sw-SVM : sensor weighting Support Vector Machines for EEG-based Brain-Computer Interfaces

Nisrine Jrad, Marco Congedo, Ronald Phlypo, Sandra Rousseau, Rémi Flamary, Florian Yger, Alain Rakotomamonjy

sw-SVM : sensor weighting Support Vector Machines for EEG-based Brain-Computer Interfaces

N Jrad[1], M Congedo[1], R Phlypo[1], S Rousseau[1], R Flamary[2], F Yger[2] and A Rakotomamonjy[2]

[1] Equipe ViBS (Vision and Brain Signal processing), GIPSA-lab, CNRS, Grenoble University, FRANCE

[2] Equipe DocApp, LITIS EA 4108 INSA/Université de Rouen, FRANCE

E-mail: nisrine.jrad@gmail.com

Abstract. In many machine learning applications, like Brain-Computer Interfaces (BCI), high-dimensional sensor array data are available. Sensor measurements are often highly correlated and Signal to Noise Ratio (SNR) is not homogeneously spread across sensors. Thus, collected data are highly variable and discrimination tasks are challenging. In this work, we focus on sensor weighting as an efficient tool to improve the classification procedure. We present an approach integrating sensor weighting in the classification framework. Sensor weights are considered as hyper-parameters to be learned by a Support Vector Machine (SVM). The resulting sensor weighting SVM (sw-SVM) is designed to satisfy a margin criterion, that is, the generalization error. Experimental studies on two data sets are presented, a P300 data set and an Error Related Potential (ErrP) data set. For the P300 data set (BCI competition III), for which a large number of trials are available, the sw-SVM proves to perform equivalently with respect to the ensemble SVM strategy that won the competition. For the ErrP data set, for which a small number of trials are available, sw-SVM shows superior performances as compared to three state-of-the art approaches. Results suggest that sw-SVM promises to be useful in event-related potentials classification, even with a small number of training trials.

Submitted to: *J. Neural Eng.*

1. Introduction

Brain-computer interfaces (BCI) are assistive technologies using brain signals to decode the **users'** intention without resorting to any muscles or peripheral nerves [1]. Some classes of BCI potentially provide motor-disabled people with a communication channel even when motricity is not preserved at all [2, 3]. More recently, BCI research has focused on improving/integrating traditional communication devices such as the keyboard and joystick, for example, in video-game applications [4, 5].

Because of its high temporal resolution, ease of use and low cost, most BCI are based on EEG (ElectroEncephaloGraphy). The EEG is a high dimensional (typically 8 to 128 sensors) scalp measurement of a smooth potential field. Whereas the potential field accurately reflects the global cerebral electrophysiological activity, the volume conduction, scalp smearing and the high spatial resolution of the sampling introduces a high correlation between the observed data at different electrodes (sensors) [6]. Moreover, the measured potentials are of low amplitude (of the order of tens of microvolts) and the measurements are very sensitive to noise of biological, environmental and instrumental origin. Such noise is of nonstationary nature and may vary considerably across sensors and along time. The poor Signal-to-Noise Ratio (SNR), which is an inherent characteristic of EEG, requires adequate processing techniques to tackle the problems of dimension reduction and noise cancelation. So far the BCI classification task has classically been solved in two steps: 1) feature extraction techniques, typically amounting to frequential, temporal and/or spatial filtering and 2) a machine learning classification task.

Concerning optimal sensor weighting or spatial filtering techniques, signal-processing criteria like the Signal-to-Noise Ratio and ratio of class variances [7, 8, 9, 10] have been often employed because of the instantaneous and approximately linear relation between the amplitudes of the generating cerebral electrophysiological current sources and the amplitude of the observed scalp potential field. The idea here is to find a linear transformation of the data (optimal spatial filters) optimizing the extraction of the relevant EEG feature and the noise suppression. The performance of such filters mainly depends on the accuracy of spatial covariance estimations and is jeopardized by the non-stationary nature of the noise. Although a relation might be found between the objective functions of [7, 8, 9, 10], yielding optimal filters, and class separability, this relation has, to the best of our knowledge, never been addressed explicitly.

Depending on the features to be extracted, some EEG sensors may not provide useful information, but, instead, add noise to the system. It is often the case for the most inferior temporal sensors (electrodes T3, T4, T5 and T6) of the international 10/20 system, which may convey more electromyographic data than EEG, due to steady or intermittent jaw contractions. In addition, temporal leads carry little information about sources generating evoked potentials such as the P300, thus for P300 detection they can usually be discarded. But, the leads affected by biological artifacts are subject- and session-dependent. For instance, some subjects tend to display more muscular

contamination on the frontal sensors (FP1, FP2) or on the occipital sensors (O1, O2) than on sensors covering the temporal region. Instrumental and environmental artifacts also may affect different leads, and again, this is subject- and session-dependent. It is thus crucial to derive data-driven criteria for sensor weighting.

Concerning the classification task, simple linear classifiers have been found to perform well in Event-Related Potential (ERP) paradigms [7, 11]. This has led to a prevailing view among BCI researchers that the effort to search for more sophisticated machine-learning approaches is irrelevant. Usually, preprocessed data are fed to a simple classifier borrowed from the machine learning literature without inquiring about possible improvements that could be done, thus resulting in classifiers that do not fully exploit the proprieties of the data. Nonetheless, a BCI is essentially a learning machine.

As mentioned above, in this work we focus on the optimal weighting of sensor data so as to improve the separability of the classes. We treat the problem within the classification problem itself. By introducing the sensor weighting as hyper-parameters in a Support Vector Machine (SVM), weights are optimized for the specific classification problem at hand. The SVM is particularly well-suited to online processing required for BCI data due to its reduced computational complexity. Indeed, SVM complexity depends far more on the number of training trials than on the number of features used to describe each one of them. The proposed algorithm has been named sensor weighting SVM (sw-SVM) and is built upon the Multiple Kernel Learning (MKL) framework [12, 13]. sw-SVM offers a very flexible approach, in that it can handle any kind of features, thus adapting to any kind of EEG-based BCI (P300, motor imagery, SSVEP, etc.) or any data selection and classification task. In this paper we focus on ERP data analysis, whereas in the discussion, we offer possible directions for its use in other contexts.

Two BCI data sets are considered to illustrate the efficiency of the proposed sw-SVM algorithm as compared to a state-of-the art SVM approach. The first is the P300 speller data set of the BCI competition III [14], for which the competition winner used an ensemble-SVM (e-SVM) approach [15]. An e-SVM constructs an ensemble of classifier decision functions on different subsets of the data and assigns a blind pattern according to the average of all decision functions. The BCI competition III data set has been chosen so as to provide a comparative element versus a state-of-the-art technique. The second data set is an Error-Related Potential (ErrP) data set and contains very few learning trials, thus no ensemble strategy is possible. Therefore, it provides an adequate base to ascertain the robustness of the proposed algorithm as compared to a spatial filter maximizing a ratio of class variances followed by an SVM classifier [16], to a spatial filter maximizing an SNR criterion followed by an SVM classifier [10] and to a classical SVM approach [17].

The remainder of this article is organized as follows. The proposed sw-SVM algorithm is introduced in section 2 where the general SVM framework is reminded. The sw-SVM optimization problem and a possible solution are presented. Section 3 accounts for the BCI data sets description and explains the preprocessing techniques

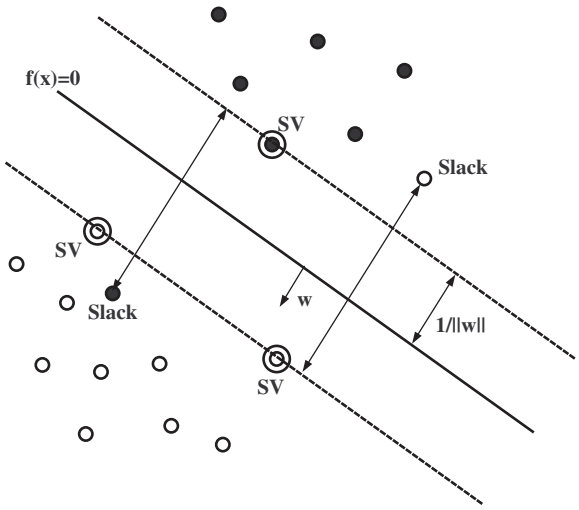

Figure 1. Schematic illustration of linear SVM. Slack variables ξ_p are observations for which classification errors are tolerated to improve generalization performance in non-linearly separable data sets. Circled points positioned on the dashed lines are called Support Vectors (SV).

used for each data set. Classification techniques used to compute comparative results are discussed and justified in Section 4. Finally, Section 5 holds our conclusions.

2. Method

In this section, the SVM primal and dual problems are firstly recalled. Secondly, the proposed sw-SVM method is detailed.

2.1. Support Vector Machine

The Support Vector Machine is a classification technique developed by Vapnik [17] which has shown to perform well in a number of real world problems, including BCI [18]. Given a set of labeled patterns $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_p, y_p), \ldots, (\mathbf{x}_P, y_P)\}$ with patterns $\mathbf{x}_p \in \mathbb{R}^d$ and labels $y_p \in \{-1, 1\}$ referring to two different classes. The central idea of SVM is to separate data by finding a hyperplane yielding the largest possible margin (a margin is the distance between nearest data points of different classes, as illustrated in Figure 1. Within this figure it is the distance between the two dashed lines.). This hyperplane is defined by a weight vector $\mathbf{w} \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$. Apart from being an intuitive idea, SVM has been shown to provide theoretical guaranties in terms of generalization ability [17].

One variant of binary linear SVM consists of solving the following primal

optimization problem:

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

where $\langle \cdot, \cdot \rangle$ stands for the inner product of two vectors. The parameters ξ_p are called slack variables and ensure that the problem has a solution in case the data is not linearly separable. The function $f(\mathbf{x}_p) = \langle \mathbf{w}, \mathbf{x}_p \rangle + b$, solution of problem (1), should correctly classify patterns along with minimizing $\|\mathbf{w}\|_2$. The trade-off between a low training error $\sum_{p=1}^{P}\xi_p$ and a large margin is controlled by the regularization parameter C. Finding a good value for C is part of the model selection procedure. If no prior knowledge is available, C has to be estimated from the training data, e.g., by using cross validation.

The dual problem of (1) can be formulated as follows :

$$\begin{aligned} &\max_{\alpha_1,\ldots,\alpha_P} \sum_{p=1}^{P}\alpha_p - \frac{1}{2}\sum_{p=1}^{P}\sum_{q=1}^{P}\alpha_p\alpha_q y_p y_q \langle \mathbf{x}_p, \mathbf{x}_q\rangle \\ &\text{subject to} \quad \sum_{p=1}^{P}\alpha_p y_p = 0 \\ &\text{and} \quad 0 \leq \alpha_p \leq C \quad \forall p \in \{1,\ldots,P\}. \end{aligned} \tag{2}$$

The linear SVM was extended to a non-linear classifier by applying the kernel trick [19] originally proposed by Aronszjan [20]. The space of possible functions $f(.)$ is now reduced to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with kernel function $K(.,.)$. Let $\phi : \mathbb{R}^d \to \mathcal{H}$ be the mapping defined over the input space. Let $\langle .,. \rangle_{\mathcal{H}}$ be a dot product defined in $\mathcal{H}$. The kernel $K(\cdot,\cdot)$ over $\mathbb{R}^d \times \mathbb{R}^d$ is defined by:

$$\forall\ (\mathbf{x}_p, \mathbf{x}_q) \in \mathbb{R}^d \times \mathbb{R}^d : \quad K(\mathbf{x}_p, \mathbf{x}_q) = \langle \phi(\mathbf{x}_p), \phi(\mathbf{x}_q)\rangle_{\mathcal{H}} \in \mathbb{R}$$

The resulting algorithm is formally similar to (2), except that every dot product is replaced by a non-linear kernel function $K(.,.)$. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation is generally non-linear and the transformed space high dimensional. Thus, though the classifier is a hyperplane in the RKHS, it is generally non-linear in the original input space. Some common kernels include Gaussian radial basis function, polynomial function, etc. For a detailed discussion please refer to [21].

2.2. Sensor Weighting procedure

The sw-SVM formulation involves sensor weights in the primal and dual optimization problem and tunes these weights as hyper-parameters of SVM. To illustrate the proposed method, let us consider time-locked evoked response potentials (ERP). Each ERP is considered in a short time period of T samples recorded over S sensors and represented as a matrix $\tilde{\mathbf{X}}_p \in \mathbb{R}^{T\times S}$. A pattern $\mathbf{x}_p$ is obtained by concatenating elements of $\tilde{\mathbf{X}}_p$

columnwise in a vector of $\mathbb{R}^{d\times 1}$, with $d = TS$. A trial $\mathbf{x}_p$ is thus a vector containing all the spatio-temporal information.

Our task consists in finding spatial weights that maximize the separation margin between two post-stimulus responses recorded on a given subject. **We assume that sensor weights for a given subject are similar across all the trials.** Thus, we aim at finding a matrix $\mathbf{D} \in \mathbb{R}^{d\times d}$ of sensor weights assigned to each of the trials $\mathbf{x}_p$ so that $\{\mathbf{D}\mathbf{x}_p\}_{p=1}^{P}$ maximize the margin of the SVM. For the application of EEG sensor weighting, time features belonging to a same EEG sensor, hereafter indexed by s, have to be dealt with in a congeneric way so that a spatial interpretation remains possible. In this work, time samples of one sensor are treated equally. The resulting matrix $\mathbf{D}$ is thus diagonal with S different unknown coefficients, each coefficient d_s is being repeated T times on the diagonal as:

$$\mathbf{D} = \begin{pmatrix} d_1\mathbf{I}_T & 0 & \cdots & 0 \\ 0 & d_2\mathbf{I}_T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_S\mathbf{I}_T \end{pmatrix}$$

where $\mathbf{I}_T$ is the identity matrix in $\mathbb{R}^{T\times T}$ and d_s are coefficients that weigh the sensors. From this context, our objective is to find the coefficients d_s that maximize the margin of a linear SVM classifier. In this sense, we are providing a method for large-margin sensor weighting.

According to the SVM definition given above, the optimization problem of the linear SVM sensor weighting problem can be stated as:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi,\mathbf{D}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{p=1}^{P}\xi_p \\ &\text{subject to} \quad y_p(\langle \mathbf{w}, \mathbf{D}\mathbf{x}_p\rangle + b) \geq 1 - \xi_p \quad \forall p \in \{1,\dots,P\} \\ &\text{and} \quad \xi_p \geq 0 \quad \forall p \in \{1,\dots,P\} \\ &\text{and} \quad \sum_{s=1}^{S} d_s^2 = 1. \end{aligned} \tag{3}$$

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

$$\begin{aligned} &\min_{\tilde{\mathbf{D}}} \max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Y}^T\mathbf{X}^T\tilde{\mathbf{D}}\mathbf{X}\mathbf{Y}\boldsymbol{\alpha} \\ &\text{subject to} \quad \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ &\text{and} \quad 0 \leq \alpha_p \leq C \quad \forall p \in \{1,\dots,P\} \\ &\text{and} \quad \sum_{s=1}^{S} \tilde{d}_s = 1, \end{aligned} \tag{4}$$

where we have used $\tilde{\mathbf{D}} = \mathbf{D}^T\mathbf{D}$ and thus $\tilde{d}_s = d_s^2$. $\boldsymbol{\alpha}$ is the vector of Lagrangian multipliers, $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_P\}$ is the matrix containing the trials, $\mathbf{y} = \{y_1, ..., y_P\}$

is the vector containing the labels and $\mathbf{Y} = \mathrm{Diag}(\mathbf{y})$ is a diagonal matrix containing the labels on its diagonal. The overall problem remains a concave problem in α and boils down to a Multiple Kernel Learning (MKL) problem where a linear kernel is used over each sensor's time series. $\{\tilde{d}_s\}$ are the positive mixing coefficients associated with the multiple kernels. According to this relationship, we propose to use a MKL algorithm based on a reduced gradient method, as in SimpleMKL [22], for solving the problem.
We proceed with an alternate optimization algorithm. For any admissible value of $\tilde{\mathbf{D}}$, the maximization problem over α is strictly concave. Noteworthy, for an admissible value of $\tilde{\mathbf{D}}$, the objective function reduces to a regular SVM optimization. Hence, we can use any SVM solver to find α for once $\tilde{\mathbf{D}}$ has been fixed [23]. For the so obtained α, the minimization problem over $\tilde{\mathbf{D}}$ is smooth and convex [24]. Hence, we can use a reduced gradient method which converges for such functions [25]. Once the gradient of the first equation in (4) is computed, $\tilde{\mathbf{D}}$ is updated by using a descent direction ensuring that the equality constraint and the non-negativity constraints on $\{\tilde{d}_s\}$ are satisfied. These two steps are iterated until a stopping criterion is reached. The stopping criterion we chose is based on a norm variation of the sensor weights.

3. Experimental Data

Experiments were performed on a P300 data set and an Error-related Potential (ErrP) data set. Experimental set up, preprocessing techniques and notations are detailed in this section.

3.1. The P300 speller data set

The P300 speller data set from the BCI competitions 2004 [14] was used to benchmark the proposed filtering algorithm and to compare it to the competition winner, where an ensemble SVM approach clearly outperformed the competitors [15]. A P300 speller paradigm allows the user to choose a character among a predefined set of alphanumeric characters [26] **(letters from A to Z, digits from 1 to 9 and _). A 6×6 matrix of characters is presented to the user and the rows and columns of the matrix are flashed (intensified) in random order. The user can select a character by concentrating on it. Since the target character is rare as compared to the others, a P300 evoked response is elicited when the target flashes. The task of the P300 speller is to guess what target the subject focuses upon by comparing responses evoked by each row/column intersection. The P300 potential is in the order of a few microvolts highly corrupted by noise and superimposed on background activity of significantly higher amplitude (as an integration over multiple ongoing activities). Thus, in order to obtain**

sufficient accuracy, the sequence of flashes must be repeated several times for each character to be spelled, typically 8 **to** 15 **times and responses should be averaged to reduce noise and enhance the signal of interest.**

3.1.1. Experimental setup and mental task EEG signals were recorded from two subjects using a 64 ear lobe-referenced scalp electrodes. Before digitization at a sample rate of 240 Hz, signals have been bandpass-filtered from $0.1 - 60$ Hz. A detailed description of the data set can be found in the BCI competition paper [27]. For each subject, the training set is composed of 85 characters and the test set of 100 characters. One spelled character corresponds to 180 post-stimulus labeled signals (12 row/column intensifications $\times$15 repetitions per letter). Only 30 post-stimuli from the 180 correspond to a target intensification yielding a P300 deflection.

Five sessions were recorded for each subject. Each session consisted of a number of runs where subjects focused attention on a series of characters. For each spelled character the matrix was displayed for a 2.5 **s period during which each character had the same intensity. This period informed the user that the previous character spelling was completed and gave instruction to focus on the next character in the word, which was displayed on the top of the screen. Subsequently, each row and column in the matrix was randomly intensified for** 100 **ms alternating with a blank period of** 75 **ms. Row/column intensifications were block randomized in blocks of** 12**.**

3.1.2. Data preprocessing In "oddball" paradigms such as the one described above, the perception of the rare stimulus typically triggers a positive low amplitude deflection approximately 300 ms following the stimulus onset, **also known as the P3b component of the P300 waveform [28, 29].** Consequently, only time-window of approximately 667 ms post stimulus onset, corresponding to 160 time samples, were retained. Before submitting the data to the feature extraction and learning algorithms the data were band-pass filtered between 0.1 Hz and 20 Hz with a 4^{th} order Tchebychev filter (type 1) and then decimated so as to retain 14 samples per sensor for each trial. **Prior to decimation, the signal is filtered with an** 8^{th} **order Chebyshev Type I low pass filter. This acts as an anti-aliasing filter suppressing frequency contents above** $\frac{0.8F_S}{2f_d}$**, where** f_d **is the decimation factor (here** $f_d = 12$**). Secondly, we downsample the so obtained signal to a sample frequency of** $\frac{F_S}{f_d}$**, retaining each** 12**th sample (**14 **samples in total).** Thus, the dimensionality of the input vector is 14×64. Let P denote the number of stimuli of the data sets ($P = 15300 = 12$ intensifications $\times$15 repetitions $\times$85 characters for the training set and $P = 18000$ for the test set) and let d denote the data dimension ($d = 14 \times 64 = 896$ for each stimulus). A trial is denoted as $\mathbf{x}_p \in \mathbb{R}^d$, $p = 1 \dots P$, with labels $y_p \in \{-1, 1\}$. A label $y = 1$ corresponds to an expected P300 post-stimulus signal and $y = -1$ corresponds to an expected absence of a post-stimulus P300 signal.

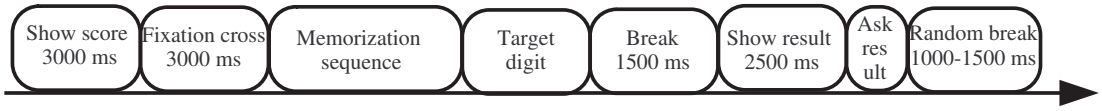

Figure 2. Temporal diagram of one ErrP trial.

3.2. The Error Related Potential data set

In 1991, Falkenstein et al. [30] reported the presence of a negative deflection in the EEG when subjects committed errors in a time-reaction task. Since then, several studies have shown the presence of Error related Potentials (ErrP) components such as error related negativity (ERN or Ne) and error-related positivity (Pe) in a variety of experimental paradigms. Error processing systems were categorized in timed reaction tasks [31, 32], feedback tasks indicating incorrect performance after a decision task [33, 34] or observation tasks following observation of errors made by an interface or someone else [35].

The experiment, described in the following, is based on a visual feedback presented on a computer screen following a memorization task.

3.2.1. Experimental setup and mental task Eight healthy volunteers (including three women) participated in this experiment. All subjects were BCI-naive at the time of the experiment. Subjects had to retain the position of an ensemble of two to nine digits. The digits were displayed as a sequence in square boxes and evenly distributed along a circle. When the sequence disappeared, a target digit was shown and subjects were asked to click on the box where it previously appeared. A visual feedback indicates whether the answer was wrong or correct. The experiment involved two sessions that lasted together approximately half an hour. Each session consisted of six blocks of six trials, for a total of $6 \times 6 \times 2 = 72$ trials.

The temporal order of each trial, illustrated in Figure 2, is detailed next. The score, initially zero, was displayed for 3000 ms followed by a fixation cross, which was in turn displayed for 3000 ms. Then the memorization sequence started with variable duration depending on the number of digits the subject had to memorize. When it ended the subject was asked to click on the box where the target digit had appeared. Once the subject had answered the interface was paused for 1500 ms and then turned the clicked box into green upon a correct answer or into red upon an erroneous answer. This feedback lasted for 2500 ms. The 1500 ms preceding the feedback was introduced to avoid any contamination of ErrP by beta rebound motor phenomena linked to mouse clicking [36]. The subject was then asked to report if the feedback (correct/error) matched his expectation by a mouse click ("yes"/"no"). Following his answer a random break of 1000 ms to 1500 ms preceded the beginning of the new trial. The number of digits was adapted with an algorithm tuned to allow about 20% errors for all subjects. The mean error rate (standard deviation) was equal to $17.87(\pm 4.64)\%$ of the trials.

Recordings of the EEG were made using 31 **sensors from the extended** 10/20 **system. Both earlobes were used as electrical reference. Connection between sensors are performed digitally by the Mitsar** 202 **DC EEG acquisition software. The ground sensor was positioned on the forehead. During acquisition, EEG was band-pass filtered in the range** $0.1 - 70$ **Hz and digitized at** 500 **Hz.**

3.2.2. Data preprocessing Raw EEG potentials were first re-referenced to the common average by subtracting from each sensor the average potential (over the 31 sensors) for each time sample. Many studies report two peaks, Ne and Pe, as the main components of Error Related Potential components [33]. Ne shows up about 250 ms after the response as a sharp negative peak and Pe shows up about 300 to 500 ms after the response as a broader positive peak. According to these knowledge, only a window of 1000 ms posterior to the stimulus has been considered for each trial, which results in 500 samples per sensor. A $1 - 10$ Hz 4^{th} order Butterworth filter was applied as error related potentials are known to be a relatively slow cortical potential. **Finally, EEG signals were decimated so as to retain** 16 **samples, with the same process as mentioned above (anti-aliasing filter followed by a decimation with factor** 32**).** Thus the dimensionality of the input vector is 16×31. No artifact rejection algorithm was applied and all trials were kept for analysis. Let P denote the number of training vectors (trials) of the data sets ($P = 72$ for all 8 data sets) and let d denote the data dimension ($d = 16 \times 31 = 496$ for all 8 data sets). A trial is denoted as $\mathbf{x}_p \in \mathbb{R}^d$, $p = 1 \dots P$, with labels $y_p \in \{-1, 1\}$. For the task used in this paper $y = 1$ denotes error trial, $y = -1$ denotes correct one.

3.3. Cross Validation

For the P300 data set, an ensemble of linear sw-SVM classifiers was used to make results comparable to those of the competition winner [15] where an ensemble of SVM (e-SVM) was learned. Data set was split, as per the competition winner method [15], into 17 subsets, each one composed of five characters or $5 \times 12 \times j$ post-stimuli, with j the number of repetitions for one character. An ensemble of classifiers system for each single subject was designed. For the ensemble of sw-SVM, 17 sw-SVM classifiers were trained for $j = 15$ repetitions on one of the 17 subsets and its regularization parameter C was chosen by validating performances on the remaining 16 subsets. To assign a test data to one of the 36 classes, 17 real-valued sw-SVM decision functions were computed for each $j = 1 \dots 15$, the most probable row and column at the j^{th} repetition was the one that maximizes the average of the 17 sw-SVM classifiers scores. **For the e-SVM,** 17 **linear SVM classifiers with backward elimination technique were trained for** $j = 15$ **repetitions on one of the** 17 **subsets and its regularization parameter** C **was chosen by validating performances on the remaining** 16 **subsets. To assign a test data to one of the** 36 **classes,** 17 **real-valued SVM decision functions with**

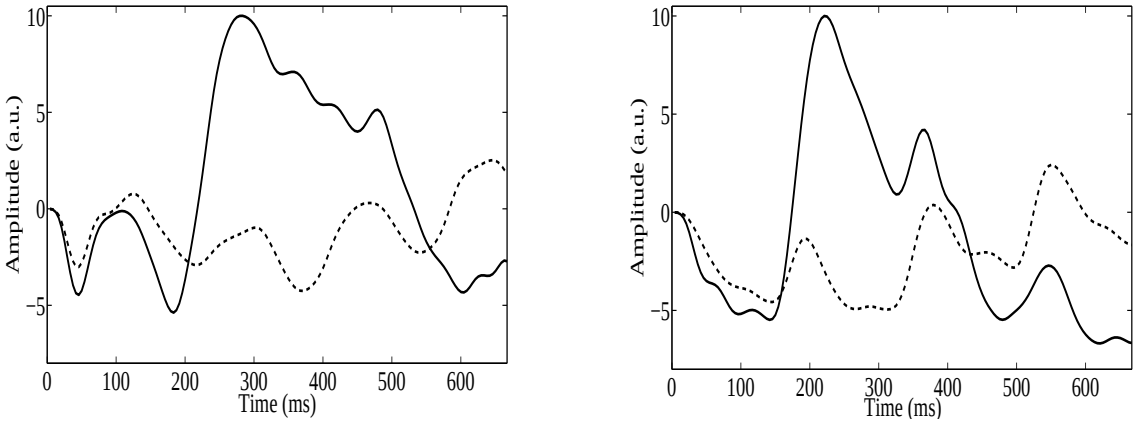

Figure 3. P300 component (solid) and non-P300 component (dotted) are illustrated for subjects A (left) and B (right). Components are the projection of the original post-stimuli (with 160 samples per post-stimulus window) according to the sw-SVM filters computed on down-sampled post-stimuli (with 14 samples per post-stimulus window).

the selected sensors were computed for each $j = 1 \dots 15$ repetitions, the most probable row and column at the j^{th} repetition was the one that maximizes the average of the 17 **SVM classifiers scores.** For the Error related Potential data set, only five subsets were considered because of the limited number of trials. An sw-SVM classifier was learned on four subsets with different regularization parameter C and performances were computed on the remaining subset. For a set of pre-defined values of C, this process was repeated five times for a given subject and averaged. The highest average accuracy was reported. Besides, five cross validation results with SVM preceded by the optimal filter obtained by xDAWN [10], SVM preceded by optimal spatial filter as proposed by Hoffmann et al. [16] and baseline SVM without previous spatial filtering, were reported for comparison.‡ It is noteworthy that for the ErrP data set, a non-linear SVM with second degree polynomial kernel was used.

4. Results

4.1. P300 experimental results

In this experiment, an ensemble of sw-SVM is compared to e-SVM [15]. For each single classifier built on one of the 17 subsets, sensor weighting has been performed based on the training set $A.k$ or $B.k$ ($k = 1 \dots 17$) and the related validation set. sw-SVM can be considered as a one-component spatial filter, and thus, a unique linear combination of sensor measurement can be computed. Figure 3 shows the average of the weighted potential for common (non-P300 in dotted line) and rare (P300 in solid line) signals for

‡ **Although a one-level cross validation performance may give optimistic results, we did not opt for a two-level cross validation performance because of the data set properties (few trials, high variability, highly unbalanced classes).**

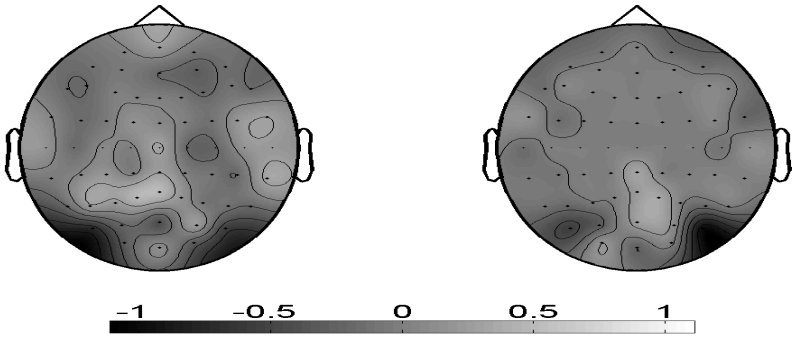

Figure 4. Topographies of sw-SVM weights for subjects A (left) and B (right).

a random subset of data, referring to subjects A (left) and B (right). For both subjects, a positive deflection in voltage with a latency of roughly 300 to 600 ms can be clearly identified. This analysis suggests that the sw-SVM provides an efficient spatial filter.

4.1.1. Sensor weighting results Sensors PO7, PO8, Pz and CPz receive consistently the highest weights, which is in line with our expectation. As compared to [15], where some frontal sensors were top ranked, the weight analysis of sw-SVM is more consistent to the midline central generation of the P300. Also, as expected, weighting varies considerably from one classifier to another and from subject to subject.

Typical topographies of sw-SVM weights are given for subjects A and B in Figure 4. These maps are in line with previous findings in P300 research (e.g., [37, 38, 26]) and confirm the ability of sw-SVM to weight sensors in such a way as to extract relevant information about the P300.

4.1.2. Classification results The character recognition rate (in %) is presented for several number of repetitions for both subject A and B in Figure 5. They are compared to the winner results of BCI competition III [15] (e-SVM) and the classical single SVM treating all sensors homogeneously.

Figure 5 illustrates that sw-SVM performs at least as good as e-SVM and simple SVM without sensor weighting, especially for small number of repetitions (less than seven repetitions for subject A and five repetitions for subject B). For 15 **repetitions all three strategies give similar results. We conclude that sw-SVM is well suited for noisy data where small number of trials is available. It enables to reduce the number of required stimulus repetitions and consequently boosts the information transfer rate.**

The e-SVM [15] witnesses in favor of the good performance of ensemble classifier averaging methods. Therefore, an ensemble of sw-SVM was used. We are of the opinion that it is interesting to reveal the advantages of sensor weighting by using a single sw-SVM.

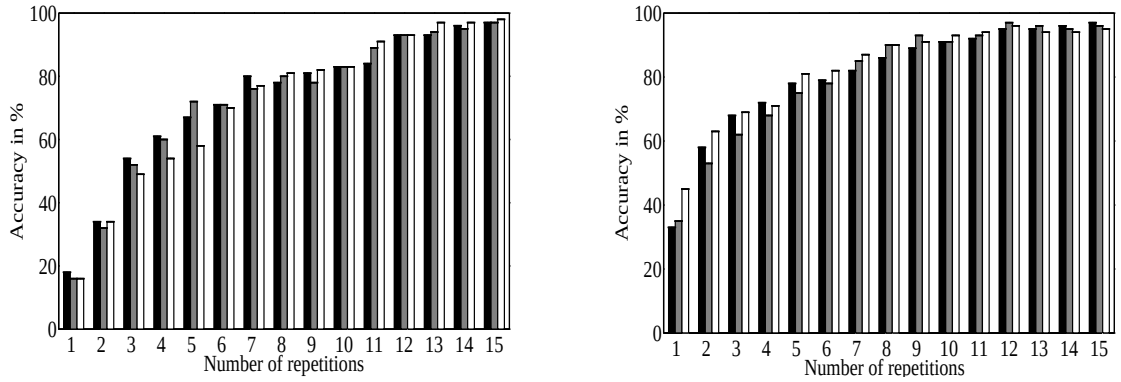

Figure 5. Percentage of correctly recognized characters for subject *A* (left) and *B* (right) for different number of repetitions. sw-SVM (black bar) results are compared with the winner results of BCI competition III [15] (e-SVM, gray bar) and a single SVM classifier (white bar).

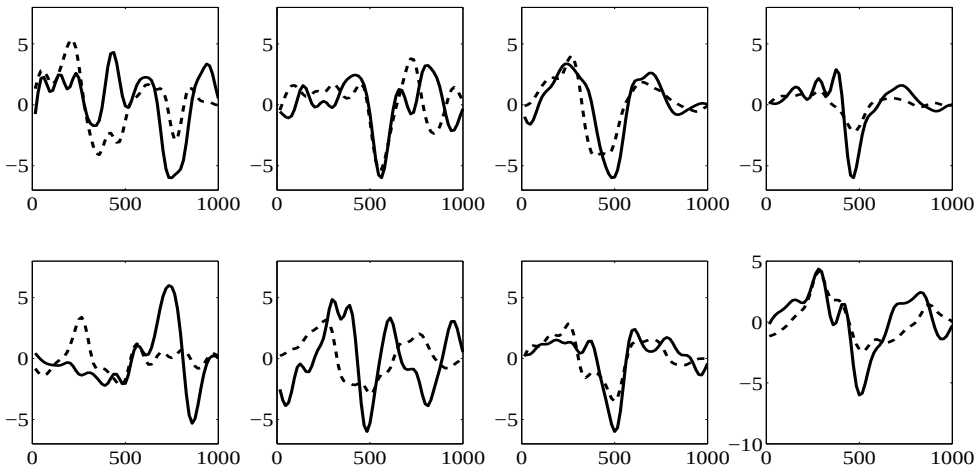

Figure 6. Average of error (solid line) and correct (dotted line) trials on channel FCz for the 8 subjects.

4.2. Error Related Potential experimental results

For ErrP data sets, the number of available trials is very low. Hence, it was not possible to use ensemble SVM strategy. Moreover, no artifact rejection whatsoever is carried out, making it a challenging classification task. The main goal of this experiment is to test robustness of sw-SVM to EEG waves of various nature and to validate the performance of a simple sw-SVM on raw data set. Eight subjects are considered to test the ability of sw-SVM to adapt to inter-subject variability. We compare the performance of sw-SVM classifier, in terms of classification accuracy, against an SVM classifying spatially filtered data as proposed by Rivet et al. [10] (xDAWN+SVM) or by Hoffmann et al.[16] (Hoff+SVM), and a baseline SVM classifier without any spatial filtering or sensor selection procedure.

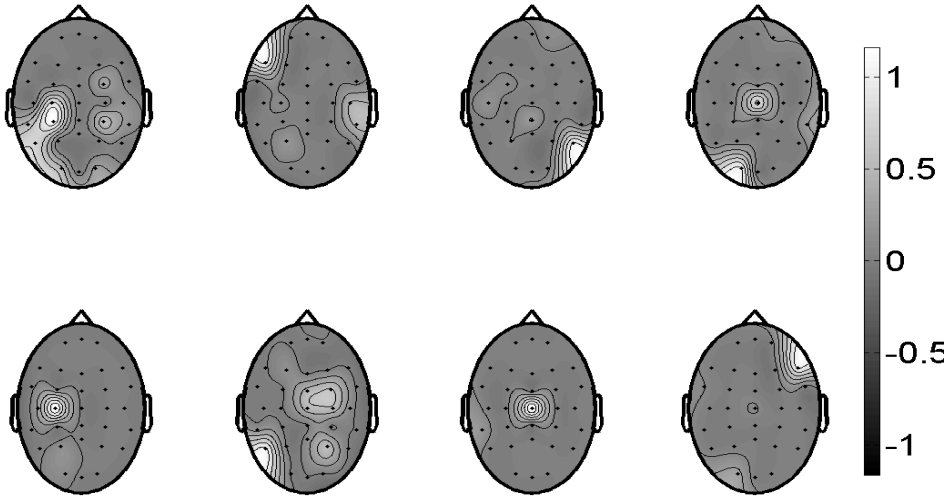

Figure 7. Topographical maps of the weights averaged across the five subsets for the ErrP dataset. Each map refers to a subject. The large variability is caused by the low number of trials.

Figure 6 shows the averages of error and correct trials for sensor FCz. As expected, and in accordance with [39], a negative deflection (Ne) can be seen after the feedback for error trials followed by a positive one for almost all the subjects. Latency and amplitudes are very different from subject to subject. Inter-subject differences are large in this data set due to the small number of trials available for averaging. For some subjects, like subjects $S3$ and $S5$, Ne and Pe do not clearly appear on sensor FCz. For subjects $S2$ and $S3$, there is no major difference to be noticed between potentials generated for correct and incorrect trials, as recorded on sensor FCz. As a consequence, the classification task promises to be hard for this data set.

4.2.1. sw-SVM sensor weighting results Results elucidate clearly the sparsity promoted by sw-SVM. Figure 7 shows the sensor weights found by sw-SVM averaged across the five subsets as topographic maps. For six out of the eight subjects central area holds the strongest weights, which is in accordance with current knowledge on error related potentials. Indeed several studies cite the Anterior Cingulate Cortex (ACC, Brodmann areas 24&32) as the main source responsible for the generation of the Ne [40, 41]. For subject $S7$, sensor Cz captures almost all necessary information. It is also remarkable that for subjects $S2$ and $S8$ weights do not have a medial central or fronto-central distributions, but rather a fronto lateral distribution, a fact that can be put in relation with studies pointing to the lateral prefrontal cortex as another possible generator of the Ne [42].

4.2.2. Comparison of classification results Figure 8 reports the recognition rates (mean and standard deviations) for the eight subjects and their average, obtained by four filtering/classification algorithms (sw-SVM, xDAWN+SVM, Hoff+SVM, baseline

SVM). Interestingly, sw-SVM shows classification accuracies between 74% and 91%, averaging to about 81%. These figures have been achieved with a relatively low number of sensors (from 1 to 11 sensors). Noteworthy, available data include a small number of trials (only 72 trails are available in total) with a small number of errors (for instance, only 10 error trials were available for subject $S4$). Thus, as expected, the cross-validation variance is elevated. Since no artifact rejection was applied as pre-processing, our results refer to a realistic situation of BCI use.

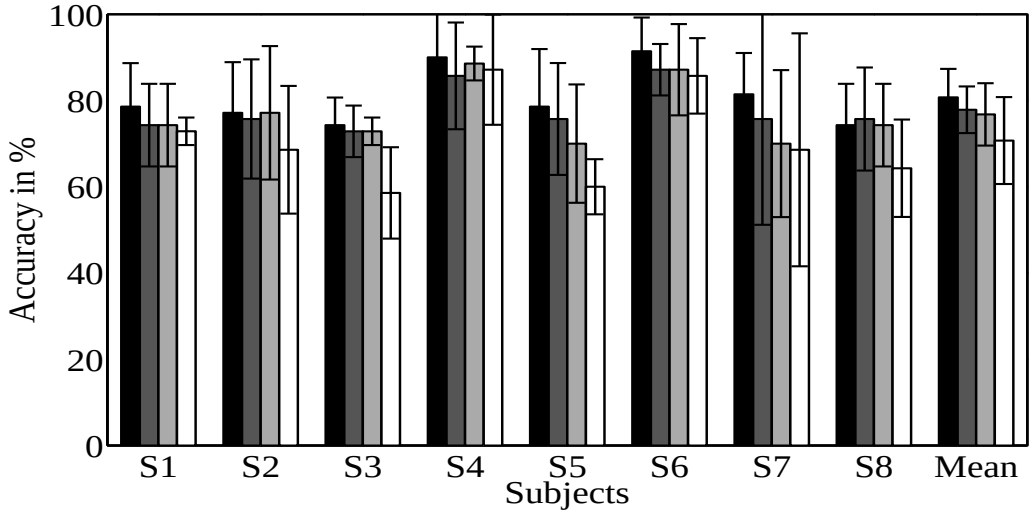

Figure 8. Accuracies (standard deviation) of the 8 subjects and their average for sw-SVM(black bar), xDAWN+SVM (dark gray bar), Hoff+SVM (light gray bar) and baseline SVM (white bar).

Mean (standard deviation) accuracies across the 8 subjects were 80.71(±6.61)% for sw-SVM, 77.85(±5.54)% for xDAWN+SVM, 76.78(±7.23)% for Hoff+SVM and 70.71(±10.77)% for baseline SVM. Three repeated-measure t-tests have been performed to test the null hypothesis of no difference in the performance of sw-SVM against xDAWN+SVM, Hoff+SVM and SVM. Pairwise comparison of means reveals that sw-SVM proves significantly and constantly superior to the other three, methods (swSVM vs. xDAWN+SVM : $t(7) = 3.5362$, p-value = 0.0095; swSVM vs. Hoff+SVM : $t(7) = 2.6720$, p-value = 0.0319 and swSVM vs. SVM : $t(7) = 5.2389$, p-value = 0.0012).

As a comparison with previous single-trial ErrP classification studies, our results are competitive in terms of accuracy. For instance, Ferrez et al. [43] reported an average detection rate of 76.2(±4.6)% for error and 81.8(±3.5)% for correct trials. Six subjects participated in their study, they used 64 sensors and 1500 trials (1125 to train the classifiers and 375 to test them) with 20% of erroneous trials, see also [44, 45]). It is important to note that for the above mentioned studies, training data sets are much larger than our sets. Considering previous studies on error-related potentials the level of performance already achieved on these small data sets appears very promising.

4.2.3. Discussion In this data set, the sw-SVM approach yields both a significant dimensionality reduction and a considerable performance improvement. sw-SVM has only one degree of freedom inherent to the regularization parameter of SVM whilst xDAWN and Hoffmann filters require, along with tuning classifier parameters, an estimation of the number of spatial components providing the highest accuracy (model order selection). For the xDAWN algorithm, a recent study [38] proposed a strategy to find the number of filters yielding the optimal classification performance. But sw-SVM extracts the component that directly optimizes the classification, while xDAWN and Hoffmann filters optimize criteria that are not explicitly related to the classification accuracy. This may lead to a suboptimal solution from the point of view of the classifier. Another reason advocating for sw-SVM is the fact that it is well suited for situations wherein one has high dimensional data with a rather limited number of trials. In these conditions, Hoffmann's method may lead to poor performance. Indeed, since it is based on an empirical estimation of scatter matrices, a proper regularization technique to handle situations of high dimensionality and small data size is needed [16, 7].

4.3. Discussion

Our experimental investigations suggest that the sw-SVM localizes relevant information from a physiological and a classification point of view and it reliably classifies ERPs. Its particularity lies in the ability to select a relatively small proportion of sensors bearing useful information, while optimally weighting them. For EEG data, in which large numbers of trials are difficult to collect, and in which each trial may contain many thousands of sample points across dozens of sensors, this is a considerable advantage.

sw-SVM is a completely data driven method not imposing any assumption regarding EEG dynamics. It appears to be a flexible technique that can be directly used in various BCI scenarios. **This flexibility is mainly due to the criterion used to weight sensors, which consists in maximizing the SVM margin (and thus the ability to generalize from the examples) and the fact that one may populate the input feature vector $\mathbf{x}_p$ with any kind of features, such as amplitude, power, coherence, etc.**

SVM margin was recently introduced as a criterion for multi-modal data filtering [46] and has proven good performance. Future work may look simultaneously to the best "spatial" configuration, as in the current paper, the best "spectral" configuration, as in [46] and best temporal configuration as well. Such attempt would consider all relevant aspects of EEG dynamics to find the best margin SVM. Such extension would also allow to exploit more than one EEG source. For instance, in ErrP data it would be possible to separate and exploit the source of Ne and Pe components.

5. Conclusion

In this paper, we have considered the problem of sensor weighting from a machine learning point of view. Sensor weights are introduced in the SVM theoretical framework and tuned as hyper-parameters of SVM. They maximize the margin between classes and optimize classification performances. The proposed sensor weighting SVM (sw-SVM) involves spatial filtering along with classification in one optimization step. Unlike usual spatial filter techniques that do not directly optimize a discrimination criterion, sw-SVM helps locating sensors which are relevant for optimal classification performance.

Experimental data on P300 and ErrP data sets illustrate the efficiency of the proposed approach. Our algorithm performs well in experimental situations as well in terms of spatial distribution as in terms of classification accuracy. For the P300 data set (BCI competition III) the sw-SVM proves equivalent performance with respect to the strategy that won the competition. For the ErrP data set, sw-SVM shows competitive performance as compared to three state-of-the art approaches (spatially filtered data using xDAWN followed by SVM, spatially filtered data using Hoffmann's method followed by SVM and baseline SVM).

We believe that sw-SVM is a promising tool for data classification that could perform well on a large variety of EEG data types, even with a small number of training trials. Besides, sw-SVM is a completely data driven strategy.

Simulations and experiments yield encouraging results motivating further research. sw-SVM may be further extended toward a spatial-temporal-spectral filtering SVM that can provide a comprehensive modeling of brain post-stimulus dynamics recorded in an EEG. It is also possible to apply sw-SVM on input data populated with various kinds of EEG feature, such as those extracted from event-related synchronization/desynchronization, etc. **In this case, weights to be found will be considered non-negative and as such the outputs will be a non negative weighting of signal power (energy). All other aspects of the method would remain unchanged.**

Acknowledgments

This research has been partially supported by ANR (Agence Nationale de la Recherche) Project Open-ViBE2, RoBIK, GAZE & EEG, and INRIA ARC MaBi.

References

[1] Wolpaw J R, Birbaumer N, McFarl D J, Pfurtscheller G, and Vaughan T M. Brain-computer interfaces for communication and control. *Clin. Neurophy.*, 113:767–791, 2002.

[2] Birbaumer N, Ghanayim N, Hinterberger T, Iversen I, Kotchoubey B, Kübler A, Perelmouter J, Taub E, and Flor H. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

[3] Hinterberger T, Kaiser J, Kübler A, Neumann N, and Birbaumer N. The Thought Translation Device and its Applications to the Completely Paralyzed. In Diebner Druckrey and Weibel, editors, *Sciences of the Interfaces*, pages 232–240. Genista-Verlag, Tübingen, 2001.

[4] Plass-Oude Bos D, Reuderink B, Laar van de B, Gürkök H, Mühl C, Poel M, Nijholt A, and Heylen D K J. Brain-Computer Interfacing and Games. In D. Tan and A. Nijholt, editors, *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*, Human-Computer Interaction Series, pages 149–178. Springer Verlag, London, July 2010.

[5] Mühl C, Gürkök H, Plass-Oude Bos D, Thurlings M E, Scherffig L, Duvinage M, Elbakyan A A, Kang S, Poel M, and Heylen D K J. Bacteria Hunt: A multimodal, multiparadigm BCI game. In *Fifth International Summer Workshop on Multimodal Interfaces*, pages 41–62, Genua, 2010. University of Genua.

[6] Nunez P L and Srinivasan R. *Electric Field of the Brain*. New York: Oxford Univ Press, 2nd edition, 2006.

[7] Blankertz B, Lemm S, Treder M, Haufe S, and Müller K. Single-trial analysis and classification of ERP components – a tutorial. *NeuroImage*, 56(2):814–825, 2010.

[8] Müller-Gerking J, Pfurtscheller G, and Flyvbjerg H. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clin. Neurophys.*, 110:787–798, 1999.

[9] Gouy-Pailler C, Congedo M, Brunner C, Jutten C, and Pfurtscheller G. Nonstationary brain source separation for multiclass motor imagery. *IEEE Trans. on Biomedical Engineering*, 57(2):469–478, 2010.

[10] Rivet B, Souloumiac A, Attina V, and Gibert G. xDAWN algorithm to enhance evoked potentials: Application to brain computer interface. *IEEE Trans Biomed Eng*, 56(8):2035–2043, 2009.

[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:R1–R13, 2007.

[12] Lanckriet G, De Bie T, Cristianini N, Jordan M, and Noble W. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, November 2004.

[13] Lanckriet G, Cristianini N, Bartlett P., El Ghaoui L., and Jordan M. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.

[14] BCI Competition Website: http://www.bbci.de/competition/iii/.

[15] Rakotomamonjy A and Guigue V. BCI Competition III : Dataset II - Ensemble of SVMs for BCI P300 speller. *IEEE Trans. Biomedical Engineering*, 55(3):1147–1154, 2008.

[16] Hoffmann U, Vesin J, and Ebrahimi T. Spatial filters for the classification of event-related potentials. In *Proceedings of ESANN 2006*, pages 47–52, 2006.

[17] Vapnik V. *Statistical learning theory*. Wiley, 1998.

[18] Blankertz B, Curio G, and Müller K. Classifying single trial EEG: Towards brain computer interfacing. In Diettrich T G, Becker S, and Ghahramani Z, editors, *Advances in Neural Inf. Proc. Systems (NIPS 01)*, volume 14, pages 157–164, 2002.

[19] Aizerman M, Braverman E, and Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[20] Aronszajn N. Theory of reproducing kernels. *Trans. American Math. Soc.*, 68:337–404, 1950.

[21] Schölkopf B and Smola A J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

[22] Rakotomamonjy A, Bach F, Canu S, and Grandvalet Y. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[23] Canu S, Grandvalet Y, Guigue V, and Rakotomamonjy A. SVM and Kernel Methods Matlab Toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[24] Lemaréchal C and Sagastizábal C. Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *SIAM J. on Optimization*, 7:367–385, February 1997.

[25] Luenberger D. *Linear and Nonlinear Programming*. Addison-Wesley, 1984.

[26] Farwell L A and Donchin E. Taking Off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-Related Potentials. *Electroencephalogr. Clin. Neurophysiol*, 70:510–523, 1988.

[27] Blankertz B, Müller K R, Krusienski D J, Schalk G, Wolpaw J R, Schlögl A, Pfurtscheller G, Millán J R, Schröder M, and Birbaumer N. The BCI competition III: Validating alternative approaches to actual BCI problems. *IEEE Trans. on Neural Systems and Rehabilitation*

Engineering, 14:153–159, 2006.

[28] Sutton S, Braren M, Zubin J, and John E R. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–8, 1965.

[29] Polich J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118:2128–2148, 2007.

[30] Falkenstein M, Hohnsbein J, Hoormann J, and Blanke L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalogr. Clin. Neurophysiol.*, 78:447–455, 1991.

[31] Falkenstein M, Hohnsbein J, Hoormann J, and Blanke L. Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In Gaillard A W K In: Brunia C H M and Kok A, editors, *Psychophysiological Brain Research*, pages 192–195. Tilburg Univesity Press, Tilburg, 1990.

[32] Falkenstein M, Hoormann J, Christ S, and Hohnsbein J. ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, 51:87–107, 2000.

[33] Miltner W H R, Braun C H, and Coles M G H. 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, 1997.

[34] Dal Seno B, Matteucci M, and Mainardi L. Online detection of P300 and error potentials in a BCI speller. *Intell. Neuroscience*, 2010, January 2010.

[35] van Schie H T, Mars R B, Coles MG H, and Bekkering H. Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, 7:549–554, 2004.

[36] Jurkiewicz M, Gaetz W, Bostan A, and Cheyne D. Post-movement beta rebound is generated in motor cortex: Evidence from neuromagnetic recordings. *NeuroImage*, 32(3):1281–1289, 2006.

[37] Cecotti H, Rivet B, Congedo M, Jutten C, Bertrand O, Mattout J, and Maby E. Suboptimal sensor subset evaluation in a P300 Brain-Computer Interface. In *Proc. European Signal Processing Conf. (EUSIPCO)*, pages 924–928, Denmark, 2010.

[38] Rivet B, Cecotti H, Phlypo R, Bertrand O, Maby E, and Mattout J. EEG sensor selection by sparse spatial filtering in P300 speller brain-computer interface. In *Proc. Int. Conf. IEEE Engineering in Medicine and Biology Society (IEEE EMBC)*, pages 5379–5382, Buenos Aires, Argentina, 2010.

[39] Holroyd C B and Coles M G H. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109:679–709, 2002.

[40] Carter C S, Braver T S, Barch D M, Botvinick M M, Noll D, and Cohen J D. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280:747–749, 1998.

[41] Dehaene S, Posner M, and Tucker D. Localization of a neural system for error detection and compensation. *Psychol. Sci.*, 5:303–305, 1994.

[42] Kiehl K, Liddle P, and Hopfinger J. Error processing and the rostral anterior cingulate: an event-related fMRI study. *Psychophysiology*, 37:216–223, 2000.

[43] Ferrez P, des Prés-Beudin A, and Millán J. EEG-based brain-computer interaction: Improved accuracy by automatic single-trial error detection. In *21st Annual Conference on Neural Information Processing Systems (NIPS)*, volume 20, pages 113–121, 2007.

[44] Bernstein P S, Scheffers M K, and Coles M G H. Where did I go wrong? A psychophysiological analysis of error detection. *Journal of Experimental Psychology: Human Perception and Performance*, 21:1312–1322, 1995.

[45] Visconti G, Dal Seno B, Matteucci M, and Mainardi L. Automatic recognition of error potentials in a P300-based brain-computer interface. In *Proceedings of the 4th International Brain-Computer Interface Workshop & Training Course*, pages 238–243, 2008.

[46] Flamary R, Labbé B, and Rakotomamonjy A. Large margin filtering for signal sequence labeling. In *International Conference on Acoustic, Speech and Signal Processing 2010*, pages 1974–1977, Dallas, Texas, USA, 2010.