109-2 5275 Brain Computer Interface

Lab 2: EEG Preprocessing and Data Cleaning

Min-Jiun Tsai
dollars9256741.sc08@nycu.edu.tw

Institute of Mathematical Modeling and Scientific Computing,
National Yang Ming Chiao Tung University
Cognitive Engineering and Computational Neuroscience Lab – CECNL

April 3, 2021

Table of Contents

Table of Contents

Instruction
Submission Policy

- **PLAGIARISM IS STRICTLY PROHIBITED. (0 point for Plagiarism)**
- For mathematical problem(s), please show your work step by step and clarify statement of theorem you use (if any). Answering without mathematical derivations will get 0 point.
- Submission deadline: **2021.04.15 09:00:00 AM**.

Late submission penalty formula

$$original\ score \times (0.7)^{\#(days\ late)}$$

Instruction
File Format

- Each group submits 1 report (.pdf and .tex file) and 1 code (.ipynb or .m).
- **Report** must contains observations, results and explanations. Please name your .pdf and .tex file as **5275_Lab2_GroupNum.pdf** and **5275_Lab2_GroupNum.tex**,respectively.
- Paper submission is not allowed. **Please use our LaTeX template to complete your report**.

Instruction
File Format

- **Code** file must contains comments to explain your code. Please name your code file as **5275_Lab2_GroupNum.ipynb/.m**
- Implementation will be graded by completeness, algorithm correctness, model description, and discussion.

Illegal format penalty

-5 points for violating each rule of file format.

Instruction

Office Hour Information

We'll have limited time to teach EEGLab and MNE on our course; therefore, if you have any question about lab 2, feel free to make an appointment or come to ask me during my office hour.

Tue.	12:20 p.m.-13:10 p.m.	EC120
Thur.	06:30 p.m.-09:30 p.m.	SC207

Note

Actually, my office hour on Thursdays is main for calculus consultation. If there are undergraduate students come to ask calculus problems, I need to teach them first and then to solve your problem during the rest of the office hour on Thursday nights.

Instruction

Prerequest

To finish programming problem, you could choose Matlab or Python base on your programming preference.

Matlab 2020a+

- NYCU installation page
- NCTU installation tutorial
- EEGLab official installation page (v2020.0+ is recommended)

Python 3.7+

- MNE official installation page (0.20.7+ is recommended)

Table of Contents

EEG Preprocessing [Cohen, 2014]

What is preprocessing?

Preprocessing refers to any transformation or recognizations that occur between collecting the data and analyzing the data.
For example,

- Extract epochs
- Remove bad electrodes or rejecting epochs
- Apply temporal filters or spatial transformations

Data analysis lecture video by Mike X Cohen

- Filter, epoch, baseline subtraction, referencing
- Trial rejection
- Independent components analysis for removing artifacts

EEG Preprocessing [Cohen, 2014]

The balance between Signal and noise

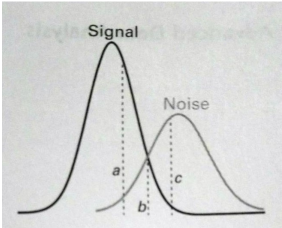


Figure: Theoretical depiction of signal and noise in EEG data as a signal detection problem.

EEG data contain signal and noise. Appropriate preprocessing will attenuate the noise in the data. Unfortunately, signal and noise are often mixed together and may be difficult to disentangle completely.

Fortunately, time-frequency-based analyses tend to increase the signal-to-noise (SNR) characteristics of the data, particularly for single-trial analyses and relatively low frequencies ($\leq 20Hz$).

EEG Preprocessing [Cohen, 2014]

Creating Epochs

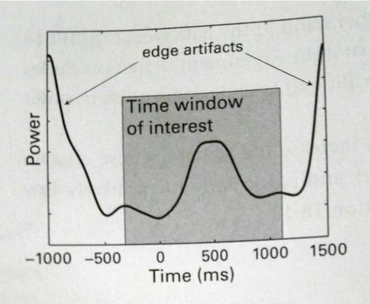
Type	**Size of data**	Example
Continuous	(channel, time points)	Resting-state datasets
Noncontinuous	(channel, time points, trials)	stimuli-related experiment response-related experiment

When epoching, one must decide... ...

1. event to use for time locking–that is, what to call $time = 0$.
2. which may have significant consequences for the quality of the TF decomposition, particularly for lower frequencies–is how much time to include before and after the $time = 0$.

EEG Preprocessing [Cohen, 2014]

Creating Epochs: Edge artifacts



Edge artifacts resulting from discontinuous breaks in the time series between trials can contaminate the result if there are insufficient buffer zones to allow those edge artifacts to subside.

If you plan on performing TF-based analyses, you should create longer epochs.

EEG Preprocessing [Cohen, 2014]

Creating Epochs: Edge artifacts

How much of a buffer zone you need depends on the frequencies that you intend to extract from the data. Edge artifacts typically last 2 or 3 cycles, although this depends on the magnitude of the edge. Thus, the lower frequency that you will extract from the data, the more buffer zone you will need to be confident that the edge artifact has subside.

EEG Preprocessing [Cohen, 2014]
Filtering

Filtering data can help remove high-frequency artifacts and low-frequency drifts.

- Applying a high-pass filter at 0.1 or 0.5Hz to the continuous data is recommended to minimize slow drift.
- High-pass filtering should be applied only to continuous data and not to epoched data.
- Notch filter at 50Hz or 60Hz help attenuate electrical line noise

EEG Preprocessing [Cohen, 2014]

Trial rejection

Removing trials that contain artifacts prior to analyses is an important preprocessing steps.

Visual inspection v.s. Automatic trial rejection

Automatic rejection is fast and free of user bias; however, it may produce both Type I and Type II error.

Preprocessing pipeline

Recommended materials

Matlab

- EEGLab wiki
- Makoto's preprocessing pipeline
- Makoto's useful EEGLAB code

Python

- MNE tutorial

Table of Contents

EEG Artifacts

Physiologic	Nonphysiologic
Eye movement	Electrode and lead
Electromyographic	Instrumental
Glossokinetic	Environmental
Electrocardiographic	Internal electric stimulators
Sweat	
Patient movement	

Table: EEG Artifacts[Schomer and da Silva, 2011]

EEG Artifacts

Eye movement

Eye blink is a "U" shape transient potential which appears in frontopolar electrodes which are closer to the eyes.

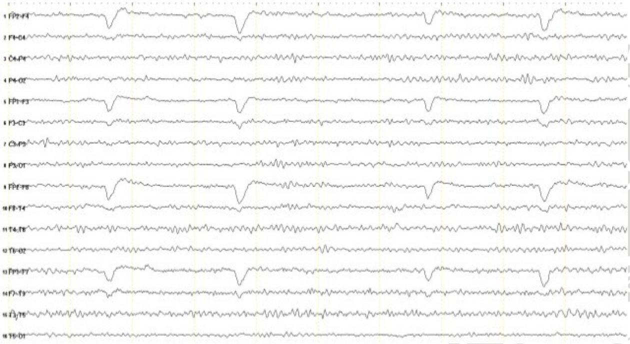


Figure: Blink [resident slides]

EEG Artifacts

Eye movement

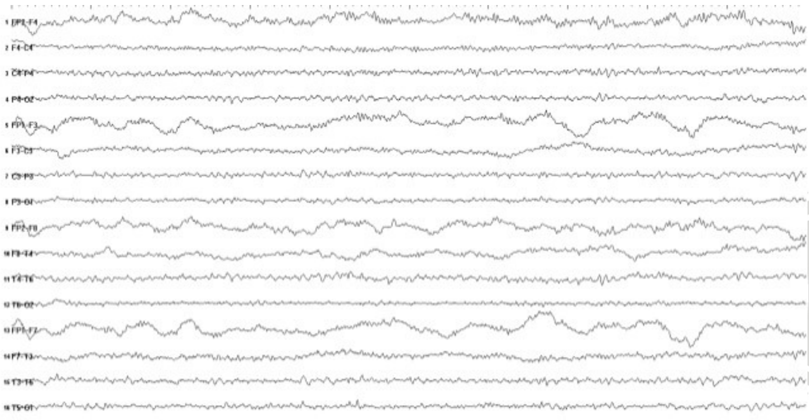


Figure: Lateral eyeball movements [resident slides]

EEG Artifacts
Eye movement

Figure: Eye opening and eye closure [resident slides]

EEG Artifacts
Eyelid flutter

Fine eyelid movements may produce rhythmic 4-8 Hz activity in the frontal leads (disappear on fixation)

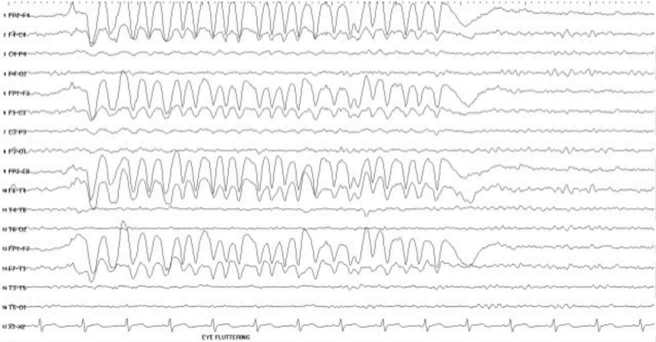


Figure: Eyelid flutter [resident slides]

EEG Artifacts
Swallowing

Swallowing of saliva usually produces a short burst activity.

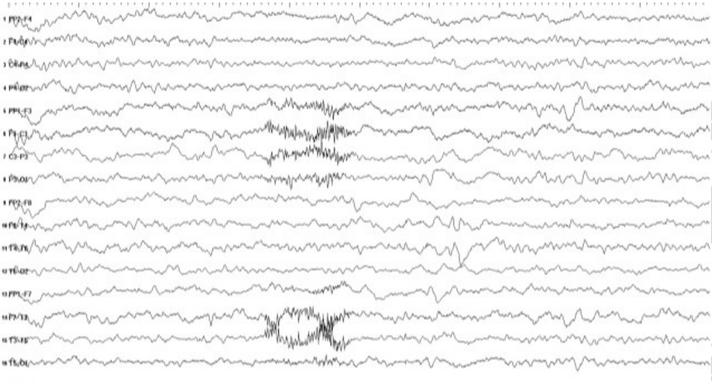


Figure: Swallowing [resident slides]

EEG Artifacts
Pulse

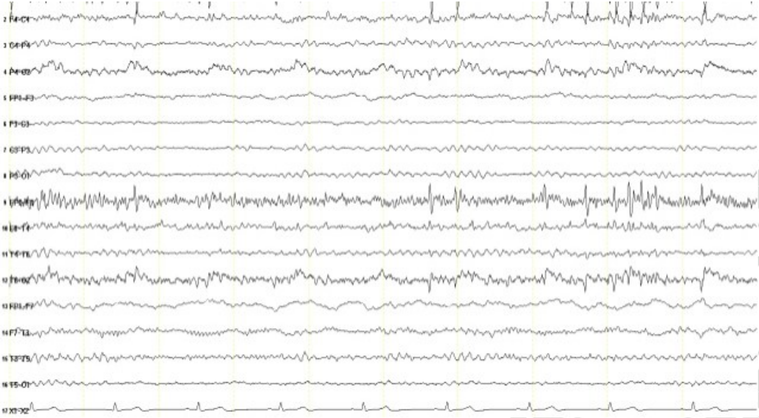


Figure: Pulse [resident slides]

EEG Artifacts

Sweat

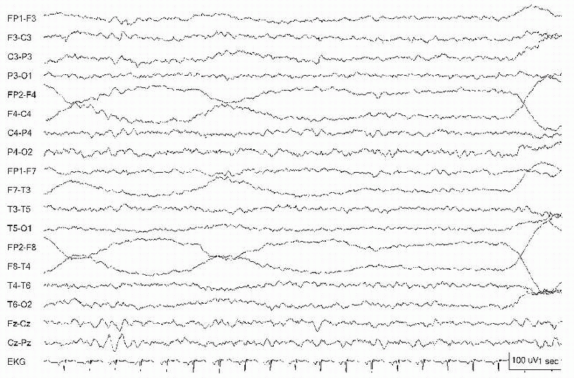


Figure: Subdelta activity is noted with crossing baselines consistent with "sweat away".[Schomer and da Silva, 2011]

EEG Artifacts

Environmental artifacts

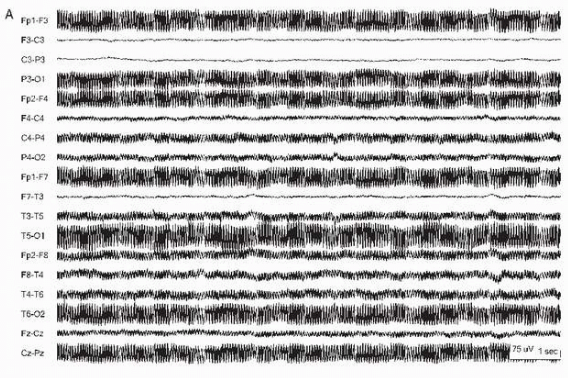


Figure: 60*Hz* artifacts (electrostatic artifact) obscuring most of the record because of the lack of a ground electrode.[Schomer and da Silva, 2011]

EEG Artifacts
Environmental artifacts

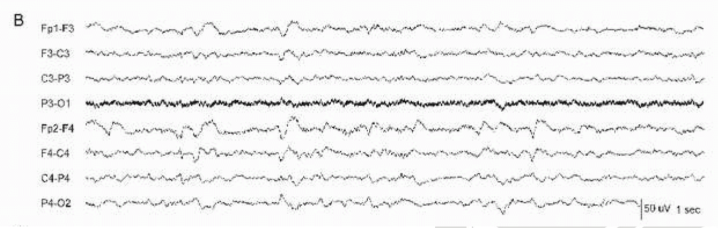


Figure: More typical appearance of 60*Hz* artifact in a high impedance O1 electrode. [Schomer and da Silva, 2011]

EEG Artifacts

Environmental artifacts

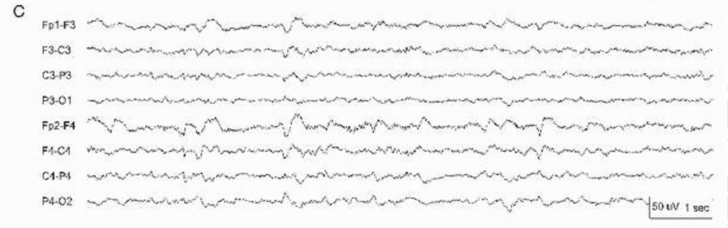


Figure: Same segment of EEG as above figure with 60Hz notch filter. [Schomer and da Silva, 2011]

EEG Artifacts
Electromagnetic Artifacts

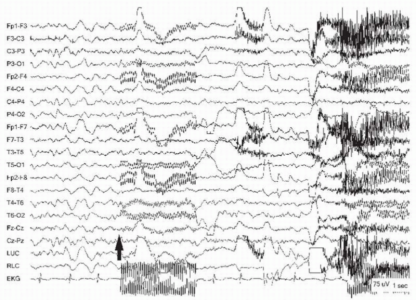


Figure: Two-second burst of 30 cycle/sec telephone artifact (electromagnetic artifact, vertical arrow) diffusely, but highest amplitude in high impedance electrode Fp2. This leads to eye movements and EMG as the phone is answered, obscuring a second "ring artifact."[Schomer and da Silva, 2011]

Table of Contents

Independent Component Analysis

Motivation: Blind Source Separation

Blind Source Separation (BSS) is a method to estimate source signals from recorded signals which consist of mixed source signals and noise.

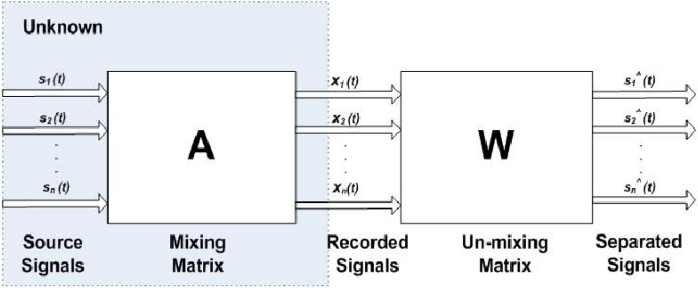


Figure: Blind source separation (BSS)[Naik and Kumar, 2011]

Independent Component Analysis

Motivation: Blind Source Separation

Model Formalization

Let n denotes number of source signal, m denotes number of channel, and d denotes dimension of signal. The matrix $S \in \mathbb{R}^{n \times d}$ denotes source signals. We assume that recorded signals $X = AS + E \in \mathbb{R}^{m \times d}$ are given by linear mixing system where $A \in \mathbb{R}^{m \times n}$ is the unknown mixing matrix and $E \in \mathbb{R}^{m \times d}$ denotes the noise.
Basically, $m \geq n$

The goal of BSS is to estimate A and S so that $\hat{S}$ provides unknown source signals as possible.

$$X = AS + E \leftarrow X = \hat{A}\hat{S}$$

Independent Component Analysis

BSS with Different constraints

Since $m \geq n$, there are a lot of combinations (A, S) satisfy $X = AS + E$. We could apply different types of constraint to solve this system:

- PCA: Orthogonal constraint
- SCA: Sparsity constraint
- NMF: Non-negative constraint
- ICA: Statistically independent constraint

Therefore, there are many methods to solve the BSS problem depending on the constraints. What we used is depended on subject matter.

In this slide, we only introduce **ICA**.

Independent Component Analysis

Model of ICA

The Cocktail Party Problem

Let X be a recorded signal and S is a source signal according to above formalization. We assume that $\{s_j \in \mathbb{R}^{d\times 1} | j \in \mathbb{Z}_n\}$ is statistically independent.

$$X = \hat{A}\hat{S} \iff \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ \vdots & \vdots & \vdots \\ - & x_m^T & - \end{bmatrix} = \hat{A}_{m\times n} \begin{bmatrix} - & \hat{s}_1^T & - \\ - & \hat{s}_2^T & - \\ \vdots & \vdots & \vdots \\ - & \hat{s}_n^T & - \end{bmatrix} \quad (1)$$

Independent Component Analysis is to estimate the independent component S from X.

Independent Component Analysis

Hypothesis

Hypothesis of ICA

- $\{s_j \in \mathbb{R}^{d\times 1} | j \in \mathbb{Z}_n\}$ statistically independent, that is, $P(s_1, ..., s_n) = \prod_{j=1}^{n} P(s_j)$
- $\{s_j \in \mathbb{R}^{d\times 1} | j \in \mathbb{Z}_n\}$ follows the Non-Gaussian distribution.
- A is regular

Therefore, we could rewrite the model as $\hat{S} = \hat{B}X$ where $\hat{B} = \hat{A}^{-1}$. It's only necessary to estimate B (compute $\hat{B}$) so that $\{s_j \in \mathbb{R}^{d\times 1} | j \in \mathbb{Z}_n\}$ is independent.

Independent Component Analysis
Whitening

Definition of White signal

White signals are defined as any $z \in \mathbb{R}^{d\times 1}$ which satisfying

- Zero mean: $E[z] = \mathbf{0} = m_z$
- Unit covariance: $C_z = E[(z - m_z)(z - m_z)^T] = E[zz^T] = I_d$

If $m_z = \mathbf{0}$, then the correlation matrix $R_z = C_z + m_z m_z^T = C_z$
Recall that recorded signals are $X = \hat{A}\hat{S}$.
ICA solve $\hat{S}$ by $\hat{S} = \hat{B}X$.

Independent Component Analysis

Whitening

From now on we assume that $m = n$ to simplify the model.
Whitening is useful for PCA and simplifies ICA problem.

Whitening simplifies ICA problem

If we denote whitening signal as

$$Z_{d\times m} = V_{d\times d} X^T_{d\times m}$$

$$\iff \begin{bmatrix} | & | & \dots & | \\ z_1 & z_2 & \dots & z_m \\ | & | & \dots & | \end{bmatrix} = V_{d\times d} \begin{bmatrix} | & | & \dots & | \\ x_1 & x_2 & \dots & x_m \\ | & | & \dots & | \end{bmatrix} \quad (2)$$

where $V \in \mathbb{R}^{d\times d}$ is a whitening matrix of $X_{m\times d}$, then model becomes... ...

Independent Component Analysis

Whitening

Whitening simplifies ICA problem

$$\hat{S}^T_{d\times m} = U_{d\times d} Z_{d\times m} = U_{d\times d} V_{d\times d} X^T_{d\times m} = \hat{B}_{d\times d} X^T_{d\times m}$$

$$\Longleftrightarrow \begin{bmatrix} | & | & \cdots & | \\ \hat{s}_1 & \hat{s}_2 & \cdots & \hat{s}_m \\ | & | & \cdots & | \end{bmatrix} = \hat{B}_{d\times d} \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & \cdots & | \end{bmatrix} \quad (3)$$

where $U \in \mathbb{R}^{d\times d}$ is an orthogonal transformation matrix.

Hence it's necessary to estimate U!

Independent Component Analysis

Whitening

[Problem 2]

Assume that an orthogonal transformation $U \in \mathbb{R}^{d \times d}$ and z is white, then

$$m_{Uz} = m_z \ \& \ C_{Uz} = C_z$$

Note

Whiteness property is preserved under orthogonal transformations.

Independent Component Analysis
Whitening

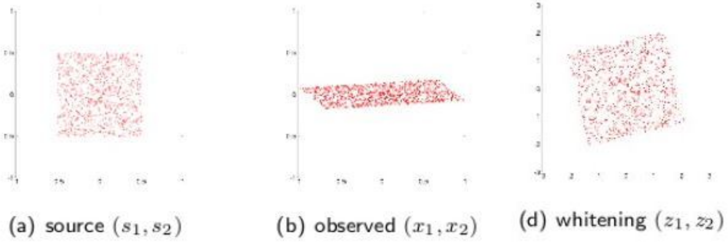


Figure: Signals [Yokota, 2012]

Independent Component Analysis

Measure of independence: Non-gaussianity

The gaussianity of X (sums of non-gaussian random variables) must be larger than S (original) according to Central Limit Theorem. Let $\{x_j \in \mathbb{R}^{d\times 1} | j \in \mathbb{Z}_m\}$ be the observed signals, we want to maximize the non-gaussianity of source signals $s_j = Bx_j$.

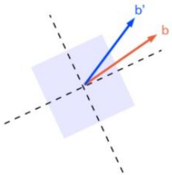


Figure: The Non-gaussianity of b is larger than b' [Yokota, 2012]

Independent Component Analysis

Kurtosis

Kurtosis is a measure of non-gaussianity.

Definition of Kurtosis

for a random variable $y \in \mathbb{R}^{d \times 1}$,

$$kurt(y) = E[y^4] - 3(E[y^2])^2$$

That is, for white signal $z \in \mathbb{R}^{d \times 1}$,

$$kurt(z) = E[z^4] - 3(E[z^2])^2 = E[z^4] - 3$$

Which means we could solve ICA problem by

$$\hat{b} = \max_b \|kurt(b^T x)\| \tag{4}$$

Independent Component Analysis
Kurtosis

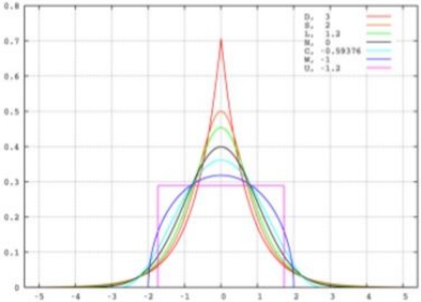


Figure: distributions [Yokota, 2012]

Independent Component Analysis

Kurtosis

We consider $z = Vx$ is a white signal given from source signal x, then we could rewrite (4) as:

[Problem 3] Solve w with constraint $w^T w = 1$

$$\max_w \|kurt(w^T z)\| \ subject \ to \ w^T w = 1 \tag{5}$$

Table of Contents

Dataset: Lane-departure paradigm

Experimental paradigm

The solid black arrows represent driving trajectory. The empty circle represents deviation onset. The double circle represents response onset. The circle with a cross represents end of response. which was sufficient for subjects to experience fatigue.

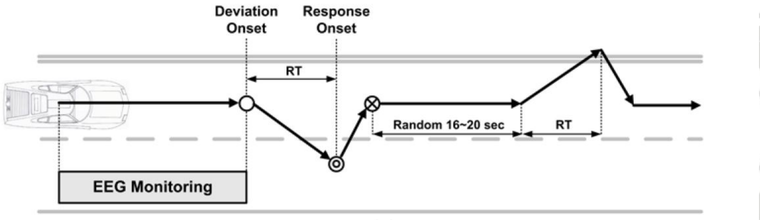


Figure: Event-related lane-departure tasks[Kuan-Chih Huang and Jung, 2016],[Chin-Teng Lin and Jung, 2010]

Dataset: Lane-departure paradigm

Problem 6

1. Plot 2D channel location map and re-reference data by $\frac{A1+A2}{2}$.
2. Down-sampling to $250Hz$.
3. Run ICA and record computational time of ICA by code.
4. Plot component map in 2D.
5. Indicate noise component(s) if it exist and explain reason why you identify this component as a noise or artifacts.
6. Plot first 10-second channel data before and after deleting noise/artifact component(s).

Dataset: Lane-departure paradigm

Problem 7

Delete vehicle position channel and then:

1. Plot 2D channel location map and re-reference data by $\frac{A1+A2}{2}$.
2. Down-sampling to $250Hz$.
3. Bandpass filtering $[1, 50]Hz$
4. Run ICA and record computational time of ICA by code.
5. Plot component map in 2D.
6. Indicate noise component(s) if it exist and explain reason why you identify this component as a noise or artifacts.
7. Plot first 10-second channel data before and after deleting noise/artifact component(s).

Dataset: Lane-departure paradigm

Problem 6

After above preprocessing steps... ...

- Compare results (e.g. component map) and try to explain your observations.
- Explain why it takes less time this time?

Dataset: Lane-departure paradigm

Problem 8

Design your own EEG preprocessing strategy:

- Describe your design idea (e.g. Apply CleanLine function in EEGLab to eliminate environmental artifacts and apply lowpass filtering to remove drift... ...)
- Compare the performance (computational time) and results with Problem 7.
- Explain potential reason(s) why performance of your preprocessing strategy is superior/inferior to performance of Problem 7?

Table of Contents

End Note for other problems

- For multiple choice problem, please give a brief explanation for option(s) you choose. Answering without any description will get 0 point.
- For problem 1, please follow the hint on lab 2 instruction to solve subproblem a.
- Good luck :)
- There is a feedback section on the bottom of instruction pdf file, please fill out the work division table (at least).

References I

Kuan-Chih Huang Chin-Teng Lin and Tzyy-Ping Jung. Tonic and phasic eeg and behavioral changes induced by arousing feedback. *Neuroimage*, 52(2):633–642, 2010.

Mike X Cohen. *Analyzing neural time series data : theory and practice*. Cambridge, Massachusetts : The MIT Press, 2014.

Chin-Teng Lin Kuan-Chih Huang and Tzyy-Ping Jung. An eeg-based fatigue detection and mitigation system. *International Journal of Neural Systems*, 26(4), 2016.

Ganesh R. Naik and Dinesh K Kumar. An overview of independent component analysis and its applications. *Informatica*, 35:63–81, 2011.

Neurology resident slides. Artifacts in eeg final. URL `https://www.slideshare.net/ranjithpolusani/artifacts-in-eeg-final`.

Donald L. Schomer and Fernando H. Lopes da Silva. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 2011. ISBN 9780781789424.

Tatsuya Yokota. Independent component analysis for blind source separation. *Remote Sensing*, 3(6): 1104–1138, 2012.

Thank you!