Development of an Automated MATLAB-Based Platform for the Analysis of Massive EEG Datasets

Daniel J. Fonseca¹, Isaac Heim²

^{1,2}*Department of Mechanical Engineering, The University of Alabama, Tuscaloosa, Alabama, USA*

Abstract— EEG studies consist of multiple sessions. Individually loading and processing each data set is repetitive and time consuming. Finding the action markers embedded into the data can be difficult since the times of these markers vary greatly between individuals. These markers indicate sections that need to be analyzed. Based on the particularities of each marker, individual data segments or sections have to be treated differently or independently from each other. Therefore, splitting an EEG data file in order to compare sections or even analyze one section independently can be beneficial. However, de-identified and uniquely named files can be difficult to run through a program since they usually follow ununiformed naming conventions. While file re-naming is an option, this adds additional steps into identifying the nature and uniqueness of the data. Also, the manually re-naming of data files is prone to error in larger studies. All of these challenges add up to considerable hurdles and difficulties when processing and filtering data originated from multiple EEG sessions. A series of MATLAB computer scripts developed by the authors of this paper address these problems.

Keywords—EEG signals, Filtering data, MATLAB scripts, Neuroscience, Neurophysiology.

I. PROBLEM BACKGROUND

Electroencephalography (EEG) is a tool used to better understand the inner workings of the human brain. By utilizing the bioelectrical potentials generated by the cortex nerve cells within the brain, it is possible to identify which areas of the brain are active [1]. EEG is a non-invasive tool. EEG is an important part of understanding neurological disorders, such as, epilepsy, tumors, locating head damages, among others [2,3]. EEG source imaging is utilized in many important applications in cognitive neuroscience and clinical neuroscience. The clinical areas include neurology, psychiatry, and psychopharmacology. Cognitive neuroscience applications mainly focus on investigating the temporal aspects by analyzing event related potentials (ERP). In this sense, neurology focuses more on the study of motor evoked potential, but in clinical applications, the main concern is localizing epileptic foci. In psychiatry and psychopharmacology, the focus tends to be towards localizing the sources in certain frequency bands.

While all of these applications aim to localize the sources, the pre- processing of the data is somewhat different for each application [4].

EEG signals are often regarded as a measurement of neural activity, but they do not give a complete understanding of the underlying neurophysiology. The neural activity of the brain is mediated by the synaptic interactions and the resulting action potential propagation. This means the EEGs signals measurements are the excitation of a neuron traveling along its axon towards the axon terminal which is where the action potentials are transmitted to other neurons, muscle cells or glands [5]. These surface electrodes do not exactly show where the electrical signals are coming from. By knowing the generators, it is possible to predict which electrodes will be activated from each given generator. When a channel is activated, it is more difficult to determine which generator in the brain caused the excitation. The channels are based on the surface of the scalp and therefore cannot convey some of the deeper activity within the brain. There are several different source localization programs and methods.

EEG studies contain numerous files of data. This type of data can have markers to indicate sections that need to be analyzed differently or separately from each other. It can be difficult to stop recording during sessions; therefore, markers are used to signal beginnings and ends. In addition to these challenges, there is also the fact that studies can contain de-identified and uniquely named files that cannot be easily organized into a list to run through a script. All of these challenges build on each other

II. THE NEED FOR MAKERS IN EEG DATA: A CASE STUDY

The MATLAB data process shown below has applicability with EEG data across disciplines. For example, the data presented in this case study are examples from two multidisciplinary research projects. The first project involved the neurological analysis of brain-wave activation patterns of college students during an immersive, mindfulness experience. The second project examined law enforcement officers (LEOs) decision-making during high threat situations. The project associated with the analysis method will be denoted in text and with the corresponding figures and tables.

For the mindfulness study, EEG data were recorded as participants completed two conditions: (a) a baseline condition and (b) virtual-reality, mindfulness condition. Each condition lasted approximately 10-minutes in length. EEG data was collected using a 64-channel mobile EEG amplifier, EEGO Sports [6]. The unit records data using EEG caps and electrodes attached to a high-performance Windows 8 tablet. The mobile EEG unit has been used to record physiological and neurological data on participants who move frequently (e.g., athletes, LEOs). It was chosen to collect data from 20 channels focused on studying neural processes while engaged in firing a weapon [7]. The reference electrode was CPz and the 19 channels utilized were Fp1, Fp2, F7, F3, Pz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T7, T8, T6, TP7, TP8, O1 and O2. Data was collected at a sampling rate of 500 Hz in an ambient temperature room.

Pre-processing of EEG data was conducted in ASA software package [6] and EEGLab toolbox for MATLAB [8]. Using ASA, a lowpass filter of 30 Hz to filter extraneous noise as first utilized. Data was then converted using EEGLab toolbox for MATLAB [8] which was used for the remainder of analysis. Continuous data as visually inspected to look for channels affected by superfluous sources of noise. Independent Component Analysis (ICA) was utilized in EEGLab [9] to re-reference problematic channels and to remove artifacts (e.g., eye blinks, muscle movements). ICA has been shown to be useful in isolating and removing eye and movement artifacts commonly found in data where some participant movement was involved, as was the case in the current study [4].

III. APPROACH

EEG studies consist of multiple sessions. Individually loading and processing each data set is repetitive and time consuming. Finding the action markers embedded into the data can be difficult since the times of these markers vary greatly between individuals. These markers indicate sections that need to be analyzed. Based on the particularities of each marker, individual data segments or sections have to be treated differently or independently from each other. Therefore, splitting the file in order to compare sections or even analyze one section independently can be beneficial.

When a marker designates the beginning of a distinctive segment of data, it becomes possible to extract the aforementioned section. Deidentified and uniquely named files can be difficult to run through a program since they usually follow ununiformed naming conventions. While file re-naming is an option, this adds additional steps into identifying the nature and uniqueness of the data.

Also, the manually re-naming of data files is prone to error in larger studies. All of these challenges add up to considerable hurdles and difficulties when processing and filtering data originated from multiple EEG sessions. A series of MATLAB computer scripts developed by the authors of this paper address these problems. The first script was programmed to quickly look through a main folder, detecting and ultimately reading subsequent subfolders inside the main one, as shown in Figure 1. It only sorts through subfolders, so data in the main folder needs to be sectioned into a subfolder. Due to Microsoft's Windows naming convention, if numbers are used while naming the files, it is then necessary to have leading zeros that match the digits of the biggest number, i.e., if the study in interest has subjects in the double digits, files should then be named 01, 02, 03, etc. For triple digits, 001, 002, 003, etc. This is due to how the '.' character interacts with the list. Normal alphabetical naming conventions should work as expected.

Figure 1. Folder Structure for Data Import

The second script developed is related to the event splitting code, created for the LEO study. An event's start and end terms are specified by the user. The script collects all markers in the file from the table created when the data is loaded, as shown in Figure 2. The script then looks for the start and end terms. It is also possible to add a shoot term, which is a third term that ends the section early. This is useful for data that becomes too noisy after a certain trigger but before the end of the data collection. The data will be collected in as shown in Figure 3 to be used later in the script. EEG data following the shooting events is called the reflection data. This is the time between the ending and the next beginning of a session. Therefore, the script's output is the scenario from start to end, or start to shoot, if it is marked to end early. Scenario data is named with an S before the data name in the output folder.

There is also an end to start reflection data set saved with an R before the data name, as shown in Figure 4. For the final reflection a “trueend” term was added for the last iteration, where the final end will then match up with the true end, giving the last reflection.

Fields	latency	type	duration
1	0	'0, Impedance'	0
2	1	'1025, 1 school library start (shoot)'	0
3	5001	'Scenario 1 shot'	0
4	8501	'1026, 1- school library end'	0
5	11501	'StartS2'	0
6	13501	'shot'	0
7	15001	'end'	0
8	16501	'Start S3'	0
9	19001	'end'	0
10	20001	'start s4'	0
11	23501	'shot'	0
12	26501	'end'	0
13	28501	'start s5'	0
14	31501	'end'	0

Figure 2. EEG Event Variable Created When a File Is Loaded

start
1
2
3
4
5
6

end
1
2
3
4
5
6

shot
1
2
3

Figure 3. (Left) Start terms including the trueend for retrieving the final reflection. (Right) End terms. (Bottom) The shot terms.

Name	Date modified	Type	Size
R1.fdt	8/23/2018 1:21 AM	FDT File	235 KB
R1.set	8/23/2018 1:21 AM	SET File	33 KB
R2.fdt	8/23/2018 1:21 AM	FDT File	118 KB
R2.set	8/23/2018 1:21 AM	SET File	30 KB
R3.fdt	8/23/2018 1:21 AM	FDT File	79 KB
R3.set	8/23/2018 1:21 AM	SET File	29 KB
R4.fdt	8/23/2018 1:21 AM	FDT File	157 KB
R4.set	8/23/2018 1:21 AM	SET File	31 KB
R5.fdt	8/23/2018 1:21 AM	FDT File	2,383 KB
R5.set	8/23/2018 1:21 AM	SET File	87 KB
S1.fdt	8/23/2018 1:21 AM	FDT File	391 KB
S1.set	8/23/2018 1:21 AM	SET File	37 KB
S2.fdt	8/23/2018 1:21 AM	FDT File	157 KB
S2.set	8/23/2018 1:21 AM	SET File	31 KB
S3.fdt	8/23/2018 1:21 AM	FDT File	196 KB
S3.set	8/23/2018 1:21 AM	SET File	32 KB
S4.fdt	8/23/2018 1:21 AM	FDT File	274 KB
S4.set	8/23/2018 1:21 AM	SET File	34 KB
S5.fdt	8/23/2018 1:21 AM	FDT File	235 KB
S5.set	8/23/2018 1:21 AM	SET File	33 KB

Figure 4 Output of the Police Split Script for One File

The third script consists of filtering and a mean-power analysis created for the mindfulness study. Mean-power analysis involves calculating the average power frequency for a relative time period. Such analyses have been utilized in prior mindfulness-based research [10]. The wave frequencies to be analyzed are set at the beginning of the script, along with other data to be initialized. In order to use the interpolation, channel locations need to be established. The process of interpolation involves replacing the power frequency of an unusable electrode with an estimate of the power frequency from adjacent and useable electrodes. Currently, this is done by a file called 'Fullset.mat' which loads up channel locations specific to the study. This can be made from saving the EEG.chanlocs from a dataset with all desired channels after setting the channel locations in the edit drop down menu. To do so, create a new variable using the following “Fullset=EEG.chanlocs;” or “Fullset=ALLEEG.chanlocs;” based on how the EEG data has been saved in the workspace. Then right click on Fullset and save as, in this case, Fullset.mat. The data should appear as in Figure 5.

Fields	Labels	type	theta	radius	X	Y	Z	sph_theta	sph_phi	sph_radius	unchan	ref
1	Fp1	0	-17.9260	0.5150	80.7840	26.1330	-4.0011	17.9260	-2.6980	85	1	0
2	Fp2	0	17.9260	0.5150	80.7840	26.1330	-4.0011	-17.9260	-2.6980	85	2	0
3	F7	0	-53.9130	0.5281	49.8714	68.4233	-7.4895	53.9130	-5.0550	85	3	0
4	F3	0	-39.9470	0.3446	57.5511	48.2004	39.8697	39.9470	27.9730	85	4	0
5	Fz	0	0	0.2534	60.7385	0	59.4629	0	44.3920	85	5	0
6	F4	0	39.8970	0.3445	57.5940	-48.1426	39.8920	-39.8970	27.9900	85	6	0
7	F8	0	53.8670	0.5281	49.9265	-68.3836	-7.4851	-53.8670	-5.0520	85	7	0
8	T7	0	-90	0.5332	5.1765e-15	84.5385	-8.8451	90	-5.9730	85	8	0
9	C3	0	-90	0.2667	3.8681e-15	63.1713	56.8717	90	41.9960	85	9	0
10	Cz	0	0	0	5.2047e-15	0	85	0	90	85	10	0
11	C4	0	90	0.2667	3.8679e-15	-63.1673	56.8761	-90	42	85	11	0
12	T8	0	90	0.5332	5.1765e-15	-84.5385	-8.8451	-90	-5.9730	85	12	0
13	P7	0	-126.0870	0.5281	-49.8714	68.4233	-7.4895	126.0870	-5.0550	85	13	0
14	P3	0	-140.0530	0.3446	-57.5511	48.2004	39.8697	140.0530	27.9730	85	14	0
15	Pz	0	180	0.2534	-60.7385	-7.4883e-15	59.4629	-180	44.3920	85	15	0
16	P4	0	140.1030	0.3445	-57.5940	-48.1426	39.8920	-140.1030	27.9900	85	16	0
17	P8	0	126.1330	0.5281	-49.9265	-68.3836	-7.4851	-126.1330	-5.0520	85	17	0
18	O1	0	-162.0740	0.5150	-80.7840	26.1330	-4.0011	162.0740	-2.6980	85	18	0
19	O2	0	162.0740	0.5150	-80.7840	26.1330	-4.0011	-162.0740	-2.6980	85	19	0
20	Oz	0	180	0.5067	-84.9812	-1.0407e-14	-1.7960	-180	-1.2040	85	20	0

Figure 5 Channel Locations From Fullset.mat

The subsequent portion of this script is the file reader which starts processing all the subfolders found, while going through all the wave frequencies specified and subjects/scenarios found in the subfolders. The first part in this process is `pop_chanedit` which assigns channel locations necessary for data interpolation. Next is the `pop_rejcont` which rejects all the data outside of the defined parameters. This may need to be changed based on how the data needs to be cleaned. Next to be triggered are `pop_runica` which runs independent component analysis (ICA) on the data followed by `pop_rejchan` which removes any faulty channels. ICA is a commonly utilized filtering process that removes data artifacts (e.g., eye blinks, muscle movements), commonly found in EEG data — particularly in studies where participants move [11]. The lost channels can be later interpolated back into the analysis with `pop_interp` by taking data from nearby channels and using them to calculate new data for the removed channel. The final part will be to re-reference the data in order to get a more accurate reference for the data. This will reference the data based on the mean of all the data rather than the current reference, which can vary between studies. This is possible by using the command `pop_reref`. These commands can be modified to fit most conventional EEG data processing studies. Typing “help (command)” displays how each of these programmed commands can be modified to meet the analyst’s particular needs. The EEG data sets are further processed to find the mean power using `spectopo` which is a version of `spectopo` edited to output the numbers in base 10 rather than in decibels (dB). This was done by editing instances of `eegspecdB = 10*log10(eegspecdB)` to `eegspecdB = eegspecdB`. When the script is done running, the command window displays the found bad channels, if interpolation was disabled.

The final numbers are then collected in `fulleasycopy` as shown in Figure 6.

	1	2	3
1	19x40 double	19x40 double	
2			
3			

Figure 6 Data Shown When fulleasycopy Variable Is Double Clicked

Figure 3 shows two folders of 19 subjects. Nineteen baselines and 19 intervention data sets when these cells are double clicked. To see the final data array, the cell can be double clicked to show data similar to that in Figure 7. Two wave analyses were run on the shown data. When more waves are analyzed, the data add onto the right side of the cell based on how the wave analysis was defined. This allows for the data to be easily copied into Microsoft Excel. The top labels should be the channels as stated in the `channel_list` variable. The left labels can be found in the `ColFilesT` variable. Alternatively, the `FolderMeanpower` array splits the data into wave analysis sections, as shown in Figure 8. These numbers can be accessed by double clicking the blue 19x20 double cells.

	1	2	3	4	5	6	7	8	9	10	11
1	3.9468	4.3709	22.0945	7.4422	5.0522	6.0097	21.8327	6.7397	4.8309	6.6905	2.9024
2	23.8292	7.7023	5.7081	12.0396	3.2276	7.4455	11.9108	5.3011	2.4866	2.5591	1.7623
3	5.7663	4.9444	13.5821	8.7341	8.6073	8.7905	11.3070	5.9854	1.2009	3.9065	2.2629
4	108.8334	55.9710	39.6639	11.0944	7.7033	7.4316	21.3473	9.3594	3.8046	3.0427	4.4247
5	95.6168	76.8766	42.6833	15.2643	8.1345	16.7222	8.3372	10.2986	7.5640	9.3598	11.3360
6	5.7796	4.5844	25.3412	14.0428	10.2522	22.7822	13.6600	14.7553	1.9366	2.1115	12.0985
7	9.4705	9.7959	2.3574	3.8597	3.5800	4.2986	2.8313	3.6322	1.8217	3.5410	1.7354
8	39.5108	41.6839	11.8123	7.5944	6.3926	5.7833	18.0867	3.1011	3.7598	2.5353	6.0586
9	23.4871	35.9805	3.1361	3.2872	3.0227	4.7717	6.3994	7.5886	1.7279	2.0106	11.0419
10	4.6109	4.5266	23.8323	5.9205	5.6291	5.7399	21.2794	9.3339	2.8487	3.8003	3.8459
11	7.2608	6.2844	22.3802	11.9421	10.4463	9.3560	29.5898	7.9433	4.1080	6.7986	6.0237
12	17.2582	90.3068	48.9411	14.8613	6.5558	4.3200	5.4354	3.8745	2.2046	2.4803	2.4162
13	5.8580	5.2097	16.6946	11.7317	6.1435	11.7796	18.4952	7.8740	4.4602	7.1006	4.0900
14	53.3036	47.8963	23.9647	5.1619	4.1572	4.2258	16.5447	4.9388	3.2170	5.0096	3.1187
15	17.2582	90.3068	48.9411	14.8613	6.5558	4.3200	5.4354	3.8745	2.2046	2.4803	2.4162
16	10.2173	44.8408	11.0234	4.4245	7.0612	5.0921	20.9892	5.6575	3.1902	2.9506	1.7021
17	109.1776	54.2163	38.8642	7.5388	7.8960	10.8633	20.9961	8.9045	3.5572	4.5327	4.8665
18	27.3404	6.5087	4.9963	12.1372	3.7555	7.7258	12.3344	5.5451	2.6254	2.8135	1.9430
19	7.0143	7.1512	23.3461	17.7667	13.1693	8.1831	32.7841	28.4559	24.7770	6.3024	8.2640
20											

Figure 7 The First 11 of 40 (for this case) Columns of Fulleasycopy{1,1} Cell Array Expanded. For 20 Channels It Will Repeat At The First Channel Starting At Column 21.

The figure shows two MATLAB workspace windows. The top window, titled 'FolderMeanpowerarray', displays a 1x2 cell array. The first cell contains a 3x4 cell array, and the second cell contains a 3x4 cell array. The bottom window, titled 'FolderMeanpowerarray{1, 1}', displays the data from the first cell, which is a 3x4 cell array. The data is as follows:

	1	2	3	4
1	0.5000	3.1000	8	12.1000
2	3	7.9000	12	24
3	19x20 double	19x20 double	19x20 double	19x20 double

Figure 8 Top: FolderMeanpowerarray Structure. Bottom: Data Displayed from Double Clicking the 3x4 Cell.

IV. RESULTS AND FINAL REMARKS

As already mentioned, the EEG data processing demands that ultimately led to the creation of the discussed MATLAB script consisted of the baselines and interventions readings for 20 channels. Interventions and baselines were both ten minutes long on average. The script used an average of 1.2 GB of memory on a 16 GB RAM desktop during the analysis for the mindfulness study. It was used to run the mean power analysis over four different waves. The waves analyzed were Delta, Theta, Alpha, and Beta. The data used for this timed trail was previously put through Independent Component Analysis (ICA) and saved. It is worth noting that ICA can be run on data sets and saved outside of running the full script. This saves time on multiple iterations using the same data sets. The full execution of this script took 27 minutes. This was fully automated, requiring no further user inputs after the start of the script. It was still necessary to extract the data from the workspace, which takes no more than a minute.

Overall, there were four separate wave analyses for each of the 38 data sets consisting of 20 channels in each. This is including the baselines for a total of 3,040 calculations. Therefore, the script accounts for over 100 calculations per minute. This far outclasses what could be done using the normal graphical user interface (GUI) approach.

In addition to the calculations, the data is collected in a way that allows for easy export and further analysis. It is also worth mentioning that the scripts created are modular. It is possible to swap out the mean power analysis for another type of analysis. It would also be possible to add code to automatically save the data from the workspace as Excel documents with channel and subject labels based on the user needs. The scripts are written and commented in a manner that shows a logical progression and facilitates the editing of sections of code for different studies.

REFERENCES

- [1] Nguyen, T., Khosravi, A., Creighton, D., and Nahavandi, S. "EEG signal classification for BCI applications by wavelets and interval type-2 fuzzy logic systems," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4370-4380, 2015.
- [2] Jatoti, M. A., Kamel, N., Malik, A. S., Faye, I., and Begum, T. "A survey of methods used for source localization using EEG signals," *Biomedical Signal Processing and Control*, vol. 11, pp. 42-52, 2014.
- [3] Rangaprakash, D., and Pradhan, N. "Study of phase synchronization in multichannel seizure EEG using nonlinear recurrence measure," *Biomedical Signal Processing and Control*, vol. 11, pp. 114-122, 2014.
- [4] Michel, C. M., Murray, M. M., Lantz, G., Gonzalez, S., Spinelli, L., and de Peralta, R. G. "EEG source imaging," *Clinical neurophysiology*, vol. 115, no. 10, pp. 2195-2222, 2004.
- [5] Jorge, J., Van Der Zwaag, W., and . Figueiredo, P. "EEG-fMRI integration for the study of human brain function," *NeuroImage*, vol. 102, pp. 24-34, 2014.
- [6] Zanow, F., and Knösche, T. R. "ASA-advanced source analysis of continuous and event related EEG/MEG signals," *Brain Topography*, vol. 16, no. 4, pp. 287-290, 2004.
- [7] Domingues, C.A., Machado, S., Cavaleiro, E.G., Furtado, V., Cagy, M., Ribeiro, P., and Piedade, R. "Alpha absolute power: Motor learning of practical pistol shooting," *Archivos de Neuro-Psiquiatria*, vol. 66, no. 2B, pp. 336-340, 2008.
- [8] Delorme, A. and Makeig, S. "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9-21, 2004.
- [9] Delorme, A., Sejnowski, T., and Makeig, S. "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *Neuroimage*, vol. 34, no. 4, pp. 1443-1449, 2007.
- [10] Lagopoulos, J., Xu, J., Rasmussen, I., Vik, A., Malhi, G.S., Eliassen, C.F., and Davanger, S. "Increased theta and alpha EEG activity during nondirective meditation," *The Journal of Alternative and Complementary Medicine*, vol. 15, pp. 1187-1192, 2009.
- [11] Thompson, T., Steffert, T., Ros, T., Leach, J., and Gruzeliier, J. "EEG applications for sport and performance," *Methods*, vol. 45, no. 4, pp. 279-288, 2008.