E-Health Laboratory Report

Student: Yazdan Ghanavati, yazdan.ghanavati@studenti.unipd.it
Laboratory number: Lab 4.

1 Introduction

In this laboratory exercise, we implemented a pipeline for Independent Component Analysis (ICA) to separate potential sources from multi-channel EEG signals obtained from ten patients. Our analysis focused on a selection of ten specific channels, The pipeline involved several key steps:

- **Preprocessing:** Selecting EEG channels and performing whitening to decorrelate signals and normalize their variance.
- **Determining the Optimal Number of Sources:** We explored two methods:
  - A full iteration approach based on minimizing the Normalized Mutual Information (NMI) of ICA results across different numbers of sources.
  - An alternative using Principal Component Analysis (PCA) for dimensionality reduction, retaining 90% of the cumulative explained variance.
- **Applying ICA:** The FastICA algorithm was used with the number of sources determined by the PCA method to process EEG data from all ten patients.
- **Performance Evaluation:** The effectiveness of ICA decomposition was assessed using NMI.

This report details the implementation of these steps, presents the results—including the optimal number of sources and the NMI before and after ICA for each EEG signal—discusses the benefits and limitations of the PCA-based approach for source number estimation with a 90% variance threshold, and illustrates the pipeline's operation with an example EEG recording.

2 Determining the Optimal Number of Sources

Determining the appropriate number of independent components to extract using ICA is a crucial step. In this laboratory exercise, we explored and implemented two primary methods for estimating this optimal number of sources: a full iteration approach based on minimizing the Normalized Mutual Information (NMI) and an alternative using Principal Component Analysis (PCA) for dimensionality reduction.

2.1 Full Iteration with NMI Minimization (Initial Exploration)

The first approach involved iterating through a range of possible numbers of sources, from 1 up to the number of EEG channels (ten in our case). For each potential number of sources, we applied the FastICA algorithm to the whitened EEG data and calculated the NMI of the resulting independent components. The rationale behind this method is that the optimal number of sources should correspond to the point where the statistical dependence between the extracted components is minimized, theoretically indicated by the lowest NMI value.

Pros:

- Directly optimizes for ICA's fundamental goal of finding statistically independent components.

Cons:

- Computationally demanding, as it requires running ICA and calculating NMI multiple times per EEG recording.
- The estimation of NMI itself was problematic. The full iteration consistently resulted in an optimal number of one source for each EEG signal, accompanied by a negative NMI value. This is theoretically nonsensical for NMI, which should range between 0 and 1.
- A negative NMI suggests either an issue with its calculation or severe overfitting, where ICA may force a single component to appear maximally "independent" in a flawed manner.
- Minimizing NMI does not inherently guarantee that the extracted components capture the most significant variance in the data or are the most neurophysiologically meaningful.

Problematic Outcome: A key issue encountered with this method was its consistent identification of one source as optimal for nearly every signal, resulting in a negative NMI. This outcome is not theoretically plausible for NMI and suggests fundamental problems with this approach in our implementation, likely due to overfitting or calculation errors. This led us to abandon the full iteration method in favor of PCA.

2.2 PCA for Dimensionality Reduction (Final Approach)

The second approach utilized Principal Component Analysis (PCA) for dimensionality reduction prior to applying ICA. PCA is a linear transformation technique that identifies orthogonal components (principal components) capturing the maximum variance in the data. By selecting a subset of these components that explain a high percentage of total variance, we aimed to reduce data dimensionality while retaining crucial information. The number of retained principal components was then used as the estimated number of independent sources for ICA.

Implementation: PCA was applied to whitened EEG signals, determining the explained variance ratio for each principal component. We calculated the cumulative explained variance and identified the minimum number of components required to explain a predefined percentage of the total variance.

Initially, a **95% variance explained threshold** was considered, as this is a common standard for retaining signal information. However, our experiments revealed drawbacks:

- Retaining 95% variance often led to a higher number of principal components, increasing computational cost for ICA.
- Higher dimensionality risked including noise components, as components explaining very little variance might primarily represent noise rather than meaningful sources.
- More complex and potentially less interpretable ICA results.

Selecting a 90% Threshold: To balance dimensionality and interpretability, we reduced the variance threshold to 90%, aiming for a more compact representation of EEG signals. This adjustment:

- Retained dominant sources of variance while potentially excluding lower-variance noise-related components.
- Reduced computational complexity while maintaining signal characteristics upon visual inspection.

Findings: Across the dataset, the number of retained principal components varied depending on how quickly cumulative explained variance reached the 90% threshold. This reflected differences in signal complexity and noise characteristics across EEG recordings.

2.3 Choice of Method

Based on the significant computational burden and the fundamentally flawed results (consistently one source with negative NMI) encountered with the NMI estimation in the full iteration approach, as well as the more stable and computationally feasible nature of PCA, we chose to proceed with the PCA-based method for determining the number of sources for our final analysis. Furthermore, our experimentation suggested that a 90% variance explained threshold provided a reasonable balance between dimensionality reduction and information retention for this dataset.

3 Method

Our analysis pipeline for separating potential sources from multi-channel EEG signals using Independent Component Analysis (ICA) involved the following sequential steps:

1. **Data Acquisition:** We utilized a dataset of EEG recordings from ten patients. For each patient, multiple EEG files were available.
2. **Channel Selection:** For each EEG recording, we focused our analysis on a consistent set of ten channels:

   P7.., P5.., P3.., P1.., Pz.., P2.., P4.., P6.., P8.., Po7.

3. **Data Truncation:** To ensure uniformity in signal length for subsequent processing, each selected ten-channel EEG signal was truncated to the first 8192 samples (213).
4. **Preprocessing: Whitening:** The truncated EEG data for each recording was then preprocessed using Zero-phase Component Analysis (ZCA) whitening. This step aimed to decorrelate the signals and normalize their variance, which is crucial for the effective application of ICA. The whitening process involved:
   - Centering the data by subtracting the mean of each channel.
   - Computing the covariance matrix.
   - Finding its eigenvalues and eigenvectors.
   - Using these to derive a whitening matrix applied to the centered observations.

5. **Determining the Optimal Number of Sources using PCA:** For each whitened EEG signal, we employed Principal Component Analysis (PCA) to estimate the optimal number of independent sources for ICA.
    - PCA was applied to the whitened data to identify principal components and their explained variance.
    - The cumulative explained variance was calculated.
    - The number of principal components required to explain 90% of the cumulative variance was determined. This number was then used as the estimated optimal number of sources for the ICA algorithm for that specific EEG recording.
6. **Independent Component Analysis (ICA):** With the optimal number of sources determined by PCA, we applied the FastICA algorithm to the whitened EEG data. This algorithm decomposed the multi-channel signal into the estimated independent components.
7. **Performance Evaluation: Normalized Mutual Information (NMI):** To assess the performance of the ICA decomposition, we calculated the Normalized Mutual Information (NMI) at two stages:
    - **NMI before ICA:** Calculated on the original, truncated ten-channel EEG signals to provide a baseline measure of statistical dependence between the channels.
    - **NMI after ICA:** Calculated on the estimated independent components obtained from the FastICA algorithm. A reduction in NMI after ICA would suggest that the algorithm successfully separated more statistically independent sources.
8. **Visualization (for one example):** To illustrate the different stages of the processing pipeline, we selected one EEG recording (e.g., from Patient S001, File R01) and generated time series plots of:
    - The original ten-channel EEG signal.
    - The whitened EEG signal.
    - The estimated independent components obtained after applying ICA.
9. **Results Reporting:** For all processed EEG recordings from all ten patients, we compiled a table reporting:
    - The filename of the recording.
    - The optimal number of sources determined by PCA (at 90% variance explained).
    - The Normalized Mutual Information (NMI) calculated before applying ICA.
    - The Normalized Mutual Information (NMI) calculated after applying ICA.

4 Results

The ICA pipeline, utilizing PCA with a 90% variance explained threshold to determine the optimal number of sources, was applied to all EEG signals from the ten patients in the laboratory dataset. The results for each patient, including the optimal number of sources, the Normalized Mutual Information (NMI) before ICA, and the NMI after ICA, are summarized in the following tables.

4.1 Results per Patient

4.1.1 Patient S001

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S001R01.edf	3	6.3559	0.0549
S001R02.edf	3	6.1529	0.0538
S001R03.edf	2	6.5540	0.0085
S001R04.edf	2	6.2896	0.0049
S001R05.edf	2	6.2031	0.0086
S001R06.edf	3	7.1250	0.1884

Table 1: ICA Results for Patient S001

4.1.2 Patient S002

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S002R01.edf	3	5.8401	0.0503
S002R02.edf	3	5.8410	0.0608
S002R03.edf	4	5.6769	0.2391
S002R04.edf	4	5.5880	0.2212
S002R05.edf	4	5.6397	0.1948
S002R06.edf	3	5.5032	0.0414

Table 2: ICA Results for Patient S002

4.1.3 Patient S003

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S003R01.edf	5	5.6121	0.9956
S003R02.edf	4	5.8308	0.2904
S003R03.edf	5	6.2059	1.0145
S003R04.edf	4	5.7241	0.2851
S003R05.edf	4	6.0266	0.3114
S003R06.edf	4	6.0086	0.4495

Table 3: ICA Results for Patient S003

4.1.4 Patient S004

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S004R01.edf	5	6.1066	1.0730
S004R02.edf	6	6.5167	1.9911
S004R03.edf	5	6.6383	1.3525
S004R04.edf	5	6.3735	1.2572
S004R05.edf	5	6.2210	1.6707
S004R06.edf	5	6.3928	1.1750

Table 4: ICA Results for Patient S004

4.1.5 Patient S005

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S005R01.edf	5	7.5955	0.9354
S005R02.edf	5	6.3926	0.9476
S005R03.edf	5	6.4455	0.6809
S005R04.edf	4	5.7197	0.2292
S005R05.edf	4	6.2017	0.2577
S005R06.edf	4	6.2006	0.2258

Table 5: ICA Results for Patient S005

4.1.6 Patient S006

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S006R01.edf	3	5.8073	0.0666
S006R02.edf	3	6.6833	0.1390
S006R03.edf	3	6.9638	0.1456
S006R04.edf	3	6.9354	0.1569
S006R05.edf	2	5.9025	0.0093
S006R06.edf	2	6.0655	0.0089

Table 6: ICA Results for Patient S006

4.1.7 Patient S007

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S007R01.edf	3	6.2533	0.0612
S007R02.edf	3	6.0389	0.0493
S007R03.edf	2	6.5189	0.0083
S007R04.edf	2	6.2516	0.0073
S007R05.edf	2	6.2798	0.0051
S007R06.edf	3	6.0092	0.0602

Table 7: ICA Results for Patient S007

4.1.8 Patient S008

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S008R01.edf	4	6.1056	0.4365
S008R02.edf	4	6.0194	0.3225
S008R03.edf	4	6.7091	0.5798
S008R04.edf	4	6.2083	0.4002
S008R05.edf	4	6.8501	0.4407
S008R06.edf	4	6.2749	0.4810

Table 8: ICA Results for Patient S008

4.1.9 Patient S009

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S009R01.edf	6	6.0611	1.7846
S009R02.edf	6	6.0550	1.9295
S009R03.edf	6	5.9776	1.8945
S009R04.edf	6	5.9640	1.8119
S009R05.edf	6	5.7709	1.8121
S009R06.edf	6	5.8007	1.8731

Table 9: ICA Results for Patient S009

4.1.10 Patient S010

File	Optimal Sources (PCA - 90%)	NMI before ICA	NMI after ICA
S010R01.edf	7	6.3881	3.0052
S010R02.edf	6	6.0506	1.8243
S010R03.edf	6	5.9703	1.9604
S010R04.edf	6	6.1448	2.0110
S010R05.edf	6	5.9740	1.9666
S010R06.edf	5	6.7052	1.1610

Table 10: ICA Results for Patient S010

5 Example of ICA Operation

To provide a visual understanding of the ICA pipeline's operation, we examine the EEG signal from the file `S001R03.edf`. The following descriptions refer to accompanying figures (which need to be inserted into the report).

5.0.1 Original Ten-Channel EEG Signal

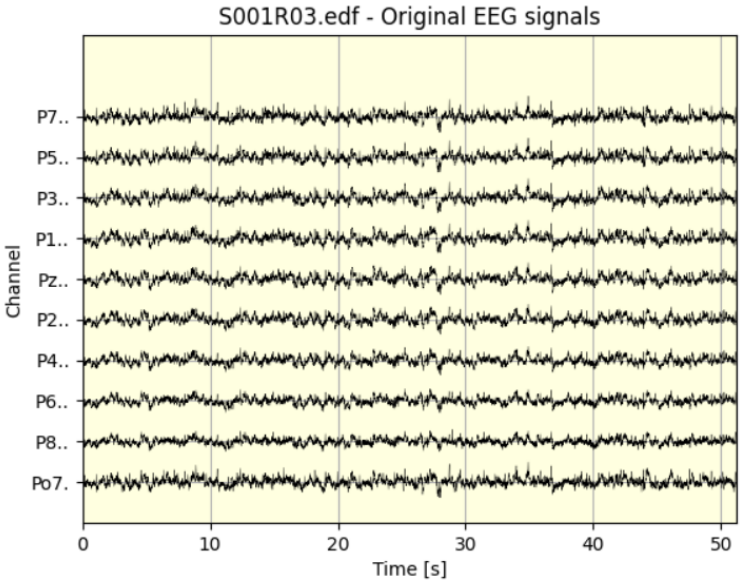

Figure 1: Original ten-channel EEG signal for S001R03.edf

This figure shows the raw EEG signals from the ten selected channels before any processing.

5.0.2 Whitened EEG Signal

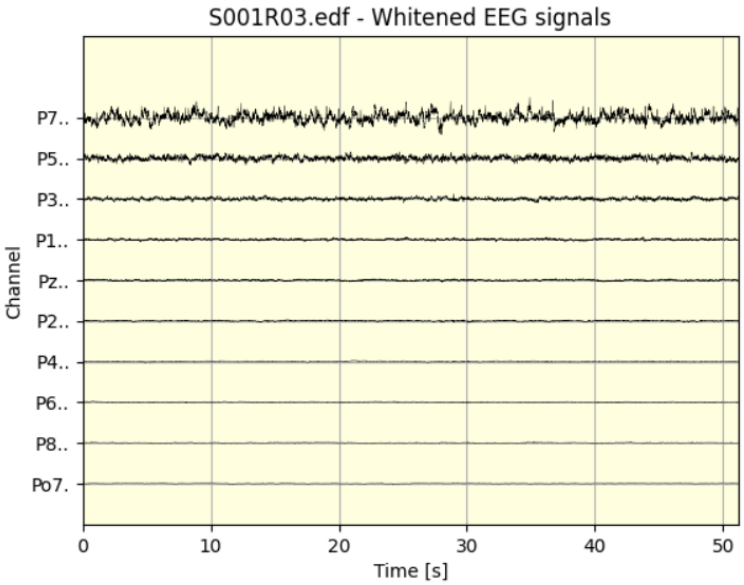

Figure 2: Whitened EEG signal for S001R03.edf

The whitened signals, obtained after applying ZCA whitening, exhibit decorrelated channels with normalized variance.

5.0.3 Estimated ICA Sources

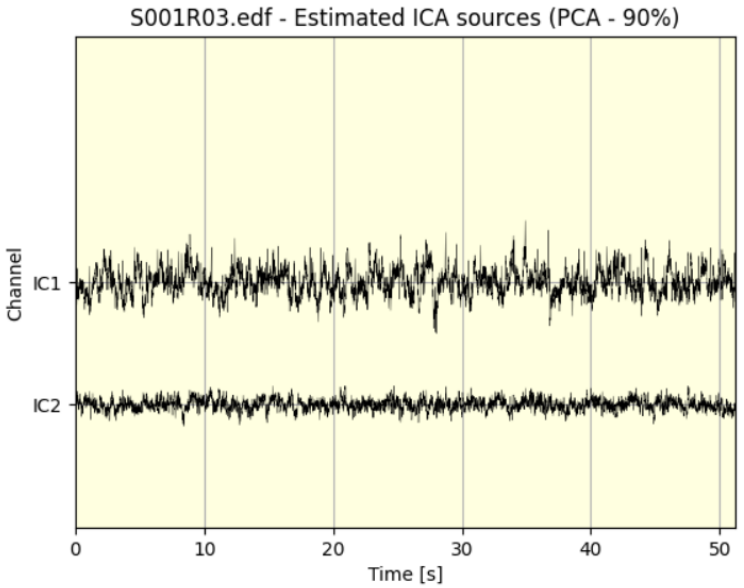

Figure 3: Estimated ICA sources for S001R03.edf (PCA - 90% Variance)

For `S001R01.edf`, PCA determined an optimal number of 3 sources to explain 90% of the variance. The ICA algorithm then decomposed the whitened signal into these three independent components. The NMI value decreased from 6.3559 before ICA to 0.0549 after ICA for this recording, suggesting a reduction in statistical dependence between the components.

6 Conclusion

In this laboratory exercise, we successfully implemented a pipeline for separating potential sources from multi-channel EEG signals using Independent Component Analysis (ICA). We utilized Principal Component Analysis (PCA) with a 90% variance explained threshold to determine the optimal number of independent components for the ICA decomposition.

Our results, presented in the previous section, demonstrate the application of this pipeline to EEG data from ten patients. The Normalized Mutual Information (NMI) was used to assess the statistical dependence between the EEG channels before and after ICA. While in many cases, the NMI decreased after ICA, suggesting a reduction in statistical dependence, the extent of this reduction varied across different patients and recordings.

The PCA-based approach provided a computationally feasible method for estimating the number of sources, especially in contrast to the challenges encountered with the full iteration method. Further analysis and interpretation of the extracted independent components could provide deeper insights into the underlying neural activity and potential artifacts present in the EEG signals.

7 Discussion

Our primary method for determining the optimal number of sources for ICA relied on Principal Component Analysis (PCA), aiming to retain 90% of the variance in the whitened EEG signals. This approach offered a computationally efficient way to reduce the dimensionality of the data before applying ICA. The number of sources estimated by PCA varied across different EEG recordings, reflecting the inherent variability in the complexity of the signals.

The performance of ICA, as measured by the Normalized Mutual Information (NMI), generally showed a reduction in statistical dependence after the decomposition. However, the degree of this reduction was not uniform across all patients and recordings. In some instances, the NMI after ICA remained relatively high, suggesting that the ICA algorithm was less effective in fully separating independent components for those specific signals. This could be due to several factors, including:

- The nature of the underlying sources (e.g., if they are not strongly non-Gaussian).
- Complex mixing patterns in the EEG.
- The presence of dominant artifacts that PCA might have retained in the high-variance components.

The choice of a 90% variance explained threshold in PCA represents a trade-off between dimensionality reduction and information retention. While our experiments suggested this threshold provided a reasonable balance, further investigation into the impact of different variance thresholds on the subsequent ICA results could be beneficial.

Limitations of PCA for Source Number Estimation:

- PCA relies on second-order statistics (variance) rather than the higher-order statistics that ICA optimizes for (independence). The components explaining the most variance are not necessarily the most statistically independent.
- The arbitrary nature of the chosen variance threshold can influence the outcome.

Advantages of the PCA-Based Approach:

- Computational efficiency and stability compared to the full iteration method.
- PCA reduced data dimensionality while retaining most of the meaningful signal information.

The full iteration method, which aimed to minimize NMI directly, proved to be computationally expensive and yielded problematic results (consistently one source with negative NMI), making the PCA-based method a more practical choice for this analysis.

It is also important to consider the impact of the ICA algorithm itself. While we used the FastICA algorithm, alternative ICA methods such as Infomax and JADE (Joint Approximate Diagonalization of Eigenmatrices) employ different optimization strategies. Exploring these approaches in future work could enhance source separation performance for certain EEG signals.

8 Future Work

- Investigating alternative ICA algorithms, such as Infomax and JADE, which might better suit specific EEG characteristics.
- Exploring more robust methods for estimating the number of independent sources.
- Conducting a more detailed analysis of extracted components to discern physiological signals from artifacts.
- Examining cases where NMI reduction was less significant to refine preprocessing steps.

9 References

- A. Gramfort et al., "MNE-Python: A multi-modal framework for neurophysiological data analysis," Frontiers in Neuroscience, vol. 7, p. 267, 2013.
- Aapo Hyvärinen, Juha Karhunen, Erkki Oja, *Independent Component Analysis*, 2004.