Spatio-temporal feature extraction and classification of Event-Related Potentials

Nisrine Jrad, Marco Congedo

Spatio-temporal feature extraction and classification of Event-Related Potentials

Nisrine Jrad[1], Marco Congedo[1]

Vision and Brain Signal processing (ViBS) Group,
GIPSA-lab CNRS (Centre National de la Recherche Scientifique), Grenoble Univ.
961 rue de la Houille Blanche, 38402 GRENOBLE Cedex, France
nisrine.jrad@gmail.com, marco.congedo@gmail.com

Résumé : Brain-Computer Interfaces (BCI) translate variations in the Electroencephalogram (EEG) into a set of particular commands, in order to control a real world machine. For this purpose, it is necessary to classify reliably EEG signals. Classifying EEG activities is a challenging task since EEG recordings exhibit distinct and individualized spatial and temporal characteristics correlated with noise and various physical and mental activities. To increase classification accuracy, it is thus crucial to enhance the Signal to Noise Ratio (SNR) and to identify relevant spatio-temporal features.

This paper presents a method for denoising Event-Related Potential (ERP) data and for identifying discriminant spatio-temporal characteristics. First, a Blind Source Separation (BSS) strategy is used to denoise data and enhance SNR. Second, a resampling procedure based on Global Field Power (GFP) automatically selects temporal windows. Third, a spatially weighted SVM (sw-SVM) learns a spatial filter optimizing the classification performance for each temporal feature. Finally, the so obtained ensemble of sw-SVM classifiers are combined using a weighted combination of all sw-SVM outputs. Results indicate that denoising and identification of spatio-temporal features of ERP enhance the classification accuracy, yield a better understanding of the underlying physiology and provide useful insight about the spatio-temporal characteristics of the ERP.

Mots-clés : **Analyse de Données Spatiales et Spatio-Temporelles, Apprentissage supervisé, Sélection, Construction et Extraction de Variables.**

1. Introduction

Many brain computer interfaces (BCI) make use of Electroencephalography (EEG) signals to categorize two or more classes and associate them to simple computer commands. Classification of brain signals is challenging because EEG records are high dimensional measurements corrupted by noise

Congedo *et al.* (2008). Thus, it is important to remove noise in order to enhance signal, and to perform feature extraction in order to feed the classification algorithm with relevant features.

Several methods based on Independent Component Analysis have been proposed to enhance the Signal to Noise Ratio (SNR) and to remove artefacts. However, these methods are not specifically designed to separate brain activities and they are supervised. Indeed, after decomposition in different components, it is necessary to select (manually or thanks to spatio-temporal prior) components containing evoked potentials. In this work, Event-Related Potentials (ERPs) are considered and an unsupervised denoising method is used. It is based on the xDAWN algorithm Rivet *et al.* (2009), which has been specifically conceived to maximize the SNR of ERPs.

Concerning feature extraction, usually spatial decomposition is performed to extract the ERP components, including Principal Component Analysis, Independent Component Analysis, etc. These methods define the decomposition in terms of statistical proprieties the components should satisfy in a specific time window. However, ERPs describe several temporal components (peaks), thus, spatial decomposition should be performed for each interesting interval occurring in the windows of interest. To this end, some algorithms have been proposed to study where the discriminative information lies into the spatio-temporal plane. They visualize a matrix of separability measures into the spatio-temporal plane of the experimental conditions. The matrix is obtained by computing a separability index for each pair of spatial electrode measurement and time sample. Several measures of separability have been used, for instance the signed-r^2 Blankertz *et al.* (2010), Fisher score and Student's t-statistic Müller *et al.* (2004), or the area under the ROC curve Green & Swets (1966). Separability matrix should be sought as to automatically determine intervals with fairly constant spatial patterns and high separability values. This proves difficult and heuristics are often employed to approximate interval borders. In addition, the three first aforementioned measures rely on the assumption that the class distributions are Gaussian, which is seldom verified.

To overcome all these drawbacks, we develop a spatio-temporal data driven decomposition technique without any a priori knowledge or any assumption regarding EEG dynamics. A two-stage feature extraction technique is proposed. First, a time feature extraction is performed based on Global Field Power (GFP) Lehmann & Skrandies (1980), defined for each time sample as the sum of the square potential across electrodes. GFP peaks are associated with maximal SNR. Second, a spatially weighted SVM (sw-SVM), proposed in Jrad *et al.* (2011), is used to learn for each time interval a sparse spatial filter optimizing directly the classification performance. Finally, the ensemble of sw-SVMs obtained on the selected temporal features are combined using a weighted average, to get a robust decision function.

The remainder of this paper is organized as follows. The proposed method is introduced in Section 2.. Section 3. accounts for data sets description and discusses the experimental results. Finally, Section 4. holds our conclusions.

2. Method

2.1. Problem description

Background brain activities, irrelevant to BCI tasks, continuously generate EEG signals that can be recorded anywhere over the scalp. These signals interfere with the EEG signals triggered by stimuli. In addition to the background EEG, there are other sources of artefact which usually affect recordings Congedo *et al.* (2008). Fortunately, post-stimulus signals present specific space and temporal characteristics, since they are generated in particular regions of the brain at a given interval of time. This section describes a method for analyzing ERPs considering the sequence :

- a signal denoising,
- a temporal feature extraction,
- a spatial feature extraction embedded in a classification scheme,
- an ensemble of classifiers learning technique.

2.2. Data denoising

A conceptual model for the elimination of noise and other undesirable components from multi-dimensional data is presented. First, a Blind Source Separation (BSS) is performed using xDAWN Rivet *et al.* (2009). xDAWN performs a signal decomposition described by a linear transformation of data as $\boldsymbol{\nu}(t) = \boldsymbol{B}\boldsymbol{x}(t)$, where $\boldsymbol{x}(t) \in \mathbb{R}^S$ is the electrodes measurement vector, $\boldsymbol{\nu}(t) \in \mathbb{R}^S$ holds the time-course of source components and $\boldsymbol{B}^{S\times S}$ is the unmixing matrix. The unmixing matrix $\boldsymbol{B}$ is computed so that ratio of the post-stimulus response to signal-plus-noise is maximised.

Due to linearity we can write $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{\nu}(t)$, where the mixing matrix $\boldsymbol{A}^{S\times S} = \boldsymbol{B}^{-1}$ is the inverse of the unmixing matrix $\boldsymbol{B}$. The entries of estimated mixing matrix indicate how strongly each electrode picks up each individual component. Denoising and back projection onto the sensor space can be written as $\tilde{\boldsymbol{x}}(t) = \boldsymbol{ARBx}(t)$, where $\boldsymbol{R}$ is a diagonal matrix with s^{th} diagonal element equal to 1 if the s^{th} component is to be retained and equal to 0 if it is to be removed.

This strategy allows us to enhance the SNR while attenuating backgroung EEG and artefacts. Since xDAWN sorts components by decreasing SNR, we retain the first half of them so as to reduce noise whilst keeping meaningfull identifiable components.

2.3. Temporal features

In the following, we consider BCI application with two classes of action. After denoising recordings, we can get a training set of labeled trials. A decision function should be learned from this training set. The decision function should correctly classify unlabeled trials. Let us denote a denoised post-stimulus trial $p(p \in \{1, \dots P\})$, recorded over electrode $s(s \in \{1, \dots S\})$ at instant $t(t \in \{1, \dots T\})$, as $\tilde{x}_s(t)$. A post-stimulus trial p recorded over S electrodes in a short time period of T samples can be considered as a matrix $\tilde{\boldsymbol{X}}_p \in \mathbb{R}^{S \times T}$. Hence, the entire available set of data is $\{(\tilde{\boldsymbol{X}}_1, y_1), ..., (\tilde{\boldsymbol{X}}_p, y_p), ..., (\tilde{\boldsymbol{X}}_P, y_P)\}$ with $y_p \in \{-1, 1\}$ the class labels. Our task consists in finding the spatio-temporal features that maximize discrimination between two classes.

To select temporal intervals in the ERP where discriminative peaks appear, the Global Field Power (GFP) Lehmann & Skrandies (1980) is computed on the difference of the grand averages of the two class post-stimulus trials such as :

$$GFP^2(t) = \frac{1}{S} \sum_{s=1}^{S} \left(\sum_{\tilde{\boldsymbol{X}}_{P/y_p=1}} \tilde{x}_s(t) - \sum_{\tilde{\boldsymbol{X}}_{P/y_p=-1}} \tilde{x}_s(t) \right)^2 \qquad (1)$$

Pronounced deflections with large peaks, denoting big dissimilarities between the two activities, are associated with large GFP values. Windows involving significant temporal features are chosen as intervals where GFP is high relative to the background EEG activity.

To select significant windows we require a statistical threshold for the observed GFP of the difference grand average trials in the two classes. Such threshold is estimated with a resampling method as the 90^{th} percentile (10% type I error rate) of the appropriate empirical null distribution. For P and Q observed single trials in classes labeled 1 and -1, respectively, we resample

P and Q trials with random onset from the entire denoised EEG recording. We compute the difference of the grand average of the P and Q random trials and retain the maximum value of GFP. The procedure is repeated 1000 times and the sought threshold is the 90^{th} percentile of such max-GPF null distribution. Taking the max-GPF at each resampling ensures that the nominal type I error rate is preserved regardless the number of windows that will be declared significant Westfall & Young (1993).

Noteworthily, contiguous samples with high GFP coincide with stable deflection configurations where spatial characteristics of the field remains unchanged Lehmann & Skrandies (1980). However, artefacts, like blinks, can also cause peaks in GFP measurement. Hence, denoising signal is recommanded before performing time interval selection. Since within each selected time window the spatial pattern is fairly constant, average across time can be calculated within these intervals. Averaging over time rules out aberrant values, reduces signal variability and attenuates noise. Besides, it reduces dramatically time dimensionality to I where I is the number of significant time features.

2.4. Spatial features and classifier : sw-SVM

Temporal filtering provides us with $\tilde{\boldsymbol{X}}'_p \in \mathbb{R}^{S\times I}$ trials. Each column vector $\tilde{\boldsymbol{x}}'_p \in \mathbb{R}^{S\times 1}$ reflects a spatial characteristic at a temporal feature $i \in \{1, ..., I\}$. Hence, I spatial filters are learned over the different time components. In this work, spatial filtering is learned jointly with a classifier. The method was proposed in Jrad *et al.* (2011) and called sw-SVM method for spatially weighted SVM. It involves spatial feature weights in the primal SVM optimization problem and tunes these weights as hyper-parameters of SVM. We denote by $\mathbf{d} \in \mathbb{R}^{S\times 1}$ the spatial filter and $\boldsymbol{D}$ a matrix with $\mathbf{d}$ on the diagonal. Matrix $\boldsymbol{D}$ is learned by solving the sw-SVM optimization problem :

$$\begin{aligned} &\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{D}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{p=1}^{P}\xi_p \\ \text{subject to}\quad & y_p(\langle \boldsymbol{w}, \boldsymbol{D}\tilde{\boldsymbol{x}}'_p\rangle + b) \geq 1-\xi_p \quad \text{and} \quad \xi_p \geq 0 \quad \forall p \in \{1,\dots,P\} \\ \text{and}\quad & \sum_{s=1}^{S} D_{s,s}^2 = 1 \quad \forall s \in \{1,\dots,S\} \end{aligned} \tag{2}$$

where $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ is the normal vector, $b \in \mathbb{R}$ is an offset, ξ_p are called slack variables that ensure the problem has a solution in case the data is not linearlyseparable, and C is the regularization parameter that controls the trade-off between a low training error and a large margin.

By setting to zero the derivatives of the partial associated Lagrangian according to the primal variables $\boldsymbol{w}$, b and ξ_p the optimization problem of the dual formulation can be written as :

$$\begin{aligned} \min_{\tilde{\boldsymbol{D}}} \max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Y}^T\tilde{\boldsymbol{X}}'^T\tilde{\boldsymbol{D}}\tilde{\boldsymbol{X}}'\boldsymbol{Y}\boldsymbol{\alpha} \\ \text{subject to} \quad \boldsymbol{y}^T\boldsymbol{\alpha} = 0 \\ \text{and} \quad 0 \leq \alpha_p \leq C \quad \forall p \in \{1, \ldots, P\} \\ \text{and} \quad \sum_{s=1}^{S} \tilde{D}_{s,s} = 1, \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}$ are the vectors of Lagrangian multipliers, $\tilde{\boldsymbol{D}} = \boldsymbol{D}^T\boldsymbol{D}$, $\tilde{\boldsymbol{X}}' = \{\tilde{\boldsymbol{x}}'_1, ..., \tilde{\boldsymbol{x}}'_P\}$, $\boldsymbol{y}^T = \{y_1, ..., y_P\}$ is the vector containing the labels and $\boldsymbol{Y} = Diag(\boldsymbol{y})$ is the diagonal matrix containing the labels. The overall problem boils down to be equivalent to a Multiple Kernel Learning (MKL) problem where a linear kernel is used over each sensor time series and $D^2_{s,s}$ are the positive mixing coefficients of the multiple kernels. Several optimization algorithms were proposed to solve MKL optimization problem. For instance, semi-infinite linear programming, gradient descent level method, etc. Bach (2008) derived equivalence between MKL problems and group lasso and Z. Xu & Lyu (2010) proposed an efficient multiple kernel learning by group Lasso. In this work, we used a gradient descent as in SimpleMKL Rakotomamonjy *et al.* (2008).

2.5. Ensemble of sw-SVM classifiers

As seen above, a way to reduce EEG variability is to perform signal averaging across time. Another way to reduce this influence, from a classification point of view, is to use an ensemble of classifiers Rakotomamonjy & Guigue (2008). According to this strategy, a multiple sw-SVM system is designed for each temporal feature. A weighted average on sw-SVM outputs is used to determine a set of significant classifiers.

Weights are set as the product of two functions growing exponentially with the accuracies of the two (positive and negative) classes, evaluated on a validation set. If TP, FP, TN and FN hold for True Positive, False Positive, True Negative and False Negative of a given sw-SVM classifier, respectively, then Positive Predective Value (PPV) and Negative Predective Value (NPV) can be defined as :

$$PPV = \frac{TP}{TP+FP}, \quad NPV = \frac{TN}{FN+TN},$$

and weights associated to the classifier are such as :

$$\begin{cases} 0 \quad \text{if} \quad PPV < 0.5 \quad \text{or} \quad NPV < 0.5 \\ \tan(2(PPV-0.5)) \times \tan(2(NPV-0.5)) \quad \text{otherwise,} \end{cases}$$

where 0.5 is the chance level. The trigonometric tangent function is used because it increases slowly around zero angles (corresponding to PPV or $NPV = 0.5$) and it increases rapidly at angles close to one (angles equal to one correspond to PPV or $NPV = 1$). Taking the product of the two tangents makes this weighting strategy ideal for unbalanced data sets since it seeks classifiers that jointly present good accuracies for both classes.

3. Experimental results

3.1. ErrP data set

The proposed method was evaluated on a visual feedback ErrP Miltner *et al.* (1997) experiment. Eight BCI-naif healthy subjects performed the experiment. They had to retain the position of a sequence of digits and to localize where a target digit previously appeared. A visual feedback indicates wether the answer was correct (green feedback) or not (red). Number of digits composing the sequences was adapted with an algorithm tuned to allow around 20% errors for all subjects. Experiment involved 2 sessions that lasted together approximately half an hour. Each session consisted of 6 blocks of 6 trials, for a total of only 72 trials. Recordings of EEG were made from 31 electrodes. Raw EEG potentials were re-referenced to the common average and filtered using a $1-10$Hz 4^{th} order butterworth filter. A window of 1000ms posterior to the feedback has been explored for each trial. All trials were kept for analysis and no supervised artifact rejection whatsoever was performed.

3.2. Results

The proposed technique was applied to the ErrP data. Single trial classification of ErrPs is assessed using a 5-Cross Validation technique. Each single training sw-SVM involves a selection procedure for setting its regularization parameter C and weights associated to its outputs. Results are those obtained with the best performance over the 5 partitions. Noteworthily, it would be more realistic to perform a two-level-5-cross validation, a cross-validation to select hyperparameters and one to compute performance. However, in the study of this dataset, with very small amount of trials, it was not possible to perform these two-levels.

Figure 1 shows the average of the difference error-minus-correct for channel FCz of subject $S7$ and the associated GFP. Only two components can be seen : a negative deflection around 150ms after the feedback and a second negative component occurring in between 400 and 480ms. Scalp potentials topographies associated with the two extracted temporal features are also shown in Figure 1. The 1^{st} negative peak seems to be occipital whereas the 2^{nd} negative peak covers a rather fronto-central area. Figure 1 shows accuracies for error and correct classes for each sw-SVM and their corresponding weights (normalized between 0 and 1). Only sw-SVM learned on the 2^{nd} deflection shows good accuracies for both classes and is thus retained.

An important question is whether time interval selection, found by GFP, are consistent across different partitions of the data and across subjects. Figure 2 shows, for each subject, and each of the 5 partitions, temporal intervals (in white) selected on the ErrP data set. Because of the very small number of trials used in each partition, some inter-partition differences can be noted in these data, but overall, the procedure appears robust and meaningfull. Latencies, thus selected time intervals, are different from subject to subject, which is not surprising. However, for almost all subjects, an important activity is noted between 400 and 600ms. These findings confirm those of Miltner *et al.* (1997) where a negative deflection, following an incorrect visual feedback of a time-production task, peaked at 330ms with a duration of 260ms. This witnesses in favor of the effectiveness and the consistency of the proposed temporal feature extraction.

Concerning classification results, figure 3 shows the 5 Cross-Validation performance provided by a classical SVM approach where all electrodes are used, the sw-SVM where only one spatial filter was used on the whole trial duration and the proposed method. The proposed method proved constantly superior to SVM and sw-SVM. A paired student's t-test was computed to compare the proposed method to the sw-SVM and SVM. Results were : $t(7) = 3.0893\,;\,(p = 0.01760)$ and $t(7) = 4.2515\,;\,(p = 0.0038)$, respectively. We conclude that, inclusion of temporal features selection after denoising, along

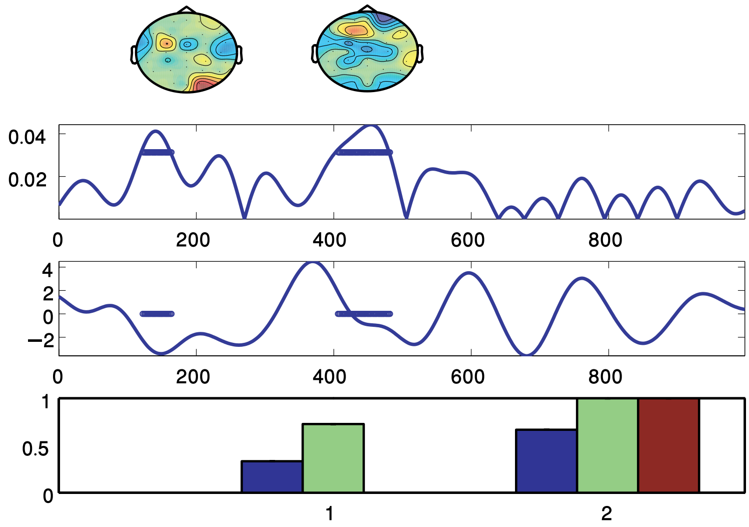

FIGURE 1: Denoised ErrP. Top : GFP computed on the difference of the grand average error-minus-correct for 1s trials, selected intervals and topographies associated. Middle : the difference computed on electrode FCz. Bottom : accuracies for error (blue bar) and correct (green bar) classes and sw-SVM associated weights (red bar, normalized between 0 and 1).

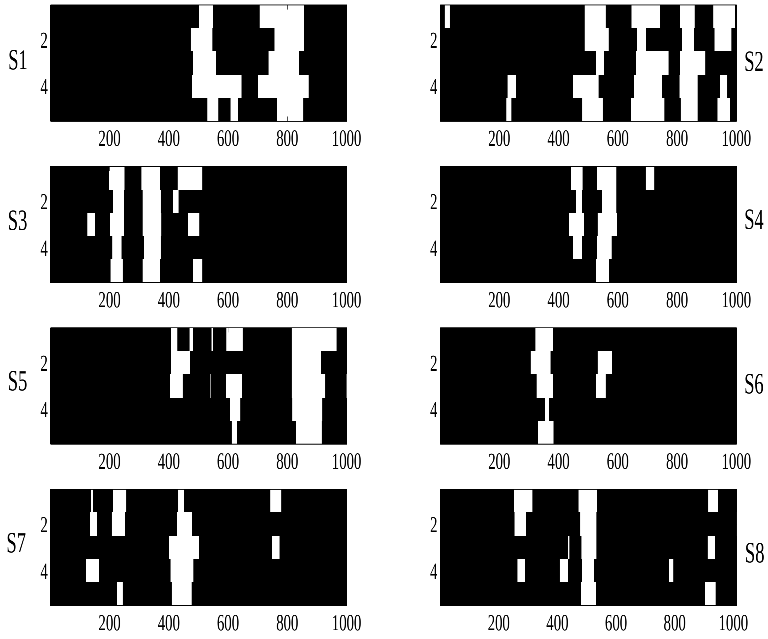

FIGURE 2: Denoised data : selected time intervals are shown in white pixels, for each of the 8 subjects and 5 partitions. Each matrix refers to a subject where columns hold time-course and rows hold partitions.

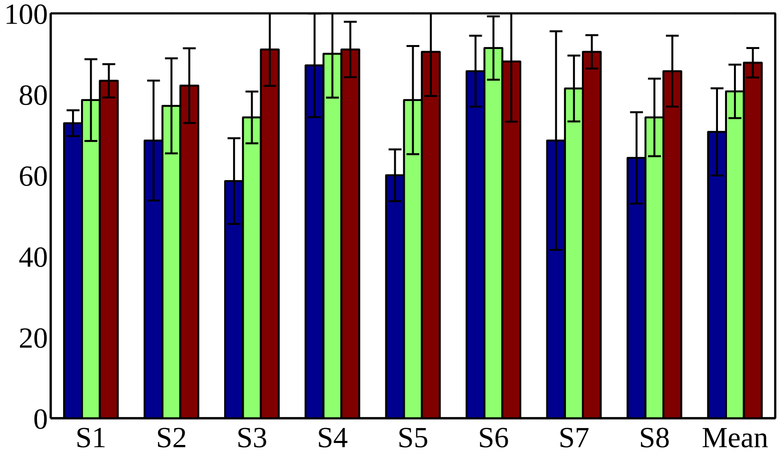

FIGURE 3: From left to right : performances of classical SVM, sw-SVM and the proposed method for the 8 subjects. Mean (std) accuracies across the 8 subjects are of $70.71(10.77)$, $80.71(6.61)$ and $87.80(3.63)$ respectively.

with learning an ensemble of classifiers, yield superior performance.

4. Conclusion

In this paper, an unsupervised EEG denoising and a spatio-temporal feature identification strategies were addressed. Denoising is based on Blind Source Separation where unmixing matrix is given by xDAWN algorithm. An analysis of Global Field Power highlighted time periods of interest where effects are likely to be robust yielding to a data-driven temporal feature extraction. For each temporal feature, a spatial filter was learned jointly with a classifier in the SVM theoretical framework. Spatial filters were learned to optimize classification performance. A weighted averaging on the so obtained ensemble of classifiers yielded to a robust final decision function. Experimental results on Error-related Potential data sets illustrate the efficiency of the method from a physiological and a machine learning points of view. The accuracies we obtained are clearly competitive against the state-of-the-art classification of Error-related Potentials. These results motivate further research that may aim to extract all relevant aspects of brain post-stimulus dynamics recorded in EEG (spatio-temporal-frequential).

Acknowledgment

This work has been supported through the project OpenViBE2 of the ANR (National Research Agency), France.

Références

BACH F. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research 9*, **9**, 1179–1225.

BLANKERTZ B., S.LEMM, TREDER M., HAUFE S. & MÜLLER K. (2010). Single-trial analysis and classification of ERP components – a tutorial. *NeuroImage*. in press.

CONGEDO M., GOUY-PAILLER C. & JUTTEN C. (2008). On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. *Clin. Neurophysiol.*, **119**(12), 2677–2686.

GREEN M. D. & SWETS J. (1966). *Signal detection theory and psychophysics*. Huntington, NY : Krieger.

JRAD N., PHLYPO R. & CONGEDO M. (2011). Svm feature selection for multidimensional eeg data. In *International Conference on Acoustic, Speech and Signal Processing 2011*.

LEHMANN D. & SKRANDIES W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalogr Clin Neurophysiol*, **48**, 609–21.

MILTNER W., BRAUN C. & COLES M. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task : Evidence for a generic neural system for error detection. *Journal of Cognitive Neuroscience*, **9**, 788–798.

MÜLLER K., KRAULEDAT M., DORNHEGE G., CURIO G. & BLANKERTZ B. (2004). Machine learning techniques for brain-computer interfaces. *Biomed Tech*, **49**(1), 11–22.

RAKOTOMAMONJY A., BACH F., CANU S. & GRANDVALET Y. (2008). SimpleMKL. *Journal of Machine Learning Research 9*.

RAKOTOMAMONJY A. & GUIGUE V. (2008). BCI Competition III : Dataset II - Ensemble of SVMs for BCI P300 speller. *IEEE Trans. Biomedical Engineering*, **55**(3).

RIVET B., SOULOUMIAC A., ATTINA V. & GIBERT G. (2009). xdawn algorithm to enhance evoked potentials : Application to brain computer interface. *IEEE Trans Biomed Eng*.

WESTFALL P. & YOUNG S. (1993). *Resampling-based Multiple Testing*. New York : Wiley.

Z. XU, R. JIN H. Y. I. K. & LYU M. R. (2010). Simple and efficient multiple kernel learning by group lasso. In *ICML*, p. 1175–1182.