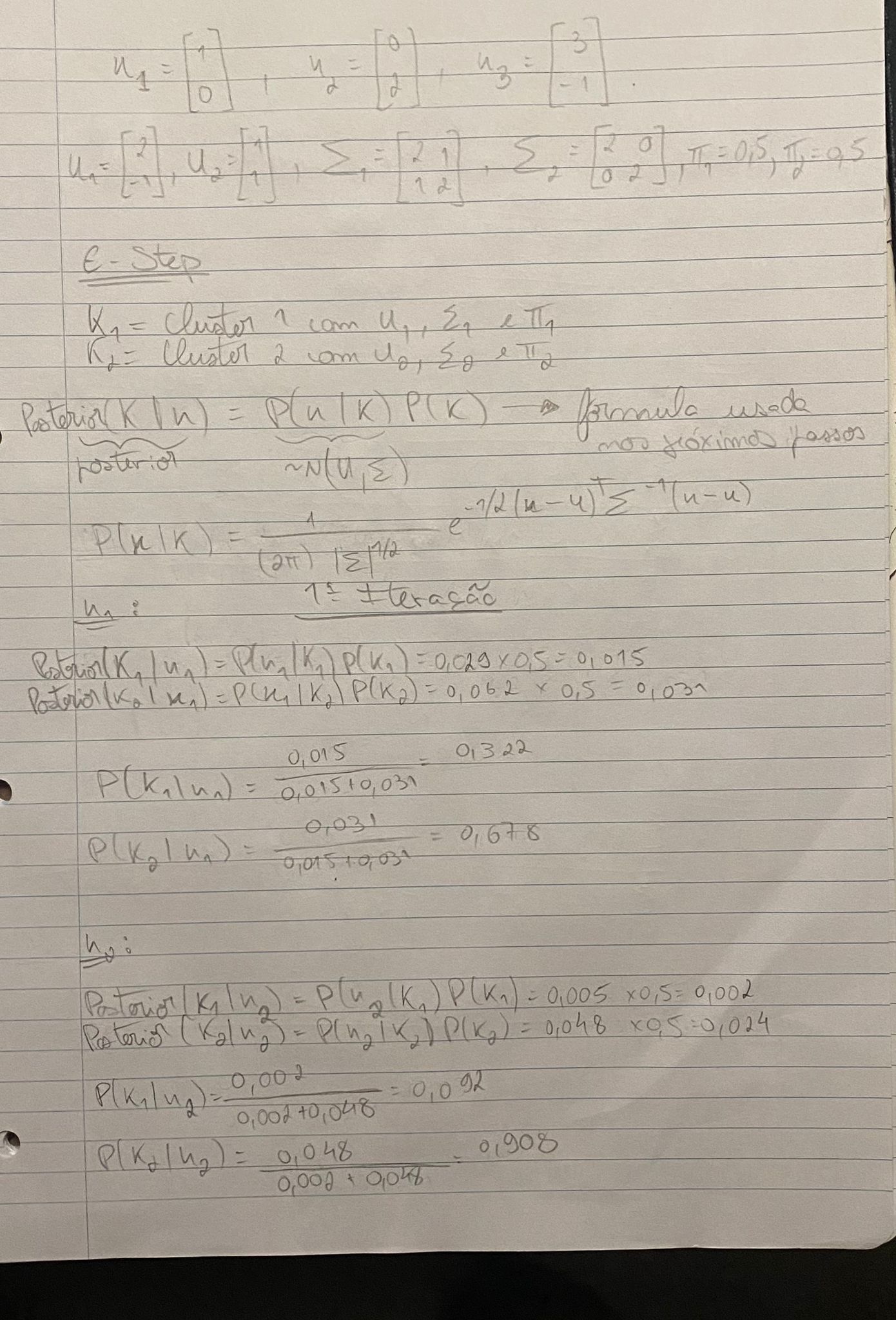
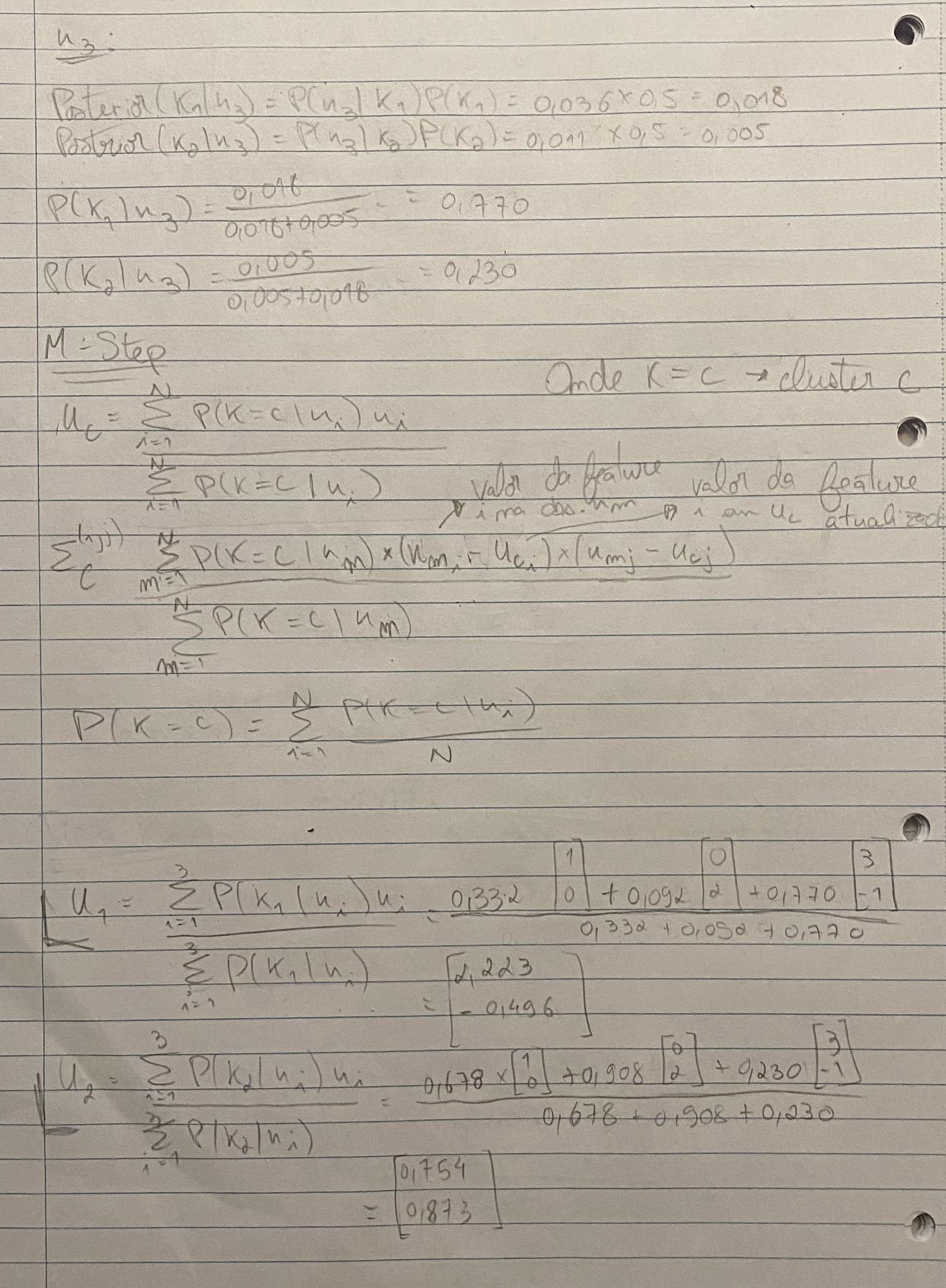
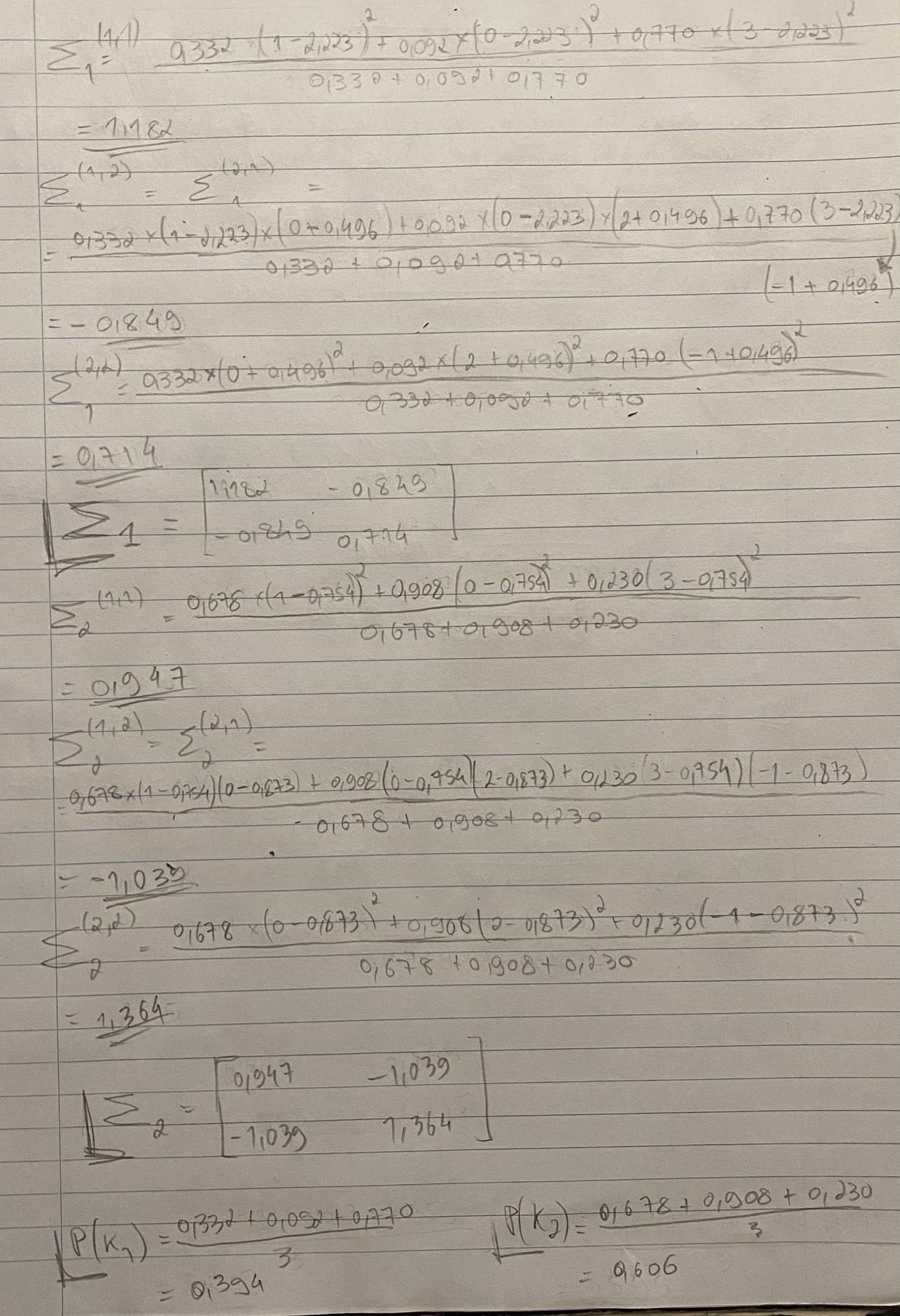
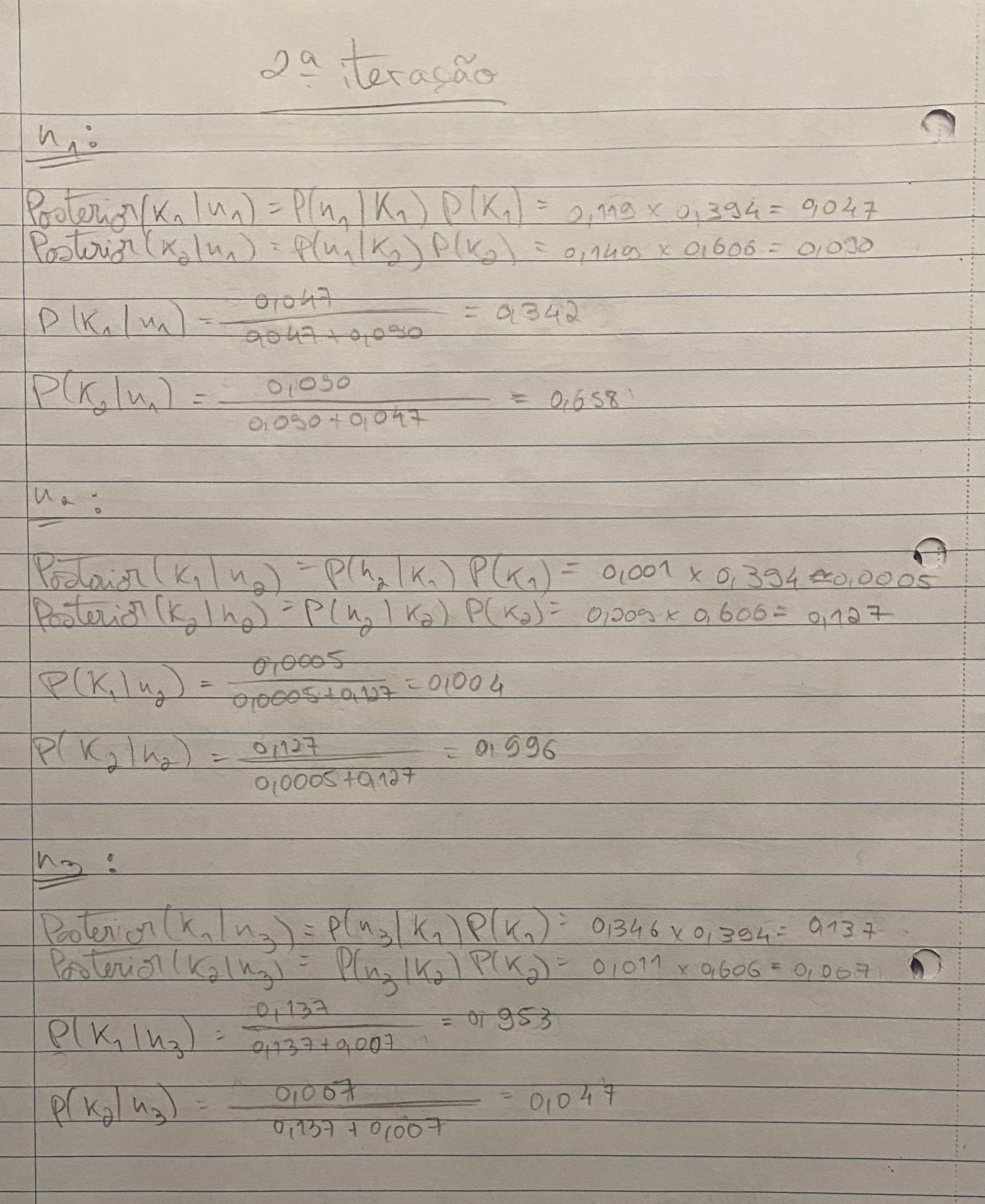
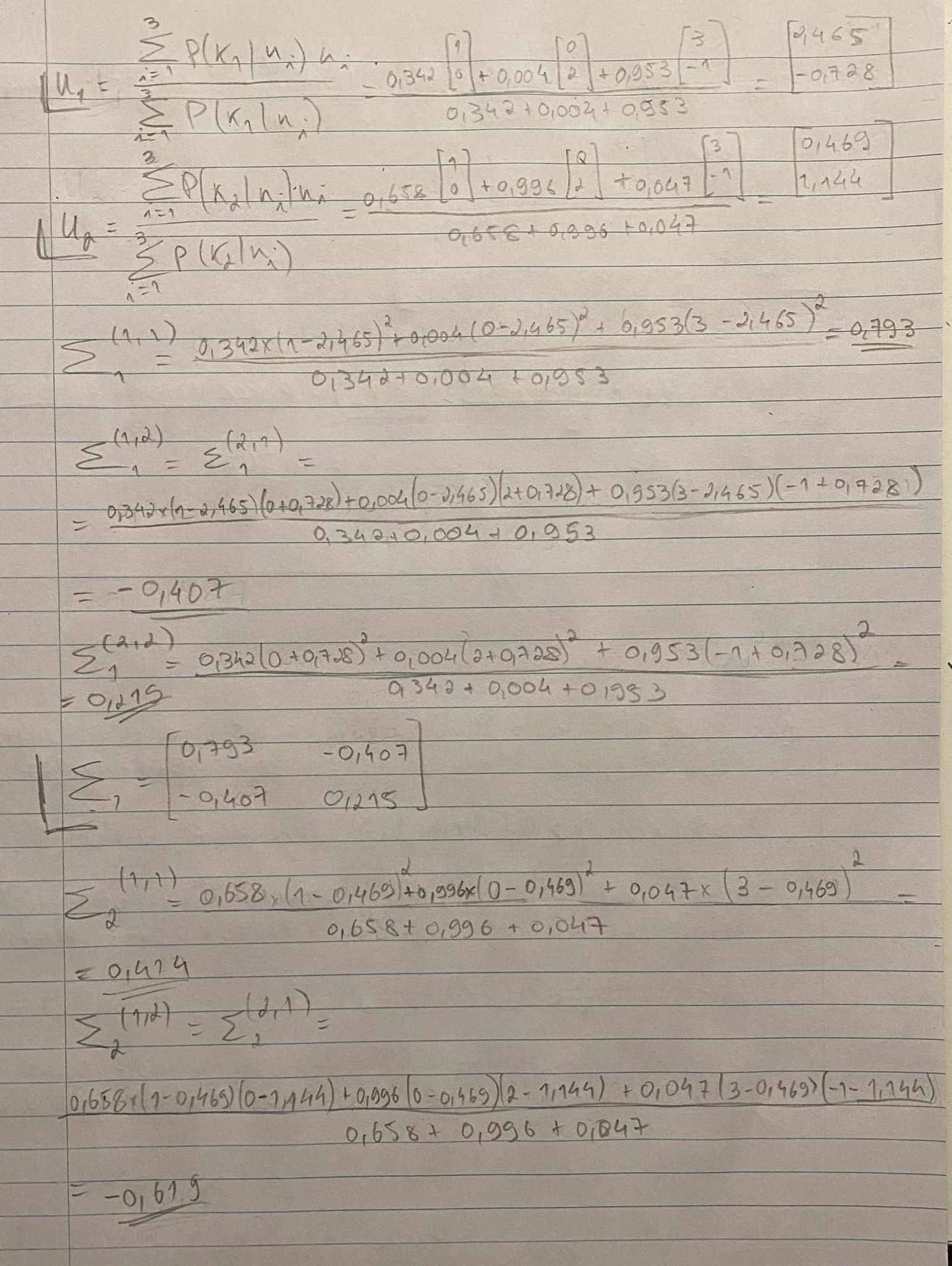
**I. Pen-and-paper**

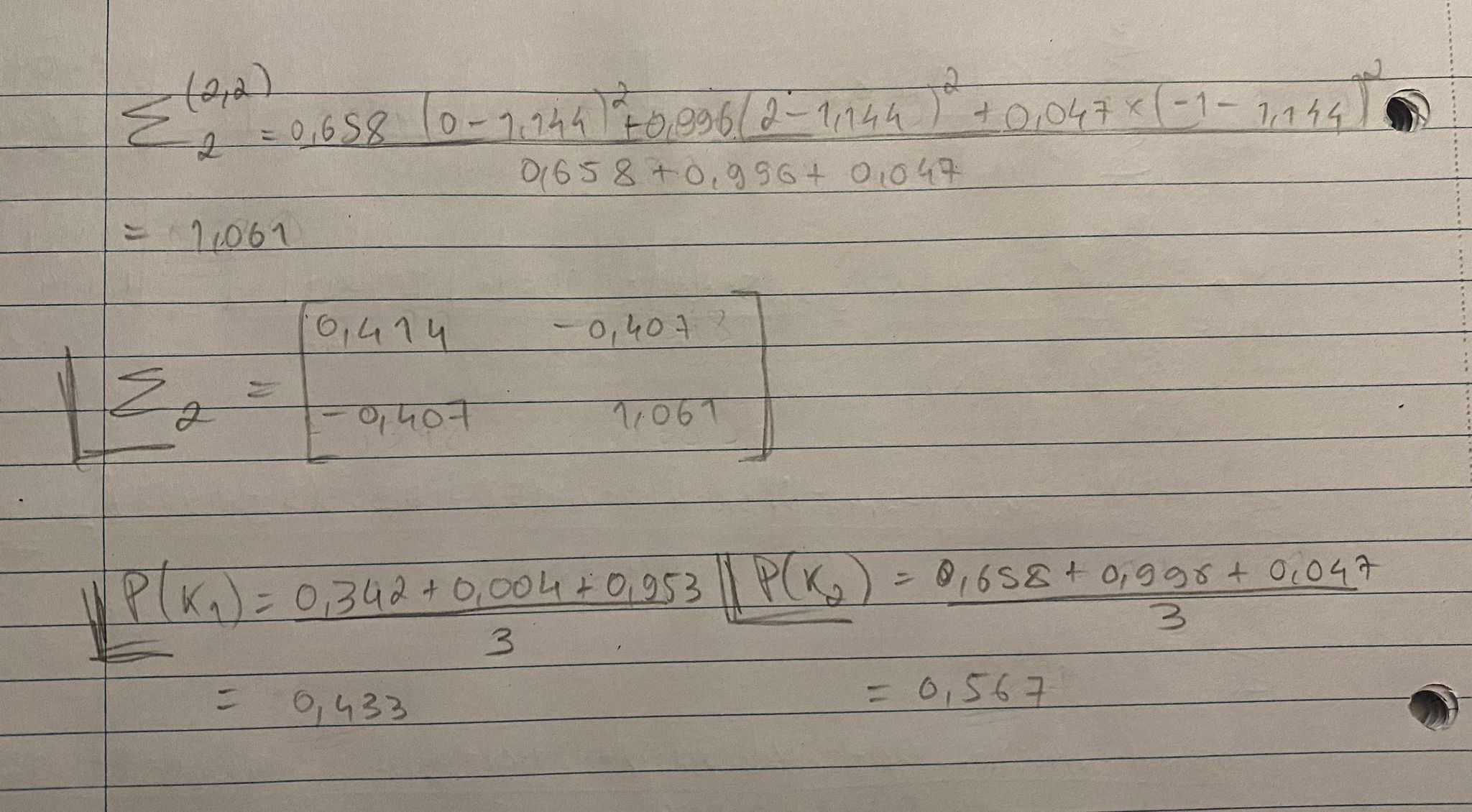




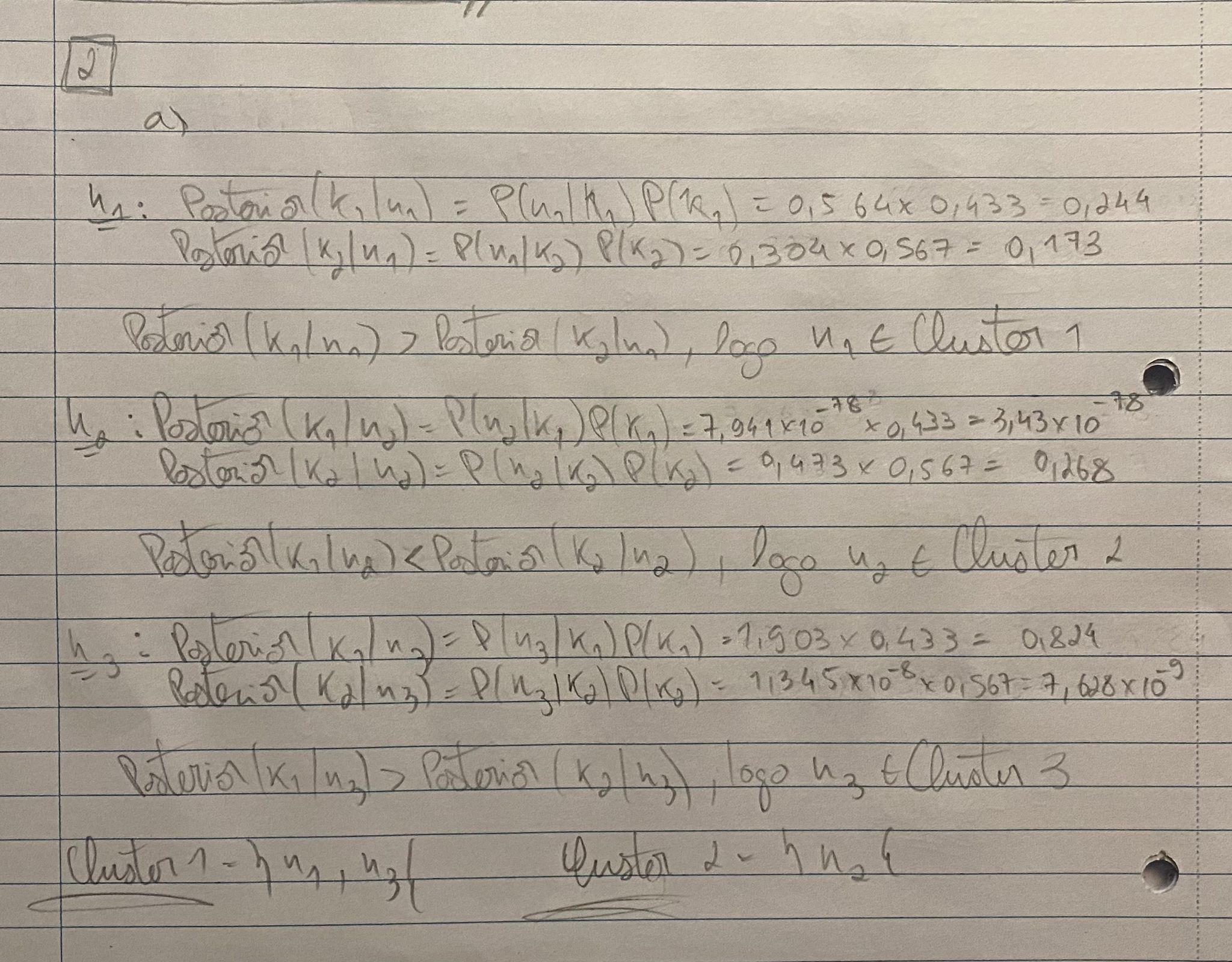




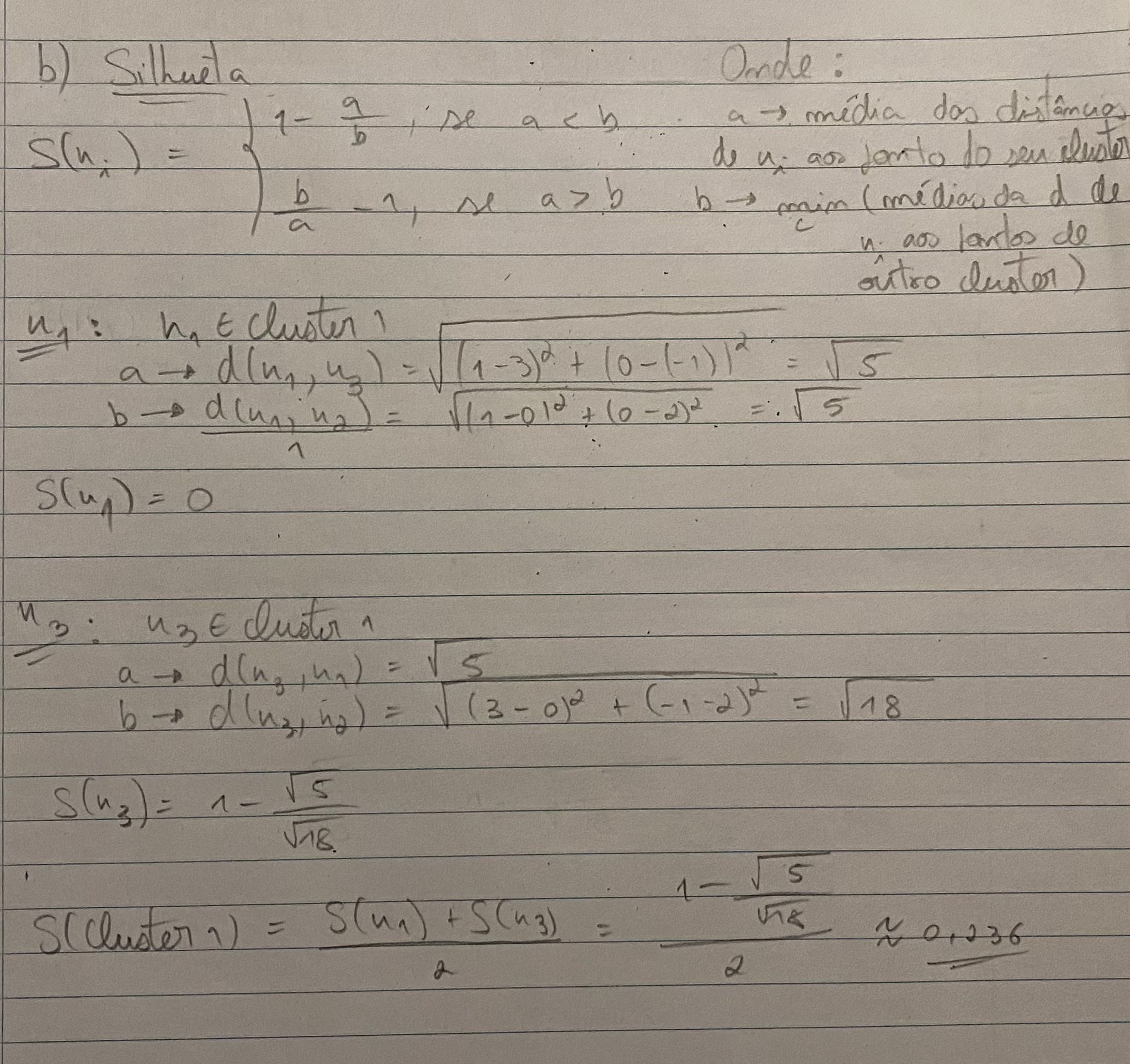




**2. a)**

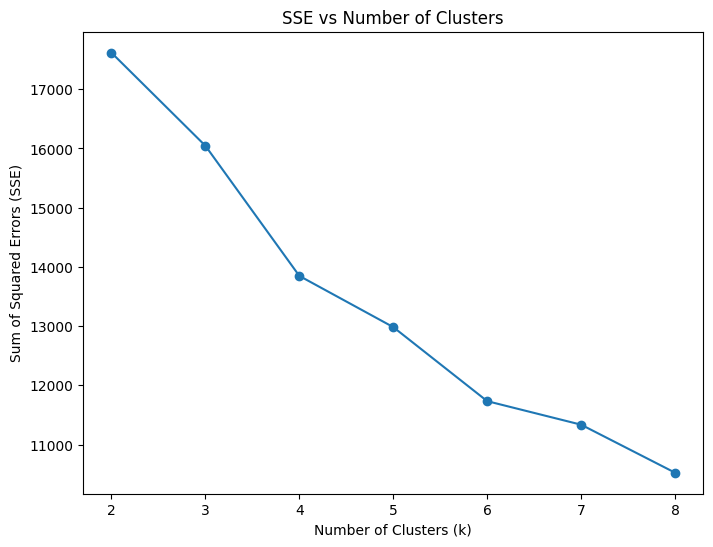


**2. b)**



**II. Programming and critical analysis**

**1.a)**



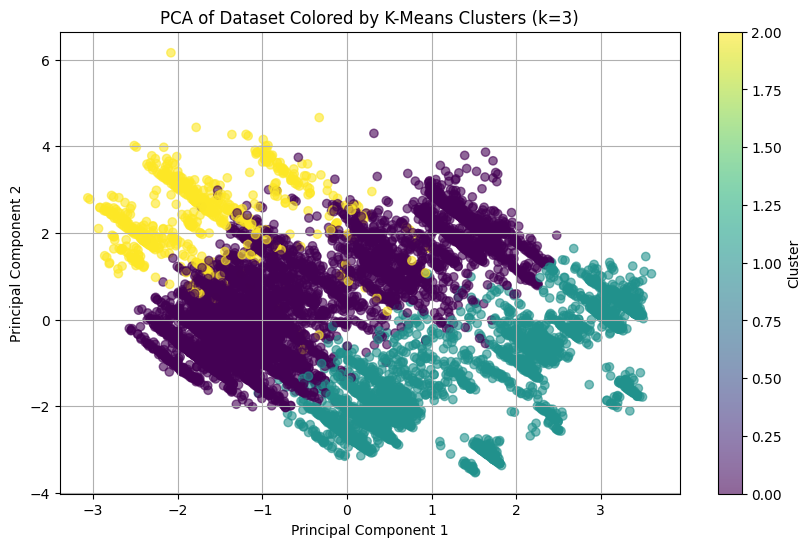
**1.b)**

Com base no gráfico, o número ideal de clusters parece ser 4, pelo facto de o declive da curva começar a diminuir (em valor absoluto). Isto indica que a adição de mais clusters não baixa suficientemente o valor do SSE para se justificar aumentar a complexidade do modelo. Assim, com 4 clusters atinge-se um equilíbrio entre divisão dos dados e simplicidade do modelo.

**1.c)**

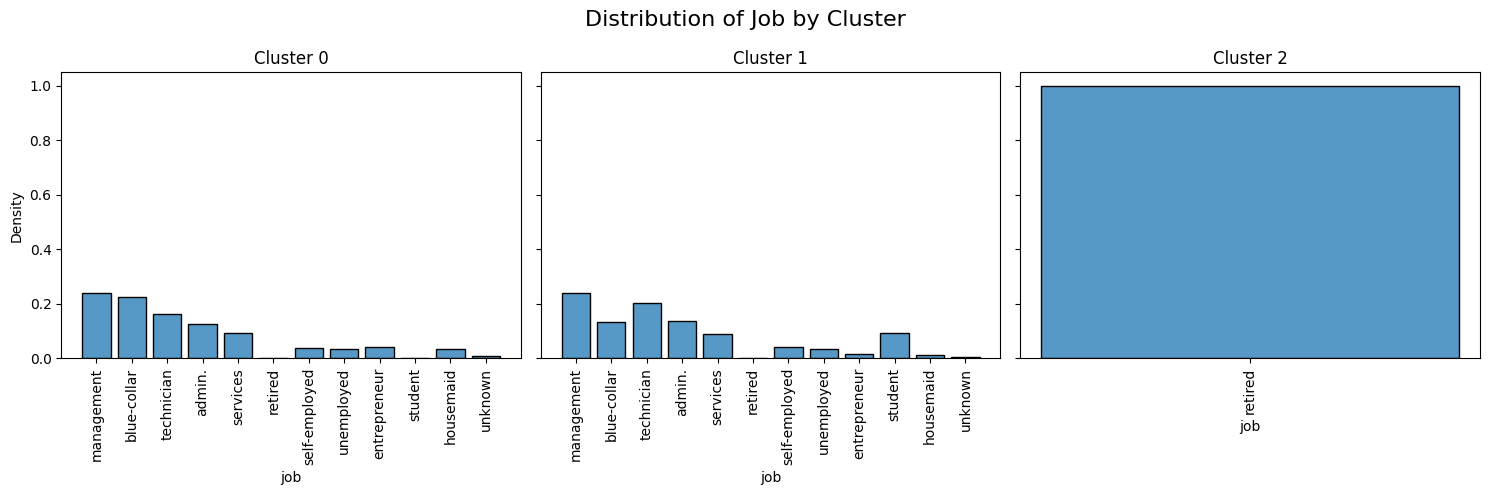
Sim, o k-modes é a abordagem mais adequada para conjuntos de dados predominantemente categóricos. Isso deve-se ao fato de que, neste caso, a utilização da moda para cálculos do algoritmo é mais apropriada do que a média, que é utilizada no k-means. A moda permite uma representação mais fiel das características dos dados categóricos, resultando num clustering mais eficaz.

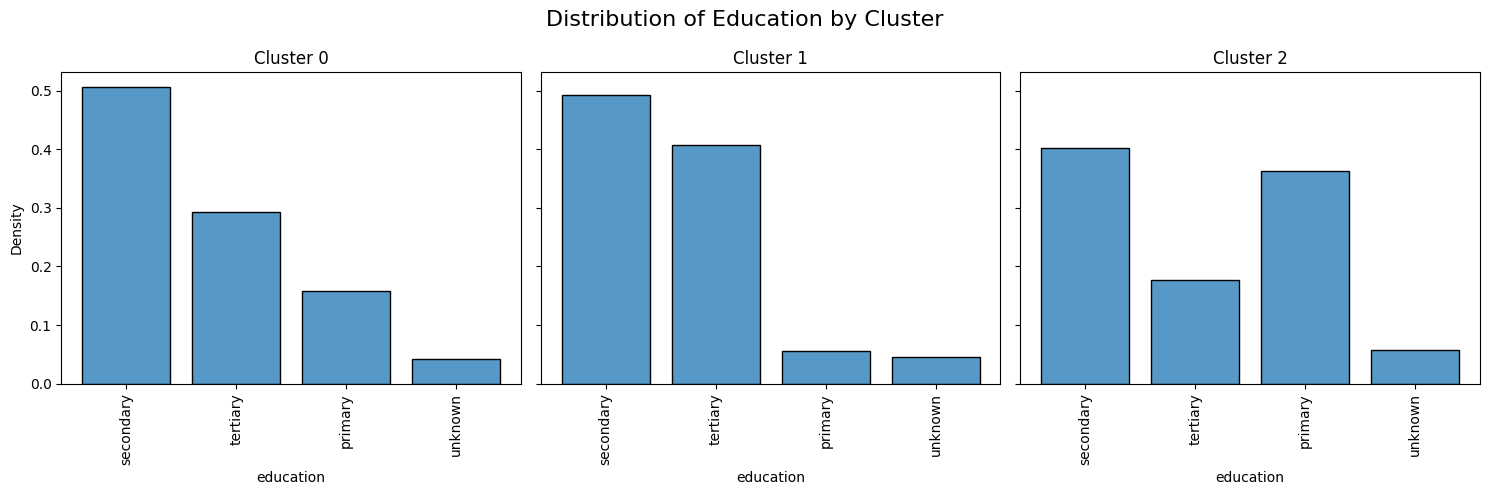
**2. b)**



Não conseguimos separar os clusters de forma clara. A mistura de pontos entre os diferentes clusters obtidos pelo k-means (com k=3) pode ser explicada pelo facto de apenas 23% da variabilidade ter sido capturada pelas duas principais componentes principais no PCA. Este valor baixo indica que as componentes não refletem adequadamente a informação total do conjunto de dados, sugerindo uma estrutura complexa que não é bem representada na projeção bidimensional. Como resultado, muitos padrões e interações relevantes entre as variáveis permanecem não capturados, comprometendo a separação entre os clusters. Assim, esta mistura deve-se à falta de distinção clara nas características que realmente importam.

**2.c)**





Os gráficos apresentam a distribuição das categorias de emprego e educação em quatro clusters (0, 1 e 2).

Relativamente à distribuição de empregos, verificamos que os clusters 0 e 1 são amplamente dominado por gestores, técnicos, operários e administradores, sendo que o resto das atividades aparece em menor quantidade e pode ser desprezado. O cluster 2 só contém reformados.

Relativamente à distribuição de educação, verificamos que o cluster 0 possui cerca de metade das pessoas com o ensino secundário, um terço com o ensino superior e uma percentagem baixa de pessoas com o ensino primário. O cluster 1 possui quase só pessoas com o ensino superior e secundário. No cluster 2, predominam pessoas com educação secundária e primária, havendo também algumas (em menor quantidade) com o ensino superior.

Podemos, portanto, concluir que os três clusters representam grupos de indivíduos com características distintas. Os Clusters 0 e 1 são compostos principalmente por pessoas em cargos de gestão, técnicos, operários e administração, com um nível de educação relativamente alto (ensino superior e secundário). O Cluster 2 é composto apenas por reformados, no qual se verifica um grau de educação mais baixo (predominantemente primário e secundário).

**END**