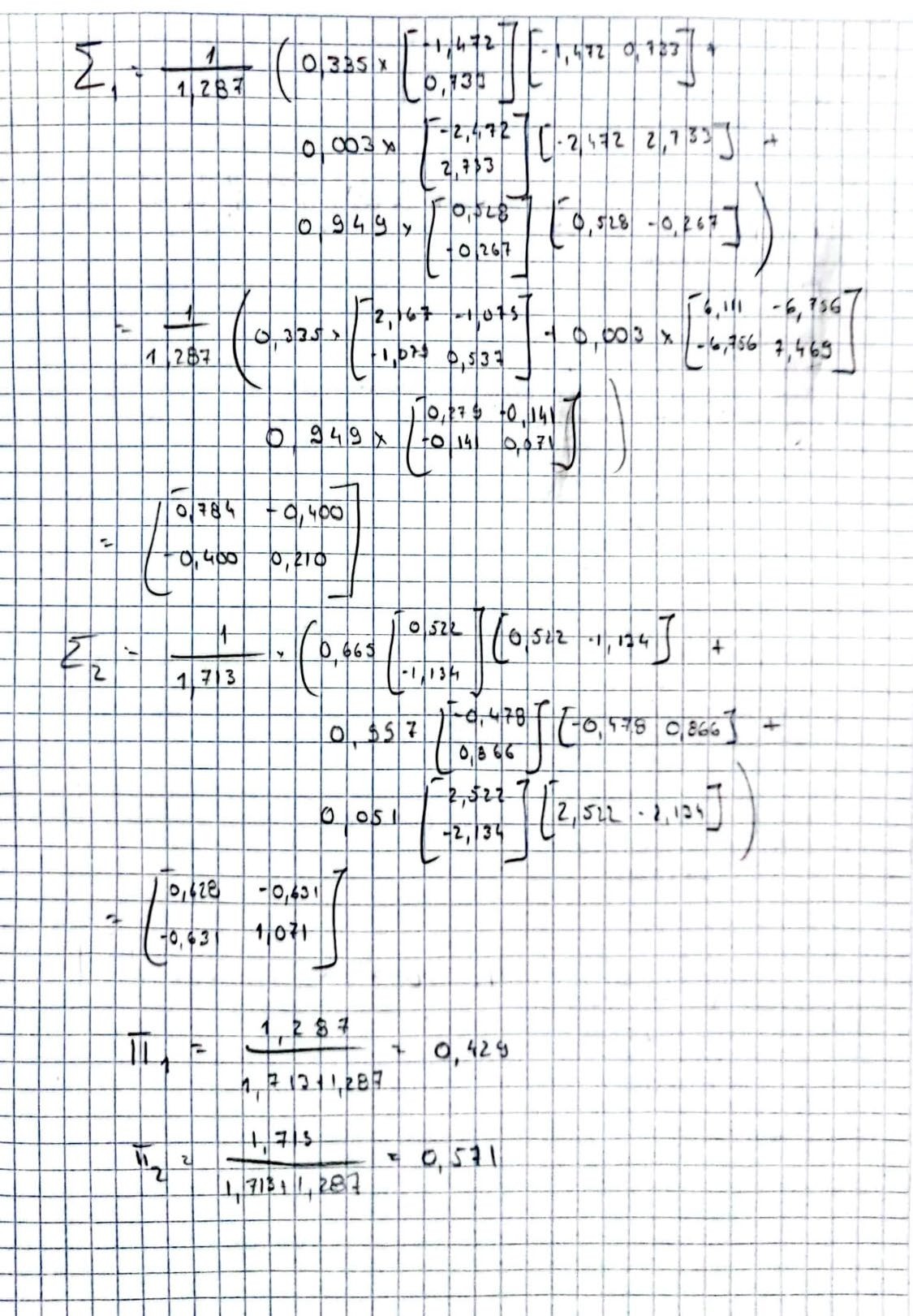
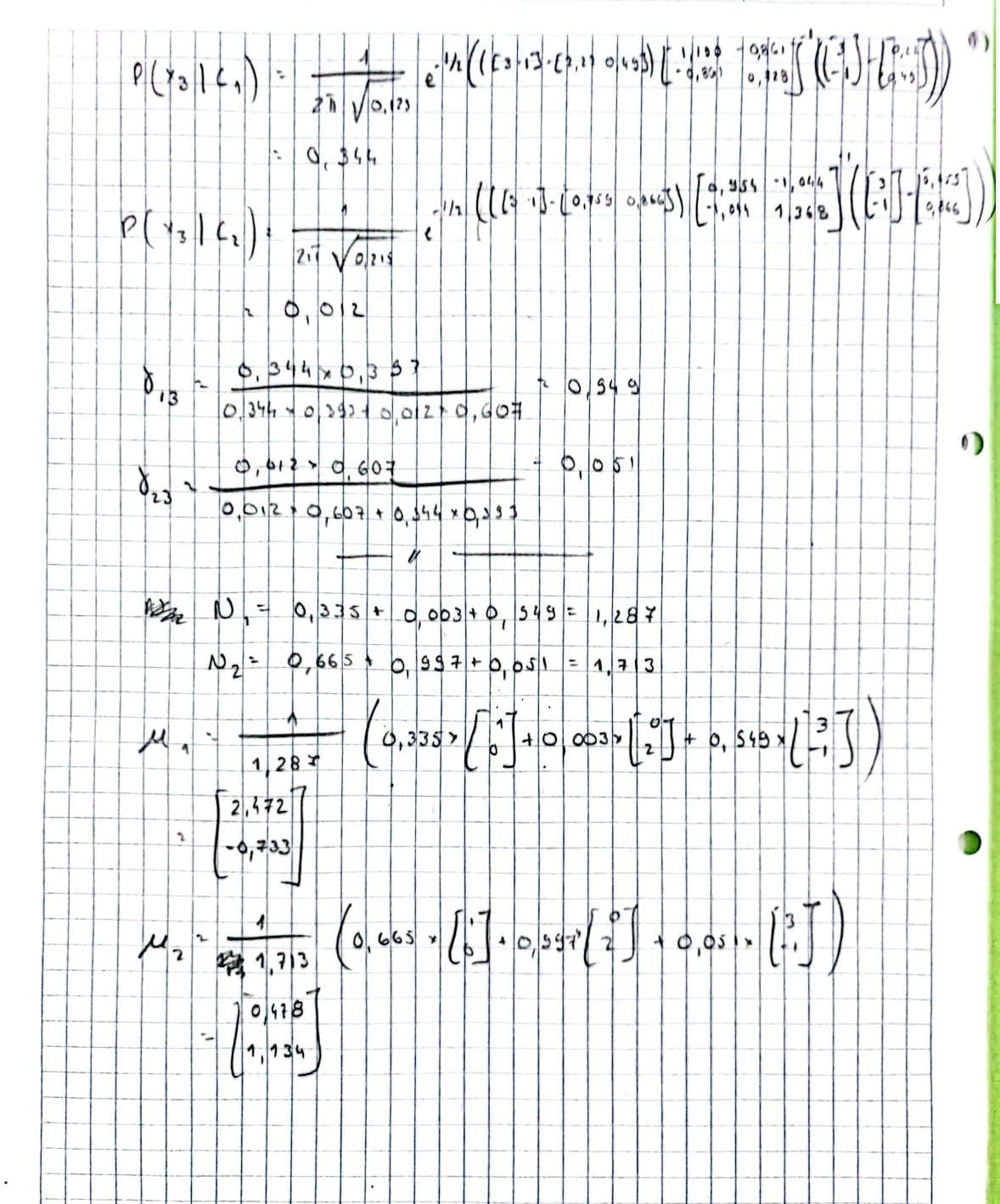
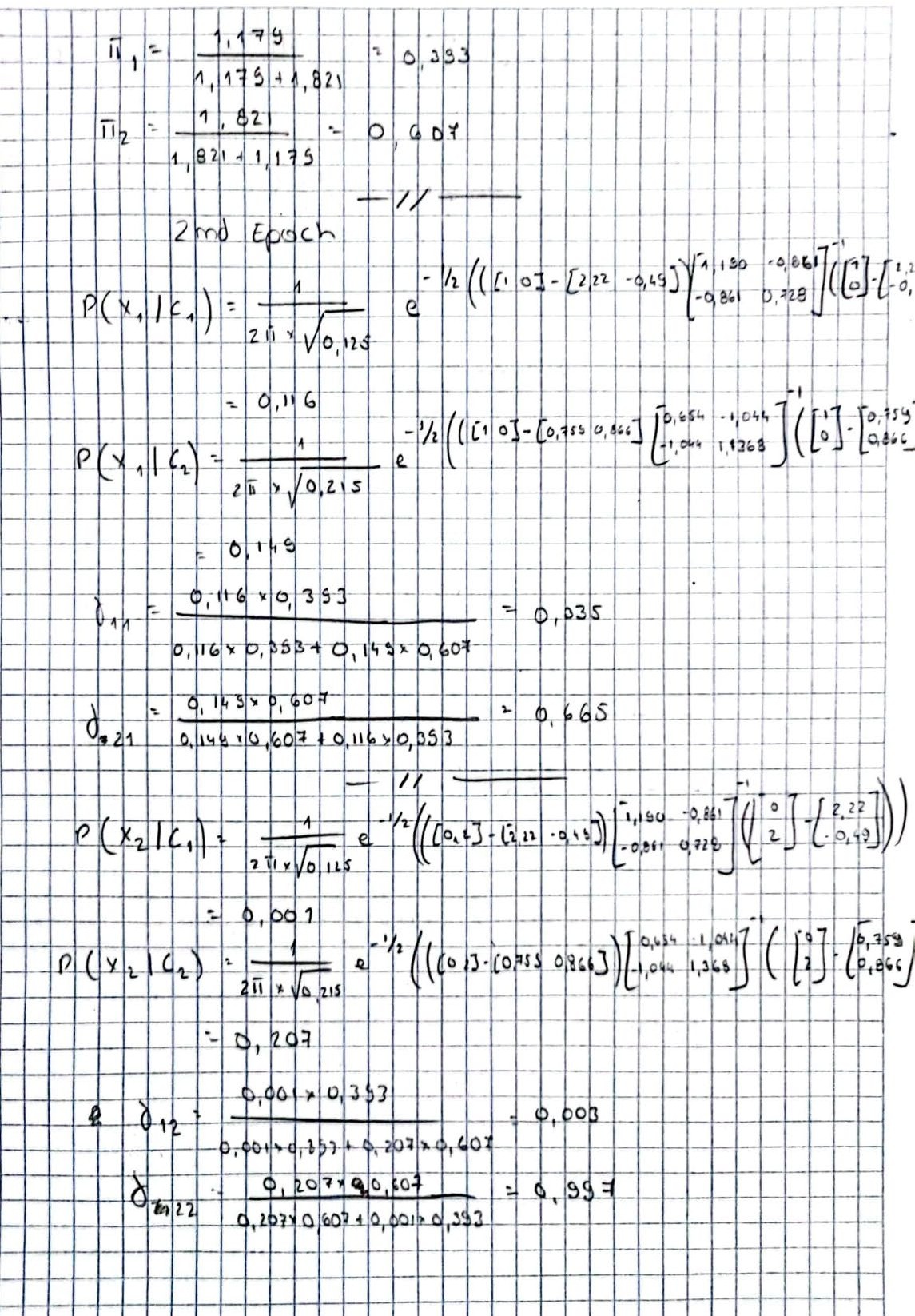
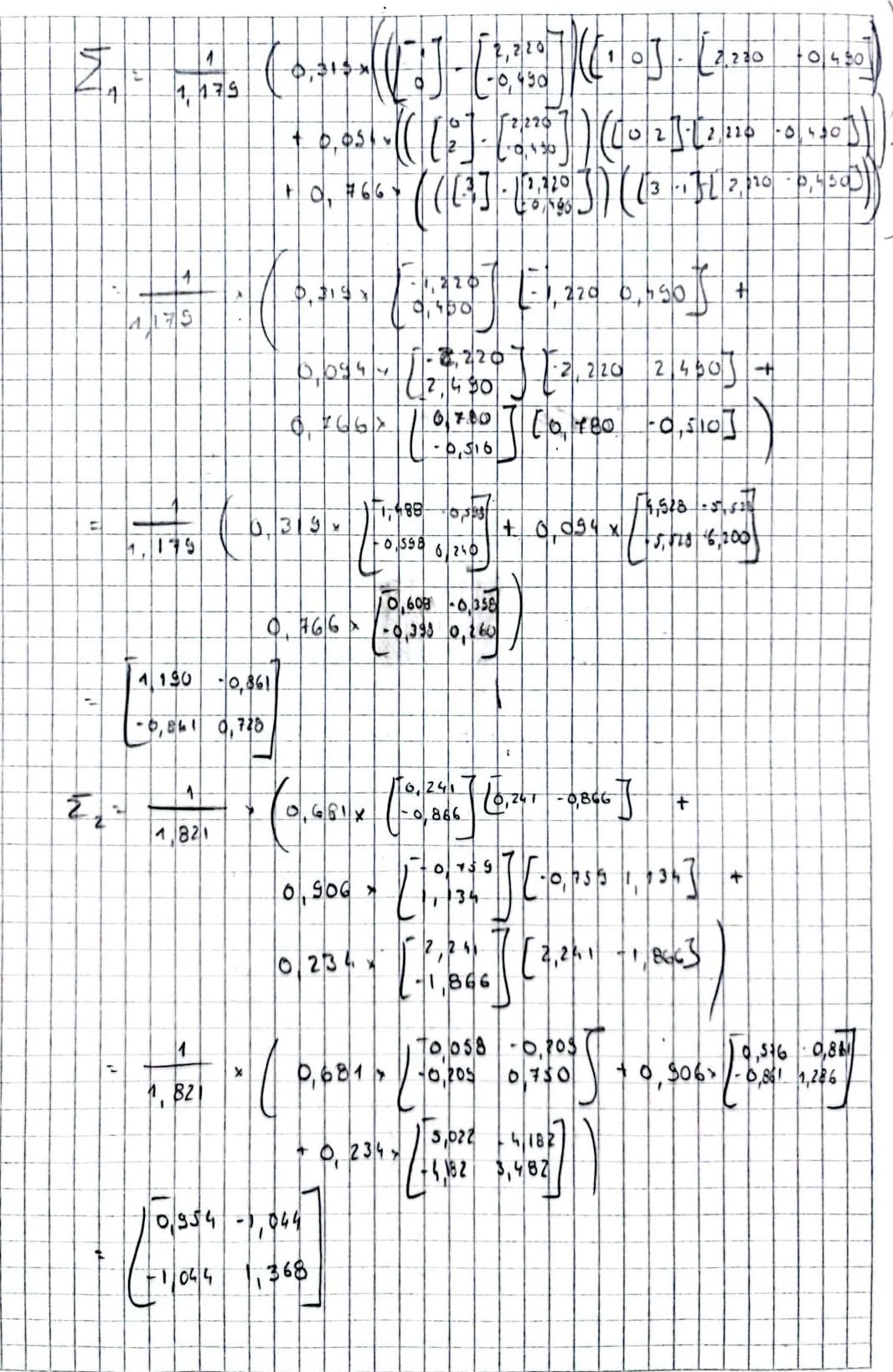
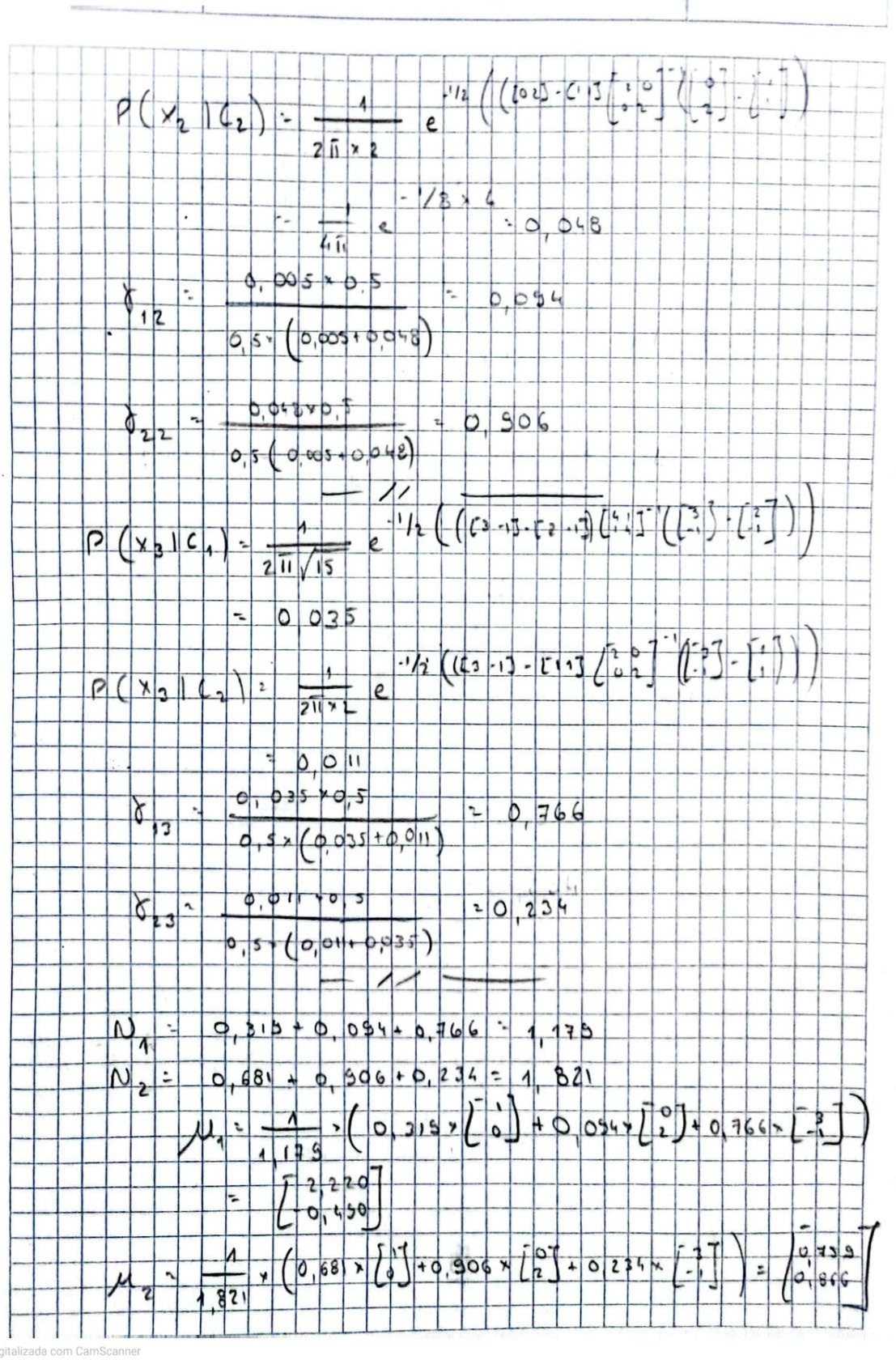
# Pen-and-paper

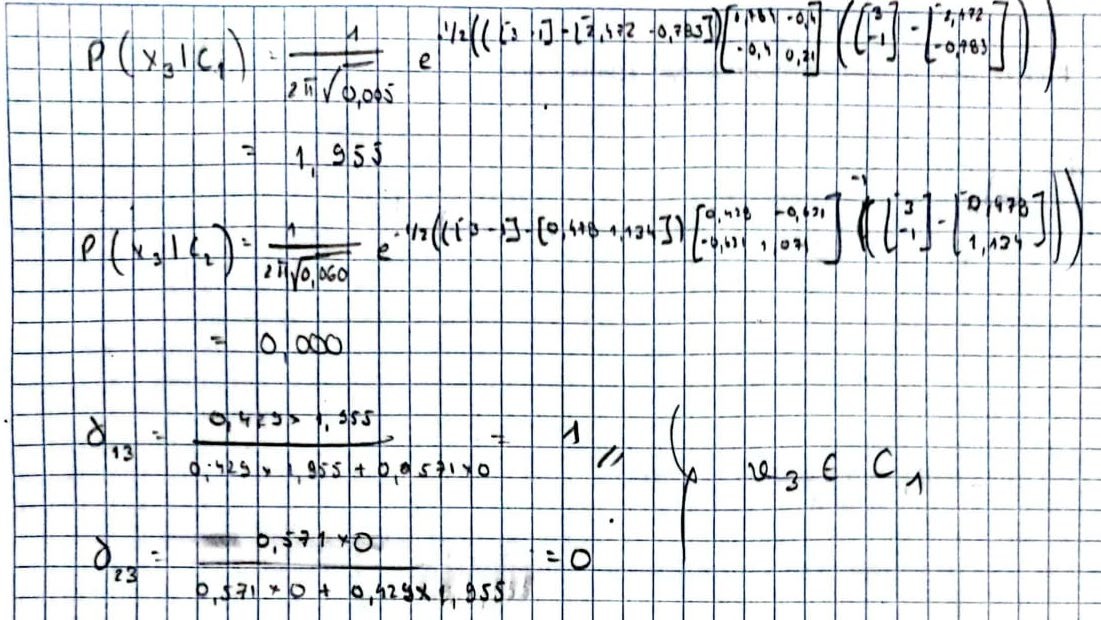
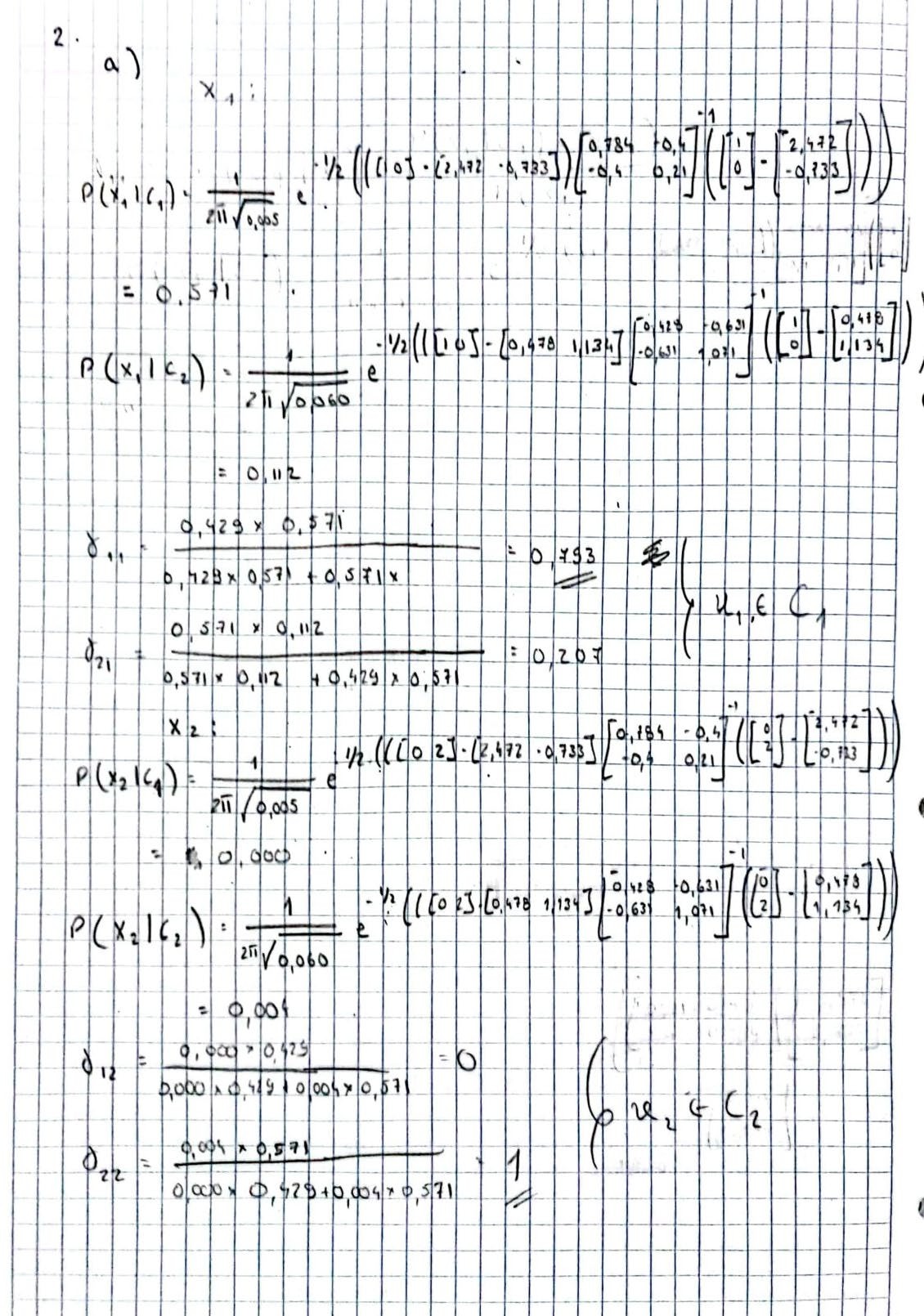
1)



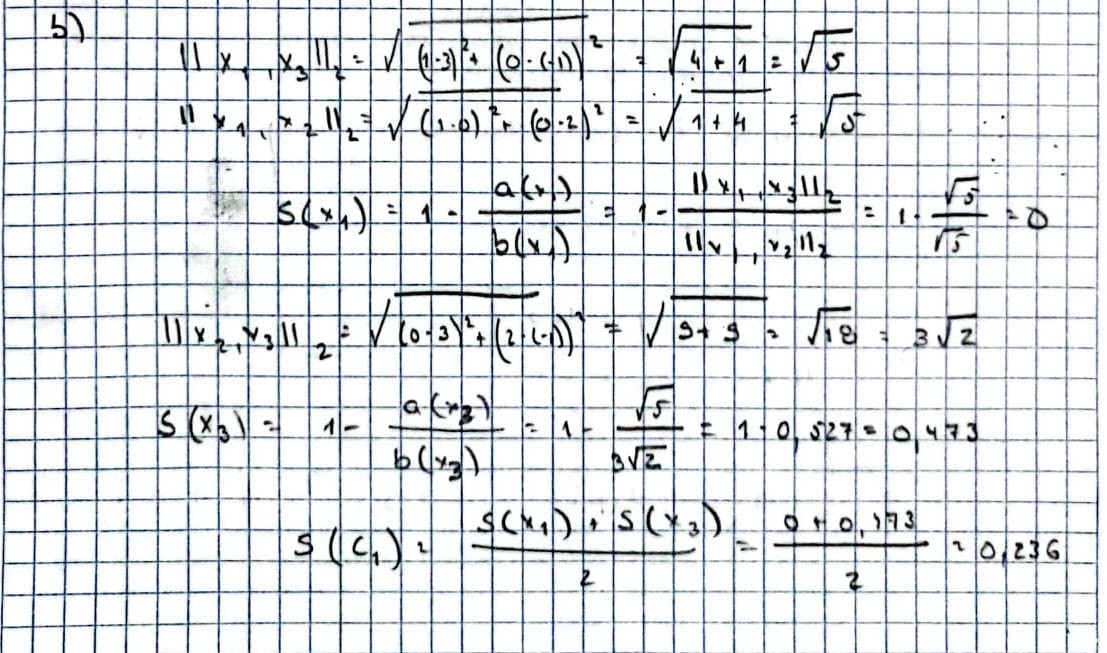


2)

a)



b)



# Programming and critical analysis

Here we have an overview of all the imports used and how we imported the dataset, which are common to all questions. Therefore, we will be omitting this code from the beginning of

each task to avoid redundancy

A screen shot of a computer program

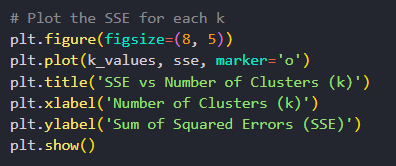
Description automatically generated

1)

1. Normalizing the data and applying k-means for each k value



Plotting the result



A graph with a line

Description automatically generatedOutput:

1. According to the plot, and using the knee/elbow finding method, we come to the conclusion that the ideal number of clusters is 6. This method consists of finding the point where the decrease in SSE becomes less accentuated with the increase of clusters.

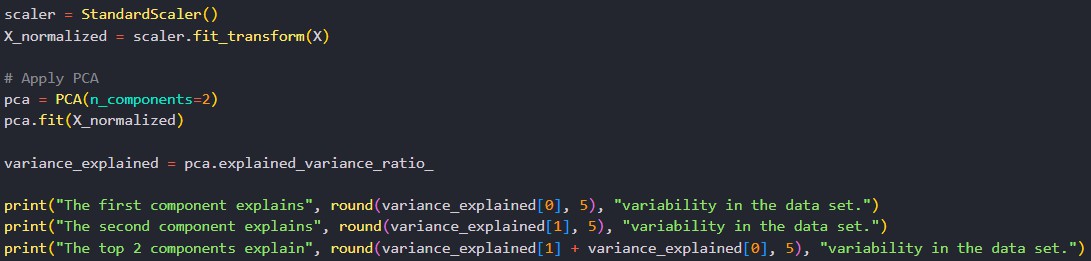
For all values before this point, the model might suffer from underfitting and for values above, the increase in the number of clusters and complexity does not provide substantial improvements to the model. Not only does it not justify, but it also contributes to overfitting the data.

1. K-modes is an adaptation of k-means used to better handle categorical features. It uses the mode (most frequent value) instead of means to represent the centroid of the cluster and uses distance based in dissimilarity (like Hamming Distance) to group the data.

Given that our dataset's features are predominantly categorical (10 categorical out of 17), in theory, k-modes would be a better clustering approach, considering the explanation above.

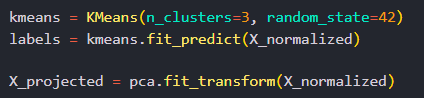
2)

1. Normalizing the data with StandardScaler, applying the PCA and getting the explained variance ratio for each component

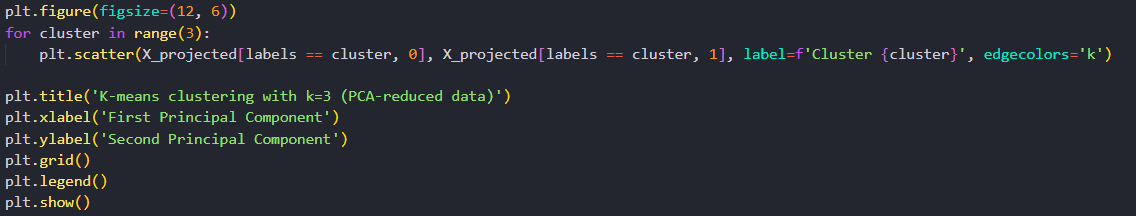


The first component explains 0.11679 variability in the data set. The second component explains 0.11076 variability in the data set. The top 2 components explain 0.22755 variability in the dataset.

1. Applying k-means clustering and PCA



Plotting the results:



Output:

A diagram of a cluster of dots

Description automatically generated with medium confidence

We can only clearly separate two of the three clusters (the yellow one and the magenta one). The third one (dark blue) is very hard to distinguish from the other two as it underlies both of them.

This happens because we are losing a lot of information on the data variance by only considering the top two components to draw the plot. The fact that there is significant overlap between some clusters suggests that using only these two components in a 2D space isn’t enough to represent a lot of the variance between the different observations of the dataset, that might become more apparent in higher dimensional spaces.

1. Plotting both graphs

A screen shot of a computer code

Description automatically generated

Output:

A graph of a number of people

Description automatically generated

A graph of different colored bars

Description automatically generated

The main differences in the obtained plots can be described as follows:

* + Jobs
    - Cluster 0: Jobs like “blue-collar”, “services”, “unemployed” and “self-employed” appear mostly in this cluster. They can be classified as less qualified positions and working-class jobs.
    - Cluster 1: Jobs like “technician”, “entrepreneur”, “housemaid” and “unknown” appear mostly in this cluster. We can’t find a significant relationship between these jobs, which makes us believe this cluster is better described by other features.
    - Cluster 2: Jobs like “admin”, “management”, “retired” and “student” appear mostly in this cluster. They describe either people who are not working or highly qualified jobs
  + Education
    - Cluster 0: Mostly represented by “primary” and “secondary” levels of education, which translates well to the conclusions we made for the job distribution.
    - Cluster 1: Mostly represented by “unknown” level of education, which can also be a good indicator that this cluster might be better represented by other features.
    - Cluster 2: Mostly represented by “tertiary” level of education, which also translates well to the conclusions we made for the job distribution.

**END**