Early Detection of Alzheimer's Disease by Blind Source Separation, Time Frequency Representation, and Bump Modeling of EEG Signals

François Vialatte[1], Andrzej Cichocki[2], Gérard Dreyfus[1], Toshimitsu Musha[3], Sergei L. Shishkin[2], and Rémi Gervais[4]

[1] ESPCI (ParisTech), Laboratoire d'Electronique (CNRS UMR 7084), 10 rue Vauquelin, 75005 Paris, France
{francois.vialatte, gerard.dreyfus}@espci.fr
[2] BSI RIKEN ABSP Lab 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
cia@brain.riken.jp
[3] Brain Functions Laboratory Inc., KSP Building E211, Sakado, Takatsu Kawasaki-shi, Kanagawa, 213-0012, Japan
musha@bfl.co.jp
[4] Equipe Neurobiologie de la Mémoire Olfactive, Institut des Sciences Cognitives (UMR 5015 CNRS UCB), 67 Boulevard Pinel, 69675 Bron Cedex, France
gervais@isc.cnrs.fr

Abstract. The early detection Alzheimer's disease (AD) is an important challenge. In this paper, we propose a novel method for early detection of AD using electroencephalographic (EEG) recordings: first a blind source separation algorithm is applied to extract the most significant spatio-temporal components; these components are subsequently wavelet transformed; the resulting time-frequency representation is approximated by sparse "bump modeling"; finally, reliable and discriminant features are selected by orthogonal forward regression and the random probe method. These features are fed to a simple neural network classifier. The method was applied to EEG recorded in patients with Mild Cognitive Impairment (MCI) who later developed AD, and in age-matched controls. This method leads to a substantially improved performance (93% correctly classified, with improved sensitivity and specificity) over classification results previously published on the same set of data. The method is expected to be applicable to a wide variety of EEG classification problems.

1 Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder. Since the number of individuals with AD is expected to increase in the near future, early diagnosis and effective treatment of AD are critical issues in neurophysiological research [1], [2]. Finding a computational method for early identification of patients who are to progress towards Alzheimer's disease (before onset of AD), but do not exhibit any clinical signs of AD at the time of the test, is thus an important challenge. Furthermore, an early detection method should be inexpensive, in order to allow mass screening of elderly patients [1]–[6]. Electroencephalography (EEG) is one of the most promising candidates in that respect.

W. Duch et al. (Eds.): ICANN 2005, LNCS 3696, pp. 683 – 692, 2005.

Due to the high complexity and variability of EEG signals, early detection of AD from EEG recordings requires the development of efficient signal processing tools [2]. In [5], Blind Source Separation (BSS) was first applied for these purposes, while standard methods were used for feature extraction and classification. In the present paper, we propose a multistage procedure employing blind source separation for filtering/enhancement of EEG, time frequency representation, subsequent bump modeling for feature generation and dimensionality reduction, and statistical feature selection. We show that it provides a further improvement in classification of AD patients and healthy subjects as compared to similar classification results obtained previously [1], [5] on the same data set.

2 Methods

2.1 Blind Source Separation for Signal Filtering

According to the currently prevailing view of EEG signal processing, a signal can be modeled as a linear mixture of a finite number of brain sources, with additive noise [5,6]. Therefore, blind source separation techniques can be used advantageously for decomposing raw EEG data to brain signal subspace and noise subspace.

In [5], the AMUSE (Algorithm for Multiple Unknown Signals Extraction [7], [8], [9], [10]) algorithm was used in order to select the five significant components of the signal that had the best linear predictability. That algorithm belongs to the group of second-order-statistics spatio-temporal decorrelation (SOS-STD) blind source separation algorithms. It relies on the idea that the estimated components should be spatio-temporally decorrelated, and be less complex (i.e., have better linear predictability) than any mixture of those sources. Therefore, the components are ranked in order of decreasing singular values of a time-delayed covariance matrix. As in PCA (Principal Component Analysis), and unlike in many ICA algorithms, all components estimated by AMUSE are uniquely defined (i.e., any run of algorithms on the same data will always produce the same components) and consistently ranked.

The algorithm can be considered as two consecutive PCAs: first, PCA is applied to input data; secondly, PCA (SVD) is applied to the time-delayed covariance matrix of the results of the previous PCA. In the first step, standard or robust prewhitening (sphering) is applied as a linear transformation

$$\mathbf{z}(t) = \mathbf{Q}\,\mathbf{x}(t) \tag{1}$$

where $\mathbf{Q} = \mathbf{R}_x^{-\frac{1}{2}}$, $\mathbf{R}_x$ is the standard covariance matrix $\mathbf{R}_x = \mathrm{E}\left\{\mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t)\right\}$ and $\mathbf{x}(t)$ is a vector of observed data at time t. Next, SVD is applied to a time-delayed covariance matrix of pre-whitened data:

$$\mathbf{R}_z = \mathrm{E}\left\{\mathbf{z}(t)\mathbf{z}^{\mathrm{T}}(t-1)\right\} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with decreasing singular values and **U, V** are matrices of eigenvectors. Then, a demixing (separating) matrix is estimated as:

$$\mathbf{W} = \hat{\mathbf{A}}^{-1} = \mathbf{U}^{\mathrm{T}}\mathbf{Q} \tag{3}$$

The algorithm is much faster than the vast majority of BSS/ICA algorithms (its computation time depends essentially on the duration of the PCA procedure) and is very easy to use, because no parameters are required. It is implemented as a part of package "ICALAB for signal processing" [11], freely available online. However, the algorithm is quite sensitive to sensor (measurement) noise; therefore, alternative ICA/BSS algorithms with suitable ranking of components can be considered in further studies [7,11].

2.2 Database

The "at-risk" state for AD is commonly referred to as Mild Cognitive Impairment (MCI) [3]. In the course of a clinical study [1], patients who complained of memory impairment only, but had no apparent loss in general cognitive, behavioral, or functional status, were recruited. Both patients and controls underwent general medical, neurological, psychiatric, and neuroimaging (SPECT, CT and MRI) investigation for more accurate diagnosis. EEG was recorded from all patients and controls, within one month of study inception; the present analysis made use of EEG recorded from the patients who later developed AD, and age-matched controls. Electrodes were located on 21 sites according to the 10-20 system, with the reference electrode on the right ear-lobe. Sampling rate was 200 Hz, analog filter bandpass 0.5-250 Hz. In [5], the first continuous artifact-free 20 s interval of each recording were used to create two datasets:

- the MCI set, featuring 22 EEG recordings of elderly patients matching the criteria of mild cognitive impairment, who developed AD within one year and a half;
- a control set, featuring 38 recordings from age-matched family members of the patients.

In the present paper, the same pre-processing method as in [5] was used on the same 60 recordings, as a baseline for assessing the efficiency of the detection obtained by the present method: a database (hereinafter referred to as *D*) containing those five components was generated (five top ranked components obtained using AMUSE).

2.3 Time-Frequency Maps and Bump Modeling for Feature Generation

In order to obtain a compact representation of the signals of *D*, suitable for automatic discrimination of MCI patients from control individuals, the signals were first analyzed in the time-frequency domain by a wavelet transformation, and the resulting time-frequency maps were modeled by bumps [12], as described below.

2.3.1 Wavelet Transformation and Time-Frequency Map Generation

EEG signals were first transformed to time-frequency maps using wavelets (see [13] for details). Complex Morlet wavelets [14] are appropriate for time-frequency analysis of electroencephalographic signals ([15], [16], [17], [18]). Complex Morlet wavelets $w(t)$ of Gaussian shape in time (deviation σ_t) are defined as:

$$w(t) = A \exp\left(-t^2 / 2\sigma_t^2\right) \exp\left(2i\pi ft\right) \tag{4}$$

where σ_t and f are appropriately chosen parameters; they cannot be chosen independently, since the product $\sigma_t f$ determines the number of periods that are present in the wavelet. In the present investigation, the wavelet family defined by $2\pi\sigma_t f = 7$ was chosen, as described in [15].

The signals present in database *D* were wavelet-transformed in the frequency range 1.5 to 31.5 Hz, discretized in 0.25 Hz frequency bins.

2.3.2 Bump Modeling

The bump modeling technique [12] is a 2-dimensional generalization of the Gaussian mesa function modeling technique that was initially designed for one-dimensional signals (electrocardiogram analysis- ECG) [19], [20]. In the present study, it was used for extracting information from the time-frequency maps. In previous investigations, it was also successfully applied to the analysis of local field potential signals, gathered from electrophysiological (invasive) measurements [21],[12]; the present paper reports the first application of bump modeling to surface EEG signals.

The main idea of this method is to approximate a time-frequency map with a set of predefined elementary parameterized functions called bumps (non-overlapping or overlapping); therefore, the map is represented by the set of parameters of the bumps, which is a very sparse encoding of the map, resulting in information compression rates that range from one hundred to one thousand (further details are given in [12], [19], [20]).

The algorithm performs the following steps on the time-frequency maps (after appropriate normalization [12]):

- (i) window the map in order to define the zones to be modeled (those windows form a set of overlapping sub-areas of the map),
- (ii) find the window that contains the maximum amount of energy,
- (iii) adapt a bump φ_b to the selected zone, and withdraw it from the original map. The parameters of the bumps are computed using the BFGS algorithm [22] in order to minimize the cost function *C* defined by:

$$C = \frac{1}{2} \sum_{t,f \in W} \left(z_{ft} - \varphi_b(f,t) \right)^2 \tag{5}$$

where the summation runs on all pixels within the window *W*, z_{ft} are the time-frequency coefficients at time *t* and frequency *f*, and $\varphi_b(f,t)$ is value of the bump function at time *t* and frequency *f*;

- (iv) if the amount of information modeled by the bumps reaches a threshold, stop; else return to (iii).

Half ellipsoids were found to be the most appropriate bump functions for the present application (Figure 2 shows a typical example of bump modeling of the time-frequency map of an EEG recording). Each bump is described by 5 parameters: its coordinates on the map (2 parameters), its amplitude (one parameter) and the lengths of its axes (2 parameters). Half ellipsoids (Figure 1) are defined by:

$$\varphi_b(f,t) = a\sqrt{1-v} \quad \text{for } 0 \le v \le 1$$
$$\varphi_b(f,t) = 0 \quad \text{for } v > 1 \tag{6}$$

where $v = \left(e_f^2 + e_t^2\right)$ with $e_f = \left(f - \mu_f\right)/ l_f$ and $e_t = \left(t - \mu_t\right)/ l_t$. μ_f and μ_t are the coordinates of the centre of the ellipsoid, l_f and l_t are the half-lengths of the principal axes, a is the amplitude of the function, t is the time and f the frequency.

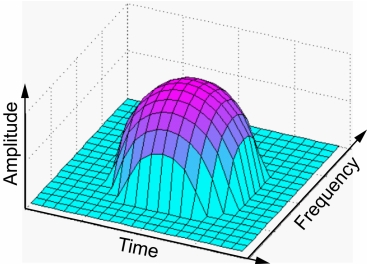

Fig. 1. Half ellipsoid function

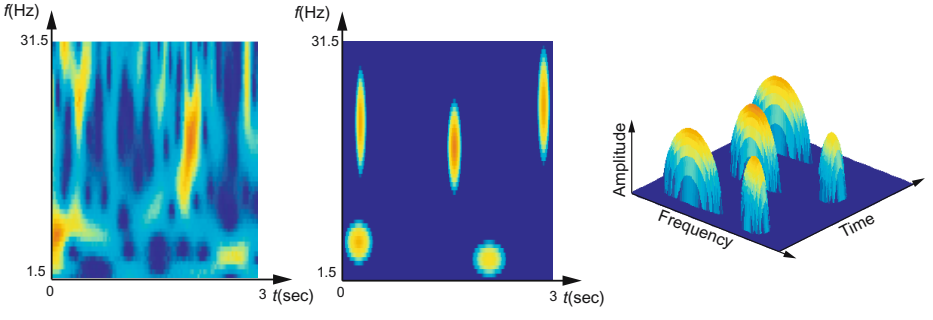

Fig. 2. Left: normalized time-frequency map of the first ICA source of an EEG recording (Control set); middle and right: 2D and 3D bump modeling of the map

After bump modeling, the parameters of the bumps are candidate features for classification. Although the model is sparse, feature selection is mandatory because of the small size of the data set.

2.4 Feature Selection

After bump modeling, the signals under investigation are represented by the set of parameters that describe the bumps. Within that set, an even more compact representation was sought, based on expert knowledge on the frequency sub-bands of interest. For ease of comparison of the results, the boundaries for five standard EEG sub-bands were defined as in the previous study [5] of the same data set: θ (3.5-7.5 Hz), α_1 (7.5-9.5 Hz), α_2 (9.5-12.5 Hz), β_1 (12.5-17.5Hz) and β_2 (17.5-25Hz). The following features were defined and computed for each sub-band:

- F_1: the number of bumps,
- F'_1: the number of high-amplitude bumps (normalized amplitude > 0.7),
- F_2: the sum of the amplitudes of the bumps present,
- F'_2: the sum of the amplitudes of the high-amplitude bumps present,
- F_3: the maximal amplitude of the bumps present.

Two groups of candidate features were defined: group A contains $\{F_1, F'_1, F_2, F'_2\}$ and group B contains $\{F_1, F'_1, F_3\}$ only. Thus, either 3 or 4 features were computed for each sub-band, depending on the group of features under consideration. Therefore, for database D (5 time-frequency maps), the number of candidate features N_f was either 75 or 100. Since the number of candidate features was still too large given the number of examples in the database (only 60), feature selection was performed by orthogonal forward regression (OFR) algorithm [23, 24] and the random probe method.

First, the candidate features are ranked in order of decreasing relevance by OFR. OFR operates in observation space, i.e. in a space whose dimension is equal to the number of observations (equal to 60 in the present work). In that space, the quantity to be modeled, and the candidate features, are represented by vectors denoted by $\mathbf{y}$ (desired outputs) and $\mathbf{u}_i$, $i = 1$ to N_f (inputs). The OFR algorithm performs the following steps:

- (i) compute the angle Θ_i between each candidate feature $\mathbf{u}_i$ and the quantity to be modeled $\mathbf{y}$ and select the candidate feature $\mathbf{u}_j$ that has the smallest angle with $\mathbf{y}$, i.e. the candidate feature that is most correlated to $\mathbf{y}$:

$$\mathbf{u}_j = \arg\max_i \{\cos^2(\Theta_i)\} \tag{7}$$

- (ii) project $\mathbf{y}$ and all the remaining candidate features onto the null space of the selected feature;

The above two steps can be iterated in subspaces of decreasing dimensions until all candidate features are ranked. Subsequently, in order to select the optimal number of features from their ranked list, the random probe method [24] is applied. One hundred "probes", i.e. realizations of random variables, are computed and appended to the feature set. A risk level P is defined [25], which corresponds to the risk that a feature might be kept although, given the available data, it might be less relevant than the probe. The following steps are performed iteratively:

- (i) obtain a candidate feature from OFR,
- (ii) compute the value of the cumulative distribution function of the rank of the probe for the rank of the candidate feature,
- (iii) if that value is smaller than the risk, select the feature and return to (i);
- (iv) else, discard the candidate feature under consideration and terminate.

In the present case, an ensemble feature ranking method [25] was used: 60 subsets were built by iteratively removing one example from the database. OFR and the random probe method were then applied to those subsets. The overall distribution of features, and the average number N_k of selected features were computed; finally, the N_k overall best features were selected.

3 Results

Each dataset was used for training and validating a neural network classifier (multi-layer perceptron model, see for instance [26]). The generalization performance was estimated using the leave-one-out cross-validation method [27] which was also used under the name of "jackknifing" in the previous study of the data set [5]. The best results, shown in Table 1, were obtained with linear classifiers (no hidden layer).

The method used in [5] to asses BSS performances was applied with a simple PCA for comparison: with the five first PCA components back-projected, relative spectral

Table 1. Number of subjects correctly and incorrectly classified by neural network models, using *D*, depending on the feature group (*A* contains $\{F_1, F'_1, F_2, F'_2\}$ and *B* contains $\{F_1, F'_1, F_3\}$). Results were obtained using the leave-one-out cross-validation method, validation set results are presented below. *P* is the risk of a false positive feature, as defined in the text.

Datasets	Misclassified		Correctly classified %		
	MCI $N = 22$	Controls $N = 38$	MCI $N = 22$	Controls $N = 38$	All $N = 60$
A group → $F = 12$, $P = 12\%$ Components 1,2,3 and 5 +Bumps	2	2	91.0	94.7	93.3
A group → $F = 6$, $P = 9\%$ Components 1-5 +Bumps	4	4	81.8	88.6	86.7
B group → $F = 11$, $P = 10\%$ Components 1-5 + Bumps	2	3	91.0	92.1	91.7
Previous study, best results [5] (without bumps)	6	6	72.7	84.2	80.0
PCA components 1-5 (without bumps)	13	10	40.9	73.7	61.7

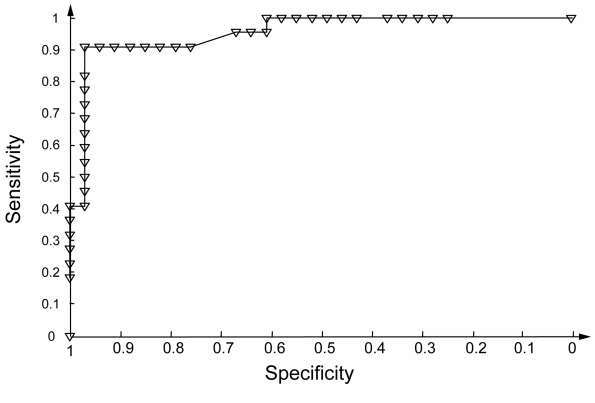

Fig. 3. R.O.C curve for the best classification results, obtained using a neural network on data set *D*, with components 1-3 and 5 found by the AMUSE algorithm (but without component 4)

powers of signals were computed by dividing the power in δ (1.5- 3.5 Hz), θ (3.5-7.5 Hz), α_1 (7.5-9.5 Hz), α_2 (9.5-12.5 Hz), β_1 (12.5-17.5Hz) and β_2 (17.5-25Hz) bands by the power in the 1.5-25 Hz band; those values were subsequently normalized using the transformation $\log \frac{p}{1-p}$, where p is the relative spectral power; finally the band power values were averaged over all 21 channels (see [5] for details). PCA did not exhibit good performance compared to BSS (Table 1, last two rows).The best results were obtained with the group of candidate features *A*, with components 1-3 and 5 found by the AMUSE algorithm, but without component 4 (component 4 was removed before OFR and the random probe method, R.O.C. curve is represented in Figure 3). For comparison, the fourth row of the Table reports previous results [5] obtained with the same EEG recordings with a different representation.

4 Discussion

In the present paper, we reported the first application of blind source separation combined with time frequency representation and sparse bump modeling to the automatic classification of EEG data for early detection of Alzheimer's disease. The developed method was applied to recordings that had been analyzed previously [5] with standard feature extraction and classification methods. With respect to that previous analysis, a substantial improvement was achieved, the overall correct classification rate being raised from 80% to 93% (sensitivity 91.0% and specificity 94.7%).

The task was the discrimination of EEG recordings of normal individuals from EEG recordings of patients who developed Alzheimer's disease one year and a half later. Therefore, the present study provides exciting prospects for early mass detection of the disease. The method is very cheap as compared to PET, SPECT and fMRI, requiring only a 21-channel EEG apparatus. Note that short intervals (20 seconds) of artifact-free recording of spontaneous EEG was already sufficient for high accuracy of classification.

PCA showed much poorer results than AMUSE algorithm, which demonstrates the significance of AMUSE (or more generally, BSS/ICA algorithms with suitable ranking and clustering) for EEG filtering/enhancement. Furthermore, sparse bump modeling appeared to be a valuable tool for compressing information contained in EEG time-frequency maps. Amplitude variations and bursts of EEG oscillations are highly related to the brain state dynamics [28]. Bump modeling can provide a good approximation of time-frequency maps; since it models appropriately important features of EEG oscillations, it is a promising tool for compact feature extraction, as demonstrated in the present paper.

Although our preliminary results are quite promising, a full validation of the method requires investigating more extensive databases. Furthermore, there is presumably a lot of information present in the recordings that is not yet exploited, such as the dynamics of the bumps and the brain functional connectivity. This will be the subject of future research.

References

1. Musha, T., Asada, T., Yamashita, F., Kinoshita, T., Chen, Z., Matsuda, H., Masatake, U., Shankle, W.R., A new EEG method for estimating cortical neuronal impairment that is sensitive to early stage Alzheimer's disease. Clinical Neurophysiology, 2002, 113(7): 1052-1058.
2. Jeong, J., EEG dynamics in patients with Alzheimer's disease. Clinical Neurophysiology, 2004, 115:1490-1505.
3. DeKosky, S.T., Marek, K., Looking backward to move forward: early detection of neurodegenerative disorders. Science, 2003, 302(5646):830-834.
4. Nestor, P.J., Scheltens, P., Hodges J.R., Advances in the early detection of Alzheimer's disease. Nature Medicine, 2004, 10 Suppl.:S34-41. Review.
5. Cichocki, A., Shishkin, S.L., Musha, T., Leonowicz, Z., Asada, T., Kurachi, T., EEG filtering based on blind source separation (BSS) for early detection of Alzheimer's disease. Clinical Neurophysiology, 2005, 116(3):729-737.
6. Cichocki, A., Blind Signal Processing Methods for Analyzing Multichannel Brain Signals. International Journal of Bioelectromagnetism, 2004, Vol. 6, N°1.
7. Cichocki, A., Amari, S., Adaptative Blind Signal and Image Processing: Learning Algorithms and Applications. New York, NY: Wiley, 2003.
8. Tong, L., Soon, V., Huang, Y.F., Liu R., Indeterminacy and identifiability of blind identification. IEEE Transactions CAS, 1991, 38:499-509.
9. Tong, L. Inouye, Y., Liu, R., Waveform-preserving blind estimation of multiple independent sources. IEEE Transactions on Signal Processing, 1993, 41(7):2461-2470.
10. Szupiluk, R., Cichocki, A., Blind signal separation using second order statistics. Proceedings of SPETO, 2001, 485-488.
11. Cichocki, A., Amari, S., Siwek, K., Tanaka, T., et al. ICALAB toolboxes [available online at http://www.bsp.brain.riken.jp/ICALAB]
12. Vialatte, F., Martin, C., Dubois, R., Quenet, B., Gervais R., Dreyfus G. A machine learning approach to the analysis of time-frequency maps, and its application to neural dynamics. Neural Networks (submitted).
13. Poularikas, A.D., The transforms and applications handbook. CRC Press, 1996.
14. Kronland-Martinet, R., Morlet, J., Grossmann, A., The wavelet Transform, in Expert Systems and Pattern Analysis, C.H Chen Edt, World Scientific, 1987, pp 97-126.
15. Tallon-Baudry, C., Bertrand, O., Delpuech, C., Pernier, J., Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. Journal of Neuroscience, 1996, 16:4240-4249.
16. Ohara, S., Crone, N.E., Weiss, N., Lenz, F.A., Attention to a painful cutaneous laser stimulus modulates electrocorticographic event-related desychronization in humans. Clinical Neurophysiology, 2004, 115:1641-1652.
17. Caplan, J.B., Madsen, J.R., Raghavachari, S., Kahana, M.J., Distinct patterns of brain oscillations underlie two basic parameters of human maze learning . Journal of Neurophysiology, 2001, 86:368-380.
18. Düzel, E., Habib, R., Schott, B., Schoenfeld, A., Lobaugh, N., McIntosh, A.R., Scholz, M., Heinze, H.J., A multivariate, spatiotemporal analysis of electromagnetic time-frequency data of recognition memory. Neuroimage, 2003, 18:185-197.
19. Dubois, R., Application de nouvelles méthodes d'apprentissage à la détection précoce d'anomalies en électrocardiographie. PhD thesis, Université Pierre et Marie Curie - Paris VI, 2004. [Available from http://www.neurones.espci.fr/Francais.Docs/dossier_recherche/bibliographie/theses_soutenues.htm#ancre41560]

20. Dubois, R., Quenet, B., Faisandier, Y., Dreyfus, G., Building meaningful representations in knowledge-driven nonlinear modeling. Neurocomputing, , 2005 (in print).
21. Vialatte, F., Martin, C., Ravel, N., Quenet, B., Dreyfus, G., Gervais, R., Oscillatory activity, behaviour and memory, new approaches for LFP signal analysis. 35 th annual general meeting of the European brain and behaviour society, 17-20 September 2003, Barcelona, Spain – Acta Neurobiologiae Experimentalis, Vol. 63, supplement 2003.
22. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., Numerical Recipes in C: The Art of Scientific Computing, 425 - 430. 1992, Cambridge Univ. Press, New York.
23. Guyon, I., Elisseef, A., An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 2003, 3:1157-1182.
24. Chen, S., Billings, S. A. and Luo, W., Orthogonal least squares methods and their application to non-linear system identification. International Journal of Control, 1989, 50:1873-1896.
25. Jong, K., Marchiori, E., Sebag, M., Ensemble Learning with Evolutionary Computation: Application to Feature Ranking. 8th International Conference on Parallel Problem Solving from Nature (PPSN). Lecture Notes in Computer Science, 2004, 1133-1142. Springer.
26. Stoppiglia, H., Dreyfus, G., Dubois, R., Oussar, Y., Ranking a Random Feature for Variable and Feature Selection. Journal of Machine Learning Research, 2003, 3: 1399-1414.
27. Haykin, S., Neural Networks – a comprehensive foundation. 2° edition, 1999, Prentice Hall.
28. Stone, M., Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society: Series B, 1974, 36:111–147.
29. Klimesch, W., EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Research Reviews 1999, 29: 169-95.