Proceedings of the
7th Graz Brain-Computer Interface Conference 2017 DOI: 10.3217/978-3-85125-533-1-37

MIXING TWO UNSUPERVISED ESTIMATORS FOR EVENT-RELATED POTENTIAL DECODING: AN ONLINE EVALUATION

D. Hübner[1], T. Verhoeven[2], P.-J. Kindermans[3], M. Tangermann[1]

[1] Brain State Decoding Lab, Cluster of Excellence BrainLinks-BrainTools, Dept. of Computer Science, Albert-Ludwigs-University, Freiburg, Germany
[2] Electronics and Information Systems, Ghent University, Ghent, Belgium
[3] Machine Learning Group, Berlin Institute of Technology, Berlin, Germany

E-mail: david.huebner@blbt.uni-freiburg.de, michael.tangermann@blbt.uni-freiburg.de

ABSTRACT

An ideal decoder in brain-computer interfaces (BCIs) would not require any calibration period and instead start with the actual online application right away. While we cannot reach this goal yet, two novel unsupervised classification methods for BCIs based on event-related potentials (ERPs) of the electroencephalogram (EEG) have recently been proposed which do not require a calibration session. The first method estimates the projection weights of the classifier heuristically using an expectation-maximization approach, while the second utilizes slight changes of the ERP paradigm and deterministically learns from label proportions. As both unsupervised methods have pros and cons, we propose to combine their strengths in a novel MIX approach. Under realistic unlabelled conditions, we compare the online performances of the mixed and the two original methods, finding that for our data recorded during visual spelling with 6 subjects, the mixed approach reveals strong performance gains. Users got perfect selection accuracy after an average of only 2 minutes of online usage.

INTRODUCTION

In Brain-Computer Interfaces (BCI) based on event-related potentials (ERP), the user is presented with a predefined set of different control commands. For example in the original P300-speller [1], a BCI for spelling text, these options are symbols of the alphabet highlighted on a screen. The user is asked to focus on the symbol that he or she wants to spell. When this target symbol is highlighted, the brain of the user elicits a different brain response compared to the case when other non-target symbols are highlighted. The decoder in the BCI has the task to classify the recorded ERP responses as target or non-target and subsequently detect the desired symbol. In this way, the user can spell words symbol by symbol, solely by attending to the symbols on the screen.

To discriminate between target and non-target responses in the brain, machine learning (ML) techniques are often used [2, 3]. With ML, previously recorded ERPs are used by the classifier to learn how to discriminate between the two classes of responses. Newly recorded ERPs are then processed by this classifier to assign them to one of these classes. A common ML technique used in BCIs is linear discriminant analysis (LDA), which searches for a one-dimensional projection $\boldsymbol{x} \cdot \boldsymbol{w}$ of the ERP response signal features $\boldsymbol{x}$ in order to assign the target label t_+ to the response when $\boldsymbol{x} \cdot \boldsymbol{w} \geq 0$ and label t_- otherwise.

It was shown that ERP-BCI data follows a Gaussian distributed with class-wise means $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$, and shared covariance matrix $\boldsymbol{\Sigma}_s$ [4]. Under this assumption, the optimal projection $\boldsymbol{w}^*$ in LDA can be computed as following [4]:

$$\boldsymbol{w}^* \quad = \boldsymbol{\Sigma}_s^{-1} \left(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_- \right). \tag{1}$$

Training the classifier comes down to estimating the values of the class-wise mean responses $\boldsymbol{\mu}_+$, $\boldsymbol{\mu}_-$ and the shared covariance $\boldsymbol{\Sigma}_s$. In a traditional supervised scenario, labelled data would be collected during a calibration session on which these three quantities can directly be estimated by using the sample statistics. In unsupervised learning, no label information are present which makes it a more challenging learning problem.

It can be shown that if the means are estimated correctly, then replacing the shared covariance by the pooled covariance $\boldsymbol{\Sigma}$, i.e. the covariance computed on all data disregarding label information, leads to the same direction of the projection $\boldsymbol{w}$. This follows from the equivalence of least square regression with rescaled outputs and LDA [5]. No label information are needed to estimate the pooled covariance matrix.

In BCI systems, the data is usually high dimensional and the amount of data recorded during calibration is low. It was shown that this makes the estimation of the covariance matrix less accurate [4] This can be compensated for by introducing a regularization term to obtain the (shrinkage)-regularized covariance matrix $\boldsymbol{\Sigma}_R$

$$\boldsymbol{\Sigma}_R = (1 - \lambda)\boldsymbol{\Sigma} + \lambda \boldsymbol{I} \tag{2}$$

where $\boldsymbol{I}$ is the identity matrix and λ is the regularization parameter.

 DOI: 10.3217/978-3-85125-533-1-37

The learning problem in the unsupervised case now boils down to estimating the class means $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ and the shrinkage parameter λ. Everything else can be computed without using label information.

To compute the class means, we recently proposed to combine two unsupervised methods [6]. The first method is an expectation-maximization (EM) algorithm which estimates the class means to maximize the likelihood of the recorded data [7]. It is a heuristic which relies on a good random initialisation to obtain accurate class estimates.

The second method is based on the learning from label proportions (LLP) concept [8]. In this approach, the train of stimuli is divided in two interleaved sequences with different proportions of targets and non-targets. The average response in these two sequences is calculated and used together with the known proportional composition, to set up two linear equations. The two unknowns are the class means. Solving the linear problem provides an estimate of these class means. We presented the application of the LLP method in BCI recently [9]. In contrast to the EM method, there is no variance in the result as there are no randomly initialized parameters in this method. Furthermore, the estimation of the mean ERP response is guaranteed to converge to the true solution as more data is recorded [9]. This convergence slowly leads to an increasing classification performance.

Two different options have been previously used to compute the regularization parameter λ. An analytical formula for λ has been presented by Ledoit and Wolf [10], see Blankertz et al. for an application in BCI [4]. Another approach directly optimized λ as part of the EM-algorithm [11].

The EM-means and LLP method for unsupervised ERP classification clearly show complementary strengths and weaknesses [6]. We proposed to exploit the different strengths by combining their individual mean estimations in a data-driven fashion. This resulted in a third method which we call MIX. We previously evaluated this approach by comparing LLP, EM and MIX by simulating an online experiment on existing visual ERP data [6]. In this previous study all three classifier used the analytic formula by Ledoit and Wolf for computing λ. To have comparable results to the original EM-method, we used the direct regularization of the EM-method in this current study, while both other methods use the analytic formula to find regularization parameter λ. The following table shows an overview of the three different methods used in this paper.

Table 1: Overview of classification methods

Method	Mean estimation	λ estimation
LLP	Using known proportions	Ledoit&Wolf
EM	Maximizing data likelihood	Direct (EM)
MIX	Combining LLP and EM	Ledoit&Wolf

Previous simulations showed that the MIX method significantly outperformed both other methods [6]. However, as simulation on previously recorded data are possibly prone to overfitting, this work presents the first online evaluation of the MIX method with 6 subjects and its comparison with the LLP and EM algorithm. The goal of this work is to compare the three methods under equal conditions on unlabelled and unseen data. With this comparison, we hope to contribute further to the integration of unsupervised classification methods in calibrationless BCIs and as such to improve the usability of these systems.

MATERIALS AND METHODS

The MIX model

In the MIX method, the estimation of the class-wise means is proposed as a mixing of the estimation found with the LLP and EM method:

$$\hat{\boldsymbol{\mu}}(\gamma) = (1-\gamma)\hat{\boldsymbol{\mu}}_{EM} + \gamma\hat{\boldsymbol{\mu}}_{LLP} \tag{3}$$

where $\hat{\boldsymbol{\mu}}$ denotes the new estimator of the mean target or non-target response, $\hat{\boldsymbol{\mu}}_{EM}$ and $\hat{\boldsymbol{\mu}}_{LLP}$ denote existing estimators and $\gamma \in [0, 1]$ is the mixing coefficient, indicating the weight given to each estimator. See our previous work about LLP [9] and the EM-algorithm [7] for more details about these two unsupervised classification methods.

To minimize the expected mean squared error between the estimator value $\hat{\boldsymbol{\mu}}$ and the unknown true parameter value $\boldsymbol{\mu}$, we proposed an analytical solution for the mixing coefficient γ^* [6]:

$$\gamma^* = \frac{1}{2}\left(\frac{\sum_d Var\left[\hat{\boldsymbol{\mu}}_{EM,d}\right] - \sum_d Var\left[\hat{\boldsymbol{\mu}}_{LLP,d}\right]}{\|\hat{\boldsymbol{\mu}}_{EM} - \hat{\boldsymbol{\mu}}_{LLP}\|^2} + 1\right) \tag{4}$$

Here, $Var\left[\hat{\boldsymbol{\mu}}_{(\cdot),d}\right]$ denotes the variance on the estimation of the d$^{\text{th}}$ entry of the estimated mean $\hat{\boldsymbol{\mu}}$. This variance is a measure for the uncertainty on the estimated value. The higher the uncertainty on the output of the EM method, the higher the weight given to the output of the LLP method in the MIX method and vice versa.

Implementation

In practice, the original EM algorithm and the one used in the mean estimation of the MIX method are different in terms of the number of parallel initialisations. It is known that the EM-algorithm requires a good random initialisation and therefore, it profits from many parallel initialisations. A number of 5 was used before in the EM algorithm [11]. In contrast, the MIX method in our previous paper only used 1 initialisation of the EM algorithm [6]. We keep these values to make the results comparable.

 DOI: 10.3217/978-3-85125-533-1-37

Experiment

Six healthy subjects (3 female, aged 22-31) performed a visual copy-spelling task. The EEG study was approved by the Ethics Committee of the University Medical Center Freiburg and the subjects gave informed written consent prior to the beginning of the session. They were compensated with 8 Euros per hour. The experiment was almost identical with the one described in [9]. A short overview is provided here. Each subject was asked to spell the 35 characters: "FRANZY JAGT IM TAXI DURCH DAS " three times. Each time, a different classifier (MIX, LLP, or EM) was trained from scratch such that each subject used each classifier exactly once. With 6 subjects, each possible order of the three different classifiers was used once to reduce order effects, see Fig. 1 for a schematic overview.

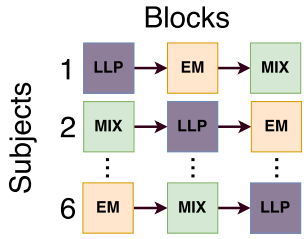

Figure 1: **Experimental structure.** Each subject performed three copy-spelling blocks with each of the three classifiers in varying order. Each block consisted of 35 characters.

The classifiers were retrained after each character utilizing the complete data set up to that point. Label information were not used for the training of the classifiers at any point in time, they were solely used to assess the performance. To spell one character, a train of 68 highlighting events with a stimulus onset asynchrony (SOA) of 250 ms was presented. An example of a highlighting event is shown in Fig. 2. Classifier outputs for each highlighting event and symbol were summed up and the symbol with the highest sum was selected and shown to the user.

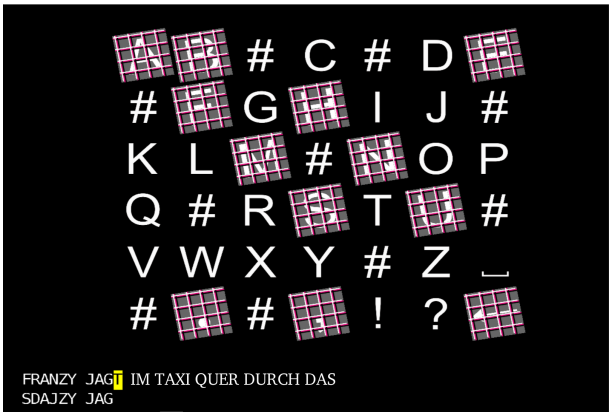

Figure 2: **Spelling Interface.** The '#' symbols serve as visual blanks, meaning that they are always non-targets, and are part of the LLP decoding strategy.

EEG and Feature Extraction

EEG signals from 31 passive Ag/AgCl electrodes (EasyCap) were recorded, which were placed approximately equidistantly according to the extended 10–20 system, and whose impedances were kept below 20 kΩ. All channels were referenced to the nose. The signals were registered by multichannel EEG amplifiers (BrainAmp DC, Brain Products) at a sampling rate of 1 kHz. The data was then bandpass filtered between 0.5 and 8 Hz and downsampled to 100 Hz. Epochs were windowed to [-200, 700] ms relative to the stimulus onset and corrected for baseline shifts observed in the interval [-200, 0] ms. Per channel, the mean amplitudes of six intervals ([50, 120], [121, 200], [201, 280], [281, 380], [381, 530] and [531, 700] ms) were finally computed as features. This resulted in a total of $6 \cdot 31 = 186$ features.

Performance estimations

Two performance metrics were used to evaluate the performance of the three different classifier during the online experiment. First, we looked at how well single targets could be discriminated from non-targets. This was assessed in terms of area under the curve (AUC) as a threshold-independent robust performance measure. The AUC values can range between 0 and 1, with a theoretical chance level of 0.5. An AUC value of 1 indicates perfect separation between the two classes, i.e. the classifier can correctly tell for each single stimulus whether it was attended or not. To compute this score during the online experiment, the unsupervised classifiers were retrained after each character and applied to the complete previous data up to the current point of the experiment. The given label information were then used to compute the AUC. Please note that overfitting is not a problem in this context, because the classifiers do not use the label information for training.

Second, we looked at the selection accuracy, i.e. to which percentage a user could spell the intended characters. To obtain more robust estimates, this metric was evaluated on sub-blocks of 5 trials, i.e. on characters 1-5, on characters 6-10, and so on.

In addition, we performed an offline analysis after the experiment to assess the overall quality of the data and to judge whether there is an interaction between classification method and classification performance. This was done by training and testing a supervised shrinkage-LDA classifier [4] in a 5-fold chronological cross-validation. This is the same classification method as described before – only that all quantities are estimated with the (supervised) sample statistics. The offline analysis was done individually for each subject and block.

RESULTS

First, we assessed the quality of the data of the online re-

 DOI: 10.3217/978-3-85125-533-1-37

cordings by looking at the grand average ERP response shown in Fig. 3. A strong early negativity with a peak around 160 ms in the occipital area is observable in target responses while the non-targets only have a very weak rhythmic response. Furthermore, a central positivity exists for targets, which is however smaller in amplitude and more washed out than the early negativity. The strongest class discriminant information comes from the early visual component. This is in accordance with earlier studies using the same highlighting scheme [9, 12].

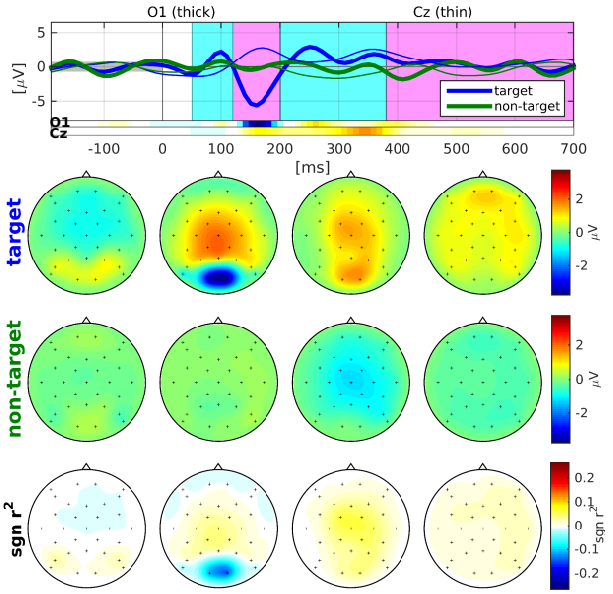

Figure 3: **Grand average (N=6) ERP plot. Top row**: Average responses evoked by visual target (blue) and non-target (green) stimuli in the occipital channel O1 (thick) and the central channel Cz (thin). The signed r^2 values for channels O1 and Cz over time are provided by two horizontal colour bars with the same scale as in bottom row scalp plot. **Middle rows**: Scalp plots visualising the spatial distribution of mean target and non-target responses within four selected time intervals marked by blue/pink shading. **Bottom row**: Scalp plots with signed r^2 values indicate spatial areas with high class-discriminative information.

Classifier influence on the data quality

To quantify the quality of the ERP responses and judge whether there is an interaction between classification method and performance, a supervised classifier was applied in an offline analysis after the experiment. The resulting target vs. non-target AUC performances were sorted according to the classifier used in each block and are shown in Fig. 4. One can see that the quality of the data is very high with all subjects having an AUC above 95 %. This is most likely due to the high saliency of the optimized stimuli [12]. In addition, a paired t-test between the supervised classification values of the three methods showed no significant differences. Hence, we cannot reject the null hypothesis that there is no interaction between classifier and performance. This means that we observed no effect of the feedback on the user performance probably due to the small sample size. Other studies did observe this effect [13, 14].

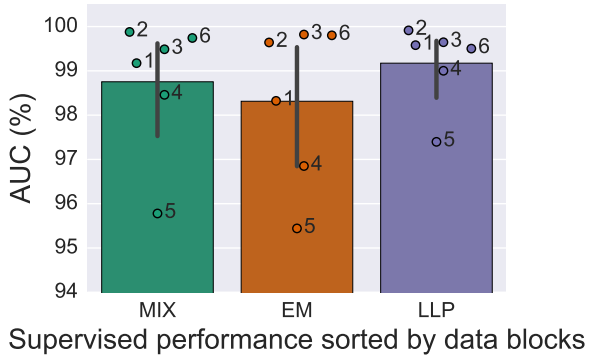

Figure 4: **Supervised offline cross-validation performance sorted by the classification method used per data block.** Bars show the mean ± std of the supervised performances for each sentence sorted according to the decoding method. Individual dots and numbers indicate the subject numbers. MIX = Mixing method, EM = Expectation-maximization, LLP = Learning from label proportions.

Next, we compared the performance of the three different classifiers in the online experiment. Fig. 5 shows the average and individual AUC performances for all subjects over time. While LLP starts at relatively high level and slowly improves over time, the EM algorithm behaves dichotomous: depending on its initialisation, it can either achieve a very high performance early on or it can fail to improve over a prolonged time period. The MIX method combines the strengths of both decoders by starting on a relatively high level and quickly finding a very good projection with almost perfect decoding performance by utilizing the complementary information of the LLP.

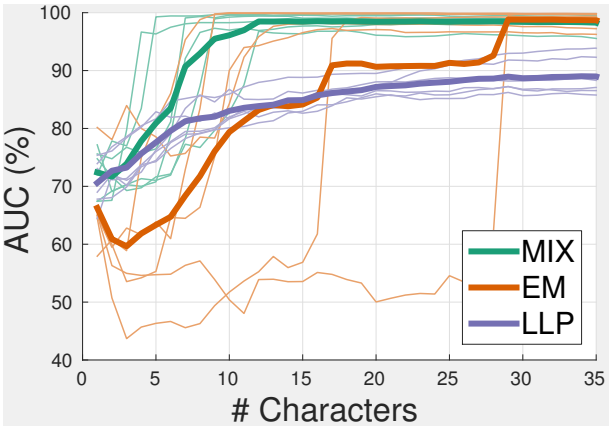

Figure 5: **Online decoding performance over time.** The y-axis shows the AUC of separating target from non-target epochs for each decoder. Thick line depicts the average performance while thin lines show results for each individual subject.

 DOI: 10.3217/978-3-85125-533-1-37

Another way of looking at the decoding performance is by considering the number of correct character selections. One can see in Fig. 6 that the MIX method slightly outperforms the LLP and that both these methods outperform the EM-method by a big margin. This is due to the two sentences in which the EM-algorithm found the right projection only relatively late, see again Fig. 5.

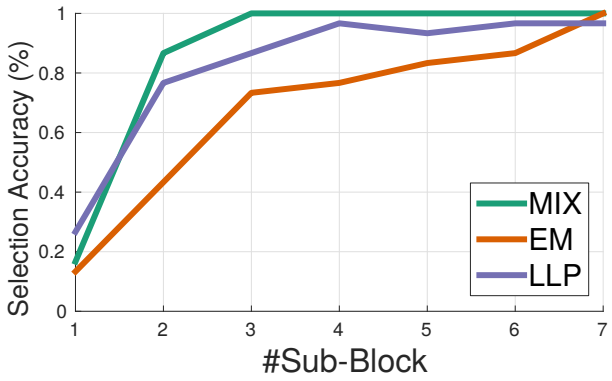

Figure 6: **Online character selection accuracy for each decoder.** The y-axis shows the percentage of correctly classified characters for sub-blocks of 5 characters each.

When looking at the correctly and incorrectly spelled characters of all three methods for each subject in Fig. 7, similar results are visible. After a short learning phase of 2-8 characters corresponding to an average of around 2 minutes of training time, users gain perfect control with the MIX method. Depending on the initialisation, the user can get very good control with the EM-algorithm at an earlier or later stage of the experiment. The LLP determines many characters correctly already after a few trials, but fails to display a very high reliability in the later stage.

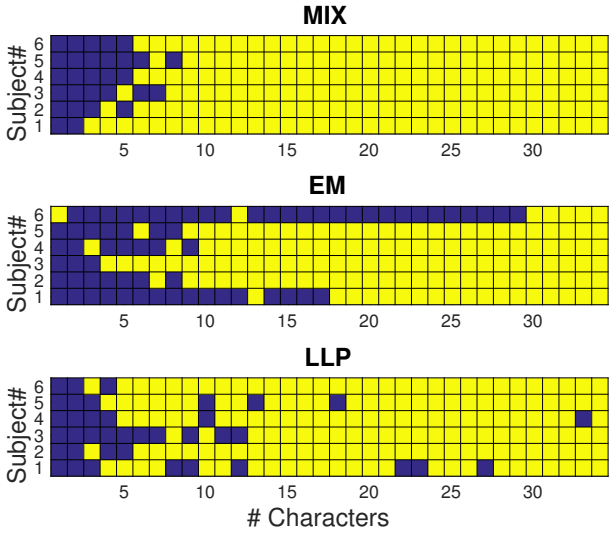

Figure 7: **Correctly (yellow) and incorrectly (blue) spelled characters of all three methods in the online experiment.**

DISCUSSION

The goal of this study was to show that the unsupervised MIX method can work in an online scenario and compare it to the EM and LLP method. We found that the MIX method could quickly and reliably decode the users' intention for all 6 subjects clearly outperforming both other methods. Remarkably, we observed almost perfect single epoch classification accuracy, meaning that the classifier could assign almost each highlighting event correctly as being attended or not. Here, the unsupervised classifier also profited from the very salient highlighting scheme.

On the other hand, spelling speed was not the focus of this work. Indeed, it was rather low with around 2.4 characters per minute after the initial training phase. This is due the high and constant number of 68 epochs per trial and long SOA of 250 ms. A moderate single epoch classification accuracy is already sufficient to correctly decode most characters with 68 highlighting events per characters. Hence, the additional performance in the MIX method is only slightly rewarded in terms of spelling speed or accuracy in this set-up. However, it could easily be boosted by implementing dynamic stopping [15], where the classifier stops a trial when he reaches a predefined certainty threshold.

Results from the online study showed that an average of around 2 minutes of online training time is sufficient to obtain perfect control over the BCI with the MIX method. Remarkably, this result was achieved without prior calibration or transfer learning. And even the data from the initial training phase can be corrected, when a more advanced classifier from a later stage of the experiment is re-applied to the initial data. In this way, initial mistakes due to limited data can be post-hoc corrected in unsupervised classifiers [11]. Hence, a potential user could directly start spelling with this MIX method when trusting the re-analysis.

CONCLUSION

The online ERP study showed that the MIX method is combining the strength of the probabilistic EM algorithm and the deterministic LLP approach. This opens the door for short ramp-up times combined with a very high reliability. Further desirable properties like the lack of calibration phase, the continuous learning, the guaranteed convergence and the possible post-hoc analysis, make this method an attractive alternative to traditional supervised methods. Future work will go towards increasing the usability of the system by increasing the information transfer per time. This can be achieved by implementing an SOA reduction, dynamic stopping, transfer learning, adaptive channel selection and using language models.

 DOI: 10.3217/978-3-85125-533-1-37

ACKNOWLEDGEMENT

DH and MT gratefully acknowledge the support by BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG), grant number EXC 1086. PJK has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement NO 657679. TV is supported by the special research fund (BOF) from Ghent University.

REFERENCES

[1] Farwell L. A and Donchin E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

[2] Lotte F, Congedo M, Lécuyer A, Lamarche F, and Arnaldi B. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1, 2007.

[3] Müller K-R, Tangermann M, Dornhege G, Krauledat M, Curio G, and Blankertz B. Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, 167(1):82–90, 2008.

[4] Blankertz B, Lemm S, Treder M, Haufe S, and Müller K-R. Single-trial analysis and classification of ERP components, a tutorial. *NeuroImage*, 56(2):814 – 825, 2011.

[5] Bishop C. M. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

[6] Verhoeven T, Hübner D, Tangermann M, Müller K-R, Dambre J, and Kindermans P.-J. Improving zero-training brain-computer interfaces by mixing model estimators. *Journal of Neural Engineering*, 14(3):036021, 2017.

[7] Kindermans P-J, Verstraeten D, and Schrauwen B. A bayesian model for exploiting application constraints to enable unsupervised training of a P300-based BCI. *PLOS ONE*, 7(4):e33758, 2012.

[8] Quadrianto N, Smola A. J, Caetano T. S., and Le Q. V. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(Oct):2349–2374, 2009.

[9] Hübner D, Verhoeven T, Schmid K, Müller K-R, Tangermann M, and Kindermans P.-J. Learning from label proportions in brain-computer interfaces: online unsupervised learning with guarantees. *PLOS ONE*, 12(4):e0175856, 2017.

[10] Ledoit O and Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[11] Kindermans P.-J, Schreuder M, Schrauwen B, Müller K-R, and Tangermann M. True zero-training brain-computer interfacing–an online study. *PLOS ONE*, 9(7):e102504, 2014.

[12] Tangermann M, Schreuder M, Dähne S, Höhne J, Regler S, Ramsay A, Quek M, Williamson J, and Murray-Smith R. Optimized stimulation events for a visual ERP BCI. *Int. J. Bioelectromagn*, 13(3):119–120, 2011.

[13] Lotte F, Larrue F, and Mühl C. Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. *Frontiers in Human Neuroscience*, 2013.

[14] Barbero Á and Grosse-Wentrup M. Biased feedback in brain-computer interfaces. *Journal of NeuroEngineering and Rehabilitation*, 7(1):34, 2010.

[15] Schreuder M, Höhne J, Blankertz B, Haufe S, Dickhaus T, and Tangermann M. Optimizing event-related potential based brain–computer interfaces: a systematic evaluation of dynamic stopping methods. *Journal of Neural Engineering*, 10(3):036025, 2013.