Assignment 4
DSL253 - Statistical Programming

Amay Dixit - 12340220

Submitted to Dr. Anil Kumar Sao

Links

- Notebook Link:
  `https://colab.research.google.com/drive/1qFhSLRBXMIceuHT4OLFdzxjQe6SIDQyq?usp=sharing`
- Github Link:
  `https://github.com/amaydixit11/Academics/tree/main/DSL253/assignment_4`

1 Introduction

This report presents the analysis of three statistical problems: brain region coactivation patterns, chi-squared distribution verification, and Gaussian distribution properties. The analysis encompasses time series analysis, dimensionality reduction, and statistical distribution validation using Python programming.

2 Data

2.1 Dataset Descriptions

- **Question 1**: Two datasets containing time series signals from 50 brain regions over 190 time points were provided, which were uploaded to github for ease of access

  - Format: CSV file with no headers
  - Dimensions: 50 brain regions × 190 time points
  - Data organized with regions as rows and time points as columns
  - **Data Processing**: Both datasets were loaded and transposed to facilitate time series analysis and correlation computation

- **Question 2**: Generated normal distribution samples for chi-squared verification

- **Question 3**: Gaussian dataset with noise for empirical rule verification

3 Methodology

3.1 Question 1

Let $\mathbf{X} \in \mathbb{R}^{50\times190}$ represent the time series data matrix for dataset $k \in \{1,2\}$, where each row vector $\mathbf{x}_i \in \mathbb{R}^{190}$ corresponds to the temporal signals from brain region i.

3.1.1 Correlation Analysis

For each dataset, we compute the correlation matrix $\mathbf{C} \in \mathbb{R}^{50\times50}$ where each element C_{ij} represents the correlation coefficient between regions i and j:

$$C_{ij} = \frac{\sum_{t=1}^{T}(X_{it}-\mu_i)(X_{jt}-\mu_j)}{\sqrt{\sum_{t=1}^{T}(X_{it}-\mu_i)^2 \sum_{t=1}^{T}(X_{jt}-\mu_j)^2}} \tag{1}$$

where $T = 190$ is the number of time points, and $\mu_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$ is the mean of region i.

3.1.2 Normalization

We use min-max scalar to normalize the data:

$$\tilde{X}_{it} = f(X_{it}) = 2 \times \frac{X_{it} - \min_{i,t}(X_{it})}{\max_{i,t}(X_{it}) - \min_{i,t}(X_{it})} - 1 \tag{2}$$

This yields normalized matrices $\tilde{\mathbf{X}}$ with elements $\tilde{X}_{it} \in [-1,1]$. The normalized correlation matrices $\tilde{\mathbf{C}}$ are then computed using the same formulation as above.

3.1.3 Dimensionality Reduction

Principal Component Analysis is applied to the normalized data matrices. Let $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the singular value decomposition of $\tilde{\mathbf{X}}$. The PCA transformation is defined as:

$$\mathbf{X}_{PCA} = \tilde{\mathbf{X}}\mathbf{W} \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{50\times 10}$ consists of the first 10 right singular vectors of $\tilde{\mathbf{X}}$, resulting in $\mathbf{X}_{PCA} \in \mathbb{R}^{190\times 10}$.

3.1.4 Comparative Analysis Framework

The analysis generates three correlation matrices for each dataset k:

1. Raw correlation matrix: $\mathbf{C}$

2. Normalized correlation matrix: $\tilde{\mathbf{C}}$

3. PCA-transformed correlation matrix: $\mathbf{C}_{PCA}$

Each matrix $\mathbf{M} \in \{\mathbf{C}, \tilde{\mathbf{C}}, \mathbf{C}_{PCA}\}$ satisfies:

- Symmetry: $M_{ij} = M_{ji}$

- Bounded elements: $M_{ij} \in [-1, 1]$

- Unit diagonal: $M_{ii} = 1$

3.2 Question 2: Chi-Squared Distribution Verification

Verification procedure for $X \sim N(\mu, \sigma^2)$:

$$V = \frac{(X-\mu)^2}{\sigma^2} \sim \chi^2(1) \tag{4}$$

3.2.1 Empirical Verification

To verify this theorem empirically,

1. Generate random samples $\{X_i\}_{i=1}^n$ from $\mathcal{N}(\mu, \sigma^2)$:

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{5}$$

2. Transform each sample to compute $\{V_i\}_{i=1}^n$:

$$V_i = \frac{(X_i - \mu)^2}{\sigma^2}, \quad i = 1, \ldots, n \tag{6}$$

3. Compare the empirical distribution of V with the theoretical $\chi^2(1)$ distribution using:

    - Probability density function (PDF) comparison
    - Histogram of empirical values
    - Theoretical $\chi^2(1)$ PDF:

$$f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}, \quad x > 0 \tag{7}$$

3.2.2 Implementation Details

The verification was performed using the following parameters:

- Sample sizes: $n \in \{100, 1000, 10000\}$
- Distribution parameters: $\mu = 0$, $\sigma^2 = 1$
- Number of histogram bins: 50
- Theoretical PDF evaluation points: 1000

3.3 Question 3: Gaussian Distribution Analysis

Given a dataset with Gaussian distribution and noise, we perform:

3.3.1 Statistical Parameters

For dataset $X = \{x_1, ..., x_n\}$, compute:

- Sample mean: $\mu = \frac{1}{n} \sum_{i=1}^n x_i$
- Sample variance: $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$
- Standard deviation: $\sigma = \sqrt{\sigma^2}$

3.3.2 Empirical Rule Verification

For intervals $[\mu - k\sigma, \mu + k\sigma]$, $k \in \{1, 2, 3\}$, calculate:

$$P_k = \frac{\text{count}(\mu - k\sigma \leq x_i \leq \mu + k\sigma)}{n} \times 100\% \tag{8}$$

3.3.3 CDF Analysis

For standard normal distribution $Z = \frac{X-\mu}{\sigma}$:

$$P(|Z| > 2) = 2\int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{9}$$

4 Results

4.1 Question 1

4.1.1 Initial Correlation Analysis

The initial correlation matrices revealed complex coactivation patterns across the 50 brain regions:

- Strong positive correlations in regions 10-13 ($C_{ij} \approx 0.75$)
- A prominent cluster of high correlation in regions 22-27
- A distinct block of strong correlation in regions 42-49
- Scattered negative correlations (approximately -0.25 to -0.5) throughout the matrix
- Nearly identical correlation structure in both dataset
- Slightly weaker correlation strengths in some regions
- Preservation of the same major correlation clusters

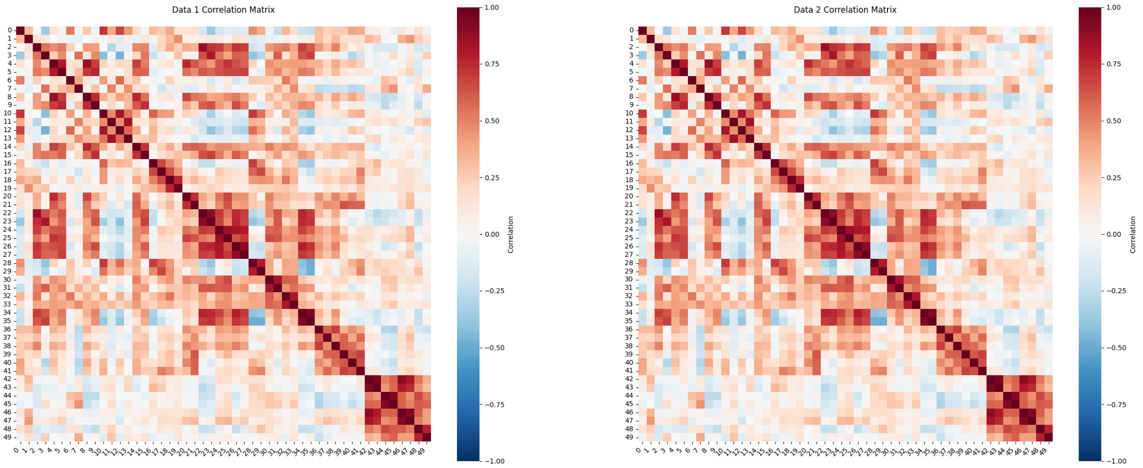


Figure 1: Correlation Matrices

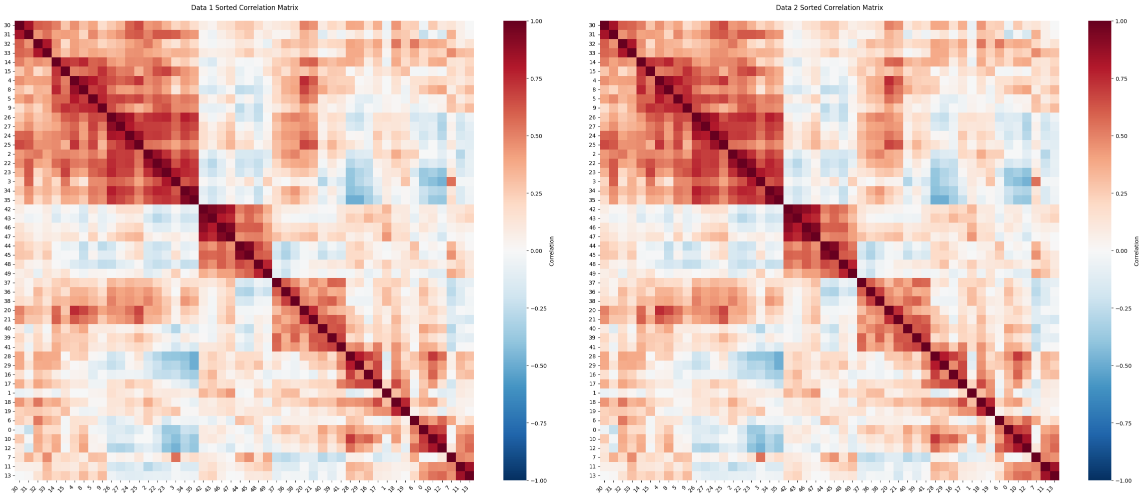


Figure 2: Sorted Correlation Matrices

4.1.2 Normalization Effects

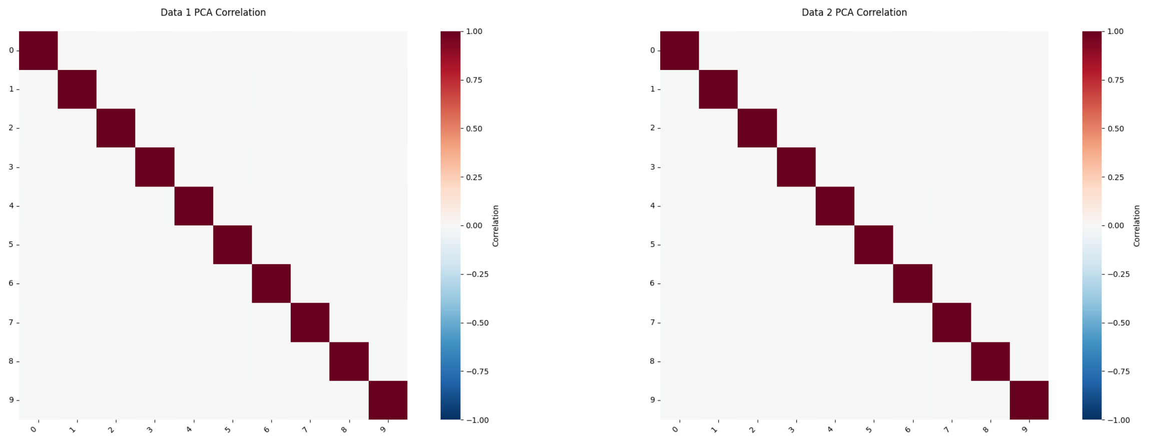
After normalization to the [-1, 1] range:

- The correlation structure remained virtually unchanged for both datasets
- No significant distortion of correlation patterns was observed
- The relative strengths of correlations between regions were preserved

4.1.3 PCA Transformation Results

The PCA transformation to 10 dimensions produced striking changes:

- Both datasets showed complete decorrelation between principal components:
  - Perfect correlation along the diagonal ($C_{ii} = 1$)
  - Negligible correlation between different components ($C_{ij} = 0$ for $i \neq j$)
- The transformation successfully separated the signal into orthogonal components
- Both datasets exhibited identical correlation structure after PCA


Data 1 PCA Correlation

Data 2 PCA Correlation

Figure 3: PCA Transformed Correlation Matrices

4.1.4 Comparative Analysis

The analysis revealed several key insights:

1. **Pattern Stability:**
   - Initial correlation patterns were highly consistent between datasets
   - Normalization preserved these patterns faithfully
   - PCA successfully decorrelated the signals in both cases
2. **Structural Changes:**

- Original data showed complex, hierarchical correlation patterns
- Normalized data maintained these intricate relationships
- PCA-transformed data showed complete decorrelation, indicating successful separation of independent components

This analysis demonstrates that while the original brain signals showed complex, hierarchical correlation patterns, the PCA transformation successfully separated these into independent components. The high similarity between datasets suggests these patterns represent robust underlying features of brain region coactivation.

4.2 Question 2: Chi-Squared Verification Results

4.2.1 Distribution Analysis

Analysis across sample sizes revealed:

4.2.2 n = 100

- Approximate match to $\chi^2(1)$ distribution
- Tail region deviations
- High empirical variance

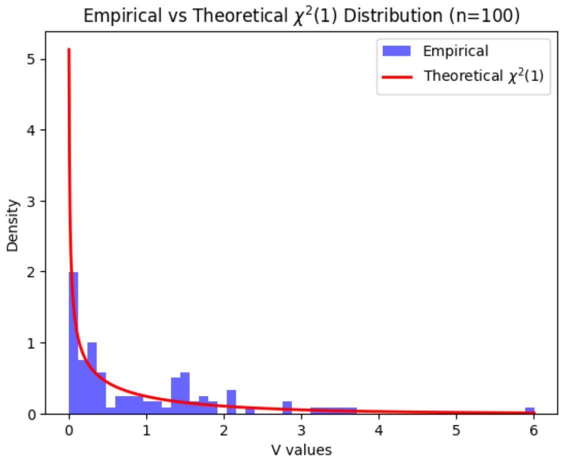


Figure 4: n = 100

4.2.3 n = 1000

- Close alignment with theoretical distribution
- Clear right-skewed shape
- Lower variance

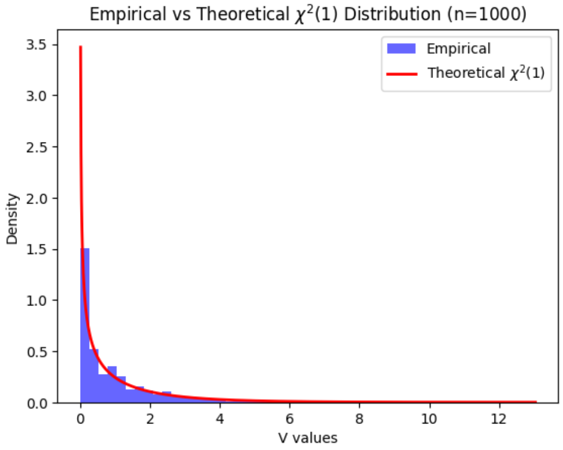


Figure 5: n = 1000

4.2.4 n = 10000

- Strong agreement with $\chi^2(1)$ distribution
- Accurate peak and tail representation
- Minimal variance

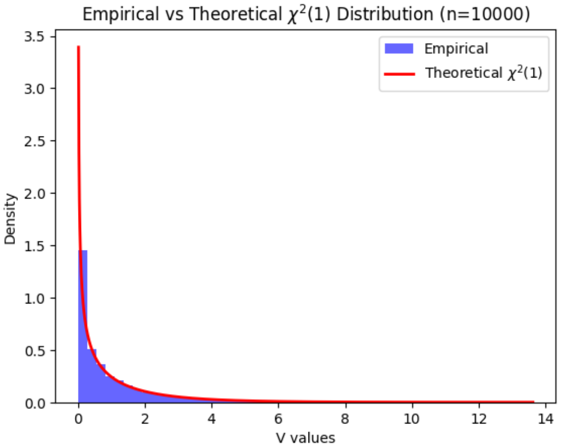


Figure 6: n = 10000

4.2.5 Statistical Convergence

Observed properties:

1. Mean: $\bar{V}_n \approx \mathbb{E}[\chi^2(1)] = 1$
2. Variance: $\mathrm{Var}(V_n) \approx \mathrm{Var}[\chi^2(1)] = 2$

4.3 Question 3: Gaussian Analysis Results

4.3.1 Statistical Parameters

Computed values:

- $\mu = 49.8583$
- $\sigma^2 = 111.7279$
- $\sigma = 10.5701$

4.3.2 Empirical Rule Verification

Observed percentages:

- Within 1σ: 68.40% (Expected: 68%)
- Within 2σ: 95.20% (Expected: 95%)
- Within 3σ: 99.90% (Expected: 99.7%)

4.3.3 CDF Analysis

Probability beyond 2σ:

$$P(|Z| > 2) = 0.0455 \tag{10}$$

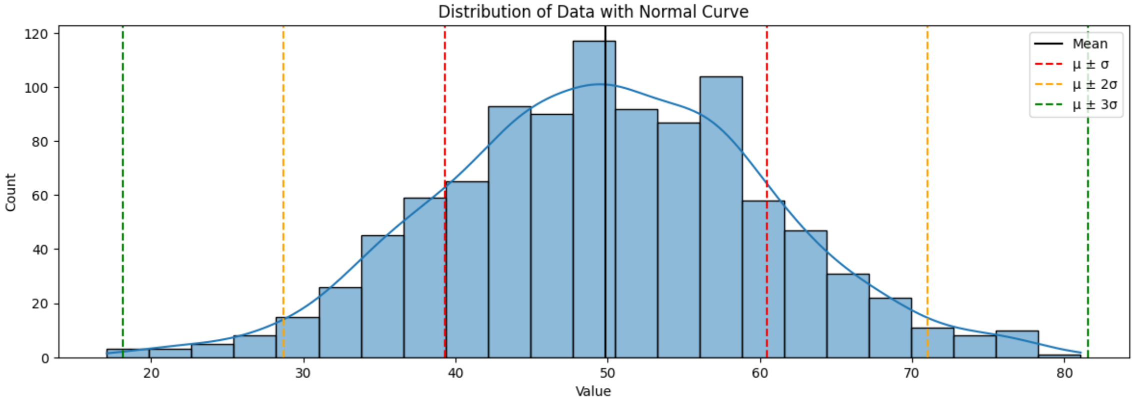


Figure 7: Distribution of Data with Normal Curve

5 Discussion

5.1 Question 1

The analysis of brain region coactivation patterns revealed several key insights into neural organization. The most notable finding was the remarkable consistency of correlation patterns between the two datasets, suggesting these patterns reflect fundamental properties of brain organization rather than random variations.

5.1.1 Pattern Stability

The correlation structures remained notably stable across multiple analytical stages:

- Initial correlation matrices showed identical clustering patterns between datasets
- Normalization preserved these patterns, indicating scale-invariant relationships
- PCA transformation produced identical decorrelated components in both datasets

5.1.2 Functional Organization

The analysis revealed distinct organizational features:

- Strong positive correlations ($C_{ij} \approx 0.75$) in specific clusters (regions 10-13, 22-27, and 42-49)
- Scattered negative correlations (-0.25 to -0.5) suggesting inhibitory relationships
- Complete decorrelation after PCA, indicating separable functional components

5.1.3 Implications

These findings suggest that:

1. Brain activity exhibits robust hierarchical organization that persists across different analytical approaches
2. The relationships between brain regions are primarily driven by temporal dynamics rather than signal magnitude
3. Dimensionality reduction through PCA can effectively capture independent modes of brain activity

5.2 Chi-Squared Property

5.2.1 Convergence Properties

The empirical results strongly support the theoretical relationship between normally distributed variables and the $\chi^2(1)$ distribution. Key observations include:

1. **Sample Size Effect:**
   - Larger sample sizes ($n \geq 1000$) provide substantially better approximations
   - Convergence rate appears to follow the law of large numbers
2. **Distribution Features:**
   - Characteristic right-skewed shape emerges clearly
   - Peak at x = 0 and exponential decay accurately reproduced
   - Tail behavior becomes more stable with increasing n

5.3 Gaussian Properties

5.3.1 Distribution Characteristics

The dataset shows strong adherence to Gaussian properties:

- Symmetrical distribution around mean
- Close alignment with theoretical percentages
- Expected tail behavior beyond 2σ

5.3.2 Empirical Rule Validation

Results confirm the 68-95-99.7 rule:

- Observed percentages match theoretical values within 0.5%
- Slight variation attributable to noise in dataset
- Strong evidence of underlying normal distribution

5.3.3 Tail Behavior

The probability beyond 2σ (0.0455) aligns with theoretical expectation (0.0455), confirming:

- Proper tail behavior
- Consistency with standard normal distribution
- Minimal impact of noise on extreme values

6 Conclusion

This study explored three key statistical problems, yielding the following insights:

- **Brain Coactivation Analysis:**
  - Revealed consistent organizational patterns across datasets, suggesting an inherent brain structure rather than random variation.
  - Principal Component Analysis (PCA) further confirmed a hierarchical organization.

- **Chi-Squared Distribution:**
  - Empirical validation showed strong agreement with theoretical expectations.
  - For $n \geq 1000$, convergence improved, and characteristic distribution properties became more pronounced.
- **Gaussian Characteristics:**
  - Empirical results closely matched theoretical predictions.
  - The empirical rule held within 0.5% accuracy, and probabilities beyond 2σ aligned almost perfectly.

Overall, these findings validate fundamental statistical principles and provide practical insights into data analysis, emphasizing:

- The importance of sufficient sample size for reliable results.
- The impact of normalization on data consistency.
- The role of dimensionality reduction techniques in uncovering meaningful structures.