MODELING EEG MICRO-STATES USING MIXTURES OF THE WATSON DISTRIBUTION

Arturo Arranz (S160412), Manuel Montoya (S162706) & Martin Simon (S163008)

Technical University of Denmark

ABSTRACT

Current state-of-the-art techniques for modeling electroencephalographs (EEG) data are based in a modified K-means clustering [1]. Such techniques lack a statistical framework that allows to find the optimal number of clusters to express the distribution. The aim of this project is to model EEG using mixture of Watson distributions (moW). To compute the parameters of the mixture we use the Expectation-Maximization algorithm (EM) with different assumption on the data samples. First we assume that the data points are independent, resulting in regular moW and later we assume a Markov Chain dependency between the samples, resulting in a Hidden Markov Model. In EEG data, the key information is believed to be scale and polarity invariant so an axially symmetric directional distribution, such as Watson, is a good candidate for the data modeling.

***Index Terms*—** electroencephalography clustering, mixture models, Watson distribution, Expectation-Maximization algorithm, Hidden Markov Models

1. INTRODUCTION

Event related brain activity can be modeled by micro-states obtained from the spatial distributions of electric potential measured by EEG. The event-related potential (ERP) micro-states, typically lasting 80-120 ms[2], oscillates in polarity and is independent of the intensity of the signal[3]. Most studies reveal the same 4 classes of micro-state topography[4, 5].

These four micro-states are believed to describe up to 80% of the variance and the mixture of duration and sequences give a rich temporal map[6]. Still, many studies come to this conclusion, there are exceptions to the rule, where, depending on the application, five[7, 8] or even seven[9] micro-states converge as optimal.

Although EEG data is a time-series, current clustering methods do not attribute a dependency between the former and prior state when classifying the micro-states. Two different models are used in the scope of this project: EM with an independence assumption between the samples and EM with a Markov Chain of order 1 assumption between the samples (Hidden Markov Model).

The EEG data used to evaluate the algorithms comes from multi-subject, multi-modal human neuro-imaging recordings. The volunteers performed a simple perceptual task on pictures of famous and scrambled faces during two visits to the laboratory[10].

2. STATE OF THE ART

Current state-of-the-art methods use K-means clustering as it can efficiently classify a large number of continuous numerical data of high-dimensions. A modified version of K-means[1] is used in an EEGlab toolbox[11].

The modified K-means generates directional clusters but it lacks a statistical framework. It can be considered a special case of the moW if we impose a hard decision on the clusters and concentration parameter. It does not guarantee to find the minimum and may need to initialize multiple times with randomly selected V_t vectors from the dataset[1].

The toolbox also includes a method to automate optimal micro-state segmentation, which is based on minimizing the modified predictive residual variance. The method was first published in the Journal of Neuroscience Methods[12], where the authors described setting a maximum of 14 micro-states and the segmentation was repeated 10 times to find the one with lowest residual variance. Although the results of this study yielded four micro-states as optimal it is clear that researchers wish to refrain from limiting the number of clusters with a convention[1].

Another recent approach is to instead of calculating the micro-state α_{kt} as a binary output, to calculate the probability $p(\alpha_{kt})$ instead[13].The method outperfomed EEGlab K-means clustering at higher numbers of clusters (i.e K=7) and was competitive in lower number of clusters (i.e K=4).

3. THE MULTIVARIATE WATSON DISTRIBUTION

The Watson distribution models data which is axially symmetric ($\pm\boldsymbol{x}$ vectors are equivalent) and scale invariant i.e. the vectors are unitary. This are two interesting properties for micro-state modeling since we are not interested in the magnitude of the measurements nor in their polarity.

Let $\mathbb{S}^{p-1} = \{\boldsymbol{x}|\boldsymbol{x} \in \mathbb{R}^p, ||\boldsymbol{x}||_2 = 1\}$ be the (p-1)-dimensional hypersphere centered in the origin. Then the Watson probability density function is

$$W_p(\boldsymbol{x}|\boldsymbol{\mu},\kappa) = c_p(\kappa)e^{\kappa(\boldsymbol{\mu}^T\boldsymbol{x})^2}, \qquad \boldsymbol{x} \in \mathbb{S}^{p-1} \tag{1}$$

where $\kappa \in \mathbb{R}$, is the *concentration parameter* and $\mu \in \mathbb{S}^{p-1}$ is the *mean direction*. The normalisation constant $c_p(\kappa)$ is given by

$$c_p(\kappa) = \frac{\Gamma(p/2)}{2\pi^{(p/2)}M(\frac{1}{2},\frac{p}{2},\kappa)}, \tag{2}$$

where M is Kummers confluent hypergeometric function defined as

$$M(a,c,\kappa) = \sum_{j\geq 0}\frac{a^{\bar{j}}\kappa^j}{c^{\bar{j}}j!}, \qquad a,c,\kappa \in \mathbb{R} \tag{3}$$

and $a^{\bar{0}} = 1, a^{\bar{j}} = a(a+1)...(a+j-1), j \geq 1$, denotes the rising-factorial.

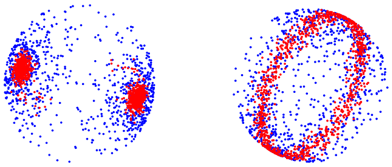
(a) $\kappa > 0$ (b) $\kappa < 0$

Fig. 1. concentration values: (a) positive and (b) negative

The Figure 1 shows a scatter plot from Watson distribution samples. Note that for $\kappa \to \infty$ the samples concentrate around the mean direction while for $\kappa \to -\infty$ the samples concentrate around a ring orthogonal to the mean direction. The uniform distributed case would correspond to $\kappa = 0$.

4. EM ALGORITHM FOR INDEPENDENT MOW

Now we turn our attention to a mixture of several Watson distributions. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_T \in \mathbb{S}^{p-1}\}$ be a sequence of i.i.d samples following a moW of K components. Each sample came from one of clusters distribution but we do not know which one. We model this information using the discrete latent variable s_t associated to each sample x_t. Let us also denote $W_p(\boldsymbol{x}|\boldsymbol{\mu}_i,\kappa_i)$ as the probability density distribution of the i-th cluster and π_i the prior probability of the cluster. Then the observation $\boldsymbol{x}_i$, which is a column vector, has the probability density

$$f(\boldsymbol{x}_t|\theta) = \sum_{i=1}^{K}\pi_i W_p(\boldsymbol{x}_t|\boldsymbol{\mu}_i,\kappa_i)$$

where θ is the set all parameters of the model $\theta = \{\pi, B\}$ being B the parameters of the Watson distribution for the different K clusters, $B = \{\mu_k, \kappa_k\}_{k=1}^{K}$. Then the log-likelihood for the entire dataset X is given by

$$\ell(\boldsymbol{X}|\theta) = \sum_{t=1}^{T} log\Big(\sum_{i=1}^{K}\pi_i W_p(\boldsymbol{x}_t|\boldsymbol{\mu}_i,\kappa_i)\Big) \tag{4}$$

In order to calculate the parameters which maximize (4) the iterative method EM [14] can be used. In every iteration two steps are solved, the *Expectation-step* where the *responsibilities* $r_t(i)$ of each sample is calculated, these $r_t(i)$ are the probability that the sample x_t came from the i-th distribution given all available data $P(s_t = i|X,\theta)$. Then we do the *Maximization-step* where the parameters are recalculated.

E-step:

$$r_t(i) = \frac{\pi_i W_p(\boldsymbol{x}_t|\boldsymbol{\mu}_i,\kappa_i)}{\sum_l \pi_l W_p(\boldsymbol{x}_t|\boldsymbol{\mu}_l,\kappa_l)} \tag{5}$$

M-step:

$$\begin{aligned} &\mu_i = s_1^j \quad if \quad \kappa_i > 0, \qquad \mu_i = s_p^i \quad if \quad \kappa_i < 0 \\ &\kappa_i = g^{-1}(1/2, p/2, r_i), \quad where \quad r_i = \mu_i^T S^i \mu_i \\ &\pi_i = \frac{1}{T}\sum_i r_t(i), \end{aligned} \tag{6}$$

where s_e^j represents the e-th eigenvector of the *weighted scatter matrix*:

$$S^i = \frac{1}{\sum_t r_t(i)}\sum_t r_t(i)\boldsymbol{x}_t\boldsymbol{x}_t^T \tag{7}$$

In order to estimate κ it is necessary to calculate first μ. However, the estimation of μ, depends on the sign of the concentration. This yield to a coupled system (9)-(10). To solve this problem both options are calculated and the solution with higher log-likelihood is chosen.

For getting $\hat{\kappa}$, the iterative Raphson-Newtons method can be used to solve the implicit formula. However, if the guess is not sufficiently good, the method may not converge. Fortunately, an analytical approximation [15] can be calculated

$$\hat{\kappa} = \frac{cr - a}{r(1-r)} + \frac{r}{2c(1-r)} \tag{8}$$

where $a = 1/2$ and $c = p/2$. The algorithm iterate (9) and (10) until the chosen convergence criteria is fulfilled.

5. EM ALGORITHM FOR MARKOV DEPENDENCY BETWEEN SAMPLES

In this version of moW, we assume that the data sequence follows a Markov Chain of order 1. The probability of a sample being generated from a given cluster, depends on the cluster that the previous sample was generated from and so on. Therefore in this model we have an initial probability vector π, and a transition probability matrix, A being a_{ij} the element in the i-th row and the j-th column meaning the probability of jumping to state j when being at state i. The model parameters now are $\theta = \{\pi, A, B\}$. In this case the data consists of N independent chains of T samples.

The Forward-Backward algorithm is used in order to compute the responsibilities $\gamma_t^n(i)$ in an efficient manner. This algorithm divides this probability into the factors $\alpha_t^n(i)$ and $\beta_t^n(i)$. In order to estimate A we also need to compute the probabilities $\xi_t^n(i,j) = P(s_t^n = j, s_{t-1}^n = i|Y^{(n)}, \theta)$.

Since the distributions of the mixture $p_i(x|B_i)$ belong to the exponential family, we can obtain its parameters B_i using moment matching and therefore we can compute them from the new weighted correlation matrix S^i. Therefore the simplified equation for the E and M steps are:

E-step:

$$\begin{aligned}
&\alpha_t^n(i) = p_i(x_t^n|B_i)\sum_{j=1}^{I} a_{ji}\alpha_{t-1}^n(j)\\
&\beta_t^n(i) = \sum_{j=1}^{I} a_{ij}p_j(x_t^n|B_j)\cdot\beta_{t+1}^n(i)\\
&\gamma_t^n(i) \propto \alpha_t(i)\beta_t(i)\\
&\xi_t^n(i,j) \propto \alpha_t^n(i)a_{ij}p_j(x_{t+1}^n|B_j)\beta_{t+1}^n(j)
\end{aligned} \tag{9}$$

M-step:

$$\begin{aligned}
&\pi_i = \frac{1}{N}\sum_{n=1}^{N}\gamma_1^n(i) \qquad N = \sum_{i=1}^{I}\sum_{n=1}^{N}\gamma_1^n(i)\\
&a_{ij} = \frac{1}{E_i}\sum_{n=1}^{N}\sum_{t=2}^{T_n}\xi_t^n(i,j) \qquad E_i = \sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{t=2}^{T_n}\xi_t^n(i,j)\\
&S^i = \frac{1}{\Gamma_i}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_t^n(i)x_t^{(n)}{x_t^{(n)}}^T
\end{aligned} \tag{10}$$

6. DATA DESCRIPTION AND PREPROCESSING

The objective of the project is analyze the mind-state of subjects which are presented with different stimulus, specifically, visual stimulus associated to see pictures of familiar faces from famous people and scramble pictures of faces. From now we will refer to each condition as ”Famous face” and ”Scramble face”.

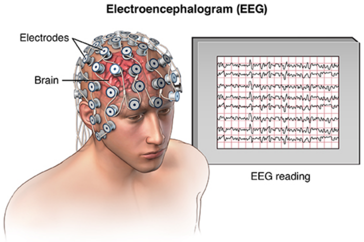

Fig. 2. EEG sampling

Then, EEG allows us to extract mind information by measuring the electric potential that the brain generate in the surface of the head. This technique is represented in Figure 2.

Our dataset contain data from 16 different subjects exposed to the aforementioned stimulus using 70 electrodes/channels. Each subject generated ∼295 trials from each of the conditions and each trial is a time series of 451 EEG samples in the span of 1 second. Every instant can also be represented as the scalp-maps in Figure 4 to have a spacial understanding of the brain activity.

6.1. Preprocessing of the data

The data has already been cleaned to some extent (for example the electric spikes produced by eye movement have being filtered), but further prepossessing is needed. The 3 applied treatments are *average several trials* from same condition, *mean removal* of channels and *projection over the unit hypersphere*.

The aim of *averaging trials* is reducing the noise and it consist on getting one single trial by averaging all of the same condition for each subject.

Mean removal consist on subtracting the mean of all the channels in every instant. This makes all the channels have the same importance since the the information is not in its mean value.

Lastly, all the samples are divided by its L2-norm to *project over the unit hypersphere* which is the constraint imposed by the Watson distribution.

Another possible preprocessing would be using a subset of the channels or a linear combination of them using techniques such as PCA.

7. MICRO-STATE MODELING RESULTS

In this section we will obtain the optimal number of clusters needed to express the data of single subject for both classes by means of 2-fold cross-validation (CV) of the data. We will compare the results obtained for the EM and HMM algorithms, then we will analyze the clusters found using the scalp maps and the time information and finally, we comment other results like inter-subjects results and single trials classification.

7.1. Optimal number of clusters

After preprocessing the data as in section 6.1, we perform CV for different number of clusters K for the EM and HMM algorithm and for each class. The training and validation likelihoods are shown in Figure 3 for the class ”Scramble face”. As we can see, the validations likelihood of both EM and HMM converge for ∼4 when characterizing ”Scramble face”. In the case of ”Famous face” it needs ∼6 clusters.

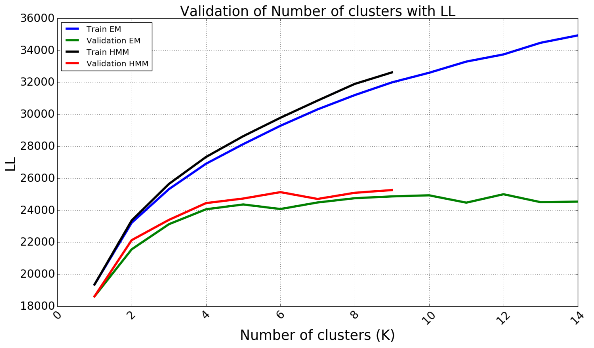

Fig. 3. LL HMM-EM for class "Scramble face"

7.2. Scalp maps of the Clusters

Figures 4 and 5 show the scalp maps produced by EM and HMM respectively. Note that due the polarity of the data, sometimes intense blue might indicate strong activity and sometimes intense red would do it.

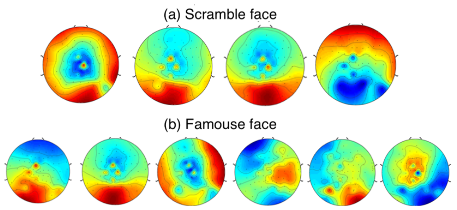

Fig. 4. Scalp maps from each class generated by EM

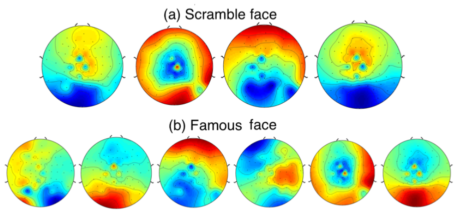

Fig. 5. Scalp maps from each class generated with HMM

All clusters have heavy reliance on the visual cortex and the frontal lobe, which indicates a visually evoked reponse. Both algorithms put more emphasis on the right-Parietal lobe for the "Famous face", which could indicate recognition. Also, the "Famous face" seems to get a stronger response from the right temporal lobe, which is connected with emotional memory. Only one cluster differs qualitatively from both models, specifically the cluster b6 in both figures.

7.3. Time analysis of the Clusters

Figure 6 shows the temporal evolution of responsibilities, $r_t(i)$, from EM algorithm for the class "Scramble face"

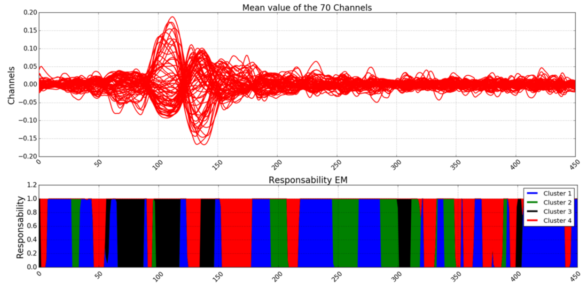

Fig. 6. $r_t(i)$ temporal evolution

From the plot we should note two tings. First, at every instant one of the clusters dominate the responsibility, which is almost like a hard decision. Secondly, each micro-state has some time duration which indicate that the assumption of independence is violated. This is even more obvious when we look at the estimated transition matrix from the HMM model

$$A = \begin{bmatrix} \mathbf{0.93} & 0.04 & 0 & 0.03 \\ 0.03 & \mathbf{0.88} & 0.08 & 0.01 \\ 0.01 & 0.08 & \mathbf{0.89} & 0.02 \\ 0.01 & 0.05 & 0.01 & \mathbf{0.93} \end{bmatrix} \tag{11}$$

the notorious higher values in the diagonal indicate that once you fall in a microstate you remain there for a short time.

7.4. Other results

Besides single subject analysis, we performed several subjects clustering. We tried to characterize each class generalizing among individuals by means of leave-one-out cross validation. The results indicate that at least 40 states are needed for the validation set to converge.

8. CONCLUSION

The main conclusions are obtained from the analysis of the experiments for 1 person using a 2-fold CV scheme to obtain the clusters. They are summarized in the following points. The moW and EM-HMM can successfully model the data better than the simpler modified K-means clustering. This model offers a bigger expressivity without causing overfitting. The number of Watson distribution clusters needed to express the EEG data for a single person is in the range 4-6. The data reflects temporal correlation indicated by the EM-HMM transition matrix and likelihood of both methods. This indicates that the EM-HMM algorithm can model the data better.

The clusters obtained for the EM and EM-HMM are very similar, since the data is sparse, the clusters are very distinguishable. The scalp map indicates high visual perception for both classes and "famous face" class has an emotional response.

9. REFERENCES

[1] Michel C. M. Lehmann D. Pascual-Marqui, R. D., “Segmentation of brain electrical activity into microstates: model estimation and validation,” *IEEE Transactions on Biomedical Engineering*, 1995.

[2] Pascual-Leone A. Khanna, A., “Reliability of resting-state microstate features in electroencephalography,” 2014.

[3] D. Lehmann, “Eeg microstates,” 2009.

[4] D.; Merlo M. C. G.; Kochi K.; Hell D.; Koukkou M. Koenig, T.; Lehmann, “A deviant eeg brain microstate in acute, neuroleptic-naive schizophrenics at rest,” *European Archives of Psychiatry and Clinical Neuroscience*, vol. 249, no. 4, pp. 205211, 1999.

[5] Prichep L.S. Lehmann D. Koenig, T., “Millisecond by millisecond, year by year: normative eeg microstates and developmental stages,” *NeuroImage*, 2002.

[6] Michel M. Christoph Lehmann, D., “Eeg-defined functional microstates as basic building blocks of mental processes,” *Clinical Neurophysiology*, vol. 122, no. 6, pp. 10731074, 2010.

[7] N. Michalopoulos, K.; Bourbakis, “Microstate analysis of the eeg using local global graphs,” *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on*, 2013.

[8] M. Hatz, F.; Hardmeier, “Microstate connectivity alterations in patients with early alzheimers disease,” *Alzheimer’s Research Therapy*, 2015.

[9] M. Gschwind, M.; Hardmeier, “Fluctuations of spontaneous eeg topographies predict disease state in relapsing-remitting multiple sclerosis,” *NeuroImage*, 2016.

[10] Henson Richard N Wakeman, Daniel G, “A multi-subject, multi-modal human neuroimaging dataset,” 2015.

[11] Pedroni A. Langer N. Hansen L. K. Poulsen, A. T., “Microstate eeglab toolbox: An introductionary guide,” 2017.

[12] Lelic D. Petrini L. Hennings, K., “An automated method for micro-state segmentation of evoked potentials,” *Journal of Neuroscience Methods*, 2008.

[13] F. Zdyb, “A probabilistic model for segmenting eeg into microstates,” 2016.

[14] Christopher Bishop, *Pattern Recognition and Machine Learning*, 1st edition, 2006.

[15] Karp Dimitrii Sra Suvrit, *The multivariate Watson distribution: Maximum-likelihood estimation and other aspects*, 2011.