IOC 5275: Brain Computer Interface **(Due: 04/15/2021)**

Lab2: EEG Preprocessing and Data Cleaning

Instructor: Chun-Shu Wei *Group Number:* group number
TA: Min-Jiun Tsai *Student Id*

Submission Policy

Read all the instructions below carefully before you start working on the assignment, and before you make a submission. For this assignment, please hand in the following your report (pdf) and code (.ipynb or .m file).

- **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:**

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

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**.
- **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.

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)

1 Mathematical problem

1.1 Find the coefficients b_n in Fourier sine series

Let us begin with the Fourier sine series

$$S(t) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi t}{T}, \textit{ in the interval } (0, T). \tag{1.1}$$

To solve b_n while $S(t)$ is given, we can use the key observation that

$$\int_0^T \sin \frac{n\pi t}{T} \sin \frac{m\pi t}{T} dt = 0 \ \forall m, n \in \mathbb{Z}, \ m \neq n$$

with $\sin a \sin b = \frac{1}{2}[\cos(a-b) - \cos(a+b)]$.

$$\begin{aligned} \int_0^T \sin \frac{n\pi t}{T} \sin \frac{m\pi t}{T} dt &= \frac{1}{2} \int_0^T \cos((n-m)\frac{\pi t}{T}) - \cos((n+m)\frac{\pi t}{T}) dt \\ &= \Big[\frac{T}{2\pi(n-m)} \sin \frac{(n-m)\pi t}{T} - \frac{T}{2\pi(n+m)} \sin \frac{(n+m)\pi t}{T} \Big|_{t=0}^{T} \Big] \\ &= 0, \ \forall m, n \in \mathbb{N}, \ m \neq n \end{aligned} \tag{1.2}$$

Let's fix m, multiply (1.1) by $\sin \frac{m\pi t}{T}$, and integrate the series (1.1) term by term to get

$$\int_0^T S(t) \sin \frac{m\pi t}{T} dt = \int_0^T \Big(\sum_{n=1}^{\infty} b_n \sin \frac{n\pi t}{T} \Big) \sin \frac{m\pi t}{T} dt = \sum_{n=1}^{\infty} b_n \int_0^T \sin \frac{n\pi t}{T} \sin \frac{m\pi t}{T} dt = b_m \int_0^T \sin^2 \frac{m\pi t}{T} dt \tag{1.3}$$

First, We can compute $\int_0^T \sin^2 \frac{m\pi t}{T} dt$:

$$\begin{aligned} \int_0^T \sin^2 \frac{m\pi t}{T} dt = -\frac{T}{m\pi} \int_0^T \sin \frac{m\pi t}{T} d \cos \frac{m\pi t}{T} &= -\frac{T}{m\pi} \Big[\sin \frac{m\pi t}{T} \cos \frac{m\pi t}{T} \Big|_{t=0}^{T} - \frac{m\pi}{T} \int_0^T \cos^2 \frac{m\pi t}{T} dt \\ &= -\frac{T}{m\pi} \Big[\sin \frac{m\pi t}{T} \cos \frac{m\pi t}{T} \Big|_{t=0}^{T} - \frac{m\pi}{T} \int_0^T 1 - \sin^2 \frac{m\pi t}{T} dt \Big] \\ &= -\frac{T}{m\pi} \Big[\sin \frac{m\pi t}{T} \cos \frac{m\pi t}{T} \Big|_0^T \Big] + T - \int_0^T \sin^2 \frac{m\pi t}{T} dt \end{aligned} \tag{1.4}$$

$$\Rightarrow \int_0^T \sin^2 \frac{m\pi t}{T} dt = \frac{1}{2} \Big\{ -\frac{T}{m\pi} \Big[\sin \frac{m\pi t}{T} \cos \frac{m\pi t}{T} \Big|_0^T \Big] + T \Big\} = \frac{T}{2} \Rightarrow \int_0^T S(t) \sin \frac{m\pi t}{T} dt = b_m \int_0^T \sin^2 \frac{m\pi t}{T} dt = b_m \frac{T}{2}$$

$$\Rightarrow b_m = \frac{2}{T} \int_0^T S(t) \sin \frac{m\pi t}{T} dt$$

This is the famous formula for the Fourier coefficients in the series (1.1).

Problem 1. Fourier Series (10+5=15 points)

Suppose that the Fourier sine series of $g(x) = x \ on \ (0, l)$ is given by

$$g(x) = \sum_{k=1}^{\infty} f_k(x) \tag{1.5}$$

with conditions : $f_k : (0, l) \to \mathbb{R}$ is integrable and $\sum_{k=1}^{\infty} f_k(x)$ is uniformly convergent.
(a) Find the **Fourier cosine series** of the function $\frac{x^2}{2}$ and the constant of integration (the 1^{st} term of cosine series).
$a_0 = \frac{2}{l}\int_0^l \frac{x^2}{2} dx = \frac{l^2}{3}$, $a_n = \frac{2}{l}\int_0^l \frac{x^2}{2}\cos\frac{n\pi x}{l} dx = \frac{2l^2}{\pi^2 n^2}\cos n\pi$
Fourier cosine series: $\frac{x^2}{2} = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{l} = \frac{l^2}{6} + 2l^2 \sum_{n=1}^{\infty} \frac{(-1)^n \cos(\frac{n\pi x}{l})}{(\pi n)^2}$
(b) Please exam whether the following series converges or not. Furthermore, find the sum of the series if it exists.

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} =? \tag{1.6}$$

Notice the result in (a): $\frac{x^2}{2} = \frac{l^2}{6} + 2l^2 \sum_{n=1}^{\infty} \frac{(-1)^n \cos(\frac{n\pi x}{l})}{(\pi n)^2}$
when $x = 0 \Rightarrow 0 = \frac{l^2}{6} + 2l^2 \sum_{n=1}^{\infty} \frac{(-1)^n}{\pi^2 n^2} \Rightarrow \frac{l^2}{6} = -2l^2 \sum_{n=1}^{\infty} \frac{(-1)^n}{\pi^2 n^2}$
divide $\frac{2l^2}{\pi^2}$ in both side $\Rightarrow \frac{\pi^2}{12} = -\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}$
We get $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}$ converges to $\frac{\pi^2}{12}$

Definition 1.1 *Uniform Convergence*
Let I be an interval on $\mathbb{R}$ and $f_k : I \to \mathbb{R}$ be a real-valued function on I. The sequence of functions $\{f_k\}_{k\in\mathbb{N}}$ is said to converge uniformly on I to the function $f : I \to \mathbb{R}$ if $\forall \epsilon > 0, \exists N = N(\epsilon) \in \mathbb{N}$ such that $\forall x \in I$ and $k > N, |f_k(x) - f(x)| < \epsilon$.

1.2 Independent Component Analysis (ICA)

1.2.1 Motivation: Blind Source Separation (BSS)

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

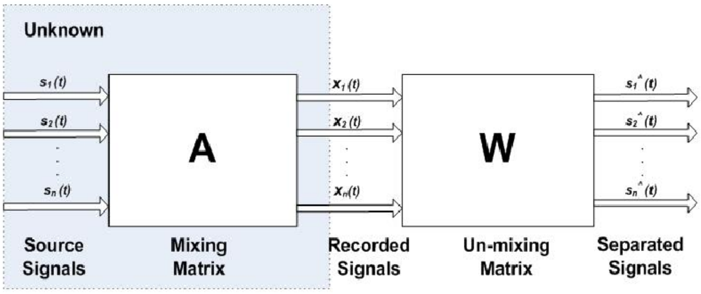


Figure: Blind Source Separation (BSS) [Naik and Kumar,2011]

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.

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

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 lab, we only introduce **ICA**.

1.2.2 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.

$$X = \hat{A}\hat{S} \iff \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ \vdots & \vdots & \vdots \\ - & x_m^T & - \end{bmatrix} = \hat{A}_{m\times n} \begin{bmatrix} - & \hat{s}_1^T & - \\ - & \hat{s}_2^T & - \\ \vdots & \vdots & \vdots \\ - & \hat{s}_n^T & - \end{bmatrix} \tag{1.7}$$

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

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.

Definition 1.2 *White signal*
White signals are defined as any $z \in \mathbb{R}^{d\times 1}$ which satisfying

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

Note: 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 $\hat{S} = \hat{B}X$.

Problem 2: Whiteness property is preserved under orthogonal transformations (5 points)

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

$$m_{Uz} = m_z \ \& \ C_{Uz} = C_z \tag{1.8}$$

We know U is an orthogonal matrix
So we can get $(Uz)^T(Uz) = z^TU^TUz = z^TIz = z^Tz$

1. Prove $m_{Uz} = m_z$:
   First we have $(Uz)^T(Uz) = z^Tz$
   $\rightarrow (Uz - m_{Uz})^T(Uz - m_{Uz}) = (z - m_z)^T(z - m_z)$
   $\rightarrow (Uz - m_{Uz})^T(Uz - m_{Uz}) = z^Tz \quad \because m_z = 0$
   $\rightarrow m_{Uz} = 0 = m_z$
2. Prove $C_{Uz} = C_z = I_d$:
   $C_{Uz} = E[(Uz - m_{Uz})(Uz - m_{Uz})^T]$
   $m_{Uz} = 0$ from previous provement $\rightarrow C_{Uz} = E[Uz(Uz)^T] = E[Uzz^TU^T]$
   $\rightarrow C_{Uz} = E[UI_dU^T] = E[UU^TI_d] = E[II_d]$
   $\rightarrow C_{Uz} = E[I_d] = I_d = E[zz^T] = C_z$

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

$$Z_{d\times m} = V_{d\times d}X^T_{d\times m} \iff \begin{bmatrix} | & | & \cdots & | \\ z_1 & z_2 & \cdots & z_m \\ | & | & \cdots & | \end{bmatrix} = V_{d\times d}\begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & \cdots & | \end{bmatrix} \tag{1.9}$$

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

$$\hat{S}^T_{d\times m} = U_{d\times d}Z_{d\times m} = U_{d\times d}V_{d\times d}X^T_{d\times m} = \hat{B}_{d\times d}X^T_{d\times m} \iff \begin{bmatrix} | & | & \cdots & | \\ \hat{s}_1 & \hat{s}_2 & \cdots & \hat{s}_m \\ | & | & \cdots & | \end{bmatrix} = \hat{B}_{d\times d}\begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & \cdots & | \end{bmatrix} \tag{1.10}$$

where $U \in \mathbb{R}^{d\times d}$ is an orthogonal transformation matrix.
Hence it's necessary to estimate U!

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 $s_j = Bx_j$.

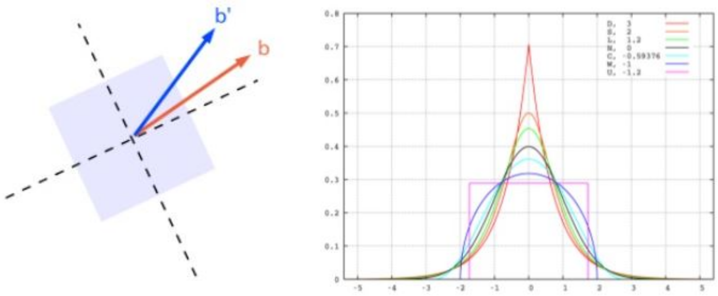


Figure: (Left) The Non-gaussianity of b is larger than $b^{'}$, (Right) Distributions [Yokota, 2012]

Kurtosis is a measure of non-gaussianity

Definition 1.3 *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

$$\hat{b} = \max_b \|kurt(b^T x)\| \tag{1.11}$$

We consider $z = Vx$ is a white signal given from source signal x, then we could rewrite (1.11) as:

Problem 3: Solving ICA problem by kurtosis (10 points)

$$\max \|kurt(w^T z)\| \; with \; w^T w = 1 \tag{1.12}$$

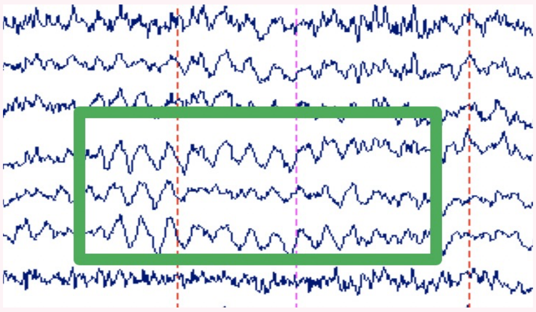
First we rewrite the kurt func: $kurt(y) = E[y^4] - 3(E[y^2])^2 = E[(y^T y)^2] - 3(E[y^T y])^2$
Def $J(W) = kurt(w^T z), F(w) = ||J(W)||$, goal is to find $max(F(w))$ with $w^T w = 1$
Solution: $J(w) = kurt(w^T z) = E[(w^T z)^4] - 3E[(w^T z)^2]^2$
Def. $y = w^T z$, then $y^T y = z^T w w^T z$ def= $h(w)$
$\rightarrow h'(w) = 2zz^T w$
$J(w) = kurt(y) = E[(y^T y)^2] - 3(E[y^T y])^2 = E[h(w)^2] - 3E[h(w)]^2$
$\rightarrow J'(w) = 2E[h(w)h'(w)] - 6E[h(w)]E[h'(w)]$
$\rightarrow J'(w) = 2 \times 2E[(z^T w w^T z)(zz^T w)] - 6 \times 2E[z^T w w^T z]E[zz^T w]$
$\rightarrow J'(w) = 4(E[(z^T w w^T z)(zz^T w)] - 3E[z^T w w^T z]E[zz^T w])$
then we can get $F'(w) = \frac{\partial ||J||}{\partial w} = 4sgn(kurt(w^T z))(E[(z^T w w^T z)(zz^T w)] - 3E[z^T w w^T z]E[zz^T w])$
then we just need to find w that can make $F'(w) = 0$ to get the $max||kurt(w^T z)||$

2 Multiple choices

Please give a brief explanation for option(s) you choose. Answering without any description will get 0 point.

Problem 4 (5 points)

The image here shows a **2-second** period of EEG from several different electrodes. What is the nature of the oscillating signal shown in the green box? **Select all correct options.**



(A) It is a 1 Hz oscillation

(B) It is a 5 Hz oscillation

(C) It is a 10 Hz oscillation

(D) It is a delta-band oscillation

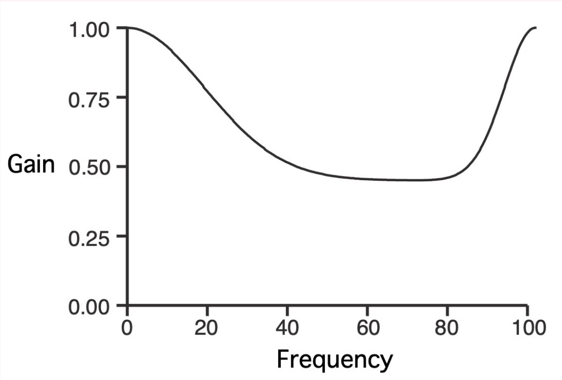
(E) It is a theta-band oscillation

(F) It is an alpha-band oscillation

Answer: (C), (F)
Explanation: We def k waves in one second as k Hz. In this EEG diagram, we can do a simply counting. In each row, there has more than 20 peaks. Hence, (A) and (B) are impossible, (C) can be one of answer. Following the def. of θ, δ, α-wave, δ-wave is under 4Hz, θ-wave is ranged in 4-8Hz, α-wave is ranged in 8-13Hz. Hence, answer should be (F).

Problem 5 (5 points)

Which of the following statements are true of the frequency response function shown below?



(A) This filter would almost completely eliminate very low frequencies (because the gain is near 1.00 for low frequencies).

(B) This filter would have very little effect on very low frequencies (because the gain is near 1.00 for low frequencies).

(C) This filter would reduce frequencies of 40-80 Hz by 50% (because the gain is near 0.50 for these frequencies).

(D) This is a high-pass filter.

(E) This is a low-pass filter.

Answer: (B), (C)
Explanation: The gain in big in low frequencies, so (A) is wrong. We can observe the phenomenon described in (B), (C) in the diagram. (D), (E) are wrong because it is a band-stop filter.

3 Programming problem

Please use the data: sXD_5678.set to answer the following problems.

3.0.1 Dataset Information

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.

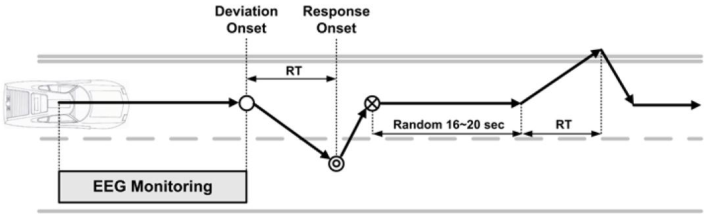


Figure: Event-related lane-departure task
[Kuan-Chih Huang and Jung,2016],[Chin-Teng Lin and Jung, 2010]

Problem 6 (2+1+1+1+4+1=10 points)

Please following the following steps:

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).

Problem 7 (0+0+0+1+1+4+1+8+5=20 points)

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).

After above preprocessing steps... ...
(a) Compare results (e.g. component map) and try to explain your observations.
(b) Explain why it takes less time this time?

Problem 8 (10+10+10=30 points)

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

References

[1] 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.

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

[3] 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.

[4] Ganesh R. Naik and Dinesh K Kumar.*An Overview of Independent Component Analysis and Its Applications*, Informatica, 35:63–81, 2011.

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

[6] Tatsuya Yokota. *Independent Component Analysis for Blind Source Separation*, Remote Sensing, 3(6):1104–1138, 2012, Molecular Diversity Preservation International.

4 Feedback for Lab 2

This part is not for grading but for understanding learning situation of each student. Please give us your feedback and comments.

4.1 Work Division

For example,

Student ID	Name	Be response for... ...
123456789	Tony	solving Problem 1 and designing preprocessing algorithm in problem 7
987654321	May	solving Problem 3 and designing preprocessing algorithm in problem 7

4.2 Suggestions and Comments

4.2.1 For instructor

4.2.2 For teaching assistant(s)

4.2.2.a For Min-Jiun

4.2.2.b For Eric

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.

Day	Time	Office
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.