IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 1, 2020, accepted May 18, 2020, date of publication May 25, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997116

Robust Motor Imagery Classification Using Sparse Representations and Grouping Structures

VANGELIS P. OIKONOMOU, (Member, IEEE), SPIROS NIKOLOPOULOS, (Member, IEEE), AND IOANNIS KOMPATSIARIS, (Senior Member, IEEE)
CERTH, Information Technologies Institute, 57001 Thessaloniki, Greece

Corresponding author: Vangelis P. Oikonomou (viknmu@iti.gr)

This work was supported by the Innovative Medicines Initiative 2 Joint Undertaking (JU) under Grant 806999.

ABSTRACT The classification of Motor Imagery (MI) tasks constitutes one of the most challenging problems in Brain Computer Interfaces (BCI) mostly due to the varying conditions of its operation. These conditions may vary with respect to the number of electrodes, the time and effort that can be invested by the user for training/calibrating the system prior to its use, as well as the duration or even the type of the imaginary task that is most convenient for the user. Hence, it is desirable to design classification schemes that are not only accurate in terms of the classification output but also robust to changes in the operational conditions. Towards this goal, we propose a new sparse representation classification scheme that extends current sparse representation schemes by exploiting the group sparsity of relevant features. Based on this scheme each test signal is represented as a linear combination of train trials that are further constrained to belong in the same MI class. Our expectation is that this constrained linear combination exploiting the grouping structure of the training data will lead to representations that are more robust to varying operational conditions. Moreover, in order to avoid overfitting and provide a model with good generalization abilities we adopt the bayesian framework and, in particular, the Variational Bayesian Framework since we use a specific approximate posterior to exploit the grouping structure of the data. We have evaluated the proposed algorithm on two MI datasets using electroencephalograms (EEG) that allowed us to simulate different operational conditions like the number of available channels, the number of training trials, the type of MI tasks, as well as the duration of each trial. Results have shown that the proposed method presents state-of-the-art performance against well known classification methods in MI BCI literature.

INDEX TERMS Motor imagery, sparse representation classification, group sparsity, collaborative representation classification.

I. INTRODUCTION

A BCI using EEG signals aims to create a communication channel between the human brain and the computer [1], [2]. An EEG based BCI system could use various components of the EEG signal to achieve its goal, such as P300 [3], Steady State Visual Evoked Potentials (SSVEP) [4], [5] and Motor Imagery (or SensoriMotor) Rhythms [6], [7]. Out of these components special interest have attracted the systems based on motor imagery (MI) due to their endogeneneous nature [2].

The functionality of an MI BCI system relies on event related desynchronization (ERD) and synchronization (ERS) [2]. The imagination of a movement produces contralateral changes in the brain activities of the motor cortex, especially in alpha and beta rhythms [2]. However, the EEG signals during a motor task are high dimensional, noisy, and present high degree of correlation, hence, their direct classification presents great difficulty. Moreover, during an MI experiment, an electrode placed on the scalp measures the signals produced by motor cortex as well as signals from other spatially neighbouring cortical regions. Thus, it is important to isolate the desired signals from other undesired signals, a requirement that has motivated the use of Spatial filters. Among the spatial filtering techniques reported in MI BCI literature, the one based on Common Spatial Patterns (CSP) is the most prominent due to its nice theoretical properties (such as low SNR, dimensionality reduction) and its experimental validation on various different datasets [6], [8].

The CSP algorithm initially computes a set of spatial filters (or a transformation matrix) that are obtained after

The associate editor coordinating the review of this manuscript and approving it for publication was Shagufta Henna.

performing a learning procedure, during which the variance of the spatially filtered signals is maximized for one class (e.g., one mental imagery task) and minimized for the other class. Then, the CSP-based features are extracted by filtering the EEG signals using the learned spatial filters and by computing the variance (or bandpower) of the resulting signals. In further improving this process, a suitable extension of the CSP algorithm using filter banks has been also proposed in [9].

Another important issue in MI BCI is how to design a sophisticated classification scheme that is able to provide good generalization ability for accurate classification. Linear Discriminant Analysis (LDA) is a widely used classifier in BCI research due to its simplicity and its efficiency in discriminating MI tasks [7]. LDA generally provides good performance under the hypothesis that the sample covariance matrices are similar between different classes. However, this might not always be the case for the classification of MI tasks due to the potential of severe noise interference. As a consequence, the overfitting problem is likely to occur, resulting in poor classification performance.

To overcome this issue, an increasing number of classification algorithms using regularization techniques have been employed for the classification of MI tasks in the past decade. One of the most prominent representatives of this category are the Support Vector Machines (SVMs), which adopt a soft margin regularization to achieve good generalization ability. In conjunction with CSP features, SVM provides state-of-the-art performance for MI tasks classification [7], [10], [11]. However, apart from the algorithms using regularization techniques, algorithms based on a bayesian version of LDA (BLDA) have been also proposed [12]. The use of a prior distribution that is inherent in these approaches helps them to avoid overfitting and provide an algorithm with sufficient generalization abilities. Under various circumstances these algorithms have shown better performance than LDA or SVM [12]. Finally, while BLDA variants predict the label of a test trial using a sparse linear combination of its features, the Sparse Representation Classification (SRC) scheme [13], [14] expresses the test trial as a sparse linear combination of the training trials, and its label is determined in terms of the minimum residual norm [13], [14].

A significant challenge in MI BCI research is to design pattern recognition systems that provide accurate performance in various operational conditions. For instance, when the amount of training data items is small, both the feature extraction, e.g., the learning of CSP transformation matrix, and the classifier are not reliable. To attack the above limitation, approaches based on regularization [6], semi-supervised learning [15], session-to-session transfer [6] and subject-to-subject transfer [16] are proposed. All these approaches aim at augmenting the available information for a particular subject by using information from other subjects or other sessions of this particular subject. Similarly, in addition to the limited amount of training trials the performance of the system can be affected by various operational conditions related to the time duration of the trials, the type of CSP filter, as well as the type of imaginary task [8], [17]. Hence, our goal in this work is to use a sparse representation scheme for designing an MI task classifier that can be robust to the aforementioned variations.

Under the sparse representation classification scheme, we represent a test signal as a sparse linear combination of the training trials. However, sparseness alone could be misleading in cases where other additional structures are present in the data. Hence, to increase the classifier's performance further constraints must be incorporated into the model. These constraints can take different forms. One particular form that rises naturally in classification problems is that of a group. In BCI applications a group can be defined with respect to the user [16] or with respect to the class [18]. However, both aforementioned approaches utilize the grouping structure by treating carefully the prior distribution over weights without being concerned about their posterior distribution.

In this work, we present a robust SRC scheme to discriminate the MI tasks in a BCI application under various operational conditions. More specifically, we propose a new Group-based Sparse Representation Scheme by imposing the group structure on the posterior distribution of weights through its factorization in the Variational Bayesian Framework. This is achieved by using a particular approximation of the posterior distribution that takes into account the grouping structure of the data. By imposing grouping constraints on the posterior and sparse constraints on the prior the proposed algorithm is able to select the most important training trials to represent a test trial by taking into account both properties, grouping and sparseness.

The rest of the manuscript is organized as follows. In Section II we review existing SRC approaches and describe how the new SRC scheme differs by defining groups based on a specific factorization of the posterior distribution. Also, in this section we discuss why the CSP algorithm is crucial to our SRC scheme. After that, in Section III we provide information about the EEG datasets that were used in this work to evaluate our method against several competing methods in the literature. Finally, in Section IV we discuss various issues related to the robustness of our work under various operational conditions, while our conclusions and future directions are presented in Section V.

II. METHODOLOGY

A. BASIC SRC SCHEME

Let C be the number of classes and p_c be the number of training EEG trials of class c. The i-th trial from class c is represented by a feature vector, $\mathbf{f}_i^c \in \Re^q, i = 1, \cdots, p_c$. Stacking all feature vectors from the same class into a matrix we obtain a class specific model:

$$\mathbf{X}_c = \left[\mathbf{f}_1^c, \mathbf{f}_2^c, \cdots, \mathbf{f}_{p_c}^c\right] \in \Re^{q \times p_c} \tag{1}$$

Given sufficient training EEG trials for class c, a test EEG trial $\mathbf{y} \in \Re^q$ of the same class will approximately lie in the linear subspace spanning from the training trials:

$$\mathbf{y} = \mathbf{X}_c \boldsymbol{\alpha}_c \tag{2}$$

where $\boldsymbol{\alpha}_c \in \Re^{p_c}$ is a coefficient vector describing the participation of each training trial to the procedure. Initially, we do not know the class of the test trial $\mathbf{y}$ hence we represent it as a linear combination of training samples from all classes:

$$\mathbf{y} = \mathbf{X}\mathbf{w} \tag{3}$$

where $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \cdots\ \mathbf{X}_C] \in \Re^{q \times m}$ is a matrix containing all training EEG trials from all classes, $m = \sum_{c=1}^{C} p_c$ is the number of training EEG trials, and $\mathbf{w}$ is the coefficient vector whose entries are zero expect those of class c, $\mathbf{w} = [0, \cdots, 0, \boldsymbol{\alpha}_c^T, 0, \cdots, 0]^T \in \Re^m$.

In the case where $q < m$ the system of equations $\mathbf{y} = \mathbf{X}\mathbf{w}$ is underdetermined and to obtain a feasible solution we need to place some constraints. A natural approach is to choose constraints based on the ℓ_2-norm, however, this approach does not take into account the structure of our data where most of the coefficients are expected to be zero (in a two classes example 50% of coefficients are expected to be zero). Hence seeking a sparse solution describes better the desired one. This solution can be obtained by the following ℓ_1-minimization problem:

$$\hat{\mathbf{w}} = \arg\max \|\mathbf{w}\|_1 \text{ subject to } \mathbf{X}\mathbf{w} = \mathbf{y} \tag{4}$$

At this point we will not argue about the choice of ℓ_1-norm versus ℓ_0-norm, more information on this subject can be found in more specialized documents [19], [20]. It is suffice to say that under some circumstances the two solutions coincide [19]. Until now, we assumed that Eq. (3) holds exactly, however, in real cases the EEG trials are noisy, hence, a more accurate model must take into account this noise. Now, the model describing the relation between the test EEG trial and the training EEG trials is given by:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e} \tag{5}$$

where $\mathbf{e} \in \Re^q$ is the noise term with bound energy $\|\mathbf{e}\|_2 \leq \epsilon$. Also, the ℓ_1-minimization problem is transformed to:

$$\hat{\mathbf{w}} = \arg\max \|\mathbf{w}\|_1 \text{ subject to } \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2 \leq \epsilon \tag{6}$$

which can be written, with the help of the ℓ_1-regularized formulation [19], [21], as:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \rho\|\mathbf{w}\|_1\} \tag{7}$$

Until now we have discussed how a test trial can be described as a linear combination of training trials. In the following we will discuss how we could use this linear combination to provide a classification rule. In an ideal scenario, the solution $\hat{\mathbf{w}}$ should have non zero coefficients in indices that correspond to training trials that belong to the same class with the test trial. However, this is not the case since EEG is a very noisy and non-stationary signal, thus nonzero coefficients could appear on the indices of other classes. In the literature different approaches have been proposed on how to deal with this [13], [14]. In our study we adopt the approach based on residuals [13]. More specifically, if we let $\delta_c(\cdot) : \Re^m \rightarrow \Re^m$ to be a function that selects the coefficients associated with the class c, while zeroing all irrelevant coefficients, then we can calculate the residuals for each class as: $r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}\delta_c(\hat{\mathbf{w}})\|_2$, $c = 1, \cdots, C$. The class for the given test trial is found by taking the minimum of the residuals $class(\mathbf{y}) = \arg\min_c\{r_c(\mathbf{y})\}$. The overall algorithm is described in Algorithm 1.

We can see that the algorithm contains two basic steps. The first step is related to the minimization problem, while the second step is related to the calculation of residuals. In our study we focus on the minimization problem. A classical approach to solve the above ℓ_1-minimization problem is the Basis Pursuit (BP) algorithm [19]. However, in Compressive Sensing (CS) literature we can found many other solvers, that presents better performance than BP [22]–[24] and they also take into account various other properties of the data such as group sparsity [22], [24], [25]. It is our intention in this work to explore other possibilities for the sparse representation of a given test trial.

In [26] it was argued that it is not only the sparse representation that enhances the accuracy of Face Recognition but also the collaborative representation of a testing sample (face image in this case) with samples from all classes (i.e. the matrix $\mathbf{X}$ contains samples from all classes). To prove their claims, the authors replaced the ℓ_1 norm of the SRC algorithm with the ℓ_2 norm proposing a new algorithm called Collaborative Representation Classification scheme. Also, they observed that the norm of coefficients $\|\mathbf{w}\|_2$ can also bring some discriminative information to the classification, hence, they normalized the residuals with this norm. The overall algorithm is provided in Algorithm 2. Results have shown that the collaborative representation of a testing sample play a significant role to the performance of the classifier.

Algorithm 1 Basic Sparse Representation Classification Scheme [13], [14]

```
Input: Training samples, $\mathbf{X}$, and one test sample, $\mathbf{y}$
  1. Solve the minimization problem:
     $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \rho\|\mathbf{w}\|_1\}$
  2. Calculate the residuals:
     $r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}\delta_c(\hat{\mathbf{w}})\|_2, c = 1, \cdots, C$
Output: $class(\mathbf{y}) = \arg\min_c\{r_c(\mathbf{y})\}$
```

B. GROUP-BASED SRC SCHEME

Both reported algorithms, SRC and CRC, have shown their efficiency in the problem of face recognition. Also, SRC has shown state-of-the-art performance in MI BCI applications [14]. In our work we propose a new representation scheme of MI tasks where a testing EEG trial is represented as a sparse representation of training EEG trials by utilizing the grouping structures of the EEG data. Our intention is to

```
Algorithm 2 Collaborative Representation Classification Scheme [26]
Input: Training samples, $\mathbf{X}$, and one test sample, $\mathbf{y}$
    1. Solve the minimization problem:
    $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \rho\|\mathbf{w}\|_2\}$
    2. Calculate the residuals:
    $r_c(\mathbf{y}) = \frac{\|\mathbf{y}-\mathbf{X}\delta_c(\hat{\mathbf{w}})\|_2}{\delta_c(\|\hat{\mathbf{w}}\|_2)}, c = 1, \cdots, C$
Output: $class(\mathbf{y}) = \arg\min_c\{r_c(\mathbf{y})\}$
```

use both representations, SRC and CRC, by adopting a group structure to represent a testing EEG trial. The group sparsity minimization problem is described by:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \rho\|\mathbf{w}\|_{1,2}\} \tag{8}$$

where $\|\mathbf{w}\|_{1,2} = \sum_{g=1}^{G}\|\mathbf{w}_g\|_2$ and $\mathbf{w}_g$ are the coefficients of the g-th group. To solve the above problem we adopt the method presented in [24]. This method has been proposed to examine the group structure in CS problems and it has shown superior performance compared to similar approaches such as the BP algorithm, see [24]. A short description of group-sparse representation classification scheme is provided in Algorithm 3.

```
Algorithm 3 Group Sparse Representation Classification scheme [18]
Input: Training samples, $\mathbf{X}$, and one test sample, $\mathbf{y}$
    1. Define groups of training trials, $\mathbf{X}$ using class labels
    2. Solve the minimization problem:
    $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \rho\|\mathbf{w}\|_{1,2}\}$
    3. Calculate the residuals:
    $r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}\delta_c(\hat{\mathbf{w}})\|_2, c = 1, \cdots, C$
Output: $class(\mathbf{y}) = \arg\min_c\{r_c(\mathbf{y})\}$
```

The above group - based algorithm has been proposed in [18], where the groups were defined using the class labels or by performing a clustering procedure to define the groups. If we consider the above algorithm under the bayesian framework, we can see that the restrictions are placed a priori on the coefficient vector $\mathbf{w}$, through the prior distribution. However, careful usage of the Variational Bayesian framework will allow us to also use restrictions over $\mathbf{w}$ by exploiting the posterior distribution.

C. SPARSE REPRESENTATION CLASSIFICATION USING A POSTERIORI GROUPING STRUCTURE

In general, prior distributions, besides incorporating prior knowledge into our problem, introduces four significant properties into the model: 1) Avoid overfitting since they restrict parameters to fit completely into the data, 2) Provide generalization capabilities due to the supressness of overfitting, 3) Avoiding numerical instabilities (or model inaccuracies) by placing constraints onto the likelihood, 4) Specific priors, such as sparse priors, favor simpler models to explain the data (Occam's razor).

In our study we adopt sparse priors since we seek sparse representations of EEG trials in order to perform the classification. A useful sparse prior is based on the combination of Normal and Gamma distributions using a hierarchical modeling approach [27], [28]. In our study we use this hierarchical prior but we also add additional parameters to the overall prior. Specific choices of these parameters could produce sparser solutions than the classical case. More specifically, the prior distribution of weights is given by:

$$p(\mathbf{w}|\mathbf{a};\boldsymbol{\lambda}) = \prod_{i=1}^{m}\mathcal{N}(w_i|0, a_i^{-1}\lambda_i^{-1}),$$

where $\mathcal{N}$ is the symbol for Normal (or Gaussian) distribution. Each parameter a_i, which controls the prior distribution of the parameters $\mathbf{w}$, follows a Gamma distribution, so the overall prior over all a_i is a product of Gamma distributions given by: $p(\mathbf{a}) = \prod_{i=1}^{m} Gamma(a_i; b_a, c_a)$. where b_a and c_a is the scale and shape of the Gamma distribution [29], respectively. Furthermore, parameters λ_i are assumed known and deterministic quantities at this point.

The overall precision (inverse variance) β of the noise follows a Gamma distribution: $p(\beta) = Gamma(\beta; b_n, c_n)$, where b_n and c_n is the scale and shape of the Gamma distribution [29], respectively. The usage of Gamma distribution is twofold: first, this distribution is conjugate to the Gaussian distribution, which helps us in the derivation of closed form solutions, and second, it places positivity constraints on the noise precision β and the parameters a_i.

So, the overall prior over model parameters $\{\mathbf{w}, \mathbf{a}, \beta\}$ is given by: $p(\mathbf{w}, \mathbf{a}, \beta; \boldsymbol{\lambda}) = p(\mathbf{w}|\mathbf{a}; \boldsymbol{\lambda})\prod_{i=1}^{m} p(a_i)p(\beta)$. The observation model is given from Eq. (5), hence, the likelihood of the data is given by:

$$p(\mathbf{y}|\mathbf{w}, \beta; \boldsymbol{\lambda}) = \frac{\beta^{\frac{m}{2}}}{(2\pi)^{\frac{m}{2}}} \cdot \exp\left\{-\frac{\beta}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})\right\} \tag{9}$$

We can observe that the true posterior $p(\mathbf{w}, \mathbf{a}, \beta|\mathbf{y}; \boldsymbol{\lambda})$ is not analytically tractable, hence an approximated approach must be utilized in order to find this posterior. In our analysis we adopt the Variational Bayesian (VB) Methodology [29]. In order to apply the VB methodology we need to define an approximate posterior based on one factorization over the parameters $\{\mathbf{w}, \mathbf{a}, \beta\}$. The factorization of posterior serves two goals: first it will provide us with closed form solutions and second it gives us the ability to place additional constraints over coefficients $\mathbf{w}$. A widely used factorization is given by: $q(\mathbf{w}, \mathbf{a}, \beta; \boldsymbol{\lambda}) = q(\mathbf{w}; \boldsymbol{\lambda})\prod_{i=1}^{m} q(a_i)q(\beta)$. Notice that the above factorization makes the coefficients dependent a posteriori. In our approach we choose a different factorization in order to utilize the (assumed) grouping structure of $\mathbf{w}$. We assume that coefficients are independent a posteriori when they belong to different group/classes and dependent when they belong to the same group/class. $q(\mathbf{w}, \mathbf{a}, \beta; \boldsymbol{\lambda}) = \prod_{c=1}^{C} q(\mathbf{w}_c; \boldsymbol{\lambda}) \times \prod_{i=1}^{m} q(a_i) \times q(\beta)$. This factorization gives us the ability to use under the same algorithm two important properties: sparsity and grouping structure. Note here that under this factorization we do not make any assumptions about sparsity between groups similar to [22], [24]. Also, this decomposition does not treat groups in a totally separate fashion, since, the noise is common to all groups/classes.

Applying the VB methodology, and taking into account the above factorization, the following posteriors are obtained:

$$q(\mathbf{w}_c) = \mathcal{N}(\hat{\mathbf{w}}_c, \mathbf{C}_c), \quad c = 1, \cdots, C \tag{10}$$

$$q(\beta) = Gamma(\beta; b', c'), \tag{11}$$

$$q(\mathbf{a}) = \prod_{i=1}^{m} Gamma(a_i; b'_{a_i}, c'_{a_i}), \tag{12}$$

The moments of each distribution are calculated by applying iteratively the following equations until convergence:

$$\mathbf{z}_c = \mathbf{y} - \sum_{j=1, j\neq c}^{C} \mathbf{X}_j \mathbf{w}_j, \quad c = 1, \cdots, C \tag{13}$$

$$\mathbf{C}_c^{(k+1)} = (\hat{\beta}^{(k)} \mathbf{X}_c^T \mathbf{X}_c + \hat{\Lambda}_c^{(k+1)})^{-1}, \quad c = 1, \cdots, C \tag{14}$$

$$\mathbf{w}_c = \mathbf{C}_c^{(k+1)} \hat{\beta} \mathbf{X}_c^T \mathbf{z}_c, \quad c = 1, \cdots, C \tag{15}$$

$$\mathbf{w}^{(k+1)} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \cdots, \mathbf{w}_C^T]^T \tag{16}$$

$$\mathbf{C}_\mathbf{w}^{(k+1)} = \begin{pmatrix} \mathbf{C}_1^{(k+1)} & 0 & \cdots & 0 \\ 0 & \mathbf{C}_2^{(k+1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{C}_C^{(k+1)} \end{pmatrix} \tag{17}$$

$$\frac{1}{b_{a_i}^{(k+1)'}} = \frac{\lambda_i^{(k+1)}}{2}((\hat{w}_i^{(k+1)})^2 + \mathbf{C}_\mathbf{w}^{(k+1)}(i,i)) + \frac{1}{b_a}, \tag{18}$$

$$c_{a_i}^{(k+1)'} = \frac{1}{2} + c_a, \tag{19}$$

$$\hat{a}_i^{(k+1)} = b_{a_i}^{(k+1)'} c_{a_i}^{(k+1)'}, \tag{20}$$

$$\frac{1}{b_\beta^{(k+1)'}} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w}^{(k+1)})^T(\mathbf{y} - \mathbf{X}\mathbf{w}^{(k+1)}) + tr(\mathbf{X}^T\mathbf{X}\mathbf{C}_\mathbf{w}^{(k+1)}) + \frac{1}{b_n}, \tag{21}$$

$$c_\beta^{(k+1)'} = \frac{N}{2} + c_n, \tag{22}$$

$$\hat{\beta}^{(k+1)} = b_\beta^{(k+1)'} c_\beta^{(k+1)'}, \tag{23}$$

In the above equations the matrix $\hat{\Lambda}_c^{(k+1)}$ is a diagonal matrix with $[\hat{a}_i^{(k)} \cdot \lambda_i^{(k+1)}]_c$ in its main diagonal, where $[\cdots]_c$ is a selection operator that selects elements of the vector belonging to class c. The Eqs. (13) - (23) are applied iteratively until convergence. For $\lambda_i^{(k+1)}$ we follow the considerations of [30] and we set them to $\frac{1}{|\hat{w}_i^{(k)}|}$. This assignment of values in parameters $\lambda_i^{(k+1)}$ provides us with sparser solutions than the classical case (i.e., $\lambda_i^{(k+1)} = 1$) [24]. The proposed algorithm (called clsSRC2) can be considered as a variant of Algorithm 1, since, it solves the minimization problem of Algorithm 1 using more sophisticated constraints over the weights $\mathbf{w}$. We say that it is a variant of Algorithm 1 due to the fact that these two algorithms use the same sparse prior over weights. Their differences are related on how they treat the approximate posterior over weights.

D. CONSTRUCTION OF DICTIONARY IN MI CASE AND REPRESENTATION OF EEG TRIALS

In the previous sections, we generally used the term "training EEG trials" without defining it explicitly. In this section we will describe how to extract features for the EEG trials and what represent in our case the training EEG trials. In MI BCI literature, a significant family of features are CSP features [6], [8]. In the following paragraphs we will describe the CSP features extraction method and we will show its connection with CS.

In MI BCIs, for each class we collect a number of multichannel EEG trials. Let us assume that $\mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_N\}$ is a set containing the multichannel EEG trials where each $\mathcal{X}_i, i \in \{1..N\}$ is a matrix of size $M \times P$ containing the samples from one EEG trial, with M being the number of channels, P the number of time samples and N the number of trials. In our analysis, we construct one feature vector for each EEG trial.

The CSP algorithm performs a decomposition of the signal though the matrix $\mathbf{Z}$, which contains the spatial filters. More specifically, this algorithm transforms the EEG signal from the original into a new domain which is occupied by the new channels,

$$\mathcal{X}_i^{(CSP)} = \mathbf{Z}\mathcal{X}_i, \tag{24}$$

where $\mathcal{X}_i^{(CSP)} \in \Re^{2Q \times P}$ is the decomposed "new" EEG trial and $\mathbf{Z} \in \Re^{2Q \times M}$ is the matrix with the spatial filters $\mathbf{z}_i$, $i = 1, \cdots, 2Q$, and Q is the number of pairs of spatial filters. The spatial filters are obtained by maximizing (or extremizing) the following function [6]:

$$J(\mathbf{z}) = \frac{\mathbf{z}^T \mathbf{C}_1 \mathbf{z}}{\mathbf{z}^T \mathbf{C}_2 \mathbf{z}} \tag{25}$$

where $\mathbf{C}_i$ is the covariance matrix of i-th class. The above maximization problem is equal to maximizing $\mathbf{z}^T\mathbf{C}_1\mathbf{z}$ subject to the constraints $\mathbf{z}^T\mathbf{C}_2\mathbf{z} = 1$. The last problem is equivalent to the generalized eigenvalue problem $\mathbf{C}_1\mathbf{z} = \lambda\mathbf{C}_2\mathbf{z}$. So, the spatial filters $\mathbf{z}_i$ are the generalized eigenvectors of the above problem. The matrix $\mathbf{Z}$ is constructed by selecting the Q pairs of eigenvectors corresponding to the largest and smallest eigenvalues respectively. It is worth to note here that in most cases after the application of the CSP algorithm for spatial filtering an additional step is performed in order to extract CSP-related features [6]. Once the spatial filters $\mathbf{z}_i$ are obtained, CSP feature extraction consists in filtering the EEG signals using the $\mathbf{z}_i$ and then computing the resulting signals variance on each "new" channel. Hence, a feature vector $\mathbf{f} = [f_1, f_2, \cdots, f_{2Q}]$ is constructed as: $f_j = \log(\text{var}(\mathcal{X}_i^{(CSP)}(j, :))), j = 1, \cdots, 2Q$.

In order to exploit the full benefits of CSP algorithm we need to acquire precise estimates of covariance matrices. However, this is a difficult procedure since the desired EEG signals are contaminated with noise. Furthermore, the number of EEG trials could be inadequate to provide us a valuable empirical estimate of the covariance. In order to avoid the above problems, regularization approaches are incorporated into the CSP algorithm. The regularization can be performed with respect to the estimation of covariance matrices $\mathbf{C}_i$ or with respect to the objective function $J(\mathbf{z})$. In the first case, the covariance estimates can be obtained as: $\mathbf{C}_i = (1-\gamma)\mathbf{C}_i^e + \gamma\mathbf{C}_i^p$ where $\mathbf{C}_i^e$ is the initial empirical covariance matrix, $\mathbf{C}_i^p$ is a generic prior covariance and γ a user defined regularization parameter. In the second case of regularized CSP, the objective function is modified by adding regularization terms in order to penalize the original solutions. More specifically, the general form of the modified objective function is given by: $J(\mathbf{z}) = \frac{\mathbf{z}^T\mathbf{C}_1\mathbf{z}}{\mathbf{w}^T\mathbf{C}_2\mathbf{z}+P(\mathbf{z})}$, where $P(\mathbf{z})$ is the penalty function. More information on regularized CSP could be found in [31].

When the CSP algorithm is used it is common practice to choose up to three pairs of CSP spatial filters [6], corresponding to the three largest and smallest eigenvalues. Hence, the CSP algorithm could be also seen from a dimensionality reduction perspective. In many cases, the CSP spatial filtering algorithm is applied in conjunction with a filter bank [9]. First, the multichannel EEG trials are bandpass filtered into multiple bands using a filter bank, and then, spatial filtering is performed on each of these bands using the CSP algorithm. The CSP - based features of the previous procedure are called FBCSP features.

In our study, each training EEG trial is represented by a feature vector $\mathbf{f}_i$ containing the CSP features. Furthermore, the dictionary matrix contains feature vectors from two classes. More specifically, the N_L feature vectors corresponding to the imagination of left hand are collected to the matrix: $\mathbf{X}_L = \left[\mathbf{f}_1^L, \mathbf{f}_2^L, \cdots, \mathbf{f}_{N_L}^L\right] \in \Re^{q\times N_L}$ while those of right hand to the matrix: $\mathbf{X}_R = \left[\mathbf{f}_1^R, \mathbf{f}_2^R, \cdots, \mathbf{f}_{N_R}^R\right] \in \Re^{q\times N_R}$ Finally, our dictionary is constructed by concatenating the two matrices: $\mathbf{X} = [\mathbf{X}_L\ \mathbf{X}_R] \in \Re^{q\times m}$. Also, we have the same number of trials of each class hence it holds that $N_L = N_R$ and $m = 2 \times N_L$. In addition, $q = N_F \times (2 \times N_{CSP})$, where N_F is the number of band-pass filters and N_{CSP} the number of pairs of spatial filters.

Current assumptions in the field of CS impose that the dictionary (or measurement matrix in CS field) matrix must have uncorrelated columns. A more formal definition of this is given by the coherence of the dictionary $\mathbf{X}$ [19]:

$$\mu(\mathbf{X}) = \max_{j<k} \frac{|<\mathbf{X}_j, \mathbf{X}_k>|}{\|\mathbf{X}_j\|_2\|\mathbf{X}_k\|_2} \tag{26}$$

where $\mathbf{X}_j$,$\mathbf{X}_k$ denotes columns of the dictionary. When μ is small we say that a dictionary is incoherent. Clearly, the above property can not be defined explicitly in MI dictionary since columns belonging to the same class tend to be coherent. However, in MI BCI applications we are not interested in mutual incoherence between all columns of the dictionary. We are interested in incoherence between columns belonging to different classes since this property will make easier the classification. As shown in [14] the use of the CSP filtering algorithm push the dictionary, constructed by using FBCSP features, into this direction. Hence, the particular dictionary $\mathbf{X}$ possess partially the property of incoherence. Closing this section, we present in Fig. 1 the general diagram on how we

TABLE 1. List of symbols with their explanation.

C	number of classes
p_c	number of training trials of class c
c	an index to represent class c, $c = 1, \cdots, C$
i	index for trial i
q	dimension of a trial vector $(q \times 1)$
$\mathbf{X}_c \in \Re^{q\times p_c}$	matrix contains training trials of class c
$\mathbf{y} \in \Re^q$	test trial
$\boldsymbol{\alpha}_c \in \Re^{p_c}$	coefficient vector describing the participation of each training trial of class c to the procedure
$\mathbf{X} \in \Re^{q\times m}$	matrix containing all trials from all classes
$m = \sum_{c=1}^{C} p_c$	number of all training trials from all classes
$\mathbf{w} \in \Re^m$	coefficients vector
$\mathbf{e} \in \Re^q$	the noise term
ϵ	bounded energy
$\delta_c(\cdot)$	indicator function
$r_c(\cdot)$	residuals function
$\mathbf{a}$	hyperparameters vector
$\boldsymbol{\lambda}$	deterministic parameter vector
a_i	element i of vector $\mathbf{a}$
λ_i	element i of vector $\boldsymbol{\lambda}$
$\mathcal{N}$	symbol for normal distribution
b_a, c_a	parameters of prior distribution $p(\mathbf{a})$
β	precision of noise
b_n, c_n	parameters of the prior $p(\beta)$
$q(\cdot)$	approximate posterior
$\mathbf{w}_c$	coefficient vector for class c
$\hat{\mathbf{w}}_c$	mean of the posterior for class c
$\mathbf{C}_c$	covariance of the posterior for class c
$b^{'}, c^{'}$	parameters of the posterior $q(\beta)$
$b^{'}_{a_i}, c^{'}_{a_i}$	parameters of the posterior $q(a_i)$
$\hat{\boldsymbol{\Lambda}}_c^{(k+1)}$	a diagonal matrix with $[\hat{a}_i^{(k)} \cdot \lambda_i^{(k+1)}]_c$ in its main diagonal
k	iteration number of algorithm
$\mathcal{S}$	set of multichannel EEG trials
$\mathcal{X}_i, i \in \{1..N\}$	matrix of size $M \times P$ containing the samples from one EEG trial
M	the number of channels,
P	the number of time samples
N	the number of trials
$\mathcal{X}_i^{(CSP)} \in \Re^{2Q\times P}$	the decomposed "new" EEG trial using CSP algorithm
$\mathbf{Z} \in \Re^{2Q\times M}$	matrix with the spatial filters
$\mathbf{z}_i, i = 1, \cdots, 2Q$	spatial filters
$J(\mathbf{z})$	objective function to be optimized, related to spatial filtering
Q	the number of pairs of spatial filters
$\mathbf{C}_i$	the covariance matrix of of CSP features of i-th class. Note: this covariance is different for that of coefficients $\mathbf{w}$. It is the covariance of features and not of coefficients
$\mathbf{f}$	CSP feature vector that is constructed as: $f_j = \log(\text{var}(\mathcal{X}_i^{(CSP)}(j,:))), j = 1, \cdots, 2Q.$
N_L	number of feature vectors corresponding to the imagination of left hand
N_R	number of feature vectors corresponding to the imagination of right hand
$\mathbf{X}_L$	matrix of left hand movement
$\mathbf{X}_R$	matrix of right hand movement
N_F	the number of band-pass filters
N_{CSP}	the number of pairs of spatial filters.

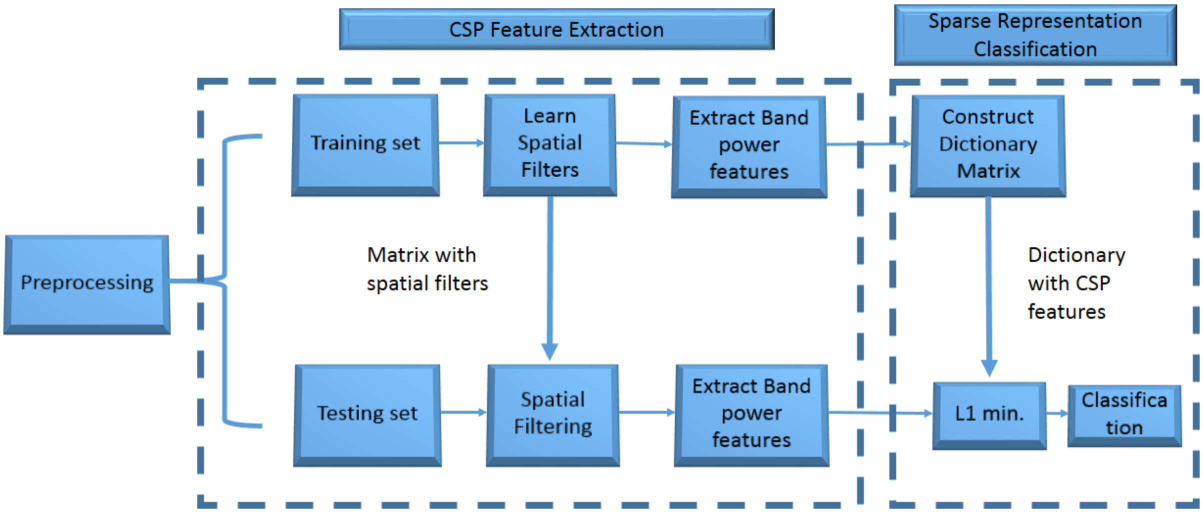
CSP Feature Extraction
Sparse Representation Classification
Preprocessing
Training set
Learn Spatial Filters
Extract Band power features
Construct Dictionary Matrix
Matrix with spatial filters
Dictionary with CSP features
Testing set
Spatial Filtering
Extract Band power features
L1 min.
Classifica tion

FIGURE 1. General workflow of the proposed analysis.

combine the classification algorithms and the CSP filters in order to analyze the EEG trials. Additionally, in Table 1 we provide a summarization of all symbols used in our algorithm as well as their meanings.

III. RESULTS - EXPERIMENTS

A. MI EEG DATASETS

1) GRAZ DATASET B

In our analysis we have used a well known motor Imagery EEG dataset, the BCI competition IV dataset 2b [32]. This dataset consists of EEG data from 9 subjects. For each subject 5 sessions are provided, whereby the first two sessions contain training data without feedback, and the last three sessions were recorded with feedback. Each session consists of six runs, and each run contains 20 trials, 10 trials for each class. Three bipolar recordings (C3, Cz, and C4) were recorded with a sampling frequency of 250 Hz. They were bandpass-filtered between 0.5 Hz and 100 Hz, and a notch filter at 50 Hz was enabled. The placement of the three bipolar recordings (large or small distances, more anterior or posterior) was slightly different for each subject. The electrode position Fz served as the EEG ground. Further information on this dataset can be acquired in [32]. This dataset consisted of two classes, namely the motor imagery of left hand (class 1) and right hand (class 2).

2) MKLab MI DATASET

This data set consists of EEG signals from 10 subjects acquired with the EbNeuro cap (64 channels based on the 10-10 international EEG system with a sampling frequency of 256 Hz). The subjects were sitting in an armchair, watching at the screen monitor placed approximately 0.6m away at eye level. For each subject two sessions were recorded, where the first session contains training data without feedback, and the second session was recorded with feedback. Each session consists of four runs, and each run contains 20 trials, 10 trials for each class. In order to acquire the EEG data the OpenVIBE platform [33] was adopted using the built in scenario of hand motor imagery based BCI. Finally, also, this dataset consisted of two classes, the motor imagery of left hand (class 1) and right hand (class 2).

B. EVALUATION PROTOCOL

For the extraction of EEG features in Graz Dataset B, we have used an approach similar to [9]. More specifically, EEG data from C3, Cz and C4 have been extracted from 3.5 sec to 5.5 sec after the beginning of each MI trial and then a band - pass filter between 8 to 40 Hz has been applied. Following, the EEG data were decomposed into multiple frequency pass bands by using a filter bank with bands: 8-12 Hz, 10-14 Hz, 12-16Hz, . . ., 36-40Hz, a total of 15 bands. Then, in each frequency band we apply the Common Spatial Filters algorithm to extract the CSP features [6]. By selecting the pair of CSP components corresponding to the maximum and minimum eigenvalues, we end-up with 30 features for each trial. Finally, these features are fed into the classifier.

For the extraction of EEG features in MKLab Dataset, we have used an approach similar to the above by taking into account the difference between the two datasets, for example the number of channels in motor cortex. More specifically, EEG data from channels: FC5, FC3, FC1, FC2, FC4, FC6, C5, C3, C1, C2, C4, C6, CP5, CP3, CP1, CP2, CP4, CP6 have been extracted from 3.5 sec to 5.5 sec after the beginning of each MI trial. After that, a band - pass filter between 8 to 40 Hz has been applied. Following, the EEG data were decomposed into multiple frequency pass bands by using a filter bank with bands: 8-12 Hz, 10-14 Hz, 12-16Hz,. . . , 36-40Hz, a total of 15 bands. Then, in each frequency band we apply the Common Spatial Filters algorithm to extract the CSP features. By selecting 1 pair of CSP components corresponding to the maximum and minimum eigenvalues, we end-up with 30 features for each trial.

TABLE 2. Classification accuracy (%) on Graz dataset using classical CSP features.

	LDA	SVM	SRC	clsSRC [18]	cstSRC [18]	SGRM [16]	clsSRC2
1	71.56	74.69	65.00	73.12	70.31	76.30	71.88
2	57.14	55.36	51.43	57.50	55.00	56.00	59.64
3	56.25	55.94	51.56	55.94	60.00	49.20	57.19
4	95.31	94.69	86.56	96.56	95.94	98.20	95.63
5	90.94	90.00	76.88	91.87	90.31	91.10	92.50
6	80.94	82.50	70.63	83.13	81.56	74.80	82.50
7	74.69	77.19	69.37	73.75	75.94	88.30	73.75
8	91.87	88.75	72.81	90.94	90.00	85.40	90.00
9	86.25	85.94	72.50	85.31	83.13	84.90	85.94
Average	78.32	78.34	68.52	78.68	78.02	78.78	78.20

We have compared the proposed approach (clsSRC2) with LDA, SVM, basic SRC scheme and the Group Sparse Representation Classification where the groups are defined in terms: of class labels (clsSRC) [18], of clusters (cstSRC) [18] and of subjects (SGRM) [16]. LDA and SVM are widely used methods in MI BCI applications and they can be considered as baseline methods with respect to the application domain. The Basic SRC scheme has limited usage in MI BCI applications, however, it has been extensively used in face recognition applications and it can be considered as the baseline approach with respect to sparse representation algorithms. The clsSRC method is an extension of the basic SRC scheme taking into account the grouping structure stemming from the class labels of the training EEG trials. The cstSRC method is, also, an extension of the basic SRC scheme, but now, the grouping structure is defined in terms of clusters. More specifically, the groups are created by performing a clustering procedure (using k-means algorithm) in the training data, while the number of clusters are defined by using the Bayesian Information Criterion (BIC). Note here, that groups could contain trials belonging to different classes, in contrast to the clsSRC algorithm, where each group is created by using trials belonging to the same class. Finally, the SGRM method defines the grouping structure with respect to the available subjects on the dataset.

In order to check the effectiveness of the above classifiers we adopt a variety of operational conditions with respect to the training procedure, to the EEG dataset, to the number of training trials and to the learning method of CSP-related features. To validate the classifiers the classical train-test scenario have been used. More specifically, for the GrazB dataset the first three sessions are used for training while the remaining two sessions are used to validate the classifiers/models. For the MKLab dataset, the first session has been used to train the classifiers while the second session has been used for the validation procedure. Note here, that the number of training (and testing) EEG trials between the two datasets is different. Carrying a considerably higher amount of trials the GrazB dataset has been preferred over the MKLab dataset, to validate the effectiveness of the classifiers using a varying number of training trials and a varying duration of the EEG trials. Finally, we examine the behaviour of classifiers with respect to the nature of CSP-related features. In our approach we use two methods to extract the CSP features. In the first approach the related covariance matrices are calculated using the sample covariance matrices. We call the CSP features extracted with the above method classical CSP features. In the second approach the covariance matrix of each class is calculated by adopting a shrinking procedure using Ledoit and Wolf method [6], and, the extracted CSP features are called regularized CSP (RCSP) features.

C. RESULTS

1) TYPE OF CSP FILTERS

In this series of experiments we examine the behaviour of examined algorithms with respect to the family of CSP filters, and hence, CSP features. Two cases have been studied: the Classical CSP features and the Regularized CSP features.

a: CLASSICAL CSP FEATURES

In Table 2 we see the obtained results for each classifiers and for each subject for the Graz dataset. We can see that the best performance is obtained for the clsSRC2 method, even if the difference from clsSRC, cstSRC, SGRM, SVM and LDA is small. However, at this point it is worth to see the difference in accuracy between the five sparse-based representation schemes (ie. SRC, clsSRC, cstSRC, clsSRC2, SGRM). On this dataset we see that the SRC scheme has the worst performance of all methods. Hence, it is necessary to include further constraints into this model in order to obtain a performance similar to LDA and SVM. By using information about the grouping structure of the EEG trials we can obtain accuracy similar to that of SVM and LDA as revealed by the group - based methods (clsSRC, cstSRC, clsSRC2, SGRM) since sparsity along it is not enough to obtain a competitive model. Further information must be incorporated into the model. Also, we can see that the clsSRC2 method presents slightly better performance than the rest group - based methods. Furthermore, we test all methods in another MI BCI dataset. In Table 3 we provide the accuracy of all classifiers for the MKLab dataset. Again the clsSRC2 method provides the best performance. However, in this case, the difference from the other approaches is considerable (ranging from 6% to 3%).

b: REGULARIZED CSP FEATURES

In the second series of experiments, we performed similar experiments to that of previous section but we have used regularized CSP features. In Tables 4 and 5 we provide the

TABLE 3. Classification accuracy (%) on MKLab dataset using classical CSP features.

	LDA	SVM	SRC	clsSRC [18]	cstSRC [18]	clsSRC2
1	58.75	62.50	67.50	70.00	67.50	75.00
2	51.25	48.75	46.25	58.75	50.00	58.75
3	60.00	56.25	53.75	58.75	50.00	53.75
4	60.00	53.75	45.00	58.75	48.75	51.25
5	53.75	50.00	47.50	43.75	51.25	46.25
6	50.00	50.00	51.25	50.00	55.00	50.00
7	46.25	45.00	48.75	47.50	55.00	51.25
8	52.50	51.25	47.50	40.00	53.75	51.25
9	50.00	50.00	52.50	60.00	61.25	67.50
10	68.33	66.67	63.33	68.33	60.00	81.67
Average	55.08	53.41	52.33	55.58	55.25	58.66

TABLE 4. Classification accuracy (%) on Graz dataset using RCSP features.

	LDA	SVM	SRC	clsSRC [18]	cstSRC [18]	clsSRC2
1	71.56	74.38	61.25	71.25	68.75	70.63
2	58.93	57.86	46.79	58.21	56.79	56.79
3	56.56	56.25	54.06	56.56	53.44	58.44
4	95.31	94.69	86.56	95.94	71.25	96.25
5	90.94	89.69	75.62	91.56	80.63	91.25
6	81.87	81.87	69.06	82.19	75.62	81.25
7	74.69	76.25	66.87	74.06	72.19	73.44
8	91.87	89.06	75.31	90.94	78.75	90.63
9	86.56	86.56	68.13	85.62	69.69	86.56
Average	78.69	78.51	67.07	78.48	69.67	78.36

TABLE 5. Classification accuracy (%) on MKLab dataset using RCSP features.

	LDA	SVM	SRC	clsSRC [18]	cstSRC [18]	clsSRC2
1	63.75	66.25	63.75	73.75	70.00	81.25
2	62.50	65.00	51.25	63.75	50.00	66.25
3	68.75	66.25	50.00	57.50	61.25	48.75
4	68.75	75.00	55.00	51.25	55.00	68.75
5	48.75	46.25	47.50	48.75	48.75	47.50
6	50.00	50.00	50.00	50.00	53.75	53.75
7	50.00	53.75	50.00	57.50	45.00	56.25
8	56.25	48.75	52.50	46.25	50.00	46.25
9	46.25	43.75	55.00	50.00	42.50	53.75
10	66.67	73.33	71.67	76.67	55.00	68.33
Average	58.16	58.83	54.66	57.54	53.12	59.08

obtained results in the two datasets. In the case where the GrazB dataset is used the results are similar to those of using classical CSP features. This is expected to some degree since the number of training trials is much larger than the number of features, and the averaging procedure involved during the calculation of the covariance matrices removes considerable part of the noise. However, the situation is different when we use the MKLab dataset. In this case we see that the use of regularized CSP features increases the performance of all classifiers except cstSRC. Concluding these series of experiments, we see that the proposed method (clsSRC2) achieves better results than other SRC-based schemes and presents the most stable behaviour under various circumstances.

2) HOW THE NUMBER OF TRAINING TRIALS AFFECTS THE ALGORITHMS?

In this series of experiments we examine the behaviour of classifiers with respect to the number of training trials. In this experiment the GrazB dataset is used due to its extensive use in the literature and the adequate number of trials. Also, the regularized CSP features were used. Furthermore, we have excluded the cstSRC method from the subsequent analysis due to its limited performance in this kind of features. The obtained results are provided in Fig. 2. Furthermore, statistical analysis, using paired-wise t-tests, was used to investigate the statistical significance of the difference in accuracy between the compared methods. We can observe that the clsSRC2 method presents the best performance compared to others approaches when we have at our disposal a small training EEG dataset. An interesting observation is that the clsSRC2 does not only outperform all other methods when using a small training dataset but exhibit also the smallest decrease in performance compared to the case where the full training dataset is used (see Table 4). Furthermore, the observed differences between the clsSRC2 and other methods are statistically significant for small number of trials (<200), except from the case of 40 trials and against the clsSRC method (see Table 6).

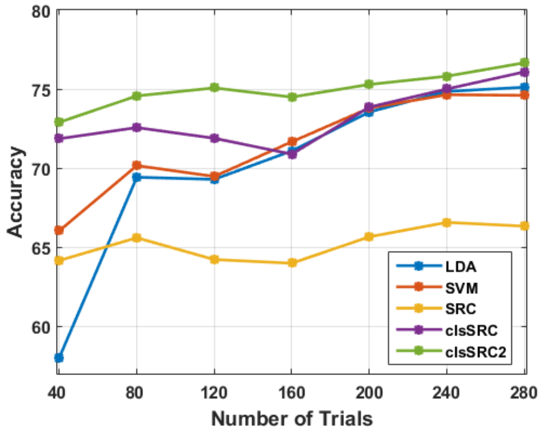

FIGURE 2. Classification accuracy (%) on Graz dataset for various number of training trials.

TABLE 6. Statistical analysis of classification accuracy between the clsSRC2 and the other approaches when various number of training trials are used.

Num of Trials	vs LDA	vs SVM	vs SRC	vs clsSRC
40	p<0.05	p<0.05	p<0.05	p=0.1127
80	p<0.05	p<0.05	p<0.05	p<0.05
120	p<0.05	p<0.05	p<0.05	p<0.05
160	p<0.05	p<0.05	p<0.05	p<0.05
200	p=0.0686	p= 0.1330	p<0.05	p<0.05
240	p=0.2889	p= 0.3215	p<0.05	p=0.3224
280	p=0.2687	p=0.1295	p<0.05	p=0.6612

3) DURATION OF EEG TRIALS

In this Section, we perform experiments with respect to the time duration of EEG trials, using the regularized CSP features from the GrazB dataset. Also, we have excluded the cstSRC method from the analysis due to its limited performance in this kind of features. The average accuracies over all subjects for various time durations of EEG trials are provided in Fig. 3. In addition, statistical analysis, using paired-wise t-tests, was used to investigate the statistical significance

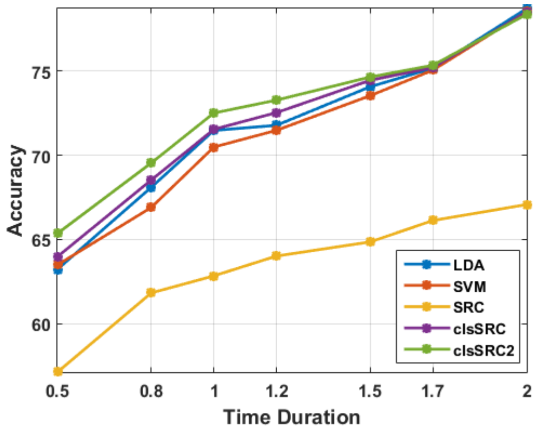

FIGURE 3. Classification accuracy (%) on Graz dataset for various time durations (in secs).

TABLE 7. Statistical analysis of classification accuracy between the clsSRC2 and the other approaches when various trial time duration.

sec	vs LDA	vs SVM	vs SRC	vs clsSRC
0.5	$p<0.05$	$p=0.0561$	$p<0.05$	$p=0.0782$
0.8	$p<0.05$	$p<0.05$	$p<0.05$	$p=0.2232$
1.0	$p=0.1224$	$p<0.05$	$p<0.05$	$p=0.1629$
1.2	$p<0.05$	$p<0.05$	$p<0.05$	$p=0.2505$
1.5	$p=0.4276$	$p=0.1175$	$p<0.05$	$p=0.7685$
1.7	$p=0.8201$	$p=0.6912$	$p<0.05$	$p=0.8589$
2.0	$p=0.4380$	$p=0.8336$	$p<0.05$	$p=0.7297$

of accuracy difference between the compared methods. We can see that all methods need a time duration of 2 secs in order to achieve their maximum accuracy. However, we can observe that for smaller time durations the clsSRC2 method consistently presents the best performance among all methods. More specifically, the clsSRC2 method outperforms all methods when the time duration is smaller than 1.5sec. Furthermore, the clsSRC2 method significantly outperforms the basic SRC scheme in all cases. Additionally, in most cases when the duration is smaller than 1.5sec, the differences between our approach and LDA, SVM and SRC are statistically significant, while, the differences between clsSRC2 and clsSRC are not statistically significant (see Table 7).

IV. DISCUSSION

An important property of the SRC schemes is that for each test sample they adaptively define the structure of the neighbourhood [34]. The number of training trials that are used to describe a test trial could be different for each test trial. This is a very important feature of the algorithm since classical methods such as k-nearest neighbour uses a global fixed parameter to determine the neighbourhoods. The above property gives us the ability not only to weight differently each training trial but to use different number of training trials for describing each test trial.

The clsSRC2 algorithm combines two important properties, grouping structure and sparseness. Each one of these two properties is used in clsSRC and SRC algorithms. More specifically, the SRC algorithm tries to express a test trial as a sparse linear combination of training trials. In this algorithm, sparseness is examined between trials. To further enhance the classification results, clsSRC incorporates into the model a grouping structure on training trials using a group sparse prior where the sparsity is examined between classes. The proposed algorithm (clsSRC2) combines the two properties under the same framework by using sparse prior over weights while at the same time restricts the approximate posterior to have a group structure using the class labels of training trials. As we have seen in our results by incorporating class label information in the SRC-scheme we achieve an overall increase of approximately 10%.

It is worth here to describe how the other methods use the underlying group structure observed in the MI EEG data, from a methodological perspective. In [18] the groups were used to define a group-sparse prior, while the bayesian framework was adopted to find their weights. In [16] the authors use a deterministic model for their SRC-based scheme that incorporates the various properties of the data through deterministic constraints. More specifically, a particular group Lasso - based method was used to estimate the optimal linear representation (or weights). This method has the significant advantage that sparseness is assumed between groups and, also, between weights belonging to the same group. However, a disadvantage of the method is that the model hyperparameters must be defined through a cross-validation procedure. Compared to these works, our proposed scheme differs in the following ways. First, the weights have sparse structure due to priors, second, grouping effects are introduced to the model by using a particular factorization of the posterior distribution, and third, there is no need for a cross-validation procedure to define the model hyperparameters.

We see that when we use the classical case (Section III-C1) to analyze the data all methods present similar performance. However, when we deviate from classical approaches (i.e., smaller number of trials, or different duration of EEG trials) we can observe the difference between the various learning methods involved in the overall procedure. When we use small number of training EEG trials classical approaches such as SVM and LDA deteriorates significantly. Also, the same phenomenon can be observed when the time duration of EEG trials is becoming smaller than 2secs. In all above cases, SRC schemes including additional information to the sparsity (i.e., clsSRC and clsSRC2) have shown more robust behaviour than the basic (or classical) SRC scheme, the LDA and the SVM.

The MKLab MI dataset can be considered a noisy dataset since the users were totally naive to BCI concepts. Furthermore, we haven't performed any subject-specific hyperparameters selection such as those reported in [17], [32]. Also, this particular dataset consists of the minimal number of required sessions, one session to train the BCI system (calibration step) and one session to train the user (feedback step). Evaluating classifiers in such noisy environments represents a great challenge. The results on this dataset have shown the usefulness of the proposed method since is outperforms all other methods in two different situations.

In this work we have presented results under different operational conditions where the robustness of clsSRC-based methods with respect to classification accuracy is validated. More specifically, we have used different MI EEG datasets, various number of training EEG trials and various feature extraction approaches. In most cases, the proposed method performed comparably well to SVM and LDA, and consistently better than SRC.

V. CONCLUSION

In this paper, a novel classification algorithm is proposed to classify MI tasks in BCI applications. The algorithm exploits the sparse representation of a test signal with the help of training signals. Besides sparsity, the algorithm exploits the dictionary structure. More specifically, the dictionary is built using CSP features that tend to produce an incoherent dictionary due to its properties. Extensive experiments in two MI EEG datasets, one publicly available and one obtained in our laboratory, have shown the usefulness of the proposed algorithm. More specifically, the comparison of the proposed algorithm with basic SRC, LDA and SVM has shown that the proposed algorithm provides us with superior performance under various operational conditions. In the future, we intend to study how SRC schemes could be modified to develop adaptive classification schemes. These schemes are very useful in BCI applications due to the time varying nature of EEG signal. In addition to the above, it is our intention to study and extend our method in cases relevant to multi-task learning and transfer learning. The above task-based learning strategies will allow us to design efficient general purpose MI BCI systems.

ACKNOWLEDGMENT

The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and SOFTWARE AG.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] B. Graimann, B. Allison, and G. Pfurtscheller, *Brain-Computer Interfaces: A Gentle Introduction*. Berlin, Germany: Springer, 2010, ch. 1.

[3] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, and K. Diserens, "An efficient P300-based brain–computer interface for disabled subjects," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 115–125, Jan. 2008.

[4] V. P. Oikonomou, A. Maronidis, G. Liaros, S. Nikolopoulos, and I. Kompatsiaris, "Sparse Bayesian learning for subject independent classification with application to SSVEP-BCI," in *Proc. 8th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2017, pp. 600–604.

[5] V. P. Oikonomou, S. Nikolopoulos, and I. Kompatsiaris, "A Bayesian multiple kernel learning algorithm for SSVEP BCI detection," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 1990–2001, Sep. 2019.

[6] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[7] V. P. Oikonomou, K. Georgiadis, G. Liaros, S. Nikolopoulos, and I. Kompatsiaris, "A comparison study on EEG signal processing techniques using motor imagery EEG data," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 781–786.

[8] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Dec. 2008.

[9] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, Mar. 2012.

[10] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 4, pp. 317–326, Aug. 2008.

[11] N. Brodu, F. Lotte, and A. Lecuyer, "Comparative study of band-power extraction techniques for motor imagery classification," in *Proc. IEEE Symp. Comput. Intell., Cognit. Algorithms, Mind, Brain (CCMB)*, Apr. 2011, pp. 1–6.

[12] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, "Sparse Bayesian classification of EEG for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2256–2267, Nov. 2016.

[13] J. Wright, A. Y. Yang, A. Ganesh, S. Shankar Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[14] Y. Shin, S. Lee, J. Lee, and H.-N. Lee, "Sparse representation-based classification scheme for motor imagery-based brain–computer interface systems," *J. Neural Eng.*, vol. 9, no. 5, Aug. 2012, Art. no. 056002.

[15] J. Meng, X. Sheng, D. Zhang, and X. Zhu, "Improved semisupervised adaptation for a small training dataset in the brain–computer interface," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1461–1472, Jul. 2014.

[16] Y. Jiao, Y. Zhang, X. Chen, E. Yin, J. Jin, X. Wang, and A. Cichocki, "Sparse group representation model for motor imagery EEG classification," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 631–641, Mar. 2019.

[17] B. Blankertz, F. Losch, M. Krauledat, G. Dornhege, G. Curio, and K.-R. Müller, "The berlin brain-computer interface: Accurate performance from first-session in BCI-naive subjects," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 10, pp. 2452–2462, Oct. 2008.

[18] V. P. Oikonomou, S. Nikolopoulos, and I. Kompatsiaris, "Motor imagery classification via clustered-group sparse representation," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2019, pp. 1–5.

[19] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing, Theory and Applications* Y. C. Eldar and G. Kutyniok, eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 1–64.

[20] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[21] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Boston, MA, USA: Birkhaüser, 2013.

[22] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2906–2921, Jun. 2014.

[23] L. He, H. Chen, and L. Carin, "Tree-structured compressive sensing with variational Bayesian analysis," *IEEE Signal Process. Lett.*, vol. 17, no. 3, pp. 233–236, Mar. 2010.

[24] V. P. Oikonomou, S. Nikolopoulos, and I. Kompatsiaris, "A novel compressive sensing scheme under the variational Bayesian framework," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–4.

[25] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[26] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 471–478.

[27] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[28] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[29] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer, 2007.

[30] G. Deng, "Iterative learning algorithms for linear Gaussian observation models," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2286–2297, Aug. 2004.

[31] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[32] R. Leeb, C. Brunnera, G. R. Müller-Putz, A. Schloogl, and G. Pfurtscheller. (2008). *Bci Competition 2008—Graz Data Set B*. [Online]. Available: https://lampx.tugraz.at/ bci/database/002-2014/description.pdf

[33] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, ''OpenViBE: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments,'' *Presence, Teleoperators Virtual Environ.*, vol. 19, no. 1, pp. 35–53, Feb. 2010.

[34] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, ''Learning with ℓ^1-graph for image analysis,'' *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

VANGELIS P. OIKONOMOU (Member, IEEE) received the Dipl., M.Sc., and Ph.D. degrees in computer science from the University of Ioannina, Ioannina, Greece, in 2000, 2003, and 2010, respectively. He is currently a Postdoctoral Researcher with the Centre for Research and Technology-Hellas (CERTH), Thessaloniki, Greece. His research interests include Bayesian methods, machine learning, medical image processing, biomedical signal processing, and brain computer interfaces.

SPIROS NIKOLOPOULOS (Member, IEEE) received the Diploma degree in computer engineering and informatics and the M.Sc. degree in computer science and technology from the University of Patras, Greece, in 2002 and 2004, respectively, and the Ph.D. degree in semantic multimedia analysis using knowledge and context from the Queen Mary University of London, in 2012. He is currently a Postdoctoral Research Fellow with the Centre for Research and Technology Hellas (CERTH), Information Technologies Institute (ITI). His scientific work has been published in peer-reviewed journals, international conferences, and book chapters. His research interests include advanced human machine interfaces, semantic multimedia analysis, and visual and augmented reality.

IOANNIS (YIANNIS) KOMPATSIARIS (Senior Member, IEEE) is currently a Research Director with CERTH-ITI, the Head of the Multimedia Knowledge and Social Media Analytics Laboratory, and a Deputy Director of the Institute. He is the coauthor of 129 articles in refereed journals, 46 book chapters, eight patents, and more than 420 papers in international conferences. Since 2001, he has participated in 59 National and European research programs including direct collaboration with industry, in 15 of which he has been the Project Coordinator and in 41 the Principal Investigator. His research interests include multimedia, big data and social media analytics, semantics, human computer interfaces (AR and BCI), eHealth, security and culture applications. He has been the co-organizer of various international conferences and workshops and has served as a Regular Reviewer, an Associate, and a Guest Editor for a number of journals and conferences, currently being an Associate Editor of the IEEE Transactions on Image Processing and *Big Data Journal*. He is a member of ACM.