Group 6: Age Prediction via Electroencephalogram Data

Zihang Wang zwang2547
Jie Sheng jsheng27
Haoran Lu hlu226
Luyang Fang lfang45
Xinran Miao xmiao27

November, 2020

1 Introduction

An electroencephalogram (EEG) records individual's brain wave signals of difference areas via small electrode sensors in a short period. Recordings of each sensor is called a channel and it looks like wavy lines with peaks and valleys over time. EEG provides information about one's brain condition, which can be used to diagnose brain diseases and make other inference. In this work, we intend to extract structure information of human brain first and try to predict ages based on the extracted signals.

2 Data Description

The EEG data set is from Kaggle [1]. It contains 1297 patients' recordings on 30 to 34 channels, each of which includes signals at more than 300,000 time points. Since the original data is high-dimensional, a dimension reduction step have to be implemented before building prediction model.

3 Statistical Models

3.1 Independent component analysis

Independent component analysis (ICA) is a computational method to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible. The resulting representation is supposed to capture the essential structure of the data in many applications, including feature extraction and signal separation [2].

Performing ICA on data matrix $\mathbf{X}_{(p\times n)}$, we can get:

$$\mathbf{X}_{p\times n} = \mathbf{S}_{p\times k}\mathbf{A}_{k\times n}$$

where k is the number of components we wish to get; **A** is the mixing matrix which stores the spatial information of brain signals; and **S** contains the signals we extracted from the original dataset.

Our goal is to find a proper way of decomposition such that the signals we extracted are as independent as possible, which can be realized by maximized the **non-Gaussianity** of those signals.[3] We use **fastICA algorithm** to maximize negative entropy to achieve the maximum non-Gaussianity. And the spatial information included in matrix $\hat{\mathbf{A}}$ could be potentially used to predict a person's age.

3.2 Aligned Similarity and k-NN

In preparation of k-NN, we want to define a similarity between two people's $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$.

By the property of ICA, denote the j-th row of$\hat{\mathbf{A}}_i$ as $\mathbf{a}_{ij}$. Orders $\{\mathbf{a}_{ij}\}_j$ in $\hat{\mathbf{A}}_i$ could be of random permutation. So we first want to align $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ to let their rows to match with each other. We simply find 3 pairs of $\mathbf{a}_{1j}$ and $\mathbf{a}_{2k}$ that have the largest absolute value of correlations and denote them as $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{12}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{22}\}$. Then we can define our aligned similarity of $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ as (1):

$$\sum_{j=1}^{3} |\mathrm{cor}(\mathbf{a}_{1j}, \mathbf{a}_{2j})| \tag{1}$$

Next, we used this similarity to build a k—NN model with which to predict age. Currently, we are using k=1 and it shows good performance.

4 Computations

Since the independent components can be disintegrated independently on each person's EEG data set, we employed CHTC to run about 1300 jobs in parallel. It took us about one hour and the matrix A of every patient were collected.

Then the dimensions of this huge EEG data is dramatically reduced. Following steps can be done on one computer. We wrote our own R codes for aligning by correlations. Finally, for each patient, we got his or her age and corresponding predictors. We built a k-NN model to predict age on test set.

5 Findings & Future Work

5.1 Results of ICA

Figure 1 and 2 shows the resulting A matrices for two patients with $k = 4$, where each sub panel indicates one extracted signal. The aligning sub panels look similar between patients, which indicates a success of ICA on extracting information.

5.2 Results of k-NN model

We are still working on this part and we plan to use mean squared error (MSE) to evaluate the models across choices of k.

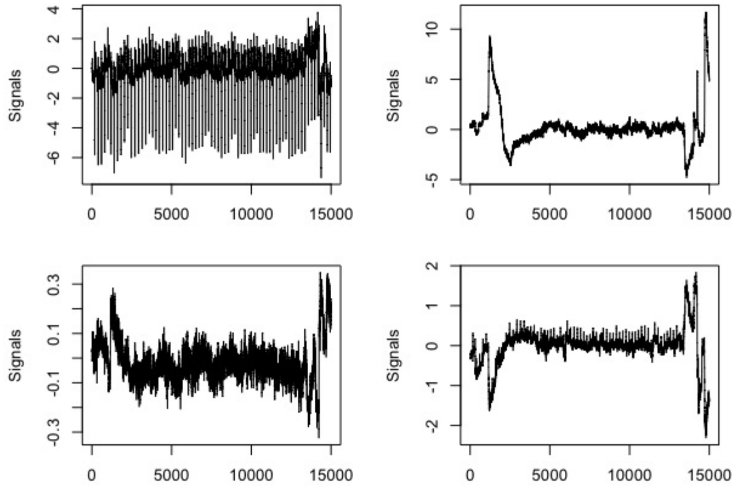

Figure 1: Example one of EEG signals after independent component analysis

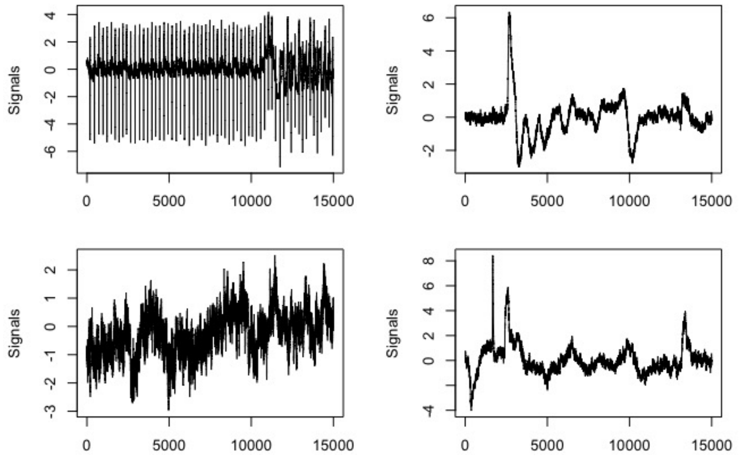

Figure 2: Example two of EEG signals after independent component analysis

References

[1] Alexander Valkovich ayurgo, Enthusiast2020. Eeg for age prediction. `https://www.kaggle.com/ayurgo/data-eeg-age-v1`, 2020.

[2] Aapo Hyvärinen. Testing the ica mixing matrix based on inter-subject or inter-session consistency. *NeuroImage*, 58(1):122–136, 2011.

[3] Aapo Hyvärinen and Oja Erkki. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.