Lab Report: Implementation of DTW- and Moment-based Curve Alignment Techniques on EEG Curves

Hüseyin Camalan, Mihail Bogojeski, Prof.Dr.Benjamin Blankertz

Abstract

In Electroencephalography (EEG) research, averaging epochs across categories is a common method to compare and classify signals, but due to the temporal variations in the brain, such averages may be inaccurate. Curve alignment techniques offer the potential to remove these variations, which may bring improvements to classification rates in EEG.

This lab rotation project consisted of the application of existing curve alignment techniques on EEG data. One template-agnostic and two template-dependent methods were implemented. Of these, the template-agnostic method (Moment-based Alignment) was only partially successful at aligning EEG curves. Template-dependent methods were based on developing an effective template for Dynamic Time Warping (DTW). Although these techniques proved to be more effective than without alignment, they failed to provide a significant advantage over other DTW-based methods.

Introduction

EEG (Electroencephalography) is a brain imaging method that allows brain signals to be recorded through electrodes placed on the scalp. It has high temporal precision; however, it is also characterized with low spatial resolution and high noise. Its noninvasive nature allows it to be applied on humans in experimental settings. Coupled with its high temporal resolution, it becomes a very useful method for many fields in Neuroscience, one of which is Brain-Computer Interaction (BCI).

BCI attempts to obtain meaningful signals from the brain so that they can be used for a variety of advanced applications (e.g. spellers that work only on brain signals, enhanced braking methods during driving, etc.). The field is relatively young and is currently more focused on classifying different categories of brain signals.

To classify signals from the brain into the right categories, it is necessary to find certain "EEG components", signals that occur repetitively and reliably after a certain event (e.g. exposure to a certain type of stimulus in an EEG experiment). The ability to recognize such components in an automated manner allows a computer to exploit the occurrence of these signals, thus making the aforementioned advanced applications possible. Often, there is a "target" signal that needs to be recognized apart from a "non-target" signal.

Unfortunately, such a classification task proves to be very difficult due to the high signal-to-noise ratio in the EEG signal. This noise originates from a variety of sources; some examples are the bad conductivity of the tissues of the scalp, signals from eye movements flooding the channels, or the high correlation levels between channels (because all EEG channels receive their signal from the entire brain alongside their local point). Thus, despite elaborate methods, often an EEG signal from a single trial in an experiment is difficult to classify through visual inspection.

A simple solution to this problem of unrecognizable EEG curves is to take the average of all trials. Given that we have two categories, averaging signals across trials in each respective category would bring noise down. However, this approach comes with an

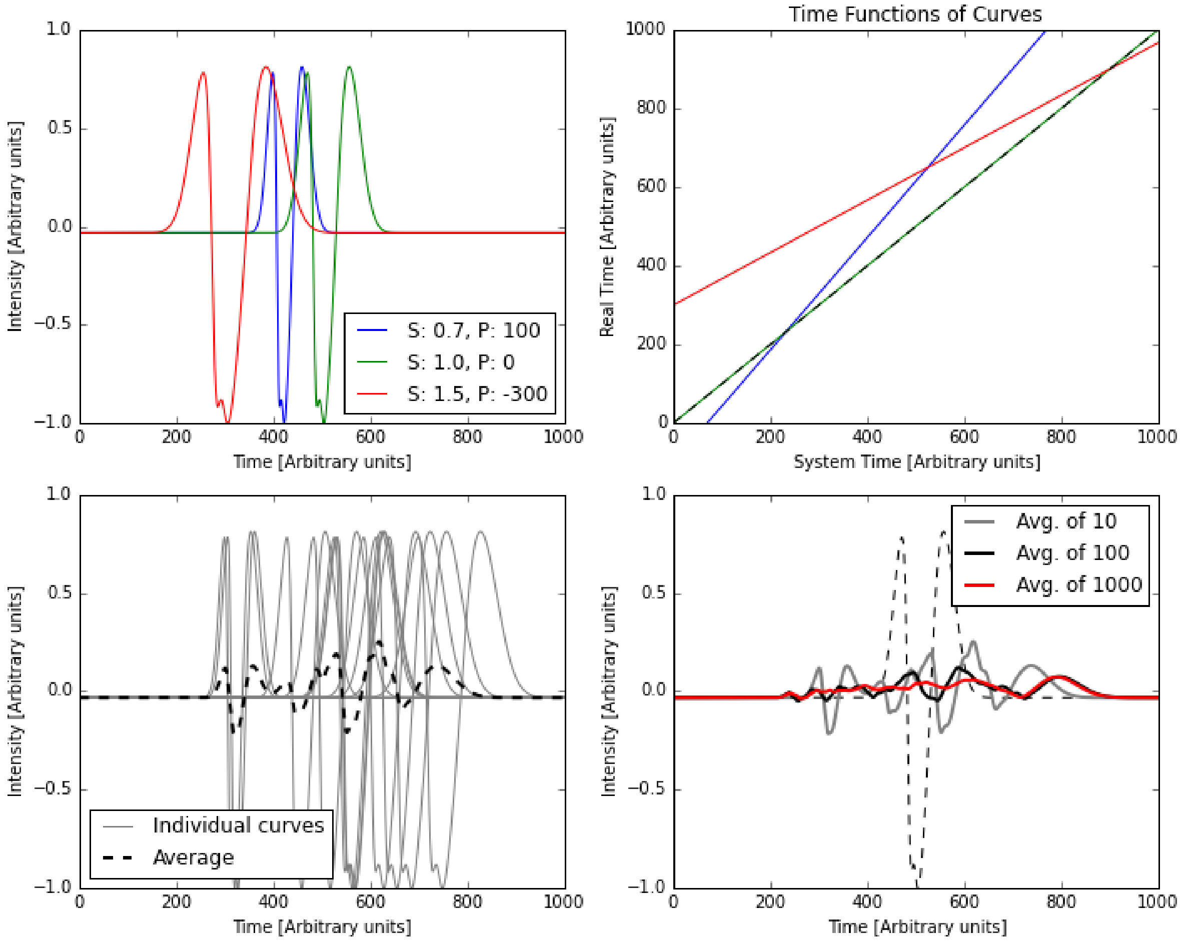

Figure 1: Demonstration of the averaging problem. Top left: 3 curves sourced from the same signal, albeit with linear time distortions. Top right: time functions of the same curves. Bottom left: demonstration of averages of many of the same curves with random phase and scale differences. Bottom right: averages of growing samples of phase-shifted curves.

assumption that the signal that we are interested in always occurs with the same onset after an event (i.e. has the same phase parameter), and that it takes exactly the same length of time (i.e. has the same time scale). Due to the nature of the brain's information processing, these two properties may indeed vary. Although taking averages are still a very effective and simple solution, classification performance is likely to be compromised since the average is flawed and this average is used to classify the EEG signals from each trial. A demonstration of this problem is shown in figure 1.

To improve the averages, it is necessary to have a procedure that removes phase and scale differences from EEG signals. Such a procedure is called "curve alignment", and there are multiple algorithms that can be used for this purpose. One distinction with regards to curve alignment algorithms is how these algorithms tend to synchronize signals. For example, one very common method, called Dynamic Time Warping (DTW), compares two signals to each other and finds an optimal, nonlinear match between their time functions. Given that EEG experiments have hundreds, if not thousands of trials, comparing two signals implies that one of these signals will be a template signal that is used to synchronize all the others. Thus, the main issue becomes the acquisition or formation of a good template. Alternatively, it is possible to calculate certain relevant parameters of curves and alter the time functions of the curves such that these parameters are equated. Hence, no template or specific landmark (certain shapes inside a curve that appear reliably) is necessary to align the signals.

Given this brief introduction, this lab rotation project can be summarized in two parts. First, a template- and landmark- agnostic algorithm called Moment-based Alignment (MBA) was implemented, where the curves are transformed to a desired function highlighting features and the moments of these functions are equated. Second, a classification procedure (Linear Discriminant Analysis, or LDA) was integrated into the template selection procedure for DTW. Normally, classification procedures like LDA are the final step of an analysis and all preprocessing is done to improve the results of LDA. In this case, this order is reversed, in the sense that LDA is used to eventually improve classification results. We implemented and tested two of such techniques, the first of which averages the best representative curves of their classes. The second one, which introduces a modification to an existing method called the Binary Tree Method, creates a prototype curve that carries information from all the curves, while giving higher weights to more successful curves.

Successful procedures were tested on data from a language experiment using EEG (described below), and classification results were obtained. These results were then compared to other results from Bogojeski (2017).

Materials and Methods

Materials

Curve alignment algorithms were implemented and tested on data obtained from an EEG experiment (Wenzel, Moreira, Lungu, Bogojeski, & Blankertz, 2015). In this experiment, participants were shown words or abstract geometrical shapes. Two categories of stimuli were presented, target (i.e. stimulus relates to a selected category, e.g. furniture/table or a shape, e.g. a pentagon) and non-target (i.e. stimulus is not relevant to the selected category, e.g. furniture/rain). The EEG signals upon presentation of the stimulus were separated and organized into trials. Both multivariate and univariate data were tested. Univariate data was obtained by subjecting the multivariate data to dimensionality reduction through the use of sliding LDA.

Prior to testing each method for classification results on EEG, their produced averages/templates were demonstrated first on an artificial EEG curve dataset (Bogojeski, 2017). The templates shown in the figures are all from this dataset.

The code for the MBA method was written in the *python 2.7* programming language, with the help of numerical/scientific computing libraries *numpy* and *scipy*, as well as the plotting library *matplotlib*. The data analysis was completed first in curves created by superimposed gaussians, then followed by the experiment dataset. DTW-based methods were implemented in the *MATLAB* programming language. Finally, the *Berlin Brain-Computer Interfacing (BBCI) Toolbox* was used both for DTW-based methods and classification procedures (Blankertz et al., 2016).

Methods

1. Moment-based Alignment (MBA)

Moment-based Alignment is a curve alignment method that equates some numerical properties of curves instead of looking for a specific landmark in a curve or a template (James, 2008). In a nutshell, "feature functions" are defined from the curves such that one or multiple desired feature(s) of the curves are emphasized. The equations for obtaining the feature functions are as follows:

$I_g^{max}(t) = g(t) - (min\{g(t)\})^r$ for emphasizing global maximum,

$I_g^{min}(t) = (max\{g(t)\} - g(t))^r$ for emphasizing global minimum, and

$I_g^{local}(t) = \begin{cases} exp(-r\frac{|g^I(t)|}{\sqrt{|g^{II}(t)|}}) & g(t) \neq 0 \\ 0 & g(t) = 0 \end{cases}$ for emphasizing local fluctuations.

We also use an updated local feature function to emphasize deviations further away from zero (assumes min-max scaling):

$I_g^{local\_scaled}(t) = (I_g^{local}(t) \circ |g(t)|)^r$

(*g(t)*: a smoothened time-series, r: parameter for adjusting the sharpness of the feature function, $\circ$: the Hadamard product or elementwise multiplication, I and II: the first and second derivatives of a function, respectively)

Because of the scaling, this feature function is henceforth referred to as the "scaled local function", or I^{local_scaled} for convenience.

The feature functions of each of the to-be-synchronized curves are obtained as per the equations above, and their moments are computed. These moments contain information about the a function (the first moment, for example, computes the center of mass of a function on the x-axis). The equation for computing moments is as follows:

$$\mu_g^{(k)} = \begin{cases} \int tI_g(t)dt & k = 1 \\ \int (t - \mu_g^{(1)})^k I_g(t)dt & k \geq 2 \end{cases}$$

(μ: the moment, k: order of the moment)

Finally, the time functions of each curve are altered such that the moments of each curve are equalized. This is done through the equations

$$\alpha_i = \mu_{Y_i}^{(1)} - \beta\widehat{\mu_z^{(1)}}, \beta_i = \sqrt{\frac{\mu_{Y_i}^{(2)}}{\widehat{\mu_z^{(2)}}}}$$

$$X_i(t) = \alpha_i + \beta_i t$$

(Y_i: unsynchronized curve of index i, X_i: the alignment function for Y_i, $\widehat{\mu_z}$: average moment)

Upon the calculation of the factors α and β; $X(t)$, the de-warping function, is obtained as a linear combination of the two. The interpolation of the unsynchronized curves with this new time function returns the synchronized curves.

A more thorough description along with proofs is available in the original paper (James, 2008).

It is argued that the main advantage of MBA is that this method doesn't rely on a template or landmarks. Clear landmarks, although seen in averages in EEG curves, are either much less clear in single curves due to the large amount of noise in the data. As for templates, finding a template that represents or approximates the original signal, is difficult; due to the very same reasons that make curve alignment a necessity.

As it can be seen from the equations above, the MBA method applied in this project was linear. This means that no local distortions to the time functions were made. The linear method was chosen over nonlinear methods because of its simplicity and the availability of an analytical solution. Given the directions in the paper, only the center masses (first moments) and the variations (second moments) within the feature functions were equated.

2. Dynamic Time Warping

Dynamic Time Warping (DTW) is a method that is used for comparing two time series. It compares each point of one time series to the other, and computes a cost that signifies the difference between the two points. The eventual result of the DTW algorithm returns a two-dimensional map, in which there is a route with minimal cost. This route also happens to be a time function that, when applied onto the time series, synchronizes it with the other. Thus, it can be used also as a curve alignment method.

DTW is nonlinear by nature, since the optimal route is often not a straight line. However, some constraints can be introduced to the algorithm, such that the time function does not overfit.

Since DTW synchronizes one time series to another, and our problem requires a large number of curves to be synchronized together, it is necessary to have a reference point, i.e. a template curve to synchronize all other curves to. Hence, the problem becomes one of obtaining the optimal template; which, ideally, should be very similar to the source signal, which is unobservable.

This lab rotation project implemented and tested two template acquirement methods to be used for DTW. In both methods, the curves fed into the algorithms were not reused for testing to avoid overfitting.

a. Average of top-LDA-scoring curves (LDA-SC)

Linear Discriminant Analysis (LDA) is commonly used for classifying EEG curves into classes. In a nutshell, given a high dimensional (2 or more) space, LDA is able to locate an axis that provides the highest between-class variance. The origin point of this line is set as the decision boundary, and each curve has a position on this line that corresponds to a rational number. The sign of this number (positive or negative) determines its predicted class, and its absolute value determines how distinctly this curve belongs to its respective class.

The underlying hypothesis behind this method is the following: if more extreme LDA scores can be theorized to be better representatives of their respective classes, a template derived from the better scoring curves will be more similar to the source signal.

Thus, LDA is applied on the dataset and a score is obtained for each curve. Then, all the curves are sorted by their score and a percentage of the highest scoring curves for each class (i.e. the largest positive and the smallest negative numbers) are picked. The averages of these picked curves for the two categories are selected as the templates.

Finally, as per the DTW procedure, the templates are applied on to the unobserved subset of the dataset for alignment, after which the classification procedure is undertaken.

b. Probabilistic Binary Tree Method (PBTM)

The Binary Tree Method for DTW was first demonstrated by Casarotto, Bianchi, Cerutti, and Chiarenza (2005). The method is self-explanatory by its name - the curves are paired up and combined together using a custom algorithm, such that the resultant curve is equally similar to both of its "parent" curves. This process is repeated with each new generation of curves until there is only one "child" curve. This curve is then picked as a template.

There are two disadvantages with this method: first, the method can only use 2**n number of curves to create a template. Thus, a dataset that has more or less curves would be unable to use all the data. Second, all curves are considered equal. It could be argued that using more distinct curves more often in the binary tree process would lead to a template that is closer to the actual source.

The solution to these problems is the following: given the number of curves inside a class, n_{curves} and the relation $n_{curves} \leq 2^k$; the Binary Tree Method is run with 2^{k+1} curves. Then, an LDA analysis is done on the existing curves, assigning a score to each one. Scores in each category are normalized to have the sum of all scores equal to 1. Then, the empty spots of the Binary Tree are filled randomly with the existing curves, with the higher LDA-scoring curves being more likely to be picked.

Upon the filling of all the empty slots, the Binary Tree Method is run as usual and the resultant final curve is picked as the template for DTW. It is coined with the adjective "probabilistic" due to the selection of parent curves with their associated probabilities.

Results

1. Moment-based Alignment

With regards to Moment-based Alignment, figure 2 demonstrates the feature functions that are obtained and the post-alignment results for the toy dataset, which is also used in figure 1. The sharpness of these functions are altered by the variable r, as discussed in the equations above. It must be noted that adjustment of this parameter is necessary to obtain proper alignment.

While the maximum and minimum feature functions tend to fare well with the alignment task, the local feature function appears to respond to small fluctuations due to the smoothed Gaussian noise in the toy dataset. Given the post-synchronization average, this function appears to be unsuccessful in retaining the original shape of the function. However, this sensitivity to noise is removed in the scaled local feature function, in which case alignment is also successful.

In figure 3, MBA is demonstrated on artificial EEG curves using the feature functions that worked on the initial toy data (minimum, maximum and the scaled local feature function). Given the averaged target curve and the individual curves, it can be seen that the minimum feature function successfully synchronized target curves. However, it can also be observed that average values of both target and non-target curves are decreased for the minimum feature function, and vice versa for the maximum feature function. As for the scaled local feature function, the sharpness of the features are reduced, as well as the area between the target and non-target averages.

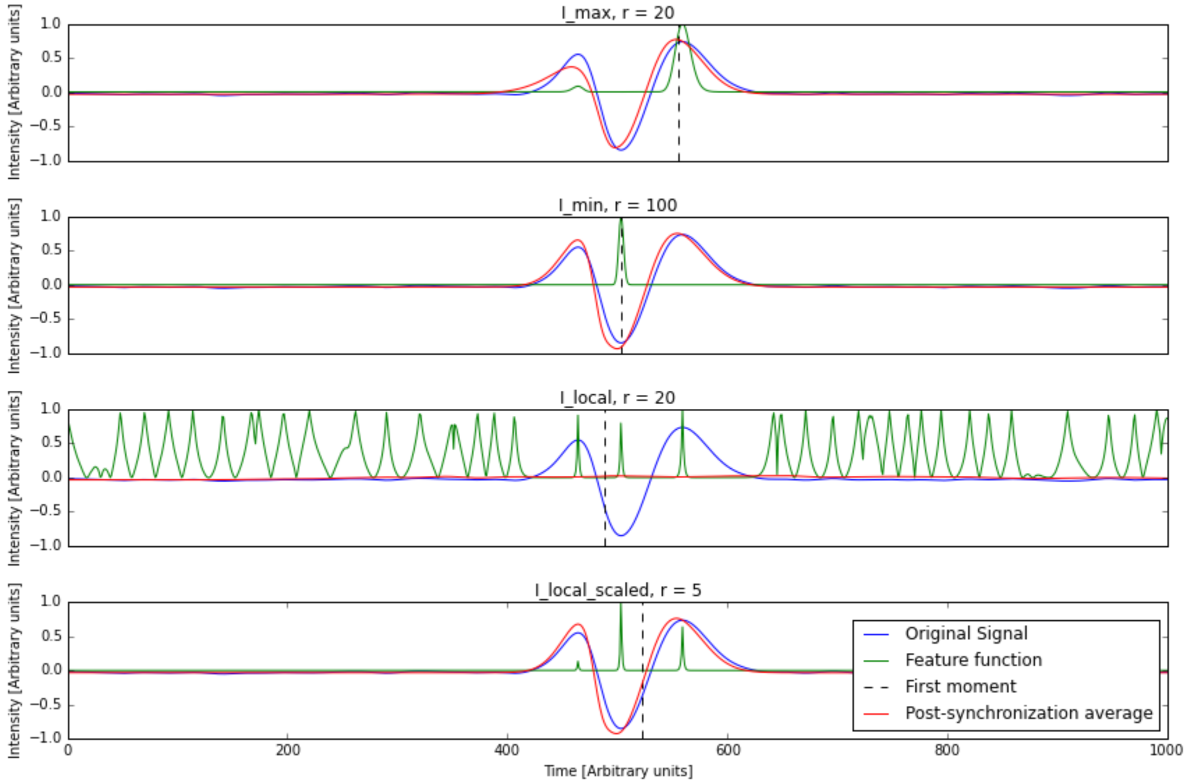

Figure 2: Demonstration of the feature functions and the alignment process. Each row shows a different feature function, the first moment of the feature function, the average of the curves obtained after synchronization, and the original signal for comparison.

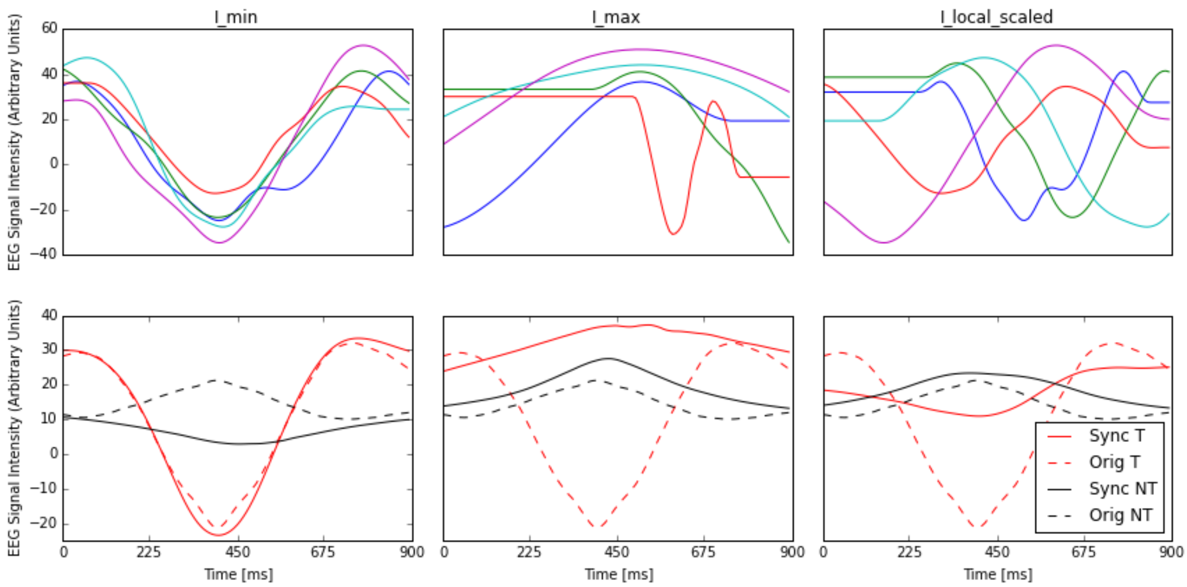

Figure 3: Results of MBA on smoothened EEG curves. Top row: five target curves after alignment. Bottom row: comparison of target and non-target averages before and after MBA. Left, middle, right: Minimum, maximum and scaled local feature functions.

Due to the non-robust results of this method and the time constraints of the lab rotation project, further work on this method was abandoned.

2. Dynamic Time Warping

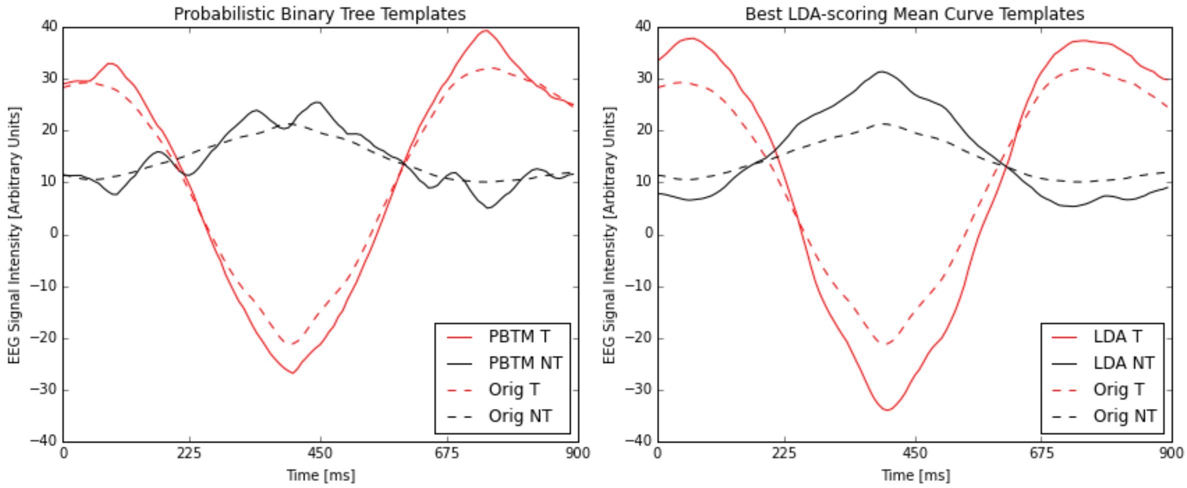

Figure 4: Effect of DTW-based template algorithms on the target and non-target averages. Dashed lines show the unsynchronized averages and solid lines show the templates for alignment. PBTM: probabilistic binary tree method, LDA: LDA-SC with top %10 of curves; T: target, NT: non-target.

Figure 4 shows the resulting templates on the artificial EEG curves upon the application of the DTW-based methods. On the left are the target and non-target templates using PBTM, and on the right are the averages of the LDA-SC. Upon visual inspection, both methods appear to produce more differentiated templates than unsynchronized averaging, given the sharpness of the features and larger area between the target and non-target templates. With regards to the difference between target and non-target averages, LDA-SC averages seem to fare better than the PBTM templates. However, PBTM templates seem to have sharper features compared to those of LDA-SC.

Figure 5 shows classification results between target and non-target curves synchronized by the LDA-SC and PBTM templates, on univariate and multivariate EEG data. It is important to note that the dataset used for classification is different than the EEG curves where the templates are shown. In the multivariate analyses, in order to reduce the computational load, the Euclidean distance of each datapoint from origin is taken. These results are compared to some benchmarks, namely; (i) an unsynchronized dataset, (ii) uni- and multivariate synchronized datasets using the best alignment method in Bogojeski (2017) and (iii) a naturally synchronized dataset whose epochs were timed retroactively based on participant key responses.

Here, multiple observations are apparent. Firstly, multivariate results tend to fare better than univariate results, albeit marginally, with an improvement around %1-2. There is a comparatively large variation of classification rate in LDA-SC with %10 compared to %50. Most importantly, all synchronized analyses appear to outperform the unsynchronized one, with a difference between %2-4. On the other hand, the methods from this lab rotation do not outperform other DTW-based methods.

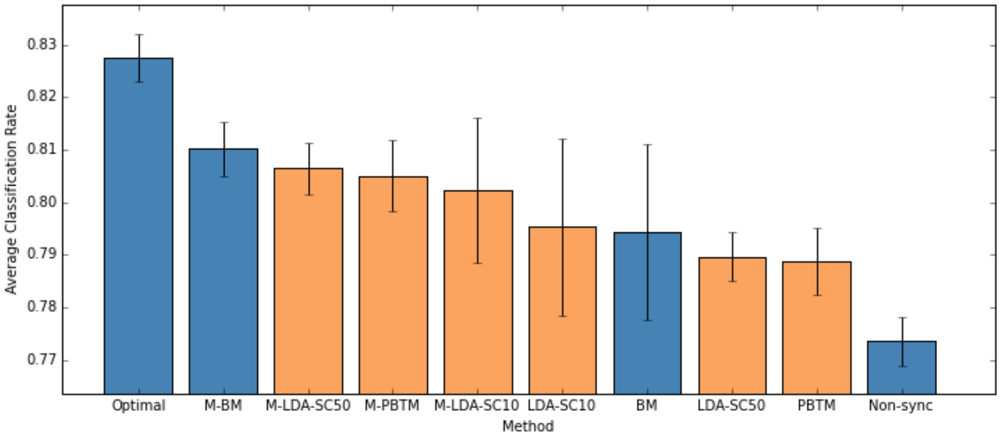

Figure 5: Comparison of classification results of various algorithms. Brown bars indicate results from methods developed during this lab rotation project, and blue bars show other results for comparison. Results for each analysis are obtained through 10-fold crossvalidation. Error bars show $\pm$ 2 standard deviations. The letter "M" in the beginning of the label indicates a multivariate dataset. LDA-SC10: LDA-SC using top scoring %10. LDA-SC50: same as before, but with %50. PBTM: Probabilistic Binary Tree Method. Non-sync: unsynchronized classification results. BM: benchmarks from Bogojeski (2017). Optimal: dataset with natural alignment (univariate dataset).

Discussion

Moment-based Alignment

To align EEG curves in the time axis, we first implemented a curve alignment method based on moments. From the results on the toy dataset as well as the EEG curves, it appeared that the method did not work very robustly. The success of the alignment method was highly dependent on the selected feature function, the time series used for alignment, the sharpness parameter and even the degree to which a curve was smoothed.

The local feature function was especially notorious, as it often resulted in division-by-zero errors due to the second derivative in the denominator in the equation. For this reason, we needed to change the dataset and introduce to it smoothed noise. However, this time, the function was too responsive to small fluctuations in the signal, which caused it to respond highly to noise, missing the important features in the toy signals. Therefore, we updated the function, where we (i) multiplied the feature function elementwise with the absolute value function of the toy signal, and (ii) took the power of the feature function by r. This ensured that small fluctuations from noise did not change the feature function. Although this function fared better on toy data, it did not synchronize EEG curves - on the opposite, the difference between target and non-target averages was reduced.

With the minimum and maximum feature functions, MBA seemed to fare better. Given adjusted sharpness parameters, the toy datasets were successfully synchronized. Additionally, with the choice of the correct feature function, EEG curves can also be synchronized (i.e. see target curves with minimum feature function in fig. 3. The downside, however, is that the opposite is also true. A curve that has a prominent peak (e.g. average non-target curve), when synchronized by a minimum feature function, tends to lose its peak upon alignment. Similarly, a curve with a prominent dip loses its dip with the maximum feature function. This suggests that application of MBA on EEG data, at least given these feature functions, is not necessarily landmark-agnostic.

Dynamic Time Warping

To obtain templates for alignment using DTW, we employed two methods. One of these was the use of a LDA to score curves (LDA-SC), the best of which would be averaged for DTW on the rest of the dataset. The other was an application of the same idea onto the Binary Tree Method (PBTM). By using the higher scoring curves more times inside the Binary Tree, we hypothesized that the end template would have stronger features, which would increase the classification rate.

As for the results, it was perhaps unsurprising that analyses with multivariate datasets performed better than those with univariate datasets. All synchronized datasets, from this project or otherwise, performed better than an unsynchronized dataset. However, they did not perform better than other DTW-based methods discussed in Bogojeski (2017), the best of which is shown in fig. 5. Also, all results were worse compared to a "naturally" synchronized dataset, where epoch times were based on participant key responses (in which case there is much less temporal variability).

With both PBTM and LDA-SC, the obtained alignment templates had more prominent features compared to class averages. Despite being a more computationally intense method, PBTM templates had a smaller difference between target and non-target templates. This might be because of the probabilistic nature of PBTM. Even though higher scoring curves have a higher probability of being selected for the template in PBTM, over a large number of repetitions, lower scoring curves also get picked for the extra slots of the binary tree. On the other hand, in LDA-SC, the lower scoring curves do not get introduced to the template at all. Since LDA finds the axis with the highest between-class variability, the highest scoring curves are those who are most distinctive of the other class. Thus, it is perhaps not surprising that the less puristic nature of PBTM causes smaller differences between target and non-target templates.

However, despite having larger between-category differences than PBTM, classification results with the datasets synchronized through the LDA-SC template did not perform better. Thus, it could be argued that larger template differences do not necessarily translate to higher classification results.

One last observation to be made concerning the classification results is the high variance of LDA-SC with top %10 of curves. This is likely a result of using a smaller sample for averaging. Given that there are no significant performance differences between the two LDA-SC analyses (top %10 and %50), it is probably arguable that there is a trade-off between using ever higher-scoring curves and using a larger sample of curves.

Future Directions

Although the results of Moment-based Alignment were not as convincing as hypothesized earlier, it is possible to obtain satisfactory classification results from it. The idea is to utilize MBA with multiple feature functions to build a linear classifier, which, in turn, predicts the classes of unobserved curves. Given this particular dataset, use of the minimum and the maximum feature function together would likely be satisfactory.

However, this conclusion can not be extrapolated to every case, as the feature functions require their specific feature to exist inside the curves.

With regards to the feature functions, it is possible to come up with more complex feature functions that satisfy the specific needs of the dataset. One possible idea is to develop a feature function where EEG curves are convolved with viable templates. However, this idea would go contradictory to the initial motivations of the developers of this method (James, 2008), who developed MBA to avoid relying on templates.

A final possible improvement on the MBA could be the implementation of a nonlinear solution. Our choice in the implementation of a linear solution was due to the existence of an analytical solution and the relative ease with which the method could be implemented. However, it is possible that the time functions of the EEG curves are warped in a nonlinear fashion. Depending on the extent of nonlinearity, a linear solution may fall short of its goals.

References

Blankertz, B., Acqualagna, L., Dähne, S., Haufe, S., Schultze-Kraft, M., Sturm, I., ... Müller, K.-R. (2016). The Berlin Brain-Computer Interface: Progress Beyond Communication and Control. *Frontiers in Neuroscience*, *10*.

Bogojeski, M. (2017). *Techniques to deal with variable response times in EEG.*

Casarotto, S., Bianchi, A. M., Cerutti, S., & Chiarenza, G. A. (2005). Dynamic Time Warping in the Analysis of Event-Related Potentials. *IEEE Engineering in Medicine and Biology Magazine*, *24*(1), 68–77.

James, G. M. (2008). Moments Based Functional Synchronization. *Los Angeles, CA: University of Southern California*.

Wenzel, M. A., Moreira, C., Lungu, I.-A., Bogojeski, M., & Blankertz, B. (2015). Neural Responses to Abstract and Linguistic Stimuli with Variable Recognition Latency. In *Symbiotic Interaction* (pp. 172–178). Springer.