BONTA AALAYA

12340530

DSL253 - Statistical Programming

INTRODUCTION

In this assignment, we observe the effect of normalization and PCA on correlation matrix to identify structural patterns in brain data, and observe the change of chi squared distribution when derived from normal distribution for different sample sizes while also computing statistics, and verifying the empirical rule (68-95-99.7 rule) for a noisy dataset that follows Gaussian distribution

DATA

Datasets for the 1st question were provided in csv format. For 2nd question, we generated random samples using inbuilt function 'randn' (for normal distribution). For 3rd question, we were provided with a noisy Gaussian dataset in csv format

METHODOLOGY

Question 1

Correlation matrix is obtained by using 'corr' function
Normalisation of raw data given can be done by using min-max normalization (because we are given both an upper limit (+1) and lower limit (-1). Formula is given by [2]

$$v' = \frac{v - min\ (A)}{max\ (A) - min\ (A)}(new\_max(A) - new\_min(A)) + new\_min(A)$$

Correlation matrix is calculated again, this time for normalised data
We use PCA to deduce dimensions and find correlation matrix for transformed signals

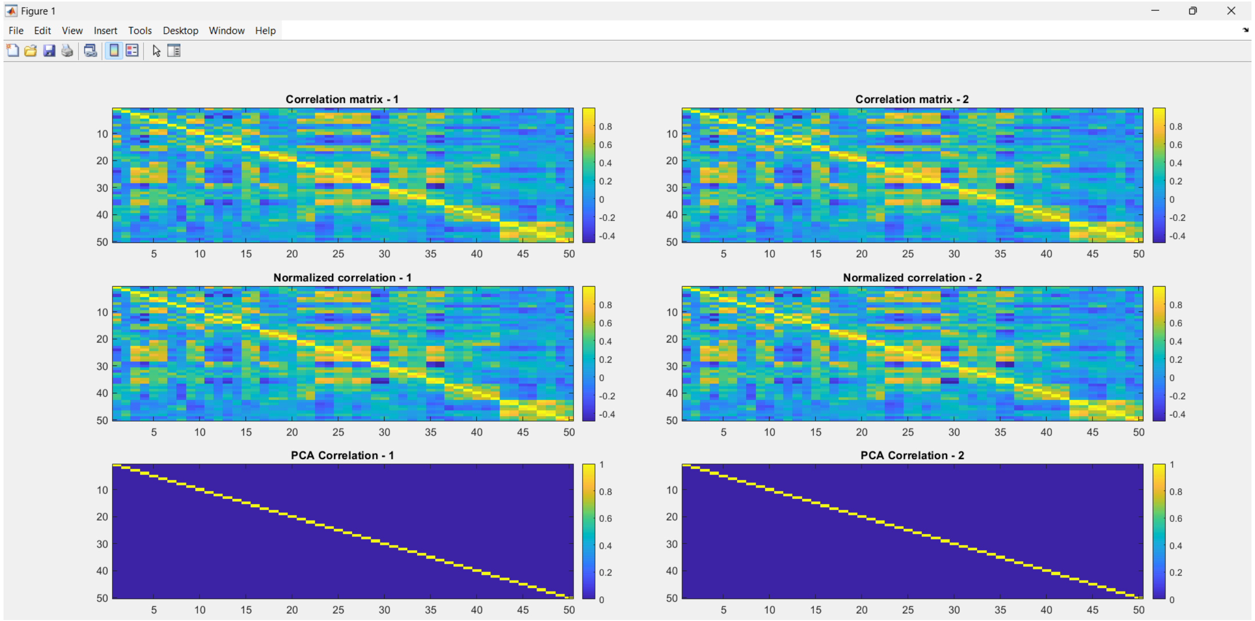
Question 2

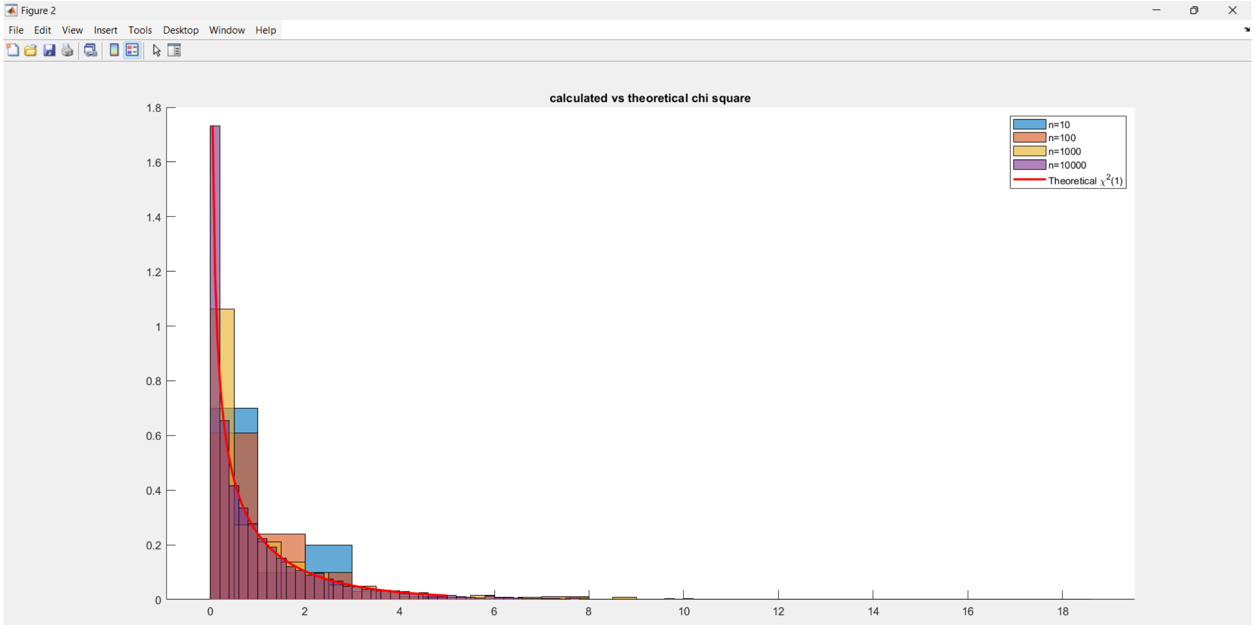
We generate normal distributions of different sizes and plot a graph to compare theoretically obtained values with calculated ones

Question 3

We skip 1st cell because it is not a numerical value
Built in function normcdf directly gives us % of data in 1σ, 2σ, 3σ but because question mentions using mean to calculate it manually, I used the formula in source [1] to obtain them

RESULTS





Mean = 49.8583

Variance = 111.8398

Region	Proportion of data points lying within
1σ	0.6840
2σ	0.9520
3σ	0.9990

Probability of exceeding 2σ: 0.05

(we can use the term probability because it is normal distribution)

DISCUSSION

The correlation matrices undergo no notable changes after normalization, and the correlation matrices post normalization are almost identical to the original correlation matrices. This suggests that scaling may not have a significant effect on correlation patterns. We can also see that the original correlation matrices are convoluted with all ranges of correlation, making the diagonal less prominent. Meaning, no single variable dominates. The real change is observed after applying PCA. The correlation matrices are nearly perfectly diagonal, meaning that each principal component is uncorrelated with the others

The histograms of initial sample sizes fail to show an accurate representation of the statement "If a random variable $X\, is\, N(\mu,\ \sigma^2)$, where $\sigma^2\ >\ 0$, then the random variable $V\ =\ \frac{(x-\mu)^2}{\sigma^2}$ follows a chi-squared distribution with 1 degree of freedom $\chi^2(1)$"
However, as sample size increases, we can observe that the histogram converges to the red-coloured curve of $\chi^2(1)$

50% of data in normal distribution will exist on the right side of the mean, and the other half lies on the left. The '68%' denotes the percentage of data that exists within 1 σ of the mean. In other words, between $\mu\ -\ \sigma$ and $\mu\ +\ \sigma$. The empirical rule is obtained by using the normal function

$$f(x)\ =\ \frac{1}{\sigma\sqrt{2\Pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Taking $\mu\ =\ 0$ and $\sigma\ =\ 1$, we have a normal curve whose mean lies on origin and pdf spreads out accordingly (standard normal distribution). Calculating the area under the curve between the required intervals, we can prove the empirical formula (refer [1] for complete proof.) We can observe that there is a slight change in the computed values, possibly due to the noise.

CONCLUSION

We observed the effect of normalization and PCA on correlation matrix of brain data, concluding that scaling doesn't have a major effect on correlation matrix. PCA on the other hand, nicely reduces the dimensions, separating the principal component from the others. From the chi square analysis, we confirmed that a standard normal variable squared follows chi square distribution with 1 dof (degree of freedom). The noisy Gaussian dataset was able to nearly validate the empirical rule, and the slight deviations from the exact values may be due to the noise

____