Contents lists available at ScienceDirect

Data in Brief

Data Article

A visual working memory dataset collection with bootstrap Independent Component Analysis for comparison of electroencephalographic preprocessing pipelines

Fiorenzo Artoni [a,b,*], Arnaud Delorme [c,d,1], Scott Makeig [c,1]

[a] *The Biorobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy*
[b] *Translational Neural Engineering Laboratory, Center for Neuroprosthetics and Institute of Bioengineering, EPFL – Campus Biotech, Geneva, Switzerland*
[c] *Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093-0559, USA*
[d] *Univ. Grenoble Alpes, CNRS, LNPC UMR 5105, Grenoble, France*

ARTICLE INFO

ABSTRACT

Here we present an electroencephalographic (EEG) collection of 71-channel datasets recorded from 14 subjects (7 males, 7 females, aged 20–40 years) while performing a visual working memory task with a T set of 150 Independent Component Analysis (ICA) decompositions by Extended Infomax using RELICA, each on a bootstrap resampling of the data. These data are linked to the paper "Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition" [1]. Independent components (ICs) are clustered within subject and thereby associated with a quality index (QIc) measure of their stability to data resampling. Sets of single ICA decompositions obtained after applying Principal Component Analysis (PCA) to the data to perform dimension reduction retaining (85%, 95%, 99%) of data variance are also included, as are the positions of the best fitting equivalent dipoles for ICs whose scalp projections are compatible

---

with a compact brain source. These bootstrap ICs may be used as benchmarks for different data preprocessing pipelines and/or ICA algorithms, allowing investigation of the effects that noise or insufficient data have on the quality of ICA decompositions.

Specifications table

Subject area	*Neuroscience, Data science*
More specific subject area	*Non-invasive Brain Imaging, Independent Component Analysis*
Type of data	*Figures, online data*
How data was acquired	*Data were collected from 71 channels (69 scalp and two periocular electrodes) at a sampling rate of 250 Hz with an analog pass band of 0.01 to 100 Hz (SA Instrumentation, San Diego) with input impedances under 5 kΩ after careful scalp preparation*
	Matlab (RRID:SCR_00162) and EEGLAB (RRID:SCR_007292) were the tools to analyze the data
Data format	*Raw, Preprocessed and Partially Analyzed*
Experimental factors	*14 subjects, 7 males, 7 females, sitting comfortably, relaxed (i.e., avoiding excessive neck muscle contraction), watching a screen*
Experimental features	*Modified Sternberg visual working memory task. For details see (Onton and Makeig, "Frontal midline EEG dynamics during working memory." NeuroImage 27:341 = 56, 2005)*
Data source location	*San Diego, USA*
Data accessibility	*Data are available with this article and online at* https://data.mendeley.com/datasets/gvt64rndnd/1 doi:10.17632/gvt64rndnd.1
Related research article	F. Artoni, A. Delorme, and S. Makeig, "Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition," *NeuroImage,* vol. 175, pp. 176–187, 2018 [1].

Value of the data

- Evaluate the effects that different preprocessing steps have on the quality of subsequent Independent Component Analysis (ICA) decompositions. Test the effects of different filtering (high-pass, low-pass) cutoff frequencies and algorithms for automatic removal of artifacts.
- The data can be used to test the quality of decompositions produced by other Blind Source Separation (BSS) algorithms. Researchers can compare the numbers of dipolar components (i.e., ICs whose scalp maps are fitted by an equivalent dipole with low residual variance) and the positions of the localized sources returned by different BSS algorithms as in [2]. Note: Analysis scripts in Matlab (The Mathworks, Inc.) applied in that paper to these datasets are available at http://sccn.ucsd.edu/eeglab/BSSComparison/
- The data can be used to compare the effects that dimensionality reduction by PCA has on subsequent decomposition by ICA [1] with that on other BSS algorithms. To this aim, the dataset includes Extended Infomax ICA decompositions obtained after applying Principal Component Analysis (PCA) with various values of retained variance (85%, 95%, 99%). The data include a full PCA decomposition (PCA-Only) that can be used to replicate the results in [1] with different amounts of retained variance or using different methods for dimensionality reduction (e.g., linear discriminant analysis – LDA, canonical correlation analysis – CCA)

- The data can also be useful to test the effects that sampling frequency (e.g., after downsampling to 50 Hz or oversampling to 1000 Hz) or channel number (e.g., by reducing the number of channels to 32 or 16) have on the quality of the ICA decompositions.

1. Data

The data are a collection of 14 electroencephalographic (EEG) datasets recorded from 14 subjects (7 males, 7 females, aged 20–40 years) performing a visual working memory task. Subjects first memorized a set of letters, then a set of probe letters was presented and participants had to press a button to indicate whether the letter was present in the memorized sequence or not. An audio cue gave them feedback on the correctness of their answer. Each subject performed 100 – 150 trials. Raw data are available at http://sccn.ucsd.edu/eeglab/BSSComparison/ with an analysis script in Matlab to compare results of linear decomposition with 18 other blind source separation methods [2]. Here the data are reformatted to include:

- PCA applied directly on the data (PCA-Only);
- Extended Infomax ICA applied after PCA with 85%, 95%, 99% retained variance respectively (PCA$_{85}$ICA, PCA$_{95}$ICA, PCA$_{99}$ICA);
- Extended Infomax ICA applied directly to the data (ICA-Only);
- Extended Infomax ICA applied 150 times to bootstrapped versions of the data – i.e., the same dataset resampled with replacement – see [3,4].

For each IC an EEGLAB `dipfit` structure [5] is provided, containing information on the equivalent best fitting dipole.

See Figs. 1–3 and data at https://data.mendeley.com/datasets/gvt64rndnd/1 doi:10.17632/gvt64rndnd.1.

2. Experimental design, materials and methods

Fourteen volunteer participants (7 males, 7 females, aged 20 – 40 years) performed a visual working memory task [6]. Each trial consisted of the following:

- A central fixation symbol is presented for 5 s.
- A series of eight single letters are presented for 1.2 s with 200 ms gaps. 3–7 are black and to be memorized, the rest are green and to be ignored.
- A memory maintenance period of 2–4 s is indicated by a presented dash
- A red probe letter is presented
- The participant presses one of two buttons with the dominant hand (index finger or thumb) to indicate whether the probe letter was part of the memorized letter set or not.
- 400 ms after the button is pressed an auditory feedback signal informs the participant whether their answer was correct or not
- The participant presses another button when ready. This signals the start of a new trial.

Each participant performed 100–150 trials.

EEG data were collected from 71 channels (69 scalp and 2 periocular electrodes), referred to the right mastoid, at a sampling rate of 250 Hz with an analog pass band of 0.01 to 100 Hz (SA Instrumentation, San Diego) and with input impedances under 5 kΩ after careful scalp preparation. Continuous data were high-pass filtered (0.5 Hz), epochs ([−700 700] ms time locked to each letter presentation) were extracted and whole-epoch mean channel ("baseline") value removed [7].

Data are presented in the form of epoched EEGLAB datasets (.set and.fdt files). To load the data it is necessary to download and start EEGLAB (http://sccn.ucsd.edu/eeglab/currentversion/eeglab_current.zip). Matlab is required unless a compiled version of EEGLAB is used – more info at
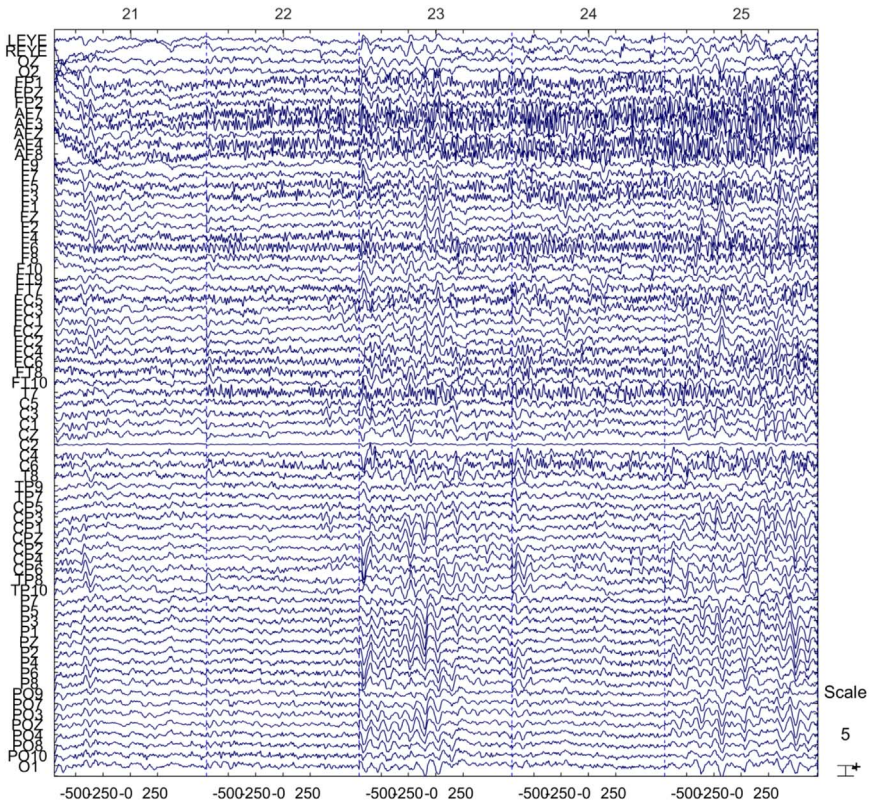
Fig. 1. : Raw epoched data, accessible via the GUI (Plot - > Channel Data) or by typing > > EEG = pop_eegplot(EEG,1,1,1).

https://sccn.ucsd.edu/eeglab/downloadtoolbox.php and in [5]. Complete documentation on how to use EEGLAB can be found at the link https://sccn.ucsd.edu/wiki/EEGLAB.

2.1. Opening and reviewing the data

To open the data it is necessary to run the EEGLAB toolbox first by typing
```
> > eeglab
```
in the matlab "Command Window" (after setting the current path to the EEGLAB main folder). Each ".set" file (Sub1.set, Sub2.set,…, Sub14.set) can be opened via the EEGLAB GUI (File - > Load existing dataset) by navigating to the folder where the datasets are. Alternatively, it is possible to type:
```
> > EEG = pop_loadset('filename','Sub2.set','filepath', 'PATH_SET');
```
Where Sub2.set is the filename of the dataset to load and PATH_SET the path to the folder containing it (e.g., C:\my_data). The data can then be perused at will by using the extensive set of functions available in the EEGLAB environment. Data are stored in the Matlab structure EEG. To plot the raw data it is possible to type (results in Fig. 1):
```
> > EEG = pop_eegplot(EEG,1,1,1)
```
Epochs are labeled in EEG.event.type as:

- B, C, D, F, G, H, J, K, etc. for black letters to be memorized
- gB, gC, gD, gF, gG, gH, gJ, gK, etc. for green letters to be ignored
- rB, rC, rD, rF, rG, rH, rJ, rK, etc. for red probe letters

Channel labels follow the International 10–20 System convention and can be retrieved by typing:
```
> > ch_labels = {EEG.chanlocs.labels}
```
Template electrode positions on the head are also contained in the `EEG.chanlocs` structure.
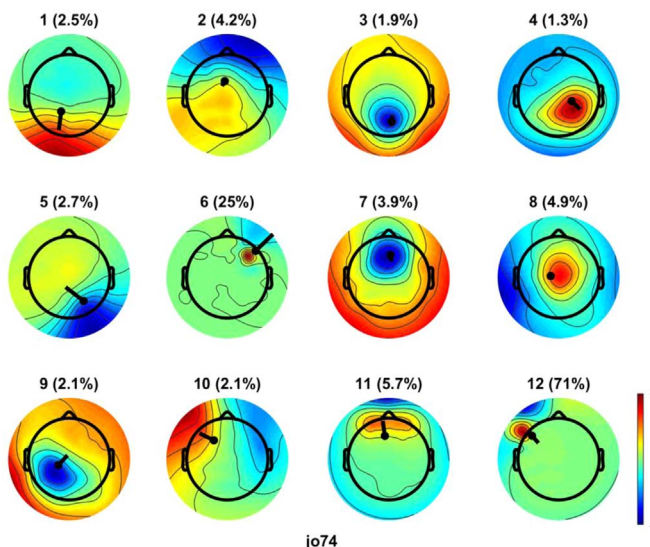
2.2. Infomax ICA with PCA

The outputs of the ICA decompositions, directly on the data (ICA only) or after preliminary PCA are collected in the following structures:

- PCA only: `EEG.etc.INFOMAX.PCA`
- ICA only: `EEG.etc.INFOMAX.ICA`
- $PCA_{85}ICA$: `EEG.etc.INFOMAX.ICA85`
- $PCA_{95}ICA$: `EEG.etc.INFOMAX.ICA95`
- $PCA_{99}ICA$: `EEG.etc.INFOMAX.ICA99`

Each structure contains the mixing matrix (A), the unmixing matrix (W) and the associated `dipfit` structure contains the position (`EEG.dipfit.model.posxyz`), orientation (`EEG.dipfit.model.momxyz`) and residual variance (`EEG.dipfit.model.rv`) of the best-fitting equivalent dipole for each IC. As an example, in order to retrieve and visualize an ICA decomposition for $PCA_{95}ICA$, type (results in Fig. 2):
```
% Load the subject (PATH_SET is the path to the dataset)
> > EEG = pop_loadset('filename','Sub2.set','filepath', PATH_SET);
% Embed the ICA95 decomposition in the data
> > EEG.icaweights = []; EEG.icawinv = []; EEG.icaact = [];
> > EEG.icasphere = []; EEG = eeg_checkset(EEG);
> > EEG.icaweights = EEG.etc.INFOMAX.ICA95.W;
> > EEG.icawinv = EEG.etc.INFOMAX.ICA95.A;
> > EEG.icasphere = eye(EEG.nbchan);
> > EEG.dipfit = EEG.etc.INFOMAX.ICA95.dipfit;
> > EEG = eeg_checkset(EEG);
> > eeglab redraw
% Plot the components
> > pop_topoplot(EEG,0);
```


Fig. 2. First 12 $PCA_{95}ICA$ IC Scalp Maps (Subject 2) with overlaid associated dipole and residual variance (percentage, out of 100%). The plot is generated after embedding the $PCA_{95}ICA$ decomposition (see "Infomax ICA with PCA").

The window that pops up allows to select the component numbers to plot, the title of the plot and whether or not to display the associated dipole. Since the ICA decomposition is now embedded in the main EEGLAB structure, it is possible to follow the steps in the EEGLAB tutorials to plot component properties, component ERPs, etc.

2.3. Boostrap ICA

Boostrap ICA weights are collected in the structure `EEG.etc.RELICA`. The ICA decomposition was performed 150 times, each time after resampling with repetition (bootstrapping) the data. ICs were clustered using a Cuvilinear Component Analysis (CCA) projection - see [1,3,4,8,9]. The clustered ICs are collected in `EEG.etc.RELICA.W_boot` (unmixing matrices) and `EEG.etc.RELICA.A_boot` (mixing matrices). The quality index of each IC is in `EEG.etc.RELICA.Iq`.

As an example, `EEG.etc.RELICA.A_boot{8}(:,5)` represents the mixing weights of the 5th IC belonging to the 8th cluster. Similarly, `EEG.etc.RELICA.W_boot{8}(5,:)` represents the unmixing weights of 5th IC belonging to the 8th cluster. The mixing weights associated to the centroid IC of the 8th cluster are in `EEG.etc.RELICA.A(:,8)`, the unmixing weights in `EEG.etc.RELICA.W(8,:)`. The following code embeds in EEGLAB the 8th cluster ICs and shows the corresponding maps (Fig. 3).

```
% Load the subject (PATH_SET is the path to the dataset)
>> EEG = pop_loadset('filename','Sub2.set','filepath', PATH_SET);
% Embed the ICs belonging to the 10th boostrap cluster
>> EEG.icaweights = []; EEG.icawinv = []; EEG.icaact = [];
>> EEG.icasphere = []; EEG = eeg_checkset(EEG); EEG.dipfit = [];
>> EEG.icaweights = EEG.etc.RELICA.W_boot{8};
>> EEG.icawinv = EEG.etc.RELICA.A_boot{8};
>> EEG.icasphere = eye(EEG.nbchan);
>> EEG = eeg_checkset(EEG);
>> eeglab redraw
% plot the components
>> pop_topoplot(EEG,0);
```
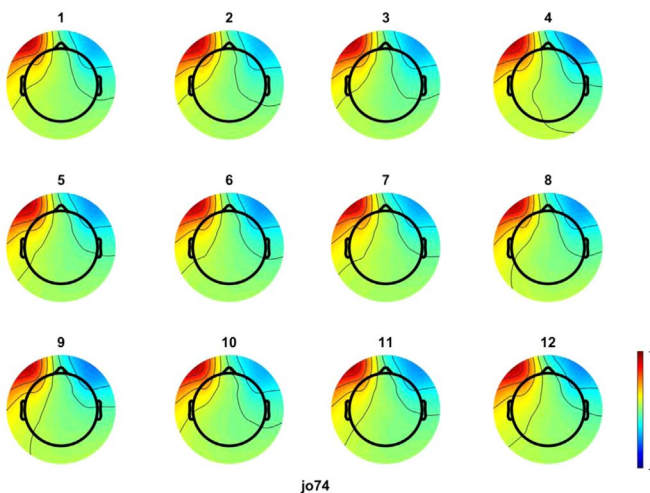

Fig. 3. First 12 ICs (Subject 2) belonging to the eye movement cluster. The plot is generated after embedding cluster 8 of the RELICA decomposition (see section "Bootstrap ICA").

Acknowledgements

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.12.022.

References

[1] F. Artoni, A. Delorme, S. Makeig, Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition, NeuroImage 175 (2018) 176–187.

[2] A. Delorme, J. Palmer, J. Onton, R. Oostenveld, S. Makeig, Independent EEG sources are dipolar, PLoS One 7 (2) (2012) e30135.

[3] L. Sebastiani, E. Castellani, A. Gemignani, F. Artoni, D. Menicucci, Inefficient stimulus processing at encoding affects formation of high-order general representation: a study on cross-modal word-stem completion task, Brain Res. 1622 (2015) 386–396.

[4] F. Artoni, D. Menicucci, A. Delorme, S. Makeig, S. Micera, RELICA: a method for estimating the reliability of independent components, Neuroimage 103 (2014) 391–400.

[5] A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, J Neurosci. Methods 134 (1) (2004) 9–21.

[6] J. Onton, A. Delorme, S. Makeig, Frontal midline EEG dynamics during working memory, Neuroimage 27 (2) (2005) 341–356.

[7] D.M. Groppe, S. Makeig, M. Kutas, Identifying reliable independent components via split-half comparisons, Neuroimage 45 (4) (2009) 1199–1211.

[8] F. Artoni, A. Gemignani, L. Sebastiani, R. Bedini, A. Landi, and D. Menicucci, "ErpICASSO: a tool for reliability estimates of independent components in EEG event-related analysis," in: Conference Proceedings of Engineering in Medicine and Biology Society, vol. 2012, 2012, pp. 368–371.

[9] D. Menicucci, et al., Brain responses to emotional stimuli during breath holding and hypoxia: an approach based on the independent component analysis, Brain Topogr. 27 (6) (2014) 771–785.