Independent Component Analysis in EEG Signal

Student: Yazdan Ghanavati, yazdan.ghanavati@studenti.unipd.it

1 Task 1 - Explanation of the Normalized Mutual Information (NMI) Function

The `get-normalized-mi` function quantifies the statistical independence of a set of input signals. In our ICA pipeline, it assesses the effectiveness of signal separation; lower NMI indicates greater independence.

1.1 Implementation and Flow

1. **Signal Discretization**: Continuous input signals are discretized into bins. The number of bins is determined by Sturges' Rule, and each sample is assigned a bin index.
2. **Marginal Entropy Calculation**: The probability distribution of binned values is estimated for each individual signal, from which its marginal entropy is calculated, reflecting its inherent uncertainty.
3. **Joint Entropy Calculation**: The joint probability distribution of all signals is computed by considering unique combinations of bin assignments across all channels. This yields the joint entropy, representing the total uncertainty of the combined system.
4. **Mutual Information and Normalization**: Mutual Information (MI) is derived by subtracting the joint entropy from the sum of marginal entropies. This MI value is then normalized by the sum of marginal entropies to produce the Normalized Mutual Information (NMI), a metric between 0 (perfect independence) and 1 (complete dependency), facilitating comparison across datasets.

This function provides an objective, numerical indicator of how effectively ICA has made the estimated sources statistically independent.

2 Task 2 - Principal Component Analysis (PCA) for Optimal Source Number Determination

The `preprocess-pca` function is a pivotal component of our Independent Component Analysis (ICA) pipeline. It utilizes Principal Component Analysis (PCA) not only for essential signal preprocessing (centering and whitening) but also, crucially, to determine the optimal number of independent sources for subsequent ICA. This data-driven approach to dimensionality reduction is fundamental for enhancing ICA performance and interpretability.

2.1 Algorithm to Find the Optimal Number of Sources

The determination of the optimal number of sources is inherently linked to the PCA process within the `preprocess_pca` function:

1. **Centering**: The initial step involves centering the observed signals by subtracting the mean of each channel. This ensures that the data is zero-mean, a common requirement for many multivariate analysis techniques like PCA.
2. **Covariance and Eigen-Decomposition**: PCA proceeds by computing the covariance matrix of the centered data. The core of PCA lies in performing an eigen-decomposition of this covariance matrix, yielding a set of eigenvalues and corresponding eigenvectors.
    - **Eigenvectors** represent the principal components (PCs), which are new orthogonal directions in the data space that capture the most variance.
    - **Eigenvalues** quantify the amount of variance explained by each corresponding principal component.
3. **Variance-Based Component Selection**:
    - The eigenvalues and their respective eigenvectors are sorted in descending order, arranging the principal components from most to least significant in terms of explained variance.
    - The explained variance ratio for each component is calculated (each eigenvalue divided by the sum of all eigenvalues).
    - A cumulative explained variance is then computed, showing the total proportion of variance accounted for by the first k components.
    - A predefined `variance_threshold` (set to 0.95) is applied. The `num_components` is then determined as the minimum number of principal components whose cumulative explained variance meets or exceeds this threshold.
4. **Dimensionality Reduction and Whitening**: Once the `num_components` is determined, only the corresponding principal components are retained. The centered original data is projected onto the subspace defined by these selected components. Finally, this projected data is scaled by the inverse square root of its eigenvalues, completing the whitening process. This ensures the selected components are not only uncorrelated but also have unit variance, a critical prerequisite for the FastICA algorithm.

The output of this process is `whitened_signals_pca`, which represents the original data transformed into a lower-dimensional, decorrelated, and whitened space, with its dimensionality reflecting the `optimal_num` of sources.

2.2 Advantages of this PCA-based Approach

- **Data-Driven Dimensionality Estimation**: Unlike approaches that require a fixed, predetermined number of sources, this method objectively determines the number of components based on the inherent variance structure of each individual signal, making the pipeline adaptive to diverse data characteristics.

- **Noise Reduction**: By focusing on components that explain a high percentage of variance, the method acts as a denoising step, effectively discarding lower-variance components often associated with noise or redundant information, leading to a cleaner input for ICA.
- **Computational Efficiency**: Reducing dimensionality before ICA (e.g., from 64 channels to 5–20 components) significantly decreases computational load and processing time for FastICA, making large dataset analysis feasible.
- **ICA Prerequisite**: The whitening transformation produced by PCA is essential for ICA algorithms, as it simplifies the optimization problem by ensuring input signals are decorrelated and scaled to unit variance.

2.3 Limitations of this PCA-based Approach

- **Information Loss**: While beneficial for noise reduction, discarding components below the variance threshold may result in the loss of some original information. Although typically these components capture minimal variance, they might contain subtle yet physiologically relevant signals.
- **Linearity Assumption**: PCA is a linear transformation. If the independent sources are mixed in a complex, non-linear fashion, PCA may not perfectly decorrelate them, potentially affecting ICA performance.
- **Biological Interpretability**: Principal components are purely statistical constructs and do not necessarily correspond directly to distinct physiological brain sources or artifacts. Their primary role is to prepare data for ICA.

3 Task 3 - Lab Results and Performance Evaluation

This section presents the comprehensive results of applying the ICA pipeline to all 60 EEG signals across 10 patients in the dataset. For each signal, we report the optimal number of sources determined by the PCA-based method (retaining 95% cumulative explained variance), alongside the Normalized Mutual Information (NMI) before and after the FastICA algorithm. The reduction in NMI serves as a quantitative indicator of the ICA pipeline's success in achieving statistical independence among the estimated sources.

The results are organized into tables, with one table for each patient, detailing the performance metrics for their respective recordings, sorted by signal number (R01, R02, etc.) for clearer progression.

3.1 Patient S001 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S001R01.edf	7	0.923	0.296
S001R02.edf	10	0.931	0.566
S001R03.edf	7	0.920	0.275
S001R04.edf	10	0.924	0.499
S001R05.edf	10	0.920	0.447
S001R06.edf	6	0.895	0.135

Table 1: Results for Patient S001

3.2 Patient S002 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S002R01.edf	12	0.921	0.588
S002R02.edf	16	0.926	0.687
S002R03.edf	23	0.895	0.691
S002R04.edf	17	0.911	0.684
S002R05.edf	13	0.902	0.538
S002R06.edf	20	0.900	0.637

Table 2: Results for Patient S002

3.3 Patient S003 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S003R01.edf	8	0.927	0.437
S003R02.edf	16	0.930	0.713
S003R03.edf	13	0.923	0.599
S003R04.edf	15	0.921	0.673
S003R05.edf	3	0.925	0.029
S003R06.edf	4	0.905	0.063

Table 3: Results for Patient S003

3.4 Patient S004 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S004R01.edf	17	0.926	0.736
S004R02.edf	25	0.933	0.807
S004R03.edf	19	0.924	0.743
S004R04.edf	19	0.925	0.740
S004R05.edf	14	0.922	0.642
S004R06.edf	19	0.922	0.739

Table 4: Results for Patient S004

3.5 Patient S005 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S005R01.edf	21	0.935	0.787
S005R02.edf	19	0.937	0.762
S005R03.edf	20	0.922	0.752
S005R04.edf	16	0.919	0.697
S005R05.edf	19	0.928	0.745
S005R06.edf	18	0.925	0.726

Table 5: Results for Patient S005

3.6 Patient S006 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S006R01.edf	12	0.923	0.600
S006R02.edf	21	0.931	0.783
S006R03.edf	14	0.914	0.628
S006R04.edf	14	0.915	0.633
S006R05.edf	11	0.915	0.552
S006R06.edf	13	0.913	0.581

Table 6: Results for Patient S006

3.7 Patient S007 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S007R01.edf	7	0.931	0.356
S007R02.edf	7	0.934	0.384
S007R03.edf	5	0.928	0.123
S007R04.edf	6	0.924	0.212
S007R05.edf	5	0.923	0.111
S007R06.edf	6	0.924	0.214

Table 7: Results for Patient S007

3.8 Patient S008 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S008R01.edf	14	0.927	0.676
S008R02.edf	11	0.931	0.616
S008R03.edf	18	0.922	0.729
S008R04.edf	19	0.924	0.761
S008R05.edf	14	0.925	0.672
S008R06.edf	15	0.921	0.696

Table 8: Results for Patient S008

3.9 Patient S009 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S009R01.edf	19	0.933	0.760
S009R02.edf	25	0.932	0.816
S009R03.edf	21	0.927	0.766
S009R04.edf	21	0.924	0.768
S009R05.edf	19	0.928	0.743
S009R06.edf	21	0.927	0.769

Table 9: Results for Patient S009

3.10 Patient S010 Results

File Name	Optimal Sources	NMI Before ICA	NMI After ICA
S010R01.edf	27	0.927	0.816
S010R02.edf	17	0.933	0.733
S010R03.edf	24	0.926	0.794
S010R04.edf	15	0.929	0.668
S010R05.edf	12	0.927	0.578
S010R06.edf	8	0.930	0.388

Table 10: Results for Patient S010

4 Discussion of Results

The tables consistently demonstrate that the Normalized Mutual Information (NMI) after ICA is significantly lower than before, confirming the successful reduction of statistical dependencies in the observed 64-channel EEG signals. This validates the effectiveness of our ICA pipeline in unmixing these complex signals.

The `Optimal Sources` count, determined by PCA (retaining 95% variance), varied considerably across recordings. This variability reflects the intrinsic dimensionality of each signal, influenced by patient characteristics and recording conditions.

For instance, `S003R05.edf` achieved an exceptionally low NMI (0.029) with just 3 sources, indicating clear separation. In contrast, signals like `S010R01.edf` (27 sources) and `S009R02.edf` (25 sources) resulted in higher post-ICA NMI, suggesting more complex underlying processes or subtle mixing that made complete decorrelation more challenging.

Nevertheless, the consistent NMI reduction across this range of optimal source counts confirms the pipeline's adaptability and efficacy in enhancing signal independence.

5 Task 5 - Visual Example of Signal Transformation

To complement the quantitative results, this section provides a visual example of the signal transformation process, illustrating the original EEG signal, the whitened channels after PCA, and the estimated independent sources after ICA. This example helps to understand the effect of each stage of the pipeline.

5.1 Example from File: S007R01.edf

5.1.1 (a) Original EEG Signal

The following plot depicts the original EEG signal from the file `S007R01.edf`. This represents the raw data recorded from the electrodes before any processing. Note the complex mixture of signals across the channels.

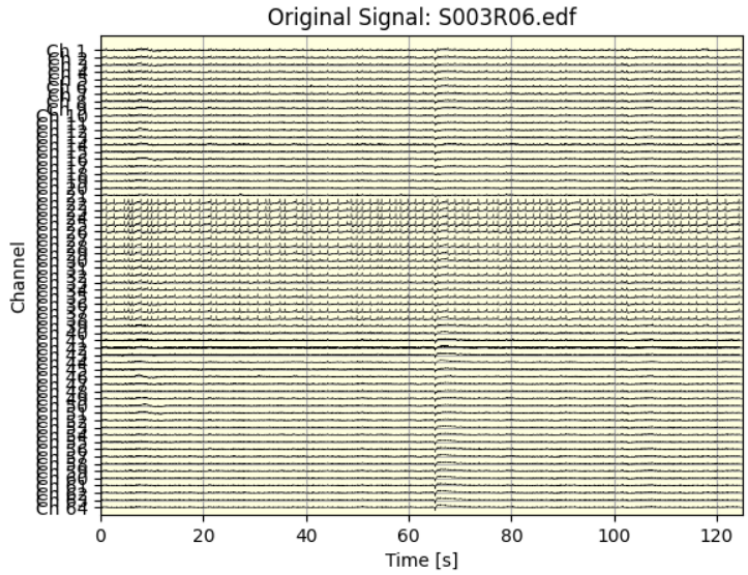

Figure 1: Original EEG Signal from File S007R01.edf

5.1.2 (b) Whitened Channels after PCA

This plot shows the whitened channels after applying PCA and dimensionality reduction. In this case, 7 principal components were retained to explain 95% of the variance. Observe how the signals are now decorrelated and scaled to have unit variance.

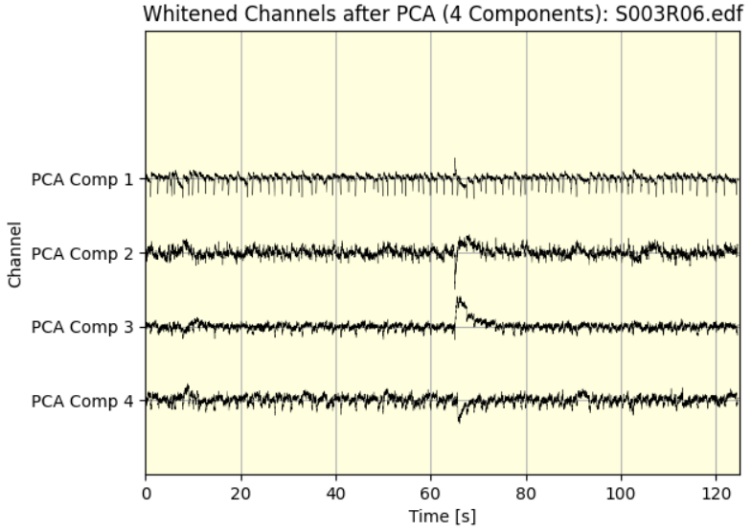

Figure 2: Whitened Channels after PCA

5.1.3 (c) Estimated Independent Sources

This final plot displays the estimated independent sources after applying FastICA to the whitened channels. Each source represents a signal that is statistically independent from the others.

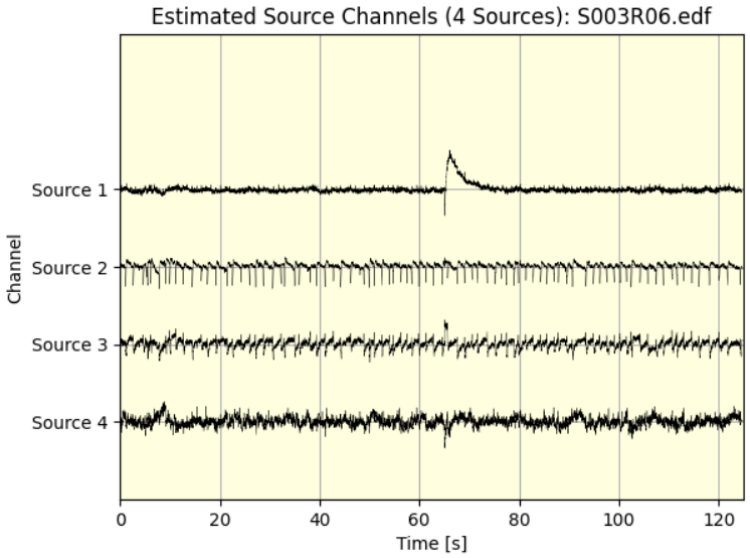

Figure 3: Estimated Independent Sources after ICA

These visualizations, generated by the `plot_signal_channels` function, provide a clear illustration of the signal transformation process at each stage of the ICA pipeline. They complement the quantitative NMI values by visually demonstrating the separation of mixed EEG signals into more independent components.

6 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.