Proceedings of the
7th Graz Brain-Computer Interface Conference 2017 DOI: 10.3217/978-3-85125-533-1-35

IMPROVING LEARNING FROM LABEL PROPORTIONS BY REDUCING THE FEATURE DIMENSIONALITY

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

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

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

ABSTRACT: Learning from label proportions (LLP) is a recently introduced unsupervised classification method for event-related potential (ERP) based brain-computer interfaces. It estimates the target and non-target means of brain signal epochs based on a known proportion of these classes in different subsets of the data, which can be generated by interleaving different stimulus sequences, e.g., in a visual ERP speller. In contrast to other unsupervised methods, estimations obtained by LLP have the theoretical property of converging to the correct class means. However, the convergence is rather slow. In this paper, we investigate the effect of varying EEG channel numbers as a simple form of regularization onto the performance of LLP classifiers by offline analyses. We found that reduced channel sets can outperform the full set both in terms of single event classification rate and symbol selection accuracy. This is especially pronounced in the initial learning phase. These findings suggest that LLP classification can be significantly improved by reducing the feature dimensionality.

INTRODUCTION

One of the fundamental tasks in brain-computer interfaces (BCI) is to tune the decoder to reliably detect a user's intention. This is a challenging problem because signals do not only differ between subjects, but also between sessions for the same subjects. While some information can be transferred from other subjects or sessions [1–3], a portion of task-specific information remains unknown prior to the start of each session. Hence, the traditional approach is to conduct a calibration session before going into the online application. While this generally works well, some problems are associated with it, namely that it requires additional time, that wrongly labelled data may be recorded when the user incorrectly follows the instructions and that the effect of feedback is not present during the calibration session, possibly leading to different data distributions between calibration and online use of the system [4].

To overcome these problems and avoid calibration sessions, unsupervised methods have been introduced to the BCI communities which can learn from scratch without requiring label informations. They have the additional benefit of continuously learning during a session, thus adapting to possible non-stationarities in the data. One example is the expectation-maximization (EM) algorithm by Kindermans et al. [5] which is applicable for BCI paradigms that use event-related potentials (ERPs) of the electroencephalogram (EEG). It optimizes a likelihood function given a probabilistic model of the data. While it generally works well in practice, it relies on a good random initialisation and has no guarantee to converge to the right solution. As an alternative unsupervised learning method, we recently introduced *learning from label proportions* (LLP) to the BCI community [6]. This method estimates the average responses to target and non-target stimuli based on ERP data. It exploits known proportions of target and non-target stimuli contained in different subsets of the data, but does not require the labels for each stimulus. In contrast to the EM approach, LLP is guaranteed to converge to the correct class means (and thus, to the corresponding decoder) given sufficient amount of data. It also has the advantages of being easy to implement, has extremely short runtime and is deterministic. However, the convergence is often slower as for the EM-algorithm. Further details on LLP are provided in the method section.

An important argument against unsupervised learning is, that these methods go through an initial learning phase in which the feedback is relatively unreliable. The major reason of this effect is the limited amount of (unlabelled) training data in this initial phase, in combination with the high dimensionality of the feature space leading to a difficult unsupervised learning problem. In this paper, we re-analyse data from a previous experiment to answer the question whether a reduced number of channels, corresponding to a reduced number of features, can improve the performance of LLP and shorten this initial ramp-up phase.

MATERIALS AND METHODS

Learning from Label Proportions

For classification, we used the recently introduced learn-

 DOI: 10.3217/978-3-85125-533-1-35

ing from label proportions (LLP) for ERP data [6] which is based on the work of Quadrianto and colleagues [7]. The main idea is to tune the stimulus presentation of a visual ERP-BCI paradigm such that the prerequisites of LLP are satisfied – in this way, the experimental paradigm and decoder are very closely linked and should be seen as a whole and not as independent steps. In order to enable LLP to estimate the target and non-target class means, it is necessary that the recorded visual ERP responses consist of at least two subsets, and that one of the sets has a higher target proportion than the other. Additionally, these proportions have to be known.

The way we created two different subsets is by randomly interleaving two sequences in each single trial (i.e. spelling one character) of a visual speller. In each highlighting event of the first sequence, 12 characters are highlighted, while an event of the second sequence only highlights 3 or 4 characters resulting in average of $3.\overline{5}$ highlighted characters per event. Because there is a total of 32 selectable characters, this leads to an average target ratio of $12/32 = 3/8$ and $3.\overline{5}/32 = 2/18$ for the first and second sequence respectively.

By manipulating the stimulus presentation in that fashion, we can write the average responses of the two sequences $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as a combination of target $\boldsymbol{\mu}_+$ and non-target responses $\boldsymbol{\mu}_-$, therefore utilizing our knowledge about the average proportions.

$$\begin{cases} \boldsymbol{\mu}_1 = \frac{3}{8}\boldsymbol{\mu}_+ + \frac{5}{8}\boldsymbol{\mu}_- \\ \boldsymbol{\mu}_2 = \frac{2}{18}\boldsymbol{\mu}_+ + \frac{16}{18}\boldsymbol{\mu}_- \end{cases} \tag{1}$$

As we are interested in the mean target and non-target ERP responses, we solve the equations for $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ which yields the following two equations.

$$\begin{cases} \boldsymbol{\mu}_+ = 3.37\,\boldsymbol{\mu}_1 - 2.37\,\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_- = -0.42\,\boldsymbol{\mu}_1 + 1.42\,\boldsymbol{\mu}_2 \end{cases} \tag{2}$$

Two steps are necessary to obtain an estimate of the average target and non-target ERP responses. In a first step, the sequence means ($\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$) are estimated. In a second step, these estimates are plugged into the equation set 2. Given independent and identically distributed (IID) data points, the estimated sequence means will converge to their true value, such that the true class means can be obtained in the limit.

An implicit assumption of LLP in this formulation is the *homogeneity assumption*, which states that the average target and non-target responses have to be the same in both subsets. The first precautionary measure to accomplish this is to randomize the order of events from both sequences within each single trial. This should guarantee that the target-to-target interval (TTI) does not depend on the sequence, which seems like a desirable characteristic given the known influence of target-to-target intervals onto e.g. the P300 amplitude [8]. The second precautionary measure we propose is to match the overall visual stimulus intensity between both sequences. It is expressed by the overall number of symbols highlighted on the screen in one point in time. To reach a balanced setting, the traditional P300-spelling matrix was extended by 10 additional blank/hash symbols ('#'). These are included only in the second highlighting sequence and should never be attended by the subject, thus they never serve as targets. This simple trick ensures that events of both sequences have the same number of highlighted symbols per event while having the predefined target and non-target ratios. An example is depicted in Fig. 1. Apart from these manipulations, the sequences were generated in a pseudo-random manner using a heuristic approach designed by Verhoeven et al. [9] with the goal to minimize double flashes and adjacency distractions. We previously showed that the homogeneity assumption holds for the data generated in the above way [6].

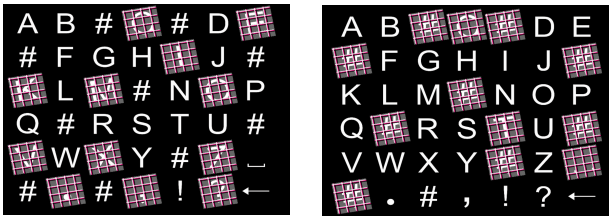

Fig. 1: **Examples of pseudo-randomized highlighting events** of sequence 1 (left) and sequence 2 (right). Sequence 1 never highlights '#' symbols, while sequence highlights 8-9 '#' symbols per event.

Classification

Given estimates of the class means, several classifier approaches are possible. It was previously observed that the ERP responses for targets and non-targets closely follow a multivariate normal distribution with the same covariance matrix [10]. Based on this assumption, linear discriminant analysis (LDA) classifier have shown to be very competitive in ERP-BCI paradigms [11]. These are linear classifiers looking for the best projection $\boldsymbol{w}$ such that samples $\boldsymbol{x}$ are assigned to the target class if $\boldsymbol{w} \cdot \boldsymbol{x} \geq 0$ and to the non-target class otherwise. It is known that the optimal projection can be found solely by knowing the shared covariance matrix $\boldsymbol{\Sigma}$ and the class-wise means $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ in the following way [10]:

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-\right).$$

In the formulation of Blankertz et al. [10], the class-wise covariance matrix is used as $\boldsymbol{\Sigma}$. However, one can show that the pooled covariance matrix, i.e. the covariance computed on the complete data-set, leads to a projection which has the same orientation, but is scaled differently [12]. As no label information are used for the pooled covariance estimation, we can replace the class-wise covariance matrix by the pooled covariance matrix and still obtain the same classifier — if the means are

 DOI: 10.3217/978-3-85125-533-1-35

correctly estimated. In addition, it was shown that regularizing the estimated covariance matrix towards an identity matrix leads to better performance. Hence, we adopt the covariance-shrinkage by Ledoit & Wolf as proposed in [10]. In summary, data is classified by combining the mean estimations derived from LLP with the regularized pooled covariance in order to obtain a linear classifier.

To compare the LLP-approach to a traditional classifier, we included a supervised classifier into our comparisons. Again, we used a LDA classifier with regularized covariance matrix [10], which is precisely the same model as described above, only that the class-wise means and covariances are estimated using the sample statistics based on the label information. As supervised classifiers are much more prone to overfitting than unsupervised approaches, we estimated the generalization performance on increasing data sets in a chronological 5-fold cross-validation.

Data

The data used in this work was previously described in our paper on LLP [6]. A short summary is given here. In an online copy spelling tasks, thirteen subjects (5 female, average age: 26 years) spelled three times the same sentence of 63 symbols. The stimulus onset asynchrony (SOA) was 250 ms. To spell an individual symbol, a train of 68 highlighting events were presented. This train is the result of randomly interleaving 32 events of sequence 1 and 36 events of sequence 2, which had been generated as described before. A very salient highlighting method proposed by Tangermann et al. [13] using a combination of brightness enhancement, rotation, enlargement and a trichromatic grid overlay was used. EEG signals were recorded at 1 KHz sampling rate with 31 passive Ag/AgCl electrodes (EasyCap) placed on the head according to the extended 10-20 system. The reference was placed on the nose. After a baseline correction of each EEG epoch, the ERP features were extracted as the average potential values in the six intervals [50, 120], [121, 200], [201, 280], [281, 380], [381, 530] and [531, 700] ms. Outlier epochs were not removed at any time during the data processing, however, visual inspection ensured, that classification was not performed based on eye movement artefacts.

The data of all 13 subjects is freely available online at: `http://doi.org/10.5281/zenodo.192684`.

Performance Measures

The selection accuracy is the percentage of characters intended to spell by a user and that were decoded correctly by the BCI system. As the selection accuracy is only based on 63 symbols per sentence, it is quite a noisy performance metric. It can be evaluated online by reporting the actual performance observed during the experiment, but it can also be estimated by a post-hoc offline analysis.

For the latter, we decided to simulate the experiment several times and on different subsets of the data. Within each sentence, the classifier was restarted multiple times to obtain a better estimate of the selection accuracy. A total of 7 classifiers were simulated per sentence, and each classifier was trained on data of 21 characters: The first one used the characters 1-21, the second one used characters 8-28, . . ., and the last one used the characters 43-68. Finally, these spelling accuracies were averaged across the classifiers, subjects and experimental blocks.

We also looked at how well single targets (attended highlighting events) can be discriminated from non-targets (not-attended highlighting events). This was quantified by the area under the curve (AUC) of the classifier outputs as a threshold-independent and robust performance metric. The AUC values range between 0 % and 100 %, with a theoretical chance level of 50 %. An AUC value of 100 % indicates perfect separation between the two classes, i.e. the classifier can correctly tell for each highlighting event whether it was attended or not. To compute these values in this paper, we simulated an online experiment with different number of channels where the LLP classifier was retrained after each character. The performance was then computed on the complete data-set up to that point by using the given label information. Please note that overfitting is not a problem in this context, because the classifiers do not use the label information for training.

Channel Selection

To determine the importance of an EEG channel in isolation, it is a common strategy to estimate its informative content with respect to the target vs. non-target classification task. We estimated this by assessing univariate informative content expressed by the AUC between the target and non-target features derived from the same six intervals as mentioned before. This was computed for each subject and channel. The AUC values were rescaled to have a theoretical chance level of 0 by applying the following formula:

$$\mathrm{AUC}_{\mathrm{rescaled}} = |(\mathrm{AUC}_{\mathrm{regular}} - 0.5) \cdot 2|$$

These values were then averaged across all subjects and intervals to obtain one relevance score per channel. Unlike in wrapper methods for channel selection [14], this simple and fast relevance score does not grasp informative content of a channel which is only expressed in combination with other channels.

RESULTS

Channels sorted according to their descending importance are shown in Fig. 2. One can observe that the occipital channels *O1* and *O2* have the highest importance which is in accordance with our expectations for a visual ERP experiment [6] that should elicit class-discriminative visual ERP components for electrodes located over the

 DOI: 10.3217/978-3-85125-533-1-35

occipital cortex. The five next top-ranked channels are mostly located over the centro-parietal cortex, showing a slight lateralization to the right hemisphere.

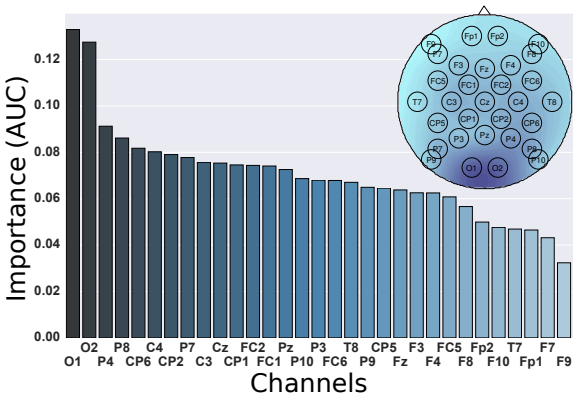

Fig. 2: **Channels sorted according to their importance.** The y-axis shows the averaged AUC value across all 13 subjects for each channel.

According to these results, we selected the n first-ranked channels for $n \in \{3, 5, 10, 31\}$ to run offline experiments with classifiers trained on smaller channel subsets. In the simulation with growing amounts of data, the classifiers were retrained after each additional character and directly applied to classify the current character. The target vs. non-target LLP performances reported in Fig. 3 overall reflect a growing amount of information provided by larger and larger training data sets. Interestingly, a reduced number of 10 channels outperforms the complete set of 31 channels, even if the full amount of data is used. The difference between channel subsets and the full channel set is especially pronounced, when data from only few epochs is available, which corresponds to the early stages of a spelling session. With more and more training data available, the configuration with 10 channels remains on top, while 5 and 31 channels are close behind. The configuration with 3 channels falls back by a good margin.

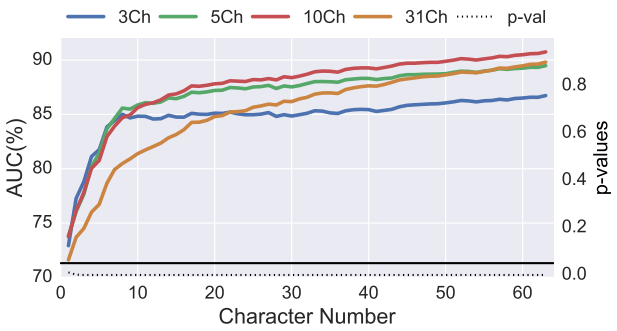

Fig. 3: **Average simulated LLP classification accuracy** in dependence of the number of channels and training points. Per trial, the p-value of a Wilcoxon signed-rank test is given, comparing the results of 31 channels and 10 channels.

When looking at the selection accuracy in the initial ramp-up phase of LLP in Fig. 4, a similar behaviour is observable. Again, the full set of channels is outperformed by a reduced number of channels. Around 70 % of the characters are already classified correctly from the third character on, when using a reduced channel number. This corresponds to less than a minute of training time. After a bit more than 2 minutes of training time, the selection accuracy already exceeds 80 % for the configuration with 5 or 10 channels.

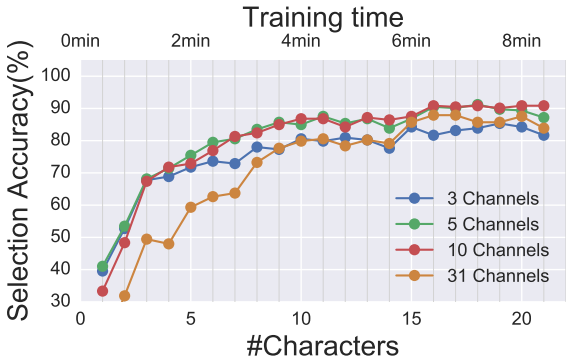

Fig. 4: **Average simulated LLP selection accuracy** as a function of training time / amount of data points and channels.

Finally, we ran a simulation of a supervised shrinkage-LDA classifier to compare the effect of reduced channel subsets with those observed for LLP. Fig. 5 shows that 10 channels significantly outperform 31 channels when few data is available. However, the supervised classifier is able to learn much quicker and utilize the additional information from all channels leading to significantly better performance when data of more than 15 characters is available.

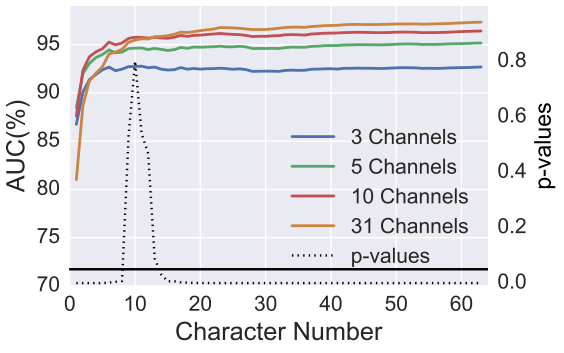

Fig. 5: **Average simulated supervised single epoch accuracy** as a function of training time / data points and channels. For each character, the complete data up to that point was divided in 5 chronological parts and a cross-validation was applied. The p-values show the outcome of a Wilcoxon signed-rank test between the configurations with 31 and 10 channels.

 DOI: 10.3217/978-3-85125-533-1-35

DISCUSSION

In theory, more information is contained in the complete channel set. Reducing the number of channels corresponds to a reduced model complexity and less free parameters and effectively regularizes the learning problem, which can be a beneficial strategy for smaller data sets. In accordance with this reduced complexity, the LLP was able to learn the essential characteristics in our simulations much faster with reduced channel sets. The best performance could be achieved with 10 channels. While this is an expected outcome, it is interesting that LLP could not reach a superior performance for the full channel set despite the full amount of data provided. To explain this, one hypothesis is that even when a lot of data is available, a reduced number of channels, which are highly correlated to the control task, could facilitate the learning of LLP in comparison to using more channels which are potentially polluted by task-irrelevant features, including noise. The alternative hypothesis is that our experiment has just not delivered a sufficient amount of data in order to make the 31 channel LLP outperform the smaller subsets. Looking at the results from the supervised classification, one can see a similar behaviour in the early stage, but the configuration with 31 channels quickly rises to the top.

The similarity of the behaviour is in accordance with our theoretical considerations in [6], where we introduced the term *noise amplification factor* (NAF). It measures how much more data is necessary for LLP to reach the same accuracy in the class mean estimation compared to the supervised case. This NAF metric depends on the target and non-target ratios of each sequence and can be influenced by designing the experimental paradigm. For the sequences used in this data-set, the NAF was around 20. We observe that – with growing data – the unsupervised LLP behaves like a slowed-down supervised method. This is expressed by a qualitatively similar behaviour of the curves, as they display the same order of intersection points in Figures 3 and 5 with growing data set sizes. Observing this similarity in the behaviour of the LLP and the supervised methods supports the second hypothesis that the LLP configuration with 31 channels will eventually outperform the smaller channel sets when a sufficient amount of data is available.

We cannot directly observe a slowing factor of 20, which would correspond to the NAF. One of the causes is that the performance of these classifiers is not only dependent on the quality of the estimated class means, but also on the quality of the estimated covariance matrix.

Realizing the slow ramp-up behaviour of LLP with large channel sets, a possible mitigation strategy is to start with a low number of channels and incrementally increase the number of channels and features over time. A similar approach was already proposed for motor imagery data classification [15, 16]. By developing new adaptive feature selection methods, other unsupervised methods, e.g. the MIX method [17], which is the combination of LLP and an expectation-maximization algorithm, could also benefit.

CONCLUSION

By re-analysing data from 13 subjects performing a copy-spelling task, we showed that the unsupervised LLP classifier can be significantly improved by simply reducing the number of utilized features. Interestingly, this is not only the case during the initial ramp-up, but even during a later stage of the experiment when more data is available. In contrast, a supervised classifier only benefits from a reduction of features in the early stage. This behaviour of the unsupervised LLP method can be understood by considering it as a slowed-down version of the supervised algorithm with the same guaranteed convergence. Future work will go towards the development of adaptive feature selection method for unsupervised classifiers utilizing the observations made in this paper.

ACKNOWLEDGEMENTS

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 thanks for the support from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement NO 657679. TV gratefully acknowledges the support by the special research fund (BOF) from Ghent University. The authors also acknowledge support by the state of Baden-Württemberg through bwHPCand the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG.

REFERENCES

[1] Lu S, Guan C, and Zhang H. Unsupervised brain computer interface based on intersubject information and online adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(2):135–145, 2009.

[2] Fazli S, Popescu F, Danóczy M, Blankertz B, Müller K-R, and Grozea C. Subject-independent mental state classification in single trials. *Neural networks: The Official Journal of the International Neural Network Society*, 22(9):1305–1312, Jun 2009.

[3] Jayaram V, Alamgir M, Altun Y, Schölkopf B, and Grosse-Wentrup M. Transfer learning in brain-

 DOI: 10.3217/978-3-85125-533-1-35

computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

[4] Vidaurre C, Kawanabe M, von Bünau P, Blankertz B, and Müller K-R. Toward unsupervised adaptation of LDA for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 58(3):587–597, 2011.

[5] 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.

[6] 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.

[7] 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.

[8] Gonsalvez C-J and Polich J. P300 amplitude is determined by target-to-target interval. *Psychophysiology*, 39(3):388–396, 2002.

[9] Verhoeven T, Buteneers P, Wiersema J, Dambre J, and Kindermans P-J. Towards a symbiotic brain–computer interface: exploring the application–decoder interaction. *Journal of Neural Engineering*, 12(6):066027, 2015.

[10] 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.

[11] 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.

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

[13] 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.

[14] Schröder M, Lal T. N, Hinterberger T, Bogdan M, Hill N J, Birbaumer N, Rosenstiel W, and Schölkopf B. Robust EEG channel selection across subjects for brain computer interfaces. *Eurasip Journal of Applied Signal Processing*, 19:3103–3112, 2005.

[15] Vidaurre C and Blankertz B. Towards a cure for BCI illiteracy. *Brain Topography*, 23(2):194–198, 2010.

[16] Sannelli C, Vidaurre C, Müller K-R, and Blankertz B. CSP patches: an ensemble of optimized spatial filters. an evaluation study. *Journal of Neural Engineering*, 8(2):025012, 2011.

[17] 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.