Learning subject-specific spatial and temporal filters for single-trial EEG classification

Dmitri Model* and Michael Zibulevsky

Technion-Israel Institute of Technology, Electrical Engineering Department, Haifa, Israel

There are a wide variety of electroencephalography (EEG) analysis methods. Most of them are based on averaging over multiple trials in order to increase signal-to-noise ratio. The method introduced in this article is a *single trial* method. Our approach is based on the assumption that the "response of interest" to each task is smooth, and is contained in several sensor channels. We propose a two-stage preprocessing method. In the first stage, we apply *spatial* filtering by taking weighted linear combinations of the sensor measurements. In the second stage, we perform *time-domain* filtering. In both steps, we derive filters that maximize a class dissimilarity measure subject to regularizing constraints on the total variation of the average estimated signal (or, alternatively, on the signal's strength in time intervals where it is known to be absent). No other spatial or spectral assumptions with regard to the anatomy or sources were made.
© 2006 Elsevier Inc. All rights reserved.

Keywords: EEG; Brain–Computer Interface; Classification; Spatial filters

Introduction

Since the discovery of electroencephalography (Berger, 1929), people have speculated that EEG might be used as an alternative communication channel that would allow the brain to bypass peripheral nerves and muscles. The first simple communication systems, that were driven by electrical activity recorded from the head, appeared about three decades ago (Vidal, 1973). In past years, it has been shown that it is possible to recognize distinct mental processes from online EEG (see, for example, Kalcher et al., 1996; Pfurtscheller et al., 1997; Anderson et al., 1998; Obermaier et al., 2001). By associating certain EEG patterns with simple commands, it is possible to control a computer, thereby creating an alternative communication channel that is usually called a *Brain–Computer Interface* (BCI) (Vidal, 1973; Wolpaw et al., 1991).

One of the most complicated problems with BCI systems is the classification of very noisy EEG signals obtained by registering the brain activity of the subject. An approach for dealing with this problem is to require extensive training, whereby the subject is taught to acquire self-control over certain EEG components, such as sensorimotor μ-rhythm (Wolpaw et al., 1991) or slow cortical potentials (Kubler et al., 1999). This ability, to create certain EEG patterns *at will*, is translated by a BCI system into cursor movement (Wolpaw et al., 1991, 1997) or into the selection of letters or words on a computer monitor (Birbaumer et al., 1999; Kubler et al., 1999).

A second approach is the development of *subject-specific* classifiers to recognize different cognitive processes from EEG signals (Pfurtscheller et al., 1997; Anderson et al., 1998; Blankertz et al., 2002). In this case, the typical BCI procedure consists of two stages. First, the person trains the system by concentrating on predefined mental tasks. Usually, two different tasks are used in the training. The BCI registers several EEG samples of each task. Then, the training data are processed in order to construct a classifier. In the second stage, the subject concentrates on one of the tasks again, and the system *automatically* classifies the EEG signals. The key for successful classification is effective preprocessing of the raw data. The objective of this paper is to develop preprocessing methods, based on spatial and temporal filtering, that improve classification accuracy.

The use of spatial filtering in order to improve classification is not a new discovery. The method introduced in Parra et al. (2002) proposed to treat each time sample individually, and to find the spatial integration weights by logistic regression. However, this approach does not take into account the time courses of EEG signals, and thus assumes, implicitly, that the coupling vector between the source and sensors is constant over the time of the response. Moreover, this method tries to maximally discriminate between signals of two classes using no regularization. Thus, the resulting spatial filter is very prone to noise.

A more advanced method for learning spatial filters is Common Spatial Patterns (CSP) (Müller-Gerking et al., 1998; Ramoser et al., 2000). It tries to find several projections that optimally discriminate based on the variances of the projected data of two classes. The Common Spatio-Spectral Patterns (CSSP) (Lemm et al., 2005) method aims to improve upon CSP by incorporating a temporal filter in addition to a spatial one. However, neither approach makes

* Corresponding author.
E-mail addresses: dmm@tx.technion.ac.il (D. Model),
mzib@ee.technion.ac.il (M. Zibulevsky).
Available online on ScienceDirect (www.sciencedirect.com).

use of the time course of the signals and, as a result, they discriminate based on some measure of the sources, rather than the sources themselves.

In this paper, we propose yet another method for spatial and temporal filtering of multi-channel EEG signals. The proposed approach is based on the assumption that the response to the mental task is *smooth* (e.g., has limited total variation) and/or is expected to be small in certain time windows, where the task is not performed. Coefficients of both spatial and temporal filters are learned by optimizing a class dissimilarity measure subject to smoothness constraints on the average signal estimate. No other information about the signals of interest is assumed to be available.

We have evaluated the proposed method on several data sets. Our experiments show that the proposed preprocessing significantly improves classification performance as compared to unprocessed data (i.e., to the simple summation of channels or to choosing the best sensor) as well as data preprocessed by CSP (Ramoser et al., 2000) and CSSP (Lemm et al., 2005) methods.

We have also developed a lower bound on signal reconstruction performance using spatial filtering that is applicable in experiments with synthetic signals. This bound applies to signal reconstruction (and hence classification) performance based on spatial integration only. In our simulation results, we find that in most cases, this bound is nearly or exactly attained. If we use *time-domain* filtering in addition to spatial integration, then even better trends in performance are observed.

Spatial integration method

Data description

In our experiments, we used several data sets, recorded with different numbers of sensors, sampling rates, etc. The details of these data sets are available in Computational experiments. Here we provide the general description of the data format that we use in our preprocessing methods.

Suppose EEG data are recorded using S channels. Single trial signals, corresponding to one of two mental tasks, are extracted from raw data, synchronized by some external stimuli or cue. Suppose each signal is T samples long. Then the signal samples from each single trial can be stored in a $T \times S$ matrix. Let us denote by X_l^1, $1 \leq l \leq L$, the trials that belong to the first class, and X_m^2, $1 \leq m \leq M$, the trials that belong to the second class.

If we average over the trials, we obtain

$$X_{\text{avg}}^1 = \frac{1}{L} \sum_{l=1}^{L} X_l^1 \qquad X_{\text{avg}}^2 = \frac{1}{M} \sum_{m=1}^{M} X_m^2$$

where X_{avg}^1 and X_{avg}^2 are $T \times S$ matrices.

The method

In our model, we assume that each sensor records the following signal:

$$x_i(t) = a_i s^j(t) + n_i(t) \tag{1}$$

where a_i is the coupling coefficient for sensor i, $n_i(t)$ denotes the noise and background activity recorded by the sensor and $s^j(t)$, $j \subset \{1,2\}$ is the response to one of the two possible mental tasks.

We will consider linear estimates of the single trial signals

$$\hat{s}_i^1 = X_i^1 w, \, \hat{s}_j^2 = X_j^2 w \tag{2}$$

where w is an $S \times 1$ *weighting vector*.

The averages of the estimated signals are:

$$\hat{s}_{\text{avg}}^1 = X_{\text{avg}}^1 w, \, \hat{s}_{\text{avg}}^2 = X_{\text{avg}}^2 w \tag{3}$$

Using the above notation, we can formulate our objective as that of finding the weighting vector w that will maximally discriminate between the average estimated signals $\hat{s}_{\text{avg}}^1$ and $\hat{s}_{\text{avg}}^2$, while keeping the single trial estimated signals $\hat{s}_l^1$ and $\hat{s}_m^2$ ($1 \leq l \leq L$, $1 \leq m \leq M$) smooth. The smoothness can be measured, for example, by the *total variation*, defined as

$$\Phi(\hat{s}) = \sum_{l=1}^{L} \sum_{t=1}^{T-1} z_l^1(t) + \sum_{m=1}^{M} \sum_{t=1}^{T-1} z_m^2(t)$$
$$\text{where } z_l^1(t) = |\hat{s}_l^1(t+1) - \hat{s}_l^1(t)|, \, 1 \leq t \leq T-1 \tag{4}$$
$$\text{and } z_m^2(t) = |\hat{s}_m^2(t+1) - \hat{s}_m^2(t)|, \, 1 \leq t \leq T-1.$$

This leads to the following optimization problem:

$$\min_w - ||\hat{s}_{\text{avg}}^1 - \hat{s}_{\text{avg}}^2||_2^2 + \mu\Phi(\hat{s})$$
$$s.t. \, ||w||_2 = 1 \tag{5}$$

where μ is a tradeoff parameter, intended to balance between signal smoothness and class discrimination. In problem (5), we constrain the norm of the weighting vector w to avoid degenerate solutions where $||w|| \to \infty$ or $||w|| \to 0$.

If we substitute the expressions for $\hat{s}_{\text{avg}}^1$, $\hat{s}_{\text{avg}}^2$ and $\Phi(\hat{s})$ from Eqs. (2), (3) and (4), then after a few simple algebraic steps, problem (5) becomes:

$$\min_w - ||X_{\text{avg}}w||_2^2 + \mu(||Yw||_1)$$
$$s.t. \, ||w||_2 = 1 \tag{6}$$

where $||\cdot||_1$ is the ℓ_1-norm, $X_{\text{avg}} = X_{\text{avg}}^1 - X_{\text{avg}}^2$ and Y is a block-matrix obtained by stacking the matrices Y_l^1, $1 \leq l \leq L$ and Y_m^2, $1 \leq m \leq M$, defined as

$$Y_l^1(t,i) = X_l^1(t+1,i) - X_l^1(t,i), \, 1 \leq t \leq T-1$$

$$Y_m^2(t,i) = X_m^2(t+1,i) - X_m^2(t,i), \, 1 \leq t \leq T-1,$$

on top of one another. Thus, if matrices Y_l^1 and Y_m^2 are of size $T-1 \times S$, then the matrix Y will be of size $(L+M)(T-1) \times S$. Note that $z_l^1 = Y_l^1 w$, $z_m^2 = Y_m^2 w$.

Eliminating the tradeoff parameter

In problem (6), there is a need to choose a value for the tradeoff parameter μ. Although we have found in our simulations that the optimization result is quite robust to changes in μ, the need to subjectively assess the tradeoff parameter is still an essential drawback. In what follows, we reformulate the problem in a way that eliminates the μ parameter.

To begin with, let us point out that the norm and the sign of the vector w have no significance. We are interested only in the *relative* values of its elements. In other words, we want to find a w that will satisfy two conditions. *Firstly*, it must minimize

the value of the second term in Eq. (6), when the value of the first term is fixed. *Secondly*, it must minimize the value of the first term, when the value of the second term is fixed. With this in mind, we rewrite the problem as follows[1]:

$$\min_w ||Yw||_1$$
$$s.t. \ ||X_{\text{avg}}w||_2^2 = 1. \tag{7}$$

A solution to the above satisfies the first condition by definition. The second condition is also satisfied. This can be proved in the following way. Suppose w_{TV} is a solution of Eq. (7). Aiming for a contradiction, assume that there exists a w_{new}, such that $||Yw_{\text{new}}||_1 = ||Yw_{\text{TV}}||_1$ and $||X_{\text{avg}}w_{\text{new}}||_2 = c^2 < 1$. In this case, $w = \frac{1}{c}w_{\text{new}}$ would satisfy the constraint, while $||Yw||_1 < ||Yw_{\text{TV}}||_1$. This contradicts the assumption that w_{TV} is a solution of Eq. (7). Thus, the second condition also holds.

Although problems (6) and (7) are not completely equivalent, problem (7) can be viewed as one that optimally (and automatically) chooses the tradeoff parameter μ. Indeed, if w_μ is a solution of Eq. (6) for some value of μ and w_{TV} is a solution of Eq. (7) then, after rescaling w_{TV}, we have proven that: $||Yw_{\text{TV}}||_1 < ||Yw_\mu||_1$ if $||X_{\text{avg}}w_{\text{TV}}||_2^2 = ||X_{\text{avg}}w_\mu||_2^2$ and $||X_{\text{avg}}w_{\text{TV}}||_2^2 > ||X_{\text{avg}}w_\mu||_2^2$ if $||Yw_{\text{TV}}||_1 = ||Tw_\mu||_1$. This is true for any value of μ. Thus, w_{TV} is really the optimal solution. In Appendix A.1, we discuss the approximate solution of Eq. (7) using numerical optimization methods.

An alternative to problem (7) can be obtained from a Basis Pursuit perspective (Chen et al., 1998). The matrix Y in Eq. (7) may contain the coefficients for representing the signal in some basis or "over-complete" dictionary matrix Ψ (e.g., Fourier or wavelet bases). Thus, we can write $Y = \Psi^T X$, where Ψ is expected to be sparse (see, for example, Zibulevsky and Zeevi, 2002). Another alternative is to construct the matrix Y from signals at predefined time windows, where their energy is expected to be small.

Learning spatial integration weights through eigenvalue decomposition

In this section, we propose to approximate the solution of Eq. (7) with the solution of

$$\max_w ||X_{\text{avg}}w||_2^2$$
$$s.t. \ ||Yw||_2^2 = 1 \tag{8}$$

The approximate problem (8) is equivalent to a version of problem (7) in which the ℓ_1-norm is replaced by the ℓ_2-norm. Put another way, it is equivalent to approximating the Total Variation measure in Eq. (4) with an ℓ_2-based non-smoothness measure.

We now normalize the constraints in Eq. (8) by making a change of variables. Let us rewrite the constraint: $||Yw||_2^2 = w^T Y^T Yw$. If the matrix Y has full column rank (which is very likely for noisy data), then the matrix $C = Y^T Y$ is positive definite and has a Cholesky factorization[2] $C = U^T U$. Now, the constraint can be written as

$||Yw||_2^2 = w^T Y^T Yw = w^T U^T Uw$. If we introduce a new variable $x = Uw$ ($w = U^{-1}x$), then the objective (8) can be written as:

$$\max_x ||X_{\text{avg}}U^{-1}x||_2^2$$
$$s.t. \ ||x||_2^2 = 1 \tag{9}$$

The mathematical program (9) is the classical problem of finding the induced ℓ_2 norm of a matrix, in this case the matrix $A = X_{\text{avg}}U^{-1}$. Its solution is $x* = v_{\text{max}}$, where v_{max} is the eigenvector corresponding to the largest eigenvalue, λ_{max}, of the matrix $A^T A = (X_{\text{avg}}U^{-1})^T X_{\text{avg}}U^{-1} = U^{-1T}X_{\text{avg}}^T X_{\text{avg}}U^{-1}$. Hence, the solution of Eq. (8) is $w = U^{-1}v_{\text{max}}$.

An important advantage of this approximation is that it does not require iterative optimization. Rather, it requires mainly Cholesky and eigenvalue decomposition steps, which can be done very efficiently. The drawback, however, is that non-smoothness is no longer measured in terms of the Total Variation. Nonetheless, our simulations show that using the approach described in this subsection, we achieve similar results to those obtained by numerically optimizing Eq. (7).

Time-domain filtering

Signals reconstructed by spatial filtering methods still suffer from noise contamination. This contamination can be reduced by a second stage of preprocessing, namely, *time-domain* filtering of the estimated signals (Eq. (2)).

The problem with applying filtering is that we do not know in advance which filter to use, because the signals of interest, as well as the background activity noise, are unknown. Thus, we propose to derive a suitable filter according to the same learning scheme on which the spatial filter was based: maximize a class dissimilarity measure, while keeping the resulting signal smooth (or, alternatively, small in predefined time windows).

With this in mind, we propose to find a filter $h[n]$, $1 \leq n \leq N_{\text{filt}}$, which will further discriminate between reconstructed signals $\hat{s}_{\text{avg}}^1[n]$ and $\hat{s}_{\text{avg}}^1[n]$ according to,

$$\max_{h[n]} ||\left(\hat{s}_{\text{avg}}^1[n] - \hat{s}_{\text{avg}}^2[n]\right)*h[n]||_2^2$$
$$s.t. \ \sum_{l=1}^{L} ||\hat{z}_l^1 * h[n]||_2 + \sum_{m=1}^{M} ||\hat{z}_m^2 * h[n]||_2 = 1 \tag{10}$$

where * denotes convolution.

Since we are considering discrete time, time-limited signals, let us construct a $(T - N_{\text{filt}} + 1) \times N_{\text{filt}}$ convolution matrix $\tilde{X}_{\text{avg}}^1$, the j-th column of which will contain $\hat{s}_{\text{avg}}^1[n]$, $j \leq n \leq (T - N_{\text{filt}} + j)$ (i.e., the j-th column of $\tilde{X}$ contains a replica of the signal $\hat{s}_{\text{avg}}^1[n]$ shifted by $(j - 1)$, which is also truncated by $(j - 1)$ taps at the beginning and $(N_{\text{filt}} - j)$ taps at the end). In the same manner, we can define convolution matrix $\tilde{X}_{\text{avg}}^2$, the columns of which will contain shifted replicas of $\hat{s}_{\text{avg}}^2[n]$. Finally, we define the matrix $\tilde{X}_{\text{avg}} = \tilde{X}_{\text{avg}}^1 - \tilde{X}_{\text{avg}}^2$.

Similarly, we can construct convolution matrices $\tilde{X}_l^1$, $\tilde{X}_m^2$ for single trials and derive from them a matrix $\tilde{Y}$ in the same manner as the matrix Y was derived from X_l^1, X_m^2 in Eq. (6).

We can now rewrite problem (10) as,

$$\max_{w_{\text{filt}}} ||\tilde{X}_{\text{avg}}w_{\text{filt}}||_2^2$$
$$s.t. \ ||\tilde{Y}w_{\text{filt}}||_2^2 = 1 \tag{11}$$

---

[1] Alternatively, we may reverse the roles of the objective and the constraint in problem (7).

[2] If the matrix Y is not full rank, we may use a regularization $C = Y^T Y + \alpha I$, where α is a small constant, and I is an identity matrix.
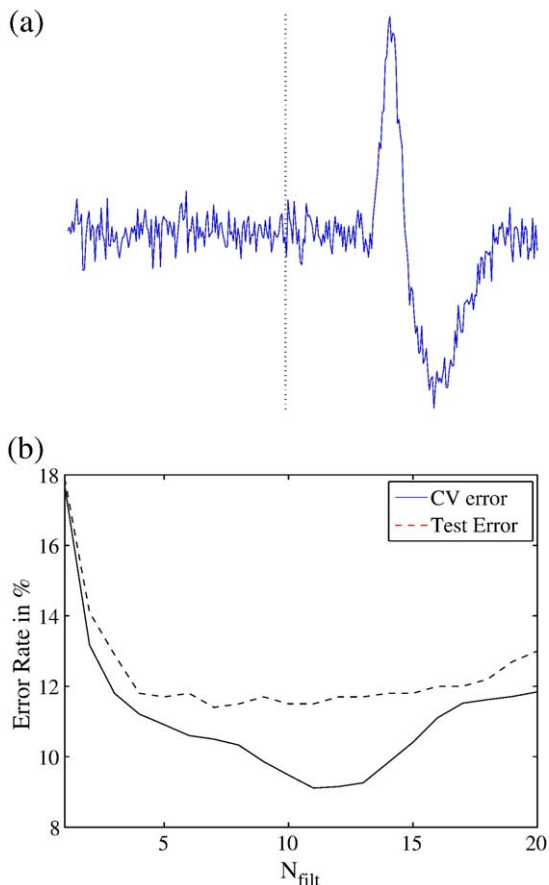
(a)

(b)

Fig. 1. (a) Illustration of a (single-channel) EEG recording, which contains background activity only. If we start to register the response well in advance then, at the beginning, we will record background activity only. The signal to the left of the dashed vertical line can be treated as background noise. The actual response appears after the dashed vertical line. (b) Cross-validation error rate for different values of N_{filt}. Note that the real test error (dashed line) is highly correlated with the CV error. This enables us to choose the optimal order for the FIR filter.
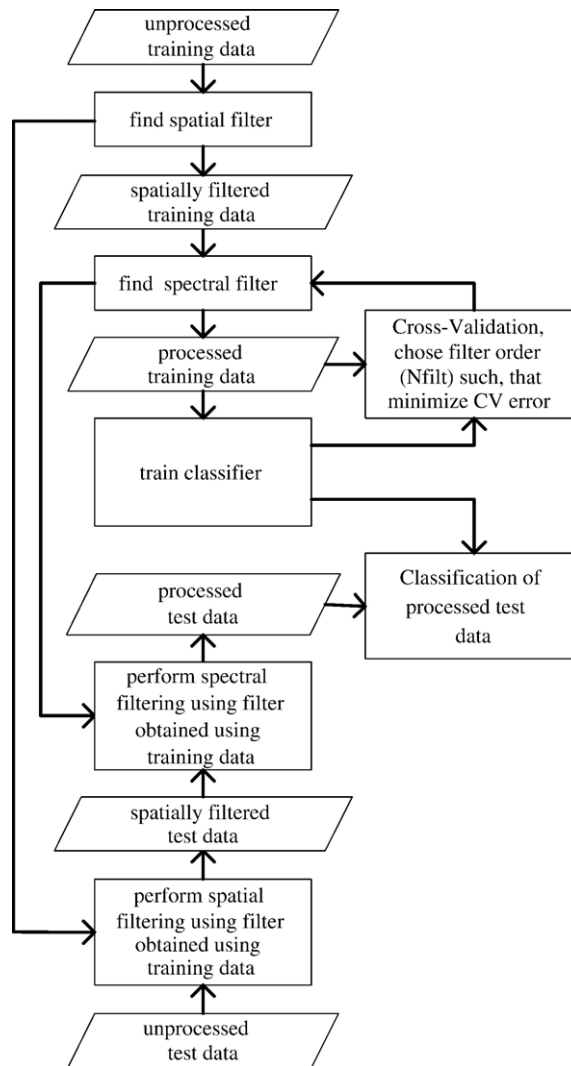

Fig. 2. Flow chart of the classification process. The temporal filtering stage is omitted for w_{TV} and w_{EVD}.

which has the same form as problem (8) Thus, the solution developed in previous section can be used to solve the problem (11). Note that the solution of problem (11), w_{filt}, is in fact a temporal filter, rather than a spatial one.

There is an alternative choice for the matrix $\tilde{Y}$. It can represent background activity noise,[3] which we also want to minimize, instead of Total Variation. This idea may even be more appealing because, with this alternative, we minimize the background noise directly, in contrast to Total Variation, which is an indirect noise measure. This is an effective approach when the background activity is stationary. Our simulations show that this alternative choice of the matrix $\tilde{Y}$ performs better for real EEG recordings, while for synthetic data, the original choice seems preferable.

The only remaining question is how to choose the optimal order N_{filt} of the FIR filter $h[n]$. We have no analytical solution for this issue. In our experiments, we chose its value based on cross-validation of the training data. We calculated the CV error for different values of N_{filt}, and then chose the one that gave the lowest

error rate. Fig. 1(b) illustrates that the CV error rate and test error rate are highly correlated.

Computational experiments

We have conducted several experiments, using both synthetic and real signals, in order to show the feasibility of the proposed

Table 1
Classification results (error rate in %) for the BCI competition 2002 data set

	CSP	CSSP	w_{TV}	w_{EVD}	$w_{TV} + w_{filt}$	$\sum$	best ch.
10-fold CV	21	12	12	12	13	51	32
	(4.1)	(3.3)	(3.8)	(3.5)	(3.9)	(11)	(10)
Test	27	21	5	5	5	53	30

Each column corresponds to a different method of preprocessing. Both *10-fold Cross-Validation* (STD is given in parenthesis) and test error results are provided. For the CSP method, we used $m = 4$. For CSSP, we used $m = 8$, $\tau = 16$. A filter length of $N_{filt} = 100$ was used for the $w_{TV} + w_{filt}$ method. All of these parameters were chosen to minimize cross-validation error.

---

[3] It can be obtained, for example, if we start to register the response well in advance (Fig. 1(a)).

Table 2
Classification results (error rate in %) for the BCI competition 2003 data set

	CSP	CSSP	w_{TV}	w_{EVD}	$w_{TV} + w_{filt}$	$\sum$	best ch.
10-fold CV	18	10	21	24	22	43	37
	(4.9)	(4.9)	(4)	(4.3)	(4)	(14)	(9)
Test	30	22	22	22	22	41	39

Each column corresponds to a different method of preprocessing. Both *10-fold Cross-Validation* (STD is given in parenthesis) and test error results are provided. For the CSP method we used $m = 6$. For the CSSP method, we used $m = 6$, $\tau = 22$. A filter length of $N_{filt} = 110$ was used for the $w_{TV} + w_{filt}$ method. All of these parameters were chosen to minimize cross-validation error.

approaches and to compare them with other spatial integration methods. These other methods included CSP (Ramoser et al., 2000) and CSSP (Lemm et al., 2005). Furthermore, the comparison included the direct use of unprocessed data, namely the simple summation of channels ("$\sum$") and the practice of choosing the best channel ("best ch."). The latter were included to emphasize the benefits of proper spatial integration. In what follows, we describe the results of our simulations.

Real EEG signals

In these experiments, we used the data obtained from two BCI competitions, held in 2002 (Sajda et al., 2003) and 2003 (Blankertz et al., 2004). The goal of these competitions was to validate signal processing and classification methods for Brain Computer Interfaces. The data set consisted of trials of spontaneous EEG activity, one part labeled (training data) and another part unlabeled (test data). The goal was to infer labels for the test set by preprocessing the training data. Inferred test labels aim to optimally fit the true test labels (which were unknown to the contestants).

BCI competition 2002

This data set (Sajda et al., 2003) consisted of EEG signals that were recorded from a single subject in sessions with a few minutes break in between. The subject was sitting in a normal chair, arms relaxed and resting on the table, and with fingers in a standard typing position at the computer keyboard (index fingers at 'f', 'j'

Table 3
10-fold Cross-Validation error rate in % (STD is given in parenthesis) for the visual stimuli data set

	CSP	CSSP	w_{TV}	w_{EVD}	$w_{TV} + w_{filt}$	$\sum$	best ch.
NM	8.9	5.8	2.2	2.2	2.3	39	26
	(2.1)	(1.5)	(0.7)	(0.6)	(0.6)	(5.5)	(4.5)
k-nn	8.3	5.6	2.2	1.7	3.3	27	16
	(2)	(1.4)	(0.7)	(0.6)	(0.7)	(6)	(5)
FLD/SVM	7.3	3.9	2.2	2.6	2.9	41	29
	(1.8)	(1.7)	(0.6)	(0.9)	(0.8)	(13)	(14)

Each column corresponds to a different method of preprocessing. Results of applying 3 different classifiers are shown: Nearest Mean (NM), k-nn (with $k = 3$) and FLD/SVM with an exponential kernel. Fisher's Linear Discriminant (FLD) was applied with the CSP and CSSP methods as proposed in previous papers. Since FLD is not applicable if signals of both classes are zero-mean, we used SVM in the remaining methods instead. For the CSP method, we used $m = 4$. For the CSSP method, we used $m = 6$, $\tau = 12$. A filter length of $N_{filt} = 40$ was used for the $w_{TV} + w_{filt}$ method. All of these parameters were chosen to minimize cross-validation error. The same parameters were used later in the test stage.

Table 4
Test error rate (in %) for the visual stimuli data set

	CSP	CSSP	w_{TV}	w_{EVD}	$w_{TV} + w_{filt}$	$\sum$	best ch.
NM	3.3	3.3	3.3	1.7	0.0	40	35
k-nn	10	8.3	0.0	1.7	3.3	35	20
FLD/SVM	6.7	5	1.7	1.7	3.3	45	42

Each column corresponds to a different method of preprocessing. The results of applying 3 different classifiers are shown: Nearest Mean (NM), k-nn (with $k = 3$) and FLD/SVM with exponential kernel. See Table 3 for further details.

and smallest fingers at 'a',';'). The task was to press two chosen keys with the corresponding fingers in a self-chosen order and at a self-chosen pace ('self-paced key typing'). A total of 516 keystrokes were made at an average speed of 1 keystroke every 2.1 s. Brain activity was measured with 27 Ag/AgCl electrodes at a 1000 Hz sampling rate using a band-pass filter from 0.05 to 200 Hz.

Additionally, windows of length 1500 ms were extracted from the continuous raw signals, each ending at 120 ms before the respective keystroke. The reason for choosing the endpoint at -120 ms is that, prior to this instant, the classification based on measurements of EMG activity only is essentially arbitrary. For the test set, 100 trials equally spaced over the whole experiment were taken leaving 413 labeled trials for training. For classification, we used only the last 700 ms of each trial. The first 800 ms was treated as background noise and used in the calculation of w_{filt}.

The learning methods derived in the previous sections were applied to the training data only. The respective weighting vectors are denoted w_{TV} (computed as outlined in Appendix A.1), w_{EVD} (computed using eigenvalue decomposition, as outlined in Section 3) and $w_{TV} + w_{filt}$ (the case where time-domain filtering is added). Next, we classified the filtered test data. Fig. 2 illustrates the classification process. We tried several classifiers using the "pr-tools" classification toolbox (Duin, 2000), including Nearest Mean, k nearest neighbor (k-nn) (Cover and Hart, 1967) and Support Vector Machines (SVM) with different kernels (Burges, 1998). All classifiers provided similar results with the BCI2002/3 data sets. Therefore, we chose to use the simplest one, namely the Nearest Mean classifier. The resulting classification errors, both of 10-fold Cross-Validation (Duda et al., 2001) and the test set error,[4] are summarized in Table 1. The best error rate reported by the competition organizers was 4%. The classification error rates using the CSP and CSSP methods with Fisher's Linear Discriminant were 27% and 21%, respectively (which was similar to the Nearest Mean classification results).

BCI competition 2003

The data set for this competition (Blankertz et al., 2004) was similar to the previous one. In particular, it was based on a self-paced key typing task. This time, the average typing rate was 1 keystroke per second. In total, there were 416 epochs of 500 ms length each ending 130 ms before a keystroke. For the training set, 316 epochs were labeled. The remaining 100 epochs were unlabeled (test set). For classification purposes, we used only the last 360 ms of data. The first 140 ms were treated as background noise and used in the calculation of w_{filt}.

---

[4] Test set error was calculated once the true labels were published by the competition organizers.
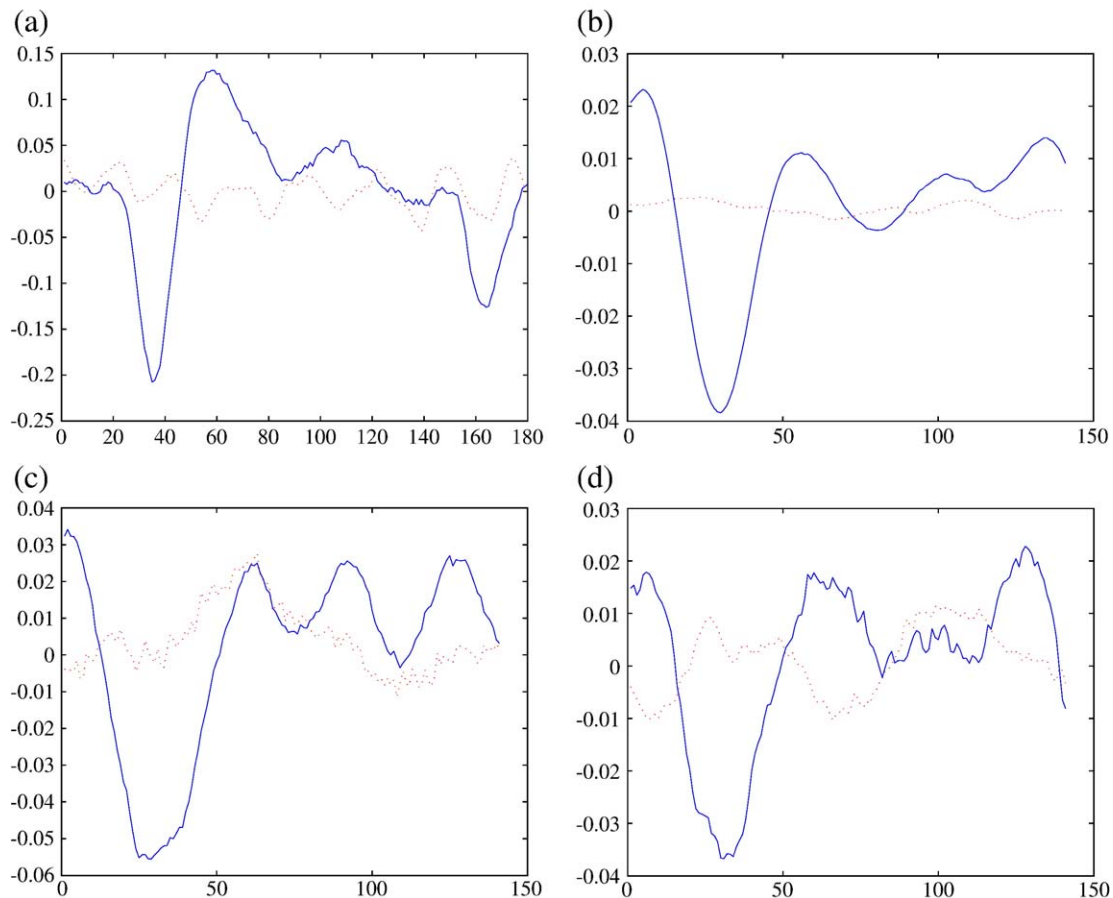
Fig. 3. Experiment with visual stimuli. The dotted line represents the signal with background activity only, while the solid line corresponds to the signal that contains the response. (a) Averaged signals reconstructed by w_{TV}; (b) averaged signals reconstructed by w_{TV} and then filtered by w_{filt}; (c, d) examples of single-trial signals reconstructed by w_{TV} and then filtered by w_{filt}. Note that the two classes are easily distinguishable even in single trials. Also, note the similarity between single-trial and averaged signals.

We used the training data for preprocessing, trying all proposed approaches (w_{TV}, w_{EVD} and $w_{TV} + w_{filt}$, ...). Again, we used a Nearest Mean classifier. The classification error of 10-fold Cross-Validation and the test set error are summarized in Table 2. The best error rate reported by the competition organizers was 16%. The classification error rates using the CSP and CSSP methods with Fisher's Linear Discriminant were 29% and 19% (which was similar to the Nearest-Mean classification results).

Response to visual stimuli

The EEG data for this experiment were acquired according to the following procedure.[5] The subject was shown a sequence of 3 different images at a preselected pace that was constant across trials. Afterwards, the subject had to respond by pressing a button. The trials were repeated with a periodicity of 7 s. The delay between the first and the second images in the sequence was 1.5 s, and 2.5 s between the second and the third image. Next, the subject was given 3 s to respond. Each session consisted of approximately 30 trials. There were several sessions with a few minutes break in between. The EEG data were recorded by 23 electrodes with a sampling rate of 256 samples per second.

Our aim here was to distinguish the response to visual stimuli from the response to the absence of stimuli, i.e., to regular background activity. We derived two classes of signals from the raw data. The first class represented the response to visual stimuli (i.e., to the display of an image) and the second class represented the response to background activity. In order to derive the first class, we extracted segments from the raw data, each of which started at the times when the first image was displayed and ran for 180 time samples. The second class was derived from segments starting 300 time samples before the third image was displayed and ending 120 time samples before the time when the third image was displayed.

We randomly chose 180 trials to be the training set. The remaining 60 trials were taken as the test set. We then applied all of our approaches to these data. We tested the classification

---

[5] The EEG data described here were recorded in the Laboratory for Evoked Potentials in the Technion-Israel Institute of Technology in association with work done in Bigman and Pratt (2004). We are grateful to Hillel Pratt for providing us with these recordings.

Table 5
Experiment with artificial signals and white Gaussian background noise: classification error rate in %

	w_{TV}	w_{EVD}	$w_{TV} + w_{filt}$	w_{th}	$\sum$	best ch.
SNR = −10 dB	0.0%	0.0%	0.0%	0.0%	43.7%	14.0%
SNR = −15 dB	3.5%	3.3%	1.4%	2.7%	44.1%	36.7%
SNR = −20 dB	20.3%	20.4%	13.7%	17.1%	49.0%	41.7%

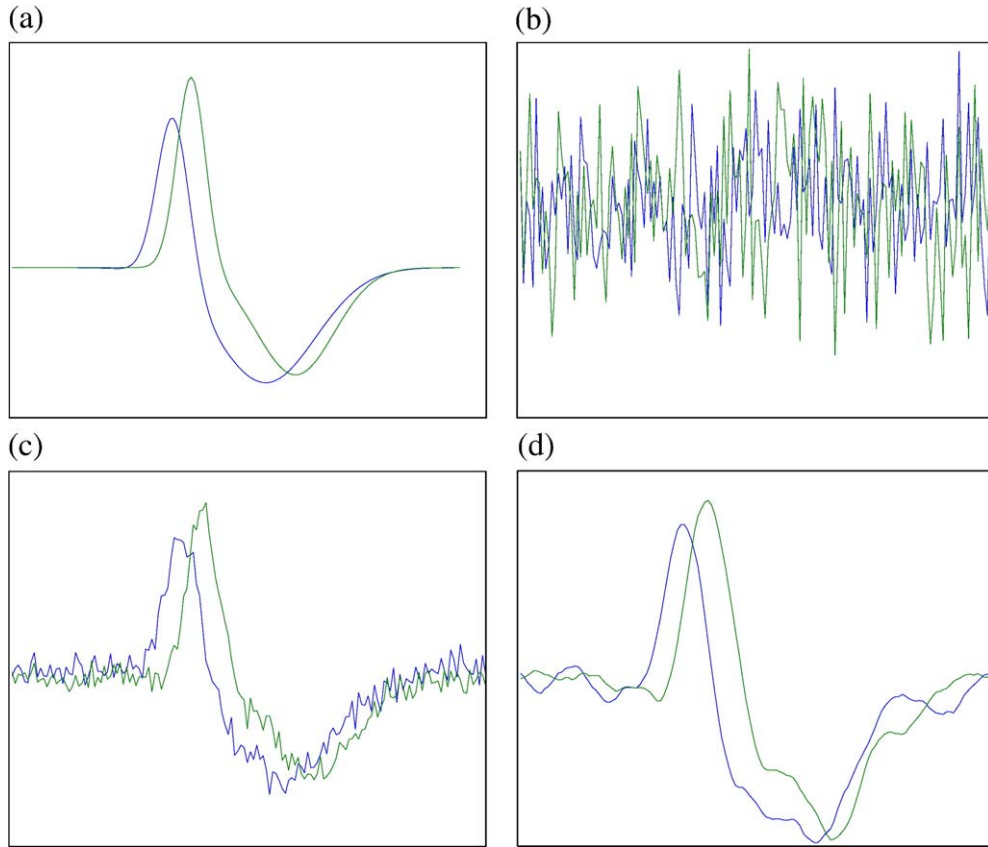The first column shows average SNR, measured at each sensor.

Fig. 4. Experiment with artificial signals and Gaussian background noise: (a) artificial signals; (b) signals restored by the simple summation of channels; (c) signals restored by w_{TV}; (d) signals restored by w_{TV} and further filtered by w_{filt}.

performance of NM, k-nn and SVM with an exponential kernel. The results of 10-fold cross-validation are summarized in Table 3. The test set error is reported in Table 4. Reconstructed signals are shown in Fig. 3.

Artificial signals

In this section, we discuss experiments on synthetic data, generated by mixing smooth signals s_{art}^1 and s_{art}^2 (each belonging to a different class) into background noise N. The data synthesis model was as follows:

$$X_l^1 = s_{art}^1 a^T + N_l$$

$$X_m^2 = s_{art}^2 a^T + N_m \tag{12}$$

Here, each s_{art} is a $T \times 1$ vector of signal time samples and a is an $S \times 1$, randomly chosen *mixing vector*, where S

denotes the number of channels. The single trial data and noise matrices (denoted X_i and N_i respectively) are each $T \times S$.

White Gaussian background noise

In this experiment, we used $T = 150$ time samples and $S = 25$ channels. Each 150×25 noise matrix N_i was generated independently in each trial from samples of a white Gaussian noise process. The elements of the mixing vector a were drawn according to a uniform distribution from the interval $[-1;1]$. We have generated 1200 trials. The first 200 trials (approximately 100 from each class) were used for preprocessing (i.e., for finding the unmixing vector w). The remaining 1000 trials were used for classification by the Nearest Mean algorithm (Duda et al., 2001).

As in previous sections, we compared the classification performance of the different methods (w_{TV}, w_{EVD} and w_{filt} and unprocessed data). In addition, baseline performance figures were obtained by applying w_{th} (Eq. (A.9)), which is derived in Appendix A.2. As discussed in the Appendix, w_{th} gives an upper bound on signal denoising performance obtainable via spatial filtering. Classification performance (based on the Nearest Mean classifier) is shown in Table 5 for different SNR.[6] Reconstructed signals are shown in Fig. 4.

Table 6
Experiment with artificial signals and real EEG recordings as background noise: classification error rate in %

	w_{TV}	w_{EVD}	$w_{TV} + w_{filt}$	w_{th}	Σ	best ch.
SNR = −20 dB	2.5%	1.3%	1.1%	0.0%	51.6%	43.3%
SNR = −25 dB	10.5%	10.3%	10.0%	0.7%	50.5%	50.1%
SNR = −30 dB	25.7%	22.6%	21.9%	1.3%	52.9%	49.5%

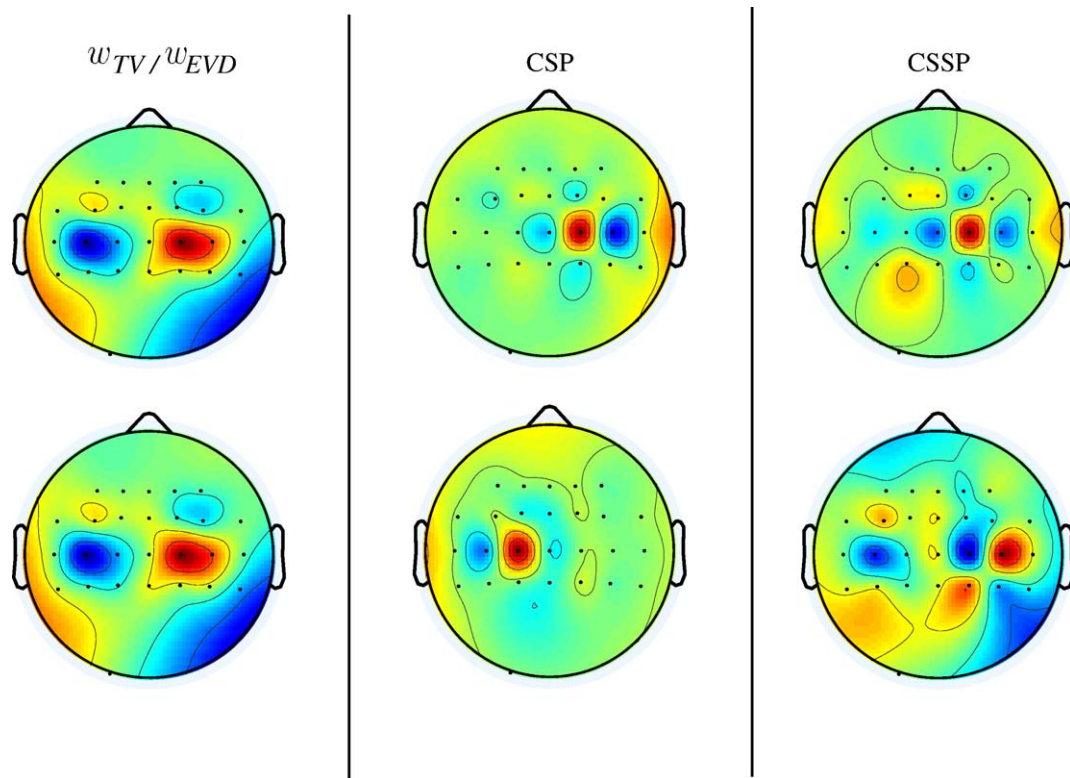The first column shows average SNR, measured at each sensor.

Fig. 5. Imagined hand movement data set (BCI competition 2002). Shown here are spatial filters corresponding to w_{TV} and w_{EVD} (left column) and CSP and CSSP methods (two patterns each). Spatial filters corresponding to w_{TV} and w_{EVD} are almost identical. Spatial filters obtained using CSP and CSSP are different, but concentrated in the same regions.

Real EEG as background noise

In this experiment, we used $T = 150$ time samples, $S = 22$ channels, and real EEG signals as the background noise N. The rest of the setup is identical to the previous experiment. Classification results are reported in Table 6.

Discussion

The classification results presented in the previous section leave no doubt that spatial integration improves classification accuracy. It is interesting to compare the spatial filters obtained from w_{TV} and w_{EVD} with the spatial patterns obtained from the CSP and CSSP methods.

Figs. 5 and 6 show the spatial patterns obtained with the different methods on imagined hand movement data sets. Note that w_{TV} and w_{EVD} are similar in both data sets. The spatial filters obtained from the CSP and CSSP methods are different, but are concentrated in the same regions. If we take into account the classification results from these data sets (the proposed methods performed at least as well as the CSP and CSSP methods), these observations suggest that w_{TV}/w_{EVD} were able to concentrate all relevant spatial information in one single filter. It may also suggest that after reconstructing the signal using a spatial filter, a better strategy for classification would be to use its time course, rather than to use some measure of it (e.g., the variance, as proposed by the CSP/CSSP methods). Indeed, the response to two different tasks can have different time courses, but the same variance.

The spatial filters for the visual stimuli data set can be observed in Fig. 7. This time, all of the methods produced very similar filters.[7] If we look at the classification results and compare w_{TV}/w_{EVD} with CSP and $w_{TV} + w_{filt}$ with CSSP,[8] we can observe that the proposed methods have similar (or even better) performance. Thus, we can again argue that w_{TV}/w_{EVD} combine the most relevant spatial information into a single filter.

Another interesting issue for discussion is the benefit of temporal filtering to classification accuracy. Note that, in the imagined hand movement data sets, the temporal filter did not improve classification accuracy. However, for the visual data set, it improved the classification accuracy considerably.

This phenomena can be explained as follows. In the imagined hand movement data sets, signals of both classes correspond to the movement task and have similar power spectra (see Fig. 8(a)). In this case, the temporal filter turns out to be a simple low-pass filter, which does not enhance the discrimination between classes. (It may, however, improve the quality of estimation by rejecting high-frequency noise.)

In the visual data set, signals of different classes have different origin. The first contains background activity, while the second is a response to the visual stimuli. As one can see in Fig. 8(b), those signals have different frequency components. In this case, the power spectra of the temporal filter have dominant peaks over those frequency regions where the power spectra of the two signals differ the most. Thus, such

---

[7] For CSP/CSSP methods, we chose the most similar filter among the first 2 m that participated in the classification.

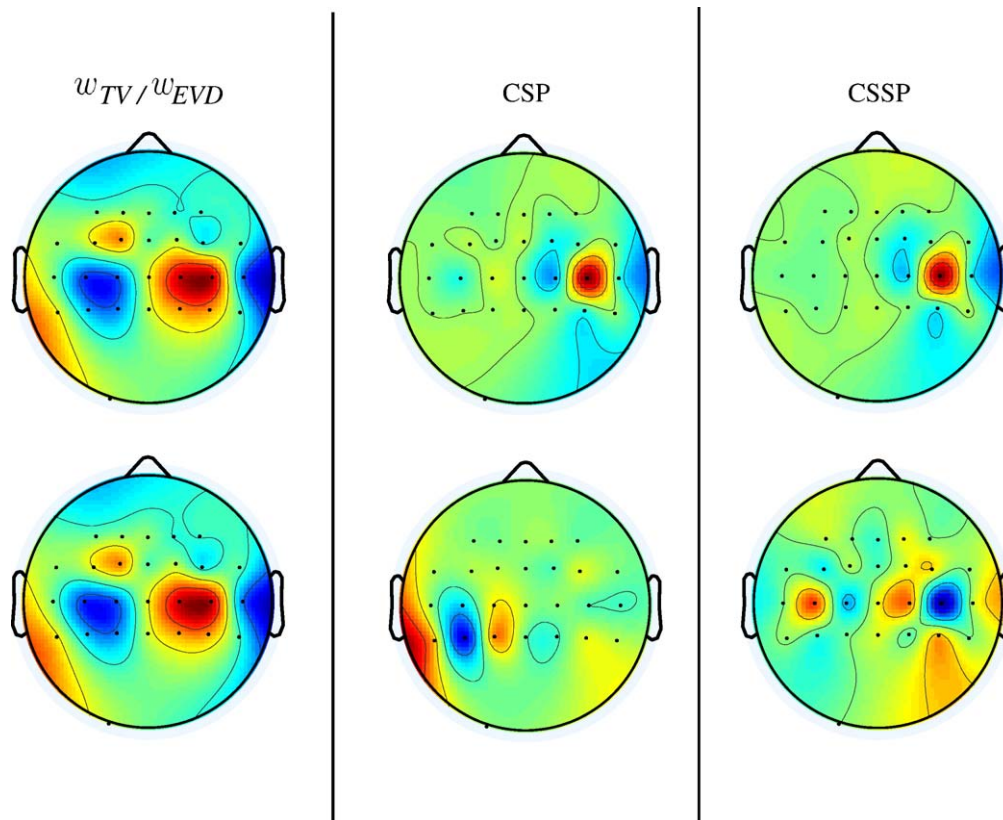[8] This comparison is the most fair since both $w_{TV} + w_{filt}$ and CSSP use spatial–temporal filtering.

Fig. 6. Imagined hand movement data set (BCI competition 2003). Shown here are spatial filters obtained from w_{TV} and w_{EVD} (left column) and from CSP and CSSP methods (two patterns each). Spatial filters obtained from w_{TV} and w_{EVD} are almost identical. Spatial filters obtained using CSP and CSSP are different, but concentrated in the same regions. Note the similarity to the filters in Fig. 5.

filtering enhances class discrimination, resulting in classification improvement.

Conclusions

We have presented a two-stage preprocessing algorithm, one that extracts the desired response from multi-channel data by means of spatial integration in the first stage and time-domain filtering in the second stage. This preprocessing is essential for effective classification. Our experiments showed that the misclassification rate achieved with the preprocessed data is significantly lower than the error rate obtained by classifying unprocessed signals (i.e., by the simple summation of channels, or choosing the best channel). In addition, in our simulations on synthetic data, we have shown that the error rate, achieved after the first stage of preprocessing, reaches (or is very close to) a lower bound developed for spatial integration methods. Moreover, when we
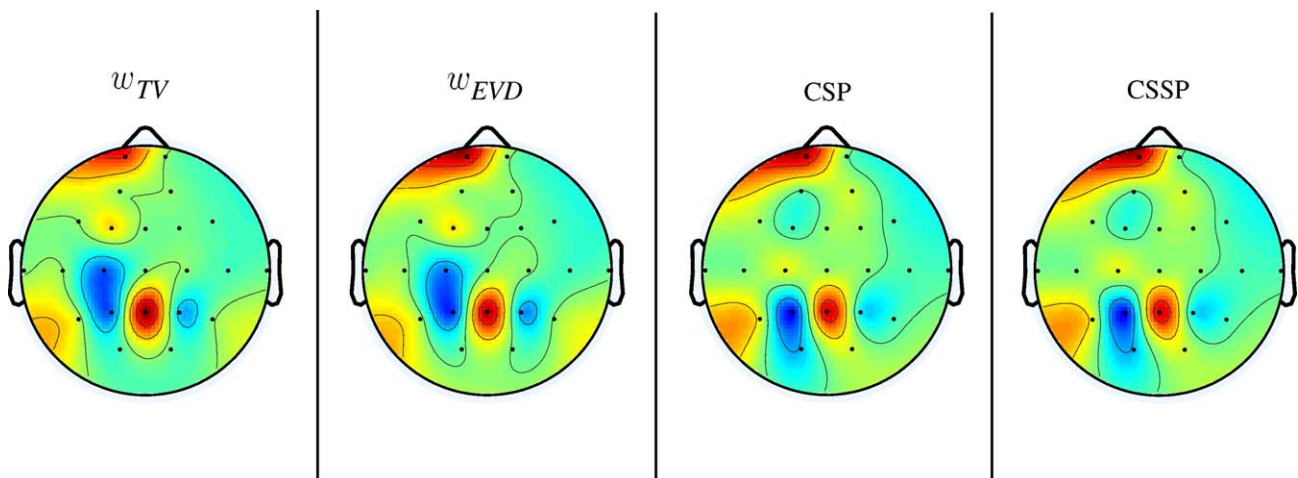

Fig. 7. Visual stimuli data set. Spatial filters obtained from w_{TV}, w_{EVD}, CSP and CSSP. Note the similarity of all spatial filters.
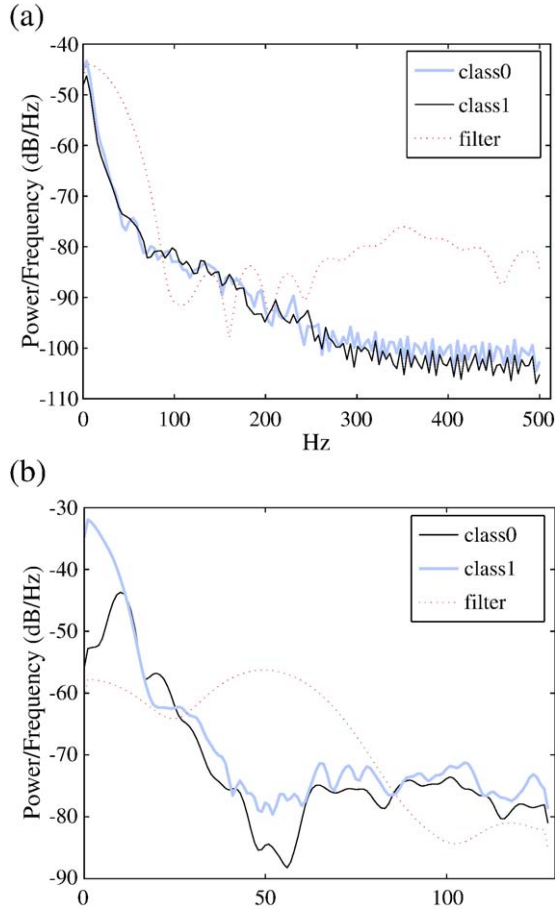
(a)

(b)

Fig. 8. Power spectra of estimated signals (solid lines) and resulting temporal filter w_{filt} (dashed line): (a) imagined hand movement data set (BCI competition 2003); (b) response to visual stimuli data set.

add time-domain filtering, we obtain an even lower error rate than given by this bound.

Appendix A

A.1. Numerical optimization

Here, we outline an approach to the (approximate) solution of the mathematical problem (7) using the method of Lagrange multipliers. An approximation is necessary because the objective function in Eq. (7) is not differentiable at all w. Accordingly, we will use the following smooth approximation of the absolute value function

$$\psi(t) = c\left(\frac{|t|}{c} - \log\left(1 + \frac{|t|}{c}\right)\right) \tag{A.1}$$

Note that $\psi(t)$ is defined at $t = 0$:

$$\psi'(t) = \frac{t}{c + |t|} \tag{A.2}$$

The approximation reaches arbitrary accuracy as $c \to 0$.

Applying this approximation to Eq. (7), we obtain

$$\min_w \sum_{l=1}^{L} 1^T \psi\left(Y_l^1 w\right) + \sum_{m=1}^{M} 1^T \psi\left(Y_m^2 w\right) \tag{A.3}$$

$$s.t. \ \|X_{\text{avg}} w\|_2^2 = 1$$

where 1 is a vector of ones and where the application of $\psi(\cdot)$ to a vector is element-wise.

Let us denote the objective function in Eq. (A.3) as $f(w)$, and the constraint as

$$g(w) = \|X_{\text{avg}} w\|_2^2 = w^T X_{\text{avg}}^T X_{\text{avg}} w.$$

The gradients of $f(w)$ and $g(w)$ are

$$\nabla f(w) = \sum_{l=1}^{L} \left(Y_l^1\right)^T \psi'\left(Y_l^1 w\right) + \sum_{m=1}^{M} \left(Y_m^2\right)^T \psi'\left(Y_m^2 w\right) \tag{A.4}$$

$$\nabla g(w) = 2 X_{\text{avg}}^T X_{\text{avg}} w. \tag{A.5}$$

Using these expressions, one can evaluate the gradient of the Lagrangian of Eq. (A.3). The problem is solved by finding a point where the Lagrangian's gradient is zero, which can be done using numerical tools.

A.2. Theoretically optimal spatial filtering

In some situations (relevant mainly to synthetic data studies), we can solve a mathematical problem that defines, in a certain sense, the optimal signal reconstruction performance possible using spatial filtering. Suppose that the 'sensor measurement' data are synthesized in the following way:

$$X = s_{\text{art}} a^T + N \tag{A.6}$$

In the above formula, the $T \times S$ 'sensor measurement' matrix X is obtained by mixing an artificial signal, represented as a $T \times 1$ vector s_{art}, with an $S \times 1$ coupling vector a, and adding a $T \times S$ noise matrix N.

If we know both the background noise covariance matrix $R = N^T N$ and the mixing vector a in Eq. (A.6), we can find the best weighting vector w solving the following problem[9]:

$$\min_w \|Nw\|_2^2 = w^T R w$$

$$s.t. \ \|s_{\text{art}} a^T w\|_2^2 = 1. \tag{A.7}$$

The artificial signal s_{art} can be normalized so that $\|s_{\text{art}}\|_2 = 1$. The constraint in Eq. (A.7) then simplifies, yielding,

$$\min_w w^T R w$$

$$s.t. \ a^T w = 1. \tag{A.8}$$

The above problem aims to find a vector w that maximally suppresses noise while constraining the norm of the average result to be constant. This task differs from the apparent task of problem (7) However, since in both cases we want to de-noise the signal of interest then, in a sense, the solution of Eq. (A.8) can be viewed as a theoretically best achievable limit, and thus a good reference point for comparison.

We can solve problem (A.8) using Lagrange multipliers:

$$\min_w w^T R w - \lambda\left(a^T w - 1\right).$$

The gradient of the Lagrangian is given by: $g(w) = 2Rw - \lambda a^T$. The solution is obtained by setting $g(w) = 0 \Rightarrow w_{\text{th}} = 0.5\lambda R^{-1} a$.

---

[9] Our aim here is to perform de-noising by taking a weighted sum of channels. Thus, the idea of subtracting the noise matrix is not relevant.

Again, though, we are indifferent, in practice, to the scale/sign of w_{th}. Hence, any real $\lambda \neq 0$ can be chosen. If we take $\lambda = 2$, the solution

$$w_{\text{th}} = R^{-1} a \qquad (A.9)$$

is obtained.

References

Anderson, C., Stolz, E., Shamsunder, S., 1998. Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. IEEE Trans. Biom. Eng. 45 (3), 277–286.

Berger, H., 1929. Uber das electrenkephalogramm des Menschen. Arch. Psychiatr. Nervenkr. 87, 527–570.

Bigman, Z., Pratt, H., 2004. Time course and nature of stimulus evaluation in category induction as revealed by visual event-related potentials. Biol. Psychol. 66, 99–128.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kubler, A., Perelmouter, J., Taub, E., Flor, H., 1999. A spelling device for the paralysed. Nature 398, 297–298.

Blankertz, B., Curio, G., Müller, K.-R., 2002. Classifying single trial EEG: towards brain computer interfacing. In: Dietterich, S.B.T.G., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems, vol. 14. MIT Press, Cambridge, MA, pp. 157–164.

Blankertz, B., Müller, K.-R., Curio, G., Vaughan, T.M., Schalk, G., Wolpaw, J.R., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schröder, M., Birbaumer, N., 2004. The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials. IEEE Trans. Biomed. Eng. 51 (6), 1044–1051 (http://ida.first.fraunhofer.de/projects/bci/competition/).

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Discov. 2 (2), 121–167.

Chen, S.S., Donoho, D.L., Saunders, M.A., 1998. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20 (1), 33–61.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, 2nd ed. John Wiley and Sons, New York, USA.

Duin, R.P., 2000. Prtools, A Pattern Recognition Toolbox for Matlab. Pattern Recognition Group, Delft University of Technology. http://www.prtools.org/.

Kalcher, J., Flotzinger, D., Neuper, C., Golly, S., Pfurtscheller, G., 1996. Graz brain–computer interface: II. Toward communication between humans and computers based on online classification of three different EEG patterns. Med. Biol. Eng. Comput. 34, 383–388.

Kubler, A., Kotchoubey, B., Hinterberger, T., Ghanayim, N., Perelmouter, J., Schauer, M., Fritsch, C., Taub, E., Birbaumer, N., 1999. The thought translation device: a neurophysiological approach to communication in total motor paralysis. Exp. Brain Res. 124, 223–232.

Lemm, S., Blankertz, B., Curio, G., Müller, K.-R., 2005. Spatio-spectral filters for improving the classification of single trial EEG. IEEE Trans. Biomed. Eng. 52 (9), 1541–1548.

Müller-Gerking, J., Pfurtscheller, G., Flyvbjerg, H., 1998. Designing optimal spatial filters for single-trial EEG classification in a movement task. Clin. Neurophysiol. 110, 787–798.

Obermaier, B., Neuper, C., Guger, C., Pfurtscheller, G., 2001. Information transfer rate in a five-classes brain–computer interface. IEEE Trans. Neural Syst. Rehabil. Eng. 9 (3), 283–288.

Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., Sajda, P., 2002. Linear spatial integration for single trial detection in encephalography. NeuroImage 17 (1), 223–230.

Pfurtscheller, G., Neuper, C., Flotzinger, D., Pregenzer, M., 1997. EEG-based discrimination between imagination of right and left hand movement. Electroencephalogr. Clin. Neurophysiol. 103 (6), 642–651.

Ramoser, H., Müller-Gerking, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehabil. Eng. 8 (4), 441–446.

Sajda, P., Gerson, A., Müller, K.-R., Blankertz, B., Parra, L., 2003. A data analysis competition to evaluate machine learning algorithms for use in brain–computer interfaces. IEEE Trans. Neural Syst. Rehabil. Eng. 11 (2), 184–185 (http://newton.bme.columbia.edu/competition.htm).

Vidal, J., 1973. Toward direct brain–computer communication. Annu. Rev. Biophys. Bioeng., 157–180.

Wolpaw, J.R., McFarland, D.J., Neat, D.J., Forneris, C.A., 1991. An EEG-based brain–computer interface for cursor control. Clin. Neurophysiol. 78 (3), 252–259.

Wolpaw, J.R., Flotzinger, D., Pfurtscheller, G., McFarland, D.F., 1997. A timing of EEG-based cursor control. Clin. Neurophysiol. 146, 529–538.

Zibulevsky, M., Zeevi, Y.Y., 2002. Extraction of a source from multichannel data using sparse decomposition. Neurocomputing 49 (1–4), 163–173.