

FINDING EEG SPACE-TIME-SCALE LOCALIZED FEATURES USING MATRIX-BASED PENALIZED DISCRIMINANT ANALYSIS

Juliette Spinnato^{*†}

Marie-Christine Roubaud^{*}

Boris Burle[†]

Bruno Torr sani^{*}

^{*} Aix-Marseille Universit , CNRS, Centrale Marseille, I2M, UMR 7373, Marseille, France

[†] Aix-Marseille Universit , CNRS, LNC, UMR 7291, Marseille, France

ABSTRACT

This paper proposes a new method for constructing and selecting of discriminant space-time-scale features for electroencephalogram (EEG) signal classification, suitable for Error Related Potentials (ErrP) detection in brain-computer interface (BCI). The method rests on a new variant of matrix-variate Linear Discriminant Analysis (LDA), and differs from previously proposed approaches in mainly three ways. First, a discrete wavelet expansion is introduced for mapping time-courses to time-scale coefficients, yielding time-scale localized features. Second, the matrix-variate LDA is modified in such a way that it yields an interesting duality property, that makes interpretation easier. Third, a space penalization is introduced using a surface Laplacian, so as to enforce spatial smoothness. The proposed approaches, termed D-MLDA and D-MPDA are tested on EEG signals, with the goal of detecting ErrP. Numerical results show that D-MPDA outperforms D-MLDA and other matrix-variate LDA techniques. In addition this method produces relevant features for interpretation in ErrP signals.

Index Terms— Matrix-based LDA, Multi-sensor signals, Discrete wavelet transforms, EEG features

1. INTRODUCTION

Analysis, interpretation and classification of EEG signals (for instance in BCI contexts) often rest on features, which are supposed to characterize brain activities of interest. Many types of features have been proposed in the literature, that depend on the type of data and the target application. In particular, spatio-temporal features [2, 12] and spatio-spectral features [5, 9, 10] allow for simple interpretation while yielding good classification performances.

In this work we are interested in designing discriminant features that can be interpreted in terms of Space-Time-Scale (STS) localization. More precisely, we seek features that can be associated with well defined localization on the scalp, as well as time-scale localization for the corresponding time course. We address this problem by introducing a new approach that combines discrete wavelet representation for the time course with LDA. Since the observations are multi-sensor signals, LDA has to be adapted to this situation. A standard practice is to turn matrices into vectors through row or column concatenation, and use classical LDA. However this approach called 1DLDA often leads to difficult covariance estimation problems. Based on an additional covariance separability assumption, we rather resort to a matrix-variate version of LDA, which extends previously proposed matrix LDA techniques. In contrast to

the MLDA technique proposed in [9], our approach features an interesting duality property similar to the duality of principal component analysis. To tackle the curse of dimensionality and enforce spatial feature smoothness, we also introduce a graph laplacian based regularization. As a result, we obtain STS-localized features, that can be interpreted and exploited further for classification purpose.

We apply this methodology to an EEG signal processing problem, namely the single trial analysis and classification of error potentials. Features are learnt from a training set, and we show that the most relevant ones can be interpreted in terms of error potentials. Their relevance is also asserted by classification results on a test set.

2. MODEL SETUP

2.1. Data structure and Preprocessing

Since our target application is EEG signal analysis, we consider multi-sensor signals $F \in \mathbb{R}^{L \times J}$, recorded as time series of length L on J sensors. Time samples being strongly correlated, a dimension reduction is performed via discrete wavelet transform followed by a projection onto a fixed subspace of most relevant wavelet and scaling coefficients. As a result, each instance of the considered signal takes the form of a matrix $X \in \mathbb{R}^{K \times J}$ wavelet and scaling coefficients (which are therefore doubly labeled by time and scale), $K = K_w + K_s$ being the number of retained wavelet and scaling coefficients. We limit here to unitary discrete wavelet transforms, which are then invertible. More details on the pre-processing are given in Section 4 below.

EEG measurements involving several repetitions (hereafter called trials), several data matrices $X_i \in \mathbb{R}^{K \times J}$ are obtained, which yields three-way arrays.

2.2. A matrix-variate Gaussian mixture model and separability assumption

In each class c , the EEG single-trial $X_i^c \in \mathbb{R}^{K \times J}$ is considered as a realization of a random matrix X^c with a matrix variate Gaussian distribution with class-dependent mean matrix μ^c and a class-independent covariance matrix Σ . As in [9] we assume that the covariance between two elements of the matrix X^c can be decomposed into an inter-row and inter-column covariances. Therefore, the following separability assumption is introduced:

$$(1) \quad \Sigma = \Sigma^L \otimes \Sigma^R,$$

where $\Sigma^L \in \mathbb{R}^{K \times K}$ is the left covariance (corresponding to rows i.e. time-scale coefficients), $\Sigma^R \in \mathbb{R}^{J \times J}$ the right covariance (columns i.e. sensors) and \otimes denotes the Kronecker product [6].

In the sequel we shall use the following notations. Given a matrix A , we denote by A' the transpose of A and $\text{vec}(A)$ the vector

This work was supported by the ANR project CO-ADAPT (ANR-09-EMER-002-05).

obtained by vertical concatenation of columns of A . We shall use the following definition [6]:

Definition 1 Let $\mu \in \mathbb{R}^{K \times J}$ and let $\Sigma^L \in \mathbb{R}^{J \times J}$ and $\Sigma^R \in \mathbb{R}^{K \times K}$ be symmetric positive definite matrices. The $K \times J$ random matrix X is said to have a matrix variate normal distribution with mean μ and covariance $\Sigma^L \otimes \Sigma^R$ if the vector $\text{vec}(X')$ is distributed according to

$$(2) \quad \text{vec}(X') \sim \mathcal{N}_{KJ}(\text{vec}(\mu'), \Sigma^L \otimes \Sigma^R).$$

We write $X \sim \mathcal{N}_{K,J}(\mu, \Sigma^L \otimes \Sigma^R)$.

Let us note that if the right covariance is equal to the identity matrix I_J , the columns of the centered matrix $X - \mu$ are independent and identically distributed (i.i.d.). Therefore in this case the K rows of $X - \mu$ are K i.i.d. Gaussian vectors with i.i.d. components and the covariance Σ^L can be estimated by empirical covariance. A similar remark can be done on the rows in the case where $\Sigma^L = I_k$. In general the columns (respectively rows) are not i.i.d. but we can get back to the previous situation as follows. Let $R \in \mathbb{R}^{J \times J}$ and $L \in \mathbb{R}^{K \times K}$ be square roots of covariances $(\Sigma^R)^{-\frac{1}{2}}$ and $(\Sigma^L)^{-\frac{1}{2}}$ respectively, multiplying the rows of X by R and the columns by L' we obtain:

$$(3) \quad XR \sim \mathcal{N}_{K,J}(\mu R, \Sigma^L \otimes I_J),$$

$$(4) \quad L'X \sim \mathcal{N}_{K,J}(L'\mu, I_K \otimes \Sigma^R).$$

In the following each test trial $X_i \in \mathbb{R}^{K \times J}$ is modeled as a realization of a matrix variate Gaussian mixture model (MGMM) of two components Gaussian matrix variate distributions with class-dependent mean matrix μ^c and a common covariance matrix $\Sigma^L \otimes \Sigma^R$. We thus write

$$(5) \quad p(X|\theta) = \sum_{c=1}^2 p_c f_{K,J}(X|\mu^c, \Sigma^L \otimes \Sigma^R),$$

where θ represents all the model parameters, p_1 and p_2 are the mixture weights, $f_{K,J}(X|\mu^c, \Sigma^L \otimes \Sigma^R)$, $c = 1, 2$, are the component matrix variate Gaussian densities and the other notations are the same as before.

For simplicity, we will take $p_c = \frac{N_c}{N}$ in the following, where $N = N_1 + N_2$ is the size of the training data set and N_c the number of trials in each class c .

3. A PENALIZED MATRIX-BASED LDA

3.1. Matrix-based LDA and row-column duality (D-MLDA)

The method is based on decoupling the analysis on row and column through Kronecker product. We search the most discriminant linear combinations of rows (resp. of columns) by optimizing suitable Fisher criterion. This boils down to find the respective eigenvectors of $(S_W^L)^{-1} S_B^L \mathbb{R}^{K \times K}$ and $(S_W^R)^{-1} S_B^R \in \mathbb{R}^{J \times J}$ where S_W^L , S_B^L and S_W^R , S_B^R are the within-class and between-class matrices in the column space (\mathbb{R}^J) and in the row space (\mathbb{R}^K) respectively.

From Section 2 the within-class matrices S_W^L and S_W^R correspond of the maximum-likelihood (ML) estimates of Σ^L and Σ^R and must satisfy

$$(6) \quad S_W^L = \frac{1}{JN} \sum_{c=1}^2 \sum_{i=1}^{N_c} (X_i^c - M^c)(S_W^R)^{-1} (X_i^c - M^c)',$$

$$(7) \quad S_W^R = \frac{1}{KN} \sum_{c=1}^2 \sum_{i=1}^{N_c} (X_i^c - M^c)' (S_W^L)^{-1} (X_i^c - M^c),$$

with

$$(8) \quad M^c = \frac{1}{N^c} \sum_{i=1}^{N^c} X_i^c.$$

The within-class matrices are calculated through two iterative steps as in [9].

With the same metrics as above, we define the between-class matrices S_B^L and S_B^R as follows:

$$(9) \quad S_B^L = \frac{N_1 N_2}{N^2} (M^1 - M^2) (S_W^R)^{-1} (M^1 - M^2)',$$

$$(10) \quad S_B^R = \frac{N_1 N_2}{N^2} (M^1 - M^2)' (S_W^L)^{-1} (M^1 - M^2).$$

An interesting consequence of these definitions is the following duality property between row and column spaces.

Proposition 1 a) The matrices $(S_W^L)^{-1} S_B^L$ and $(S_W^R)^{-1} S_B^R$ have the same Q non-zero eigenvalues where $Q \leq \min(K, J)$.

b) If $u \in \mathbb{R}^K$ is an eigenvector of $(S_W^L)^{-1} S_B^L$ associated to the eigenvalue $\lambda > 0$, then $v = (S_W^R)^{-1} (M^1 - M^2)' u$ is an eigenvector of $(S_W^R)^{-1} S_B^R$ associated to the same eigenvalue λ .

Let us denote by $U \in \mathbb{R}^{K \times Q}$ and $V \in \mathbb{R}^{J \times Q}$ respectively the eigenvector matrices of $(S_W^L)^{-1} S_B^L$ and $(S_W^R)^{-1} S_B^R$ sorted in decreasing order of associated non-zero eigenvalues. From Proposition 1, the following relation holds:

$$(11) \quad V = (S_W^R)^{-1} (M^1 - M^2)' U.$$

Remark 1 The difference with the MLDA method proposed by Mahanta et al. [9] is the calculation of the left and right between-class matrices. We introduce a column withening through the matrix $\hat{R} = (S_W^R)^{-\frac{1}{2}}$ to find the discriminant linear combinations of rows (time-scale filters) and respectively a row withening through the matrix $\hat{L} = (S_W^L)^{-\frac{1}{2}}$ to obtain discriminant linear combinations of columns (spatial filters). This yields duality between the two analyses and simplifies the calculation of eigenvectors. For this reason this method will be called D-MLDA.

3.2. Penalized D-MLDA (D-MPDA)

In some situation it can be useful to introduce prior information into the feature construction. A relevant example of such prior information is spatial smoothness of the features which will be used in the application below. Following the penalized discriminant analysis proposed by Hastie et al. [7], this can be done by replacing the spatial within-class covariance S_W^R by a regularized version

$$(12) \quad \widetilde{S}_W^R = S_W^R + \lambda \Omega,$$

where Ω encodes the prior spatial information and λ is an hyperparameter. After this replacement the procedure is exactly the same as before and yields penalized STS features. These features obviously depend on the hyperparameter λ . The determination of λ requires an additional validation step. In the application developed below the latter is estimated through cross-validation. This penalized version of D-MLDA method, will be called D-MPDA.

3.3. Application to classification

The above analysis naturally leads to introduce the following space-time-scale (STS) features.

We recall that $U \in \mathbb{R}^{K \times Q}$ and $V \in \mathbb{R}^{J \times Q}$ are respectively the eigenvector matrices of $(S_W^L)^{-1} S_B^L$ and $(S_W^R)^{-1} S_B^R$ whose columns are generically denoted by u_k and v_j .

Definition 2 a) For each trial i the corresponding STS feature matrix is

$$(13) \quad Y_i = U' X_i V.$$

The matrix elements $Y_i(k, j)$ are the features.

b) For a given feature $Y_i(k, j)$, its spatial and time-scale signatures $\sigma_i^{(s)}$ and $\sigma_i^{(ts)}$ are defined by

$$(14) \quad \sigma_i^{(s)} = u_k' X_i \in \mathbb{R}^J \quad \text{and} \quad \sigma_i^{(ts)} = X_i v_j \in \mathbb{R}^K.$$

The time-signature $\sigma_i^{(t)}$ is obtained by inverse DWT from the time-scale signature.

From Definitions 1 and 2, the STS feature matrix Y_i^c of any training trial is a realization of a matrix variate normal distribution :

$$(15) \quad Y_i^c \sim \mathcal{N}_{K,J} \left(U' \mu^c V, \tilde{S}_W \right), \text{ where}$$

$$(16) \quad \tilde{S}_W = (U \otimes V)' \Sigma (U \otimes V).$$

The model for STS feature matrix Y^i of any test trial is a matrix variate Gaussian mixture model. After vectorization $y_i = \text{vec}(Y_i')$, we end up with a standard multivariate Gaussian mixture decision problem in the case of equal covariances. It is known that in this situation LDA provides a Bayes optimal classification rule.

Remark 2 (Feature selection) All coefficients from the STS feature matrix are not equally relevant, and various feature selection strategies could be considered, for example selecting those coefficients whose associated eigenvalues are the largest, as proposed in [9]. In the Application section below, we limited ourselves to visualizing signatures associated to the features with maximally different averages on the training dataset.

4. APPLICATION: ERROR POTENTIALS ANALYSIS AND CLASSIFICATION

We now evaluate D-MLDA and D-MPDA in terms of features selection and classification accuracy and compare to other approaches. We investigate Error-related EEG potentials from a previously published dataset [13]. We first focus on the interpretability of the selected features, before comparing correct classification rate (CCR) performances using D-MLDA and D-MPDA with CCR performances of 1DLDA and MLDA.

4.1. EEG dataset and preprocessing

The dataset includes ten participants who had to respond with either the left hand or right hand as fast as possible to visual stimulations that were conflicting, hence inducing erroneous responses (ErrP or errors). The signal was recorded using 64 scalp electrodes (Active-two system, Biosemi, Amsterdam) and acquired at a sampling rate of 1024 Hz. After preprocessing (artifacts removal...), data were downsampled to 256 Hz. The selected trials were segmented into epochs of 800 ms from -400 ms to $+400$ ms, where zero corresponds

to the participant's response. Two classes of trials are considered: errors and correct responses.

As described in Section 2, time-domain reduction was achieved using DWT (for which we used the Wavelab package [3]). DWT was performed on zero-padded signals with Daubechies filter D6 [11] on 5 decomposition levels. Dimension reduction was achieved by removing the first 3 (smallest scale) decomposition levels, which is equivalent to an orthogonal projection onto the complementary subspace [1]. Finally, remaining coefficients that are strongly affected by boundary effects (resulting from zero-padding) were not accounted for in the statistical analysis. After this step, $K = 24$ coefficients are selected: $K_s = 6$ scaling coefficients and $K_w = 6 + 12$ wavelet coefficients for levels 4 and 5 respectively.

4.2. Laplacian penalization

In the penalized version of the method, we enforce spatial smoothness by adding to the within-class matrix a Laplacian penalization, based upon a discrete surface Laplacian [8]. The graph Laplacian penalization is calculated from a first-order neighbourhood matrix of the electrodes as follows. Given two electrodes e_i and e_j , $i, j = 1, \dots, 64$, the Laplacian matrix Ω is defined by

$$(17) \quad \Omega_{ij} = \begin{cases} \deg(e_i) & \text{if } i = j \\ -1 & \text{if } e_i \text{ and } e_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

where $\deg(e_i)$ is the number of e_i 's first order adjacent electrodes.

The determination of the hyperparameter λ in (12) requires an additional validation step. For the sake of simplicity λ is chosen using a cross-validation procedure as the optimizer of the CCR on a validation data set.

4.3. Space-time-scale features analysis

To analyse and interpret the features provided by D-MLDA and D-MPDA, we compare the corresponding signatures with the known relevant spatial and temporal components of the Error-related EEG potentials as discussed in the literature (see [13], [4] and references therein). The latter are mainly fronto-central brain activities that are expected to peak around 100 ms after participant's response.

To this end we analysed the features obtained from the two classes (Error and Correct responses). For the sake of simplicity we limit the current discussion to D-MPDA and to mean features namely features averaged over trials within each class (and therefore the corresponding mean signatures). We illustrate in Fig. 1 Space-Time-Scale (STS) mean features for Subject A. The two most discriminant features, corresponding respectively to $(k, j) = (1, 1)$ and $(k, j) = (8, 1)$ were selected, following the rule described in Remark 2. For both features $j = 1$, therefore the temporal signatures are the same, whereas spatial signatures differ ($k = 1$ and $k = 8$). Let us note that this time signature corresponds to the time course of the optimal spatial filter (linear combinations of electrodes) averaged over trials within each class.

The first selected feature $(1, 1)$ corresponds to a scaling coefficient in the wavelet decomposition and contains low frequency information of the signal. For each class its spatial signature (see Definition 2) is displayed in Fig. 1 (topographies 1 and 3 on the top panel). In agreement with a large literature on error potentials, the latter largely loads on fronto-central electrodes while the temporal signature presents a significant negative activity shortly after the response (time 0 on bottom panel of Fig. 1) followed by positive one. The second selected feature corresponds to an higher frequency

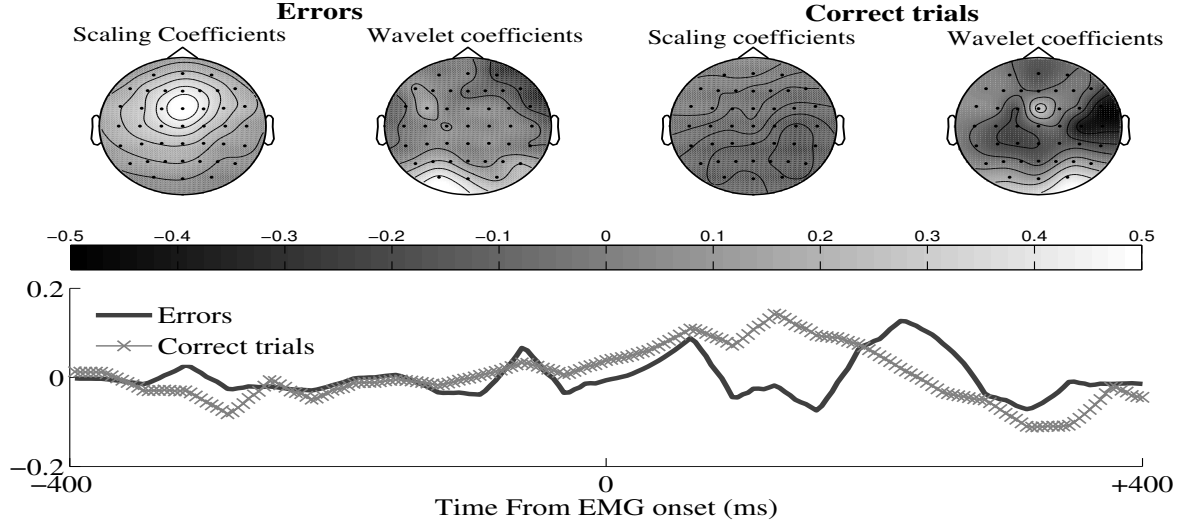


Fig. 1. Display of the two relevant STS features from the mean feature matrix in both EEG classes (errors and correct trials). Top: mean spatial signatures of the selected features represented in gray scale (from black to white). Each pattern corresponds to the top view of a head. Nose and ears give the direction of the head. The dots on the head give the location for each electrode. Bottom: mean temporal signatures of the selected features for Errors and Correct trials; 0 corresponds to the electromyographic (EMG) onset that triggered the participant's response.

	Methods			
	1DLDA	MLDA	D-MLDA	D-MPDA
Subject A	82.0	53.5	77.5	82.9
Subject B	69.4	49.2	71.8	69.7
Subject C	78.5	52.7	79.2	80.3
Subject D	72.4	50.8	71.9	73.6
Subject E	81.4	52.0	83.2	84.8

Table 1. Correct classification rate (%CCR) for the studied methods averaged over 100 iterations. The bold numbers are the highest CCR for each subject.

activity (within the frequency band 8 – 16 Hz). For each class its spatial signature is mainly located on the back of the head. Both spatial localization and frequency band seem unusual for interpretation in terms of error potential. Nevertheless we stress that discarding this feature significantly degrades classification results which tends to indicate the relevance of this feature.

4.4. Classifier performance

We now compare the performance of D-MLDA and D-MPDA with 1DLDA and MLDA for which the same preprocessing had been done (see Subsection 2.1). We implemented the following protocol: classifiers were trained over 80 trials (40 in each class) and methods are tested on the same amount of trials for the test step (30 in each class). For D-MPDA the hyperparameter λ had been selected on a validation data set (20 trials in each class). It must be noted that we are in a small dataset situation: some subjects turned out to produce few errors during the experiment (less than 70) and were excluded from the present study. We performed 100 iterations of cross-validation on 5 subjects.

Results are displayed in Table 1. The proposed D-MPDA and D-MLDA methods outperform very significantly the existing MLDA

method, and to a smaller extent 1DLDA. For subjects A, C, D and E, the penalized version significantly improves D-MLDA. It should be noted that MLDA, initially implemented for spatio-spectral features, does not provide consistent results applied to a space-time-scale feature extraction problem. Our results seem to indicate that the introduced spatial penalization improves classification results. Further validations on larger data sets will be necessary to draw definite conclusion.

5. CONCLUSIONS AND PERSPECTIVES

A new approach for constructing discriminant space-time-scale features for two-class multi-sensor signals has been proposed and studied in this paper. The approach combines discrete wavelet representation for representing time courses, with an original matrix-variate linear discriminant analysis, involving spatial penalization. This produces features defined in a space-time-scale representation domain, which have been shown to be relevant when applied to the problem of detecting error related potentials (ErrP) in EEG signals.

The current contribution essentially focused on the description of the main features of the proposed approach and the application to ErrP detection. Several aspects are still to be investigated further, among which the choice of relevant features (for which we stuck here to a very simple strategy, as our goal was mainly illustrative), the tuning of the number of needed trials in training and validation datasets, and the actual influence of the wavelet filters. In this respect, let us stress that the choice of time-scale (i.e. wavelet) representation was motivated by the ErrP application, other types of data and problems will presumably call for different types of transformations, such as Fourier or short time Fourier transformations, as discussed in [10].

Acknowledgements

This work was supported by the ANR project CO-ADAPT (ANR-09-EMER-002-05).

6. REFERENCES

- [1] F. Abramovich, T. C. Baillye, and T. Sapatinas. Wavelet analysis and its statistical applications. *Royal Statistical Society : Series D (The statistician)*, 49(1):1–29, 2000.
- [2] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller. Single-trial analysis and classification of ERP components - a tutorial. *Neuroimage*, 56:814–825, 2011.
- [3] J. B. Buckheit and D. L. Donoho. *WaveLab and Reproducible Research*. Springer-Verlag, 1995.
- [4] B. Dal Seno, M. Matteucci, and L. Mainardi. Online detection of p300 and error potentials in a bci speller. *Comput Intell Neurosci*, 2010:307254, 2010.
- [5] J. Farquhar. A linear feature space for simultaneous learning of spatio-spectral filters in BCI. *Neural Networks*, 22:1278–1285, 2009.
- [6] A. Gupta and D. Nagar. *Matrix variate distributions*. Chapman & Hall/CRC, Boca Raton, Florida, USA, 2000.
- [7] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.
- [8] G. Huiskamp. Difference formulas for the surface laplacian on a triangulated surface. *Journal of Computational Physics*, 95:477–496, 1991.
- [9] M. S. Mahanta, A. S. Aghaei, and K. N. Plataniotis. A bayes optimal matrix-variate LDA for extraction of spatio-spectral features from EEG signals. In *International conference of the IEEE Engineering in Medicine and Biology society (EMBC)*, 2012.
- [10] M. S. Mahanta, A. S. Aghaei, and K. N. Plataniotis. Regularized LDA based on separable scatter matrices for classification of spatio-spectral EEG patterns. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 2013.
- [11] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.
- [12] Y. Meir-Hasson, A. Zhdanov, T. Hendler, and N. Intrator. *Biomedical Engineering Systems and Technologies*, chapter A Robust and Efficient Spatio-Temporal Feature Selection for Interpretation of EEG Single Trials, pages 219–232. Springer-Verlag, 2013.
- [13] C. Roger, C. G. Bénar, F. Vidal, T. Hasbroucq, and B. Burle. Rostral cingulate zone and correct response monitoring: ICA and source localization evidences for the unicity of correct- and error-negativities. *NeuroImage*, 51:391–403, February 2010.