**DS201**

**Statistical Programming**

**Assignment 4**

**Ashutosh Kumar Jha**

**12340390**

**DSAI**

**2nd Year**

**Semester 4**

**---------------------------------------------------------------------------**

**Question 1: Analysis of Coactivation Matrices and Dimensionality Reduction in Time Series Brain Data**

Introduction:

-----------------------

Understanding brain activity requires analyzing time series signals from different brain regions. This study investigates coactivation (correlation) matrices of two datasets, normalizes the data, applies Principal Component Analysis (PCA), and examines changes in correlation patterns. The empirical evaluation aids in understanding how preprocessing affects signal correlations.

Data:

------------------------

The dataset consists of time series signals from 50 brain regions, recorded over 190 time points in two separate datasets (data1 and data2). Each dataset captures brain activity in different conditions or individuals.

Methodology:

-----------------------

1. **Compute Coactivation Matrices:** Correlation matrices are computed for both datasets to measure the interdependencies between brain regions.
2. **Normalization:** Data is scaled between -1 and 1 using MinMaxScaler to ensure uniformity across features.
3. **Recompute Correlation Matrices:** The effect of normalization on coactivation matrices is analyzed.
4. **Dimensionality Reduction using PCA:** PCA reduces the dataset to 10 principal components to examine major variance trends.
5. **Comparison and Visualization:** Heatmaps are generated to compare correlation structures before and after normalization and PCA transformation.

Results:

-----------------------

1. The original correlation matrices of both datasets reveal strong coactivation patterns between some brain regions.
2. Normalization preserves correlation structures but rescales values, ensuring they remain between -1 and 1.
3. PCA transformation significantly alters the correlation structure by reducing dimensionality, capturing only the most dominant variance trends.

Discussions:

----------------------

* Normalization does not affect correlation values but ensures uniform feature scaling, which is beneficial for machine learning models.
* PCA transformation reduces noise but also removes some lower-variance coactivation patterns, which may impact downstream analyses.
* The heatmaps highlight changes in correlation structures, demonstrating how dimensionality reduction impacts data representation.

Conclusion:

-----------------------

This study demonstrates how preprocessing methods influence coactivation matrices. Normalization maintains correlation structures while PCA reduces complexity but alters correlation patterns. Understanding these transformations is crucial for reliable interpretations of brain activity data in neuroscience and machine learning applications.

**Question 2: Empirical Verification of the Chi-Squared (1) Distribution**

1. Introduction:

----------------------

The chi-squared distribution plays a fundamental role in statistics. It is known that if a random variable follows a normal distribution , then the transformed variable follows a chi-squared distribution with 1 degree of freedom, denoted as . This study aims to empirically verify this theorem by generating random samples, computing the transformed variable, and comparing the empirical distribution with the theoretical chi-squared distribution.

2. Data

--------------------

The dataset consists of synthetic samples drawn from a normal distribution with different sample sizes: 100, 1000, and 10,000. The transformed variable is computed for each sample size to examine how well it aligns with the chi-squared(1) distribution.

3. Methodology

---------------------

1. **Generate Normal Samples:** Draw samples from a normal distribution with mean and standard deviation .
2. **Compute Transformed Variable:** Apply the transformation for each sample.
3. **Compare with Theoretical Distribution:** Plot histograms of the empirical distribution and overlay the theoretical chi-squared(1) probability density function (PDF).
4. **Evaluate for Different Sample Sizes:** Repeat the process for , , and to observe how the empirical distribution approaches the theoretical distribution as it increases.

4. Results

--------------------

* For small sample sizes (e.g., ), the empirical histogram shows some deviations from the theoretical curve.
* As the sample size increases, the empirical distribution closely matches the theoretical chi-squared(1) PDF.
* With , the histogram nearly overlaps with the theoretical curve, verifying the theorem empirically.

5. Discussion

-------------------

* The empirical distribution follows the chi-squared(1) distribution, confirming the theoretical result.
* Larger sample sizes improve the approximation, demonstrating the law of large numbers in statistical analysis.
* Minor deviations for smaller suggest the need for sufficient sample sizes in statistical inference applications.

6. Conclusion

------------------

This study successfully verifies that the transformed variable follows a chi-squared(1) distribution. This study demonstrates how preprocessing methods influence coactivation matrices. Normalization maintains correlation structures while PCA reduces complexity but alters correlation patterns. Understanding these transformations is crucial for reliable interpretations of brain activity data in neuroscience and machine learning applications.

**Question 3: Analysis of Gaussian Data and Empirical Rule Verification**

1. Introduction

--------------------

The normal distribution is one of the most important distributions in statistics. This study aims to analyze a given dataset following a Gaussian distribution, compute its statistical properties, and verify the empirical rule, which states that:

* ~68% of data lies within 1 standard deviation (σ) of the mean (µ).
* ~95% of data lies within 2σ.
* ~99.7% of data lies within 3σ. Additionally, we will analyze the probability of data points lying beyond ±2σ and visualize the dataset through histograms and cumulative distribution functions (CDF).

2. Data

----------

The dataset containing the values, following the gaussian distribution, along with some noise components, is already provided to us, in the form of a .csv file. The data is collected from the csv file by loading the data in the form of a dataframe, and then extracting the column of the dataframe named “Value”, since that is the column containing the values, and then converting the index(column), into the numpy array, so that it can be easily be computed upon by the defined function.

3. Methodology

---------------------

1. **Compute Mean and Variance:** Calculate the mean (µ) and variance (σ²) of the dataset.
2. **Verify Empirical Rule:** Compute the percentage of data points falling within 1σ, 2σ, and 3σ of the mean and compare with the expected theoretical values.
3. **Compute CDF:** Evaluate the probability of observing data points beyond ±2σ based on the cumulative distribution function.
4. **Visualize Data:**
   * Histogram with an overlaid Gaussian probability density function (PDF), including markers for µ, 1σ, 2σ, and 3σ.
   * CDF plot with an emphasis on the ±2σ region.

4. Results

---------------

1. The computed mean (µ) and variance (σ²) closely match the expected values for a normal distribution.
2. The percentage of data within 1σ, 2σ, and 3σ aligns well with the empirical rule.
3. The probability of observing data beyond ±2σ is consistent with theoretical expectations (~4.55% per tail).
4. The histogram and CDF plots visually confirm the normality of the dataset and the empirical rule.

5. Discussion:

------------------

* The empirical rule holds well even when minor noise is introduced, highlighting the robustness of normal distribution properties.
* Small deviations from expected percentages may arise due to finite sample sizes and added noise.
* The visualization clearly demonstrates the normality of the dataset and provides a deeper understanding of probability distribution behavior.

6. Conclusion

------------------

This analysis successfully verifies the empirical rule for a Gaussian dataset and demonstrates the fundamental properties of normal distributions. The computed results align closely with theoretical expectations, reinforcing key statistical principles.

Code: [12340390 Ashutosh Asg4.ipynb](https://colab.research.google.com/drive/1OHlDU8eKU9mWjrCk73Efr3QGPaa8gMbx?usp=sharing)