Kernel Alignment for Identifying Objective Criteria from Brain MEG Recordings in Schizophrenia

Mario Martín[1], Javier Béjar[1], Gennaro Espósito[1], Neus Català[2], Ulises Cortés[1], Ferran Viñas[3], Josep Tarragó[3], Emilio Rojo[3], and Rafal Nowak[4]

[1]Knowledge Engineering and Machine Learning Group. Universitat Politècnica de Catalunya
[2]TALP Research Center. Universitat Politècnica de Catalunya
[3]Hospital Benito Menni CASM
[4]Clínica Teknón

Abstract

The current wide access to data from different neuroimaging techniques has permitted to obtain data to explore the possibility of finding objective criteria that can be used for diagnostic purposes. In order to decide which features of the data are relevant for the diagnostic task, we present in this paper a simple method for feature selection based on kernel alignment with the ideal kernel in support vector machines (SVM).

The method presented shows state-of-the-art performance while being more efficient than other methods for feature selection in SVM. It is also less prone to overfitting due to the properties of the alignment measure. All these abilities are essential in neuroimaging study, where the number of features representing recordings is usually very large compared with the number of recordings.

The method has been applied to a dataset in order to determine objective criteria for the diagnosis of schizophrenia. The dataset analyzed has been obtained from multichannel magnetoencephalogram (MEG) recordings, corresponding to the recordings during the performance of a mismatch negativity (MMN) auditory task by a set of schizophrenia patients and a control group. All signal frequency bands are analyzed (from δ (1-4Hz) to high frequency γ (60-200Hz)) and the signal correlations among the different sensors for these frequencies are used as features.

1 Introduction

According to [Institute of Mental Health, 2013], schizophrenia affects approximately 1% of the world population. Despite the huge development of several techniques to analyze brain's structure and function (*i.e.* EEG, fMRI, MEG, MRI, PET, SPECT), no specific biomarkers are available for a diagnostic purpose. While there is vast evidence of anatomical changes and abnormalities in neurotransmitter systems in schizophrenia, psychiatry research in recent years has focused in functional disconnectivity of cortical circuits as a core feature of the illness. Abnormal oscillatory activity throughout different neural populations and its aberrant synchronization may account for the symptoms observed in schizophrenia. Thus, defining specific patterns of synchrony in individuals with schizophrenia should allow us to make a reliable diagnosis.

Early pioneering work using quantitative EEG laid the foundation for the application of discrimination and classification of electrophysiological brain patterns in health and disease. [John et al., 1977] introduced the concept of neurometrics, to provide quantitative information about brain activity related to anatomical integrity, developmental maturation, and mediation of sensory, perceptual, and cognitive processes. However, as already indicated by [Hinkley et al., 2009], there are significant advantages of using MEG over EEG to study neural processes, especially when modeling how activity within a cortical field can influence and interact with activity in other parts of the brain. As well as EEG, MEG is able to reconstruct neural activity on the order of milliseconds. However, MEG sensors allows investigators to access oscillatory neural activity in higher frequency ranges (e.g. α, β, γ) than those attainable in both fMRI/PET.

In addition, the sampling frequency of data acquisition in MEG (generally greater than 100Hz) is not limited by electrode impedance (as in EEG), permitting the examination of ultra-high frequency brain activity in this modality. Moreover, volume conduction artifacts commonly found in other imaging modalities are significantly reduced in MEG, as structures such as the skull and CSF do not interfere with the propagation of the magnetic fields ([Leahy et al., 1998]).

Hence, MEG arises as a useful method to study how neural oscillations behave and relate to clinical manifestations in schizophrenia.

1.1 Previous work

Several methods using measures of brain activity and brain anatomy have been suggested for discriminating schizophrenia patients from healthy subjects. These methods are based on features extracted from different types of signals. We will focus in the works that use MEG approach to characterize subjects.

In [Georgopoulos et al., 2007] a classifier for separating several conditions (six conditions including schizophrenia) from healthy subjects is described. The dataset used was obtained from around one minute of MEG recordings from 142 subjects without engaging any specific task during the recording. This dataset was obtained by prewhitening the signals, computing the cross-correlation of all the pairs of sensors and normalizing the feature values. An initial set of 30,628 features described each recording, so the authors decided to perform feature selection using a genetic algorithm. Linear Discriminant Analysis (LDA) on the complete dataset was used to assess the fitness function of each feature set in the genetic algorithm. In the end, they reported that a set of 16 features, when used for building LDA classifiers, returned 78.9% percent average accuracy for all conditions (without specifying the accuracy for the different conditions, specially for the schizophrenic condition we are studying) using the leave one out (LOO) testing procedure, and 86.4% accuracy for 10 runs of 80/20% random split cross-validation. The genetic algorithm was parameterized to only consider up to 20 features in the feature set. This reduces the overfitting effect that is prone to happen in small datasets described with a lot of features.

This paper shows that MEG recordings carry relevant information for condition discrimination. However, from the methodological point of view, the set of 16 features selected to build the classifiers was obtained applying the genetic algorithm and LDA procedure to the complete set of 142 individuals (not to the 80% of individuals from the training set in the crossvalidation step), which biases the results of the validation.

In different papers [Ince et al., 2007, Ince et al., 2008, Ince et al., 2009] describe their analysis of MEG oscillatory patterns when subjects engaged in a working memory task during the recordings. They studied a number of other works and reported that a higher accuracy can be obtained from features extracted from recordings of functional brain activity (77 – 94%) than from resting state brain activity (67 – 76%).

The goal of their study was to discriminate schizophrenic patients from control subjects. The dataset had 15 patients and 23 control subjects. From the MEG recordings, event related synchronizations (ERS) and event related desynchronizations (ERD) patterns were extracted, accounting for an increase and decrease of amplitude in the rhythmic activity in different frequency bands and time points. Data is filtered between 1 and 48Hz and separated in 8 not standard frequency bands: 4 with 4Hz wide and 4 with 8Hz wide. The patterns obtained were used as features for training different machine learning classifiers. Specifically, decision trees ([Ince et al., 2007]), support vector machines (SVM) ([Ince et al., 2008]) and LDA ([Ince et al., 2009]) were used. For the last two methods, feature selection strategies were also used to reduce the number of features of the final classifier. In particular, for LDA the ROC curve for each feature was used to filter the set of attributes, and for SVM the recursive feature elimination algorithm (RFE [Guyon et al., 2002]) was applied. The highest average accuracy using the LOO testing procedure was reported for the LDA classifier, being of 94.5%. It is important to notice that, when selecting a classifier method, the results in LOO help to find the best parameters of the learning algorithm, but they are not a true indication of the generalization ability of the classifier. An unseen validation test set has to be used in order to measure the accuracy on unseen data. So, results are not indicative of percentage of success in clinical application. In all three cases, data was obtained assuming a set of hypothesis on the schizophrenic conditions, for instance that it can be discriminated in ERS/ERD episodes, or the separation in bands.

[Escudero et al., 2013] used resting state MEG recordings to obtain a classifier of schizophrenic patients. The dataset was composed by 15 patients and 17 control subjects. The sensors were divided in five anatomical meaningful areas and the average of the power spectral density, Median Frequency, Spectral Entropy, and

Relative Power of the sensors of each area was computed and used as features. Logistic Regression (LR) was used for building the classifier. It is reported that subset feature selection was used to reduce the number of features, but the methodology used was not specified. The mean accuracy reported for the final classifier was of 71,3%. Again, accuracy is computed as the average of a 5-fold cross-validation testing procedure, not on a validation data set.

1.2 Plan of the paper

In the current paper, we suggest how to take profit of MEG information for diagnostic purposes, and how to identify some MEG characteristics in the brain relevant to discriminate individuals among a sample of patients with chronic schizophrenia (stable compensate or acute exacerbation state) and healthy controls. In order to achieve these goals, we propose the use of state-of-the-art classification procedures from the machine learning area such as support vector machines (SVM) together with a proposal for feature selection to use with support vector machines that shows some advantages over other methods in terms of reduction of overfitting, speed and need of extra set of samples for stopping the feature selection procedure.

The paper is organized as follows: In §2 we describe the materials and the methods to prepare data used in this research. In §3 we describe how Support Vector Machines (SVM) can be used to discriminate between patients and control individuals in the data gathered for this task. In the same section we explain how we set the parameters for the SVM and the initial results obtained. In §4 we describe our novel procedure for feature selection in support vector machines. In §4.1 we analyze and discuss the results obtained using this new feature selection procedure. In §5 we present the validation of the approach. Finally, in §6 we discuss the experimental results and give our conclusions and propose our future lines of research.

2 The dataset

A total of 30 subjects were recruited over a 12 month period: 20 patients meeting DSM-IV diagnostic criteria for a chronic psychotic disorder (schizophrenia or schizo-affective disorder) and 10 healthy control subjects. All patients were receiving care at Granollers Mental Health Center from Vallès Oriental Area (Barcelona) and their diagnosis was confirmed by clinical psychiatrist in charge at least 2 years before recruitment, claiming a longitudinal diagnosis. Their psychopathological status was assessed on the day of the experiment by a psychiatrist by means of the Positive and Negative Syndrome Scale (PANSS).

On the other hand, healthy volunteers were picked from the local area, after screening by the structured Clinical Interview for DSM-IV Non-Patient version (SCID-NP) to exclude any history of DSM-IV Axis I/II diagnosis or substance abuse, with an additional exclusion criterion of any first or second-degree biological relatives with psychotic disorder diagnosis.

In order to obtain a good representativeness of the heterogeneity of chronic psychotic patients, we aimed to recruit a sample including patients through different stages of the illness (acute vs stable patients). Therefore, the sample was divided in:

1. Ten patients with a chronic psychotic disorder diagnosis, suffering an acute psychotic episode at the moment of the study, scoring PANSS >27 (acute patients).

2. Ten patients with a chronic psychotic disorder with no current relevant positive psychotic symptomatology at the moment of the study (PANSS <14) (stable patients).

3. Ten healthy subjects (control group).

Exclusion criteria for all participants included neurological diseases, substance dependence, a history of head injury or a full-scale IQ estimation of less than 70. All subjects proved to be right-handed according to modified version of the Edinburgh Handedness Questionnaire ([Oldfield, 1971]), which asks subjects to demonstrate hand use on various actions.

Prior to the study, all subjects were fully informed and gave written consent to participate. This study was approved the Ethics commission of Clinical Research (CEIC) of Hospitals of Hermanas Hospitalarias in Barcelona in accordance with the Declaration of Helsinki protocols.

Table 1: Demographic and clinical data of study groups.

	Stable Disorder	*Acute Exacerbation Disorder*	*Healthy Control*
Gender	9 Male/ 1 Female	8 Male/ 0 Female	9 Male/ 1 Female
Age (years)	36.7 (±10.7)	37.5 (±13.3)	36.7 (±10.7)
Education (years)	10 (±1.7)	9.60 (±2,2)	11 (±1.8)
Distribution Diagnosis axis I (DMS-IV)	7 Schizophrenia disorder 3 SchizoAffective disorder	6 Schizophrenia disorder 2 SchizoAffective disorder	Not applicable
Duration of illness (years)	13.8 (±7.5)	17.5 (±13.3)	Not applicable
PANSS-P	10.5 (±2.6)	30.3 (±4)	Not applicable
PANSS-NN	15.4 (±4.8)	19.9 (±3)	Not applicable

All subjects were Spanish natives or bilingual Catalan/Spanish speakers and they were matched by gender, age and educational degree through the three groups. To guarantee the homogeneity of the groups regarding these characteristics, subjects were recruited according to a sequential design, starting by the acute psychotic patients, following by the stable group, and finally enrolling the control group. Demographic and clinical data for all subjects are presented in Table 1.

At the time of MEG recordings acquisition, all patients were using atypical antipsychotic medication. Two patients from the acute psychotic group could not complete the MEG recording due to psychomotor agitation during the tasks, so their recordings were not considered for the purpose of the study.

The data used in the experiments corresponds to MEG recordings acquired with a 148-channel whole-head magnetometer (MAGNES 2500WH) in a magnetically shielded room (Vacuumschmelze GmbH, Germany) at the Teknon Hospital in Barcelona, Spain. These were acquired while the subjects were lying on a patient bed with their head inside the helmet-like MEG device. Activity was recorded during an auditory oddball paradigm to assess the mismatch negativity (MMN) response. Acoustic stimuli were generated using the E-Prime software (Psychology Software Tools, USA) and were presented binaurally through Etymotic ER-30 (Etymotic Research, Inc. USA) non-magnetic earpieces. Subjects were instructed to relax and ignore the auditory stimuli during the task. The oddball paradigm was based on sequences of two tones, each starting on a pseudorandom basis, either with a 1000-Hz standard (p=0.8) tone followed by a 800-Hz deviant (p=0.2) tone. Tone duration was 100 ms, with rise and fall times of 10 ms, and the inter-stimulus interval was 350 ms. The recording frequency was 678.19Hz (band-pass 0.1-250Hz). Five position coils were attached to the forehead and to the periauricular points in order to determine the position of the head and to track any head movement during the recording. Data sets in which the relative position of the head changed by >0.5cm throughout the recording session were discarded from further analysis. For each subject, the headshape including the forehead, the nose, and the location of the sensor position coils were digitized using a digitizer wand (Polhemus Fastrak, Polhemus Inc., USA).

From the MEG data recordings, six different frequency bands were extracted, using an 8^{th}-order Butterworth digital IIR filter: δ [1-4 Hz], θ [4-8 Hz], α [8-13 Hz], β [13-30 Hz], *lower* γ [30-60 Hz], and *upper* γ [60-200 Hz].

The *upper* γ band was included given that several studies indicate that synchronizations at this frequency range can be linked to schizophrenia (see for example [Uhlhaas et al., 2011]). Studies without band separation were also performed.

2.1 Data Preparation

One objective of this study is to find out if there is a set of features, in terms of MEG recording of the brain, for determining an objective, efficient and safe way whether an individual has schizophrenia or not. This set of attributes has to be easy to interpret in terms of brain areas and their relationships. In order to explore this hypothesis, first we must extract from the raw data a suitable group of features that allow the separation of subjects suffering schizophrenia from other ones.

We want these features to represent the connectivity of different areas of the brain. They also have to capture the global behavior of the signals independently of their magnitude. Different feature extraction methods that extract synchronization information can be used for MEG signals, being the most common those oriented to the frequency domain like coherence ([Srinivasan et al., 2007]) or to the time domain like

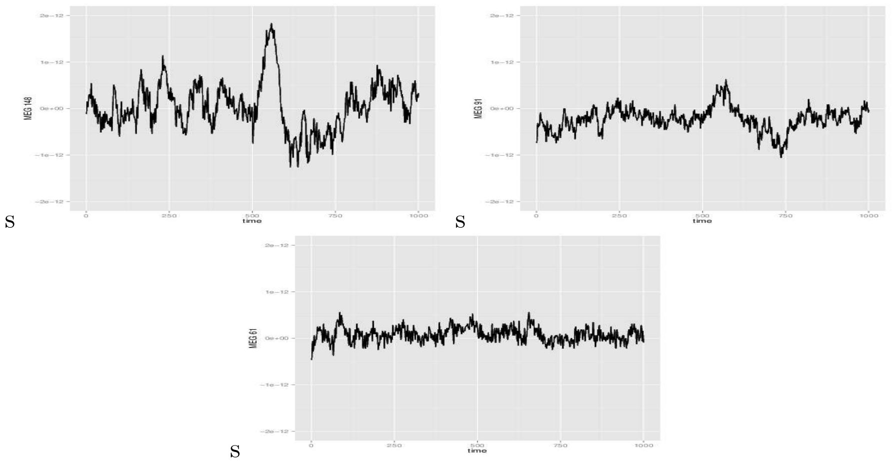

Figure 1: Signals from sensors MEG148, MEG91 and MEG61 (from left to right) with 1000 time points length. Correlation is higher for more synchronized signals (corr(MEG148,MEG91)=0.66, corr(MEG148,MEG61)=-0.008). MEG148 and MEG91 correspond to right side temporal sensors and MEG61 corresponds to an occipital sensor.

covariance ([Shenoy et al., 2006]). Given that one of the goals of was to study separately different frequency bands, we judged more adequate to extract the characteristics from the time domain.

Among the different alternatives, the decision was to use the Pearson product-moment correlation coefficient because it is a simple way to measure global similarity among the different sensors and normalizes the contribution of each pair of sensors. The coefficients capture this effect as a coincidence of the statistical distribution of the signals as can be seen in figure 1.

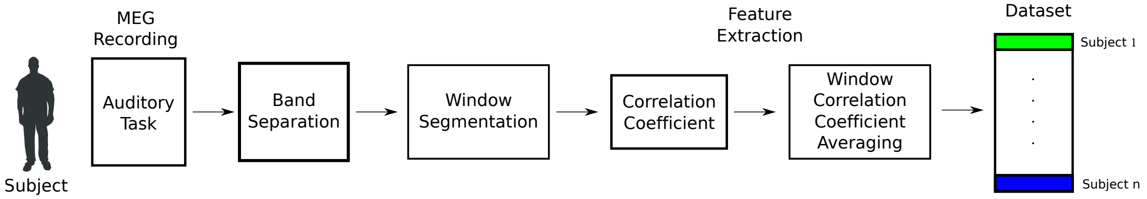

Figure 2: Data preparation process.

We wanted to use relatively large time periods of the signal (around a minute), so the synchronizations among different areas are captured by the features, capturing a fingerprint of the brain activity. Also, It was needed to reduce the variability of the measures and to obtain characteristics that were not affected by artifacts. Thus, an additional preprocess was performed, consisting in the segmentation of the 10 minute recordings in several windows of one minute to obtain measures at different time segments of the experiment. We proceeded to average the correlation coefficient matrices of these windows to smooth the behavior along the time, obtaining this way a better estimate of the mean behavior of the sensor synchronizations. This procedure also reduces the effects of possible artifacts during the recording, allowing working directly with the raw data.

Data preparation for a specific frequency band was performed in the following way (see figure 2):

1. Extract a specific band to study from a file of a subject.
2. Separate the total length of the signal in $k = 10$ windows of the same size.
3. Compute the correlation coefficient among *all pairs* of sensors for each window

$$\rho_{xy} = \frac{cov_{xy}}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

4. Average the k correlation matrices of all windows building the global correlation matrix for the whole signal

In the resulting dataset, separated for each band, each subject file is described by a number of continuous attributes in the range $[-1, 1]$ as $\frac{n^2-n}{2}$ measures, where n is the number of sensors used in the recordings.

For the learning task a random training subset of the data was selected, consisting in 7 control subjects, 7 compensated patients and 6 non compensated patients, for a total of 20 examples. On this data we will study the bands and parameters for the learning algorithm using the standard leave-one-out procedure. As we are concerned with overfitting, we will finally validate the approach selected using the remaining 8 recordings: 3 control subjects, 3 compensated patients and 2 non compensated patients.

Given that the number of individuals in training and testing is very small, we explore bootstrapping methods for validation and to obtain a valid p-value for the statistical point of view (see section 5.1).

3 Experimental procedure

As stated in §2, preliminary experiments are directed to test whether the dataset has enough information to separate control individuals from the ones with schizophrenia diagnostic. Machine learning literature offers a plenty of tools to perform this task. Nevertheless, the problem under study shows a key feature which reduce possible choices. On one hand, recordings in the data set are described by correlations among 146 (of 148, due to two reported malfunctioning sensors) pairs of sensors entailing 10,585 correlations. On the other hand, we only have 20 recordings. So, we have an extremely small number of recordings described by a lot of features. We have to choose a method that, while showing a state-of-the-art accuracy in most benchmarks, it is also able to deal with such kind of dataset. Support Vector Machines (SVMs) seem a promising approach because they have shown the ability to deal with large dimension spaces compared with the number of instances in several domains, for instance in document classification ([Joachims, 1998]).

3.1 Classification using Support Vector Machines

In this section we define the notation that we will be using to describe SVMs and related expressions. Suppose having a set of n data points $\mathbf{x}_i \in \mathbb{R}^d$ along with each point's classification can take on one of two possible values $y_i \in \{\pm 1\}$. The linear SVM is defined considering the best hyperplane $\langle \mathbf{x}, \mathbf{w} \rangle + b = 0$ separating the points in two different classes. It is often useful to consider the SVMs in its dual formulation:

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j K(\mathbf{x}_i \mathbf{x}_j)\alpha_i \alpha_j - \sum_{j=1}^{n} \alpha_j \\ s.t. & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad 0 \le \alpha_i \le C \end{array} \tag{1}$$

where K is the selected kernel and C is a parameter usually determined experimentally specifying the trade-off between the empirical error and the complexity term.

Once this problem has been solved the coefficients (α_i, b) can be used to classify the test point $\mathbf{x}$ based on separating two classes with the hyperplane trough

$$f(\mathbf{x}) = sign(\sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}) + b) \tag{2}$$

which classifies $\mathbf{x}$ as either $+1$ or -1.

Interpreting the Equation 2, one understands that the dual variables α_i represent the importance of a particular point (Support Vectors) within the model. A high value of α_i associated with a particular support vector $\mathbf{x}_i$ indicating that the kernel value calculated using this support vector will have more influence on the final value for $f(\mathbf{x})$. We will consider in this paper the α values of the resulting solution of the SVM, because it is know that the number of supports (α different of zero) can be used as a bound to the empirical LOO error (see next section).

Table 2: Accuracy, Recall and Precision in separation from control to patient diagnosed for each band. Last column shows number of support vectors and proportion of support vectors. See §3.3 for a complete discussion.

Band	*Accu.*	*Recall*	*Prec.*	*Sups. (perc.)*
Full Band	80	57	80	13 (0.68)
δ	**90**	**86**	**86**	**13 (0.66)**
θ	50	14	20	12 (0.61)
α	60	29	40	14 (0.75)
β	55	29	33	15 (0.78)
lower γ	**80**	**57**	**80**	**14 (0.74)**
upper γ	**80**	**57**	**80**	**13 (0.68)**

3.2 Parameters selection and initial results

SVMs performance depends on the selected kernel, its parameters and, eventually, on some data preparation. Within our experiments using MEG data, the only pre-process is a standard data normalization, common in the learning of SVMs.

On the one hand, selecting a kernel can be done experimentally or using some heuristics, and usually requires some cross-validation technique, to study the possible kernels and its parameters. In order to avoid overfitting in our small data set, we limit our choices by using a linear kernel, which is the simplest approach and does not require further parameters to be tuned. Intuitively, it should be enough for our discriminative task for our dataset, due to the large number of features used to describe the data when compared with the number of recordings.

Although SVMs often produce effective solutions for balanced datasets, they could be sensitive to unbalanced datasets and might produce sub-optimal models. A simple method to overcome this problem consists in modifying the SVM objective function by assigning two different misclassification costs, such that the cost of misclassifying training instances of the minority class becomes higher than the penalty for misclassifying instances of the majority class ([Cawley and Talbot, 2001]). A rule of thumb working reasonably well for classification tasks is to set penalty C_- for the negative class (assuming that negative class is the majority class) and $C_+ = f_c C_-$ for the positive class, where f_c is equal to the number of negative examples in the training set divided by the number of positive ones. The effect of this procedure is equivalent to building a SVM with an oversampled minority class to match the number of examples of both classes. Our dataset is imbalanced, having 7 control and 13 patient individuals. We set $C_- = 1$ for negative (patient) and $C_+ = (13/7)C_-$ for positive (control) examples. We tested different values of the C_- parameter in the interval $[0.1, \infty]$ essentially leaving unchanged the obtained results.

Finally, in order to evaluate the classification performance we apply a standard *Leave one out* (LOO) procedure which gives an almost unbiased estimate of the expected generalization error. LOO consists in removing one element from the training data set, constructing the decision rule on the basis of the remaining training data and then testing on the removed element. The procedure is repeated n times (where n is the number of instances of the data set), each time leaving a different individual for testing. Our data set, with 20 examples, allow to repeat this procedure 20 times. Results of the 20 runs were averaged.

The LOO error (E_{LOO}) is related with the number of supports in SVMs. It is known that the proportion of support vectors with respect to the total number of instances is a theoretical bound of the LOO error. That is, $E_{LOO} \leq \frac{n_{SV}}{n}$ where n_{SV} is the number of support vectors (see for instance Remark 7.57 in [Shawe-Taylor and Cristianini, 2004]). As a consequence SVMs with a reduced set of support vectors show a lower leave on out error, which means that we expect from such machines a higher accuracy on data unseen in the learning process (better generalization).

3.3 Classification results

Results obtained for the correlation computation on the original data, and on data after filtering each band, are reported in table 2. These results cannot be understood only in terms of accuracy. Note that, being this an imbalanced dataset, a classification method saying that *all individuals are schizophrenic* leads to an accuracy of 65%. Hence, to avoid this kind of misinterpretation, fairer evaluation scores on the smaller class (in this the control case) are also displayed:

- *Recall:* number of control individuals correctly classified with respect to the total of control individuals. It estimates the probability of detecting the control cases.
- *Precision:* number of the cases classified as control that were actually control instances with respect to the number of cases classified by the system as control. Complementary to recall, it estimates the probability of success in determining the class of the cases classified as control.

Table 2 shows that using the correlation of sensors. Considering the full spectrum band, we obtain SVMs with an accuracy of 80% to distinguishing between control and patients individuals. It seems that the correlation of sensors carries information useful for the discrimination task. Separating the signal in different frequency bands shows that θ, α and β do not help to discriminate control from patients. On the other hand, the δ and γ bands have a higher accuracy, specially in the case of the δ band, that recognizes almost all patients and only fails in 2 cases from 20. However, from the point of view of the robustness of the results, it is worth noting that all classifiers show a high rate of support vectors with respect to the number of examples, most of them around 70% (see last column in table). We have mentioned in section 3.1 that LOO error (E_{LOO}) is theoretically bounded by proportion of supports in the SVM. The higher the proportion of supports vectors needed to build the SVM is, the higher are the expectations on error on unseen cases. So, such machines could show in the worst case unreliable predictions on test data.

This is a sign that we may have ended with overfitted SVMs for our data set. Given the small number of examples that we have available and the large number of features describing them, this is an issue that we have to check carefully.

Also, being encouraging results, the problem within this approach is that they are not easy to interpret in terms of the visualization of areas of the brain possibly related to schizophrenia. The SVM describes the hyperplane separating positive and negative examples with a set of 10,585 weights in the primal form that are not easily interpretable. Nevertheless, most of them may not be necessary to build an effective hyperplane and we will try to simplify the dataset by using feature selection.

4 Feature Selection by incrementing alignment

Fortunately, the interpretation of the classification results is actually possible by studying the effect of feature selection on the dataset. Feature selection methods reduce the feature set in the data collection by removing noisy or redundant features, and thus allowing a better understanding of the classification procedure, which in medical datasets is desirable.

Several methods have been proposed for this task in the SVM framework. [Weston et al., 2000] proposes to find, via gradient descent, those features which minimize the leave-one-out error bounds, that depend on the radius which includes all vectors and the margin of the learned SVM. [Guyon et al., 2002] presents the method *Recursive Feature Elimination* (RFE) that from the whole set of features and the learned SVM, removes the feature that shows the smaller weight $\mathbf{w}$ in primal after learning (and so decreases less the margin). Features are removed iteratively until a given threshold, in terms of weight, is achieved.

The main problem with all these approaches is their computational cost. A feature is only detected after learning takes place. This means that many SVMs have to be learned with the associated cost. Also, those methods rely on cross-validation to adjust the parameters for stopping the removing of features.

We propose a different approach that, while taking advantage of the SVM peculiarities, does not require the previous building of SVMs and has also a natural threshold for stopping the feature removal. This approach is based on a measure known as *empirical kernel-target alignment* ([Cristianini et al., 2002]).

SVMs find the maximum margin hyperplane separating data with different labels. To do that, they only need the dot product of the data in the original space or in the feature space. For this, they use the kernel function and the kernel matrix over the data set. A way to measure *how* a kernel helps to separate the data is by comparing it with the so called *ideal kernel*: $T = yy'$. This is a $n \times n$ matrix that by definition presents $T_{i,j} = +1$ when x_i and x_j have the same label, and -1 otherwise.

Empirical kernel alignment measures the fitness of a kernel K to training labels [Cristianini et al., 2002]:

$$A(K,T) = \frac{\langle K,T\rangle_F}{\sqrt{\langle K,K\rangle_F \langle T,T\rangle_F}} \tag{3}$$

where $\langle M,N\rangle_F$ is the Frobenius product of matrices M and N.

Alignment is a measure ranging in $[0,1]$ (the higher the better), having a number of convenient properties, specially (1) it can be evaluated before learning and (2) it is a *sharply concentrated* statistic around its expected value (stable to different splits of the training data and not prone to overfitting). The last feature is specially relevant in our case because, given the small number of examples and large number of features in our dataset, we want to avoid an overfitting of the results.

Alignment has been used for *a priori* kernel selection (choosing the best kernel from a set of candidates) and for kernel adaption ([Lanckriet et al., 2004]). We will use this *a priori* model selection procedure to find a suitable set of features, facilitating the learning process. We define kernel $K_{\mathcal{F}}$ from the set of features $\mathcal{F}$ as the kernel matrix obtained by applying the selected kernel function on the data represented only with features in set $\mathcal{F}$. The idea is to implement feature selection using the empirical kernel-target alignment, using the fact that choosing one set of features or another is equivalent to choosing one kernel or another. *Hence, we select the set of features $\mathcal{F}$ defining the kernel $K_{\mathcal{F}}$ with the highest empirical kernel-target alignment.*

In order to deal with unbalanced data sets, we used a modified version of alignment ([Kandola et al., 2002]) in which labels to calculate the ideal kernel are changed from $+1$ to $+1/m$ and from -1 to $-1/n$, where m and n are the number of positive and negative examples respectively. The learning of the SVM is done with the original set of labels.

For feature selection, we implemented an iterative procedure that, starting from the empty set, incrementally adds the feature increasing the most the alignment with the ideal kernel. This stops when no increment in alignment is found. The advantages over other approaches as RFE ([Guyon et al., 2002]) are (1) that our method avoids the costs of building SVMs to consider relevance of features, and (2) that the addition of features naturally stops when alignment is not increased, that is, no validation dataset is required to determine when to stop.

However, the cost of the method is still expensive. The cost of computing the linear kernel $K_{\mathcal{F}}$ is $O(n^2 f)$, where n is the number of examples and f is the number of features in $\mathcal{F}$. The kernel has to be calculated for each set of features before computing its alignment. In addition, the number of features f in our dataset is very large, so removing the f term is desirable. A careful examination of the problem shows that there is an efficient incremental way to compute the kernel after adding a new feature from the same kernel without that feature. This reduces the cost of computing the new kernel for our incremental algorithm from $O(n^2 f)$ to $O(n^2)$. For the linear kernel, the following equation shows how to quickly calculate the new kernel from the old kernel *when only adding a new feature i*:

$$K_{\mathcal{F}\cup\{i\}}(x,y) \;=\; \sum_{j\in\mathcal{F}\cup\{i\}} x_j y_j \;=\; \sum_{j\in\mathcal{F}} x_j y_j + x_i y_i \;=\; K_{\mathcal{F}}(x,y) + x_i y_i$$

with constant cost for one pair of examples. Hence computing the kernel $K_{\mathcal{F}\cup\{i\}}$ from $K_{\mathcal{F}}$ only has a cost of $O(n^2)$. We will take advantage of this fact in our algorithm.

This procedure can be easily extended for non linear kernels, like polynomial or the RBF kernel. In this way, we can reduce also to $O(n^2)$ the cost of computing the alignment $A(K_{\mathcal{F}\cup\{i\}},T)$ from $K_{\mathcal{F}}$. For large data sets this could still be a problem. However, alignment is a sharply concentrated measure ([Cristianini et al., 2002]) and could be accurately estimated from a sample of the whole training set.

With this in mind, the method of *incremental Feature Selection by Alignment* can be detailed as shown in algorithm 1.

The algorithm starts from the empty set of features $\mathcal{F}$. At each iteration it always looks for the feature which increases the alignment the most. This feature is added to $\mathcal{F}$ and the procedure is repeated until no more features remain in the set of candidate features $\mathcal{C}$ increasing the alignment. Note that after selecting one feature, that feature is removed from set $\mathcal{C}$.

The computational time complexity of line 7 is n^2. Thus, loop of lines 6-8 has complexity $n^2|\mathcal{C}| \leq n^2 f$, and thus, the outer loop (lines 5-16) has complexity $n^2 f f_a$ (where f_a is the final number of features of set $\mathcal{F}$). So, the final computational complexity of the feature selection procedure is then $O(n^2 f f_a)$.

Algorithm 1 Feature Selection by incrementing Alignment procedure (*FSiA*)

```
$\mathcal{C}$ = Complete set of features
$\mathcal{F}= \emptyset$
$K^1_{\mathcal{F}} = 0$ for all x,y
$CurrentAlign = 0$
repeat
    for each $i$ in $\mathcal{C}$ do
        $IncrAlign(i) = A(K^1_{\mathcal{F}\cup\{i\}},T) - CurrentAlign$
    endfor
    $add$ = Feature $i$ with higher $IncrAlign(i)$
    if $IncrAlign(add) > 0$ then
        Calculate incrementally $K^1_{\mathcal{F}\cup\{add\}}$ from $K^1_{\mathcal{F}}$
        $\mathcal{F}= \mathcal{F} \cup \{add\}$
        $\mathcal{C}= \mathcal{C}\backslash\{add\}$
        $CurrentAlign= A(K^1_{\mathcal{F}},T)$
    endif
until $IncrAlign(add) \leq 0$
```

Table 3: Accuracy, Recall and Precision in separation from control to patient diagnosed for each band using feature selection. Last column shows number of support vectors and proportion of support vectors.

Band	*Accu.*	*Recall*	*Prec.*	*#SV (prop.)*
Full Band	75	57	67	12 (0.64)
δ	**85**	**71**	**83**	**9 (0.45)**
θ	65	57	50	13 (0.68)
α	45	43	30	12 (0.62)
β	75	43	75	12 (0.64)
lower γ	**90**	**86**	**86**	**9 (0.50)**
upper γ	**90**	**86**	**86**	**9 (0.47)**

The method described above was implemented in `matlab` and applied on our dataset to do feature selection. The running time of the algorithm was less than 2 seconds for finding a reduced set of features for the task on hands. After that we used the `libsvm` ([Chang and Lin, 2011]) implementation of SVMs to obtain classification accuracy on test set.

4.1 Results using feature selection

Results obtained for each band of recordings are shown in table 3. Best results appear now in the *upper* γ and *lower* γ bands. Table shows that for those bands of frequencies the method is able to separate with 90% of accuracy recordings for control individuals from recordings for diagnosed schizophrenia individuals. Only 2 errors have been produced: 1 control individual was classified as patient and 1 patient was classified as control.

Compared with results in table 2 we may confirm that β, α and θ bands do not contain enough information to separate control from patient recordings. Notice that the feature selection method increases the accuracy for the γ bands from 80% to 90%. On the other hand, δ band reduces the accuracy from 90% to 85%. So, feature selection improves results for two bands while reducing accuracy in another one. It is also worth noting that the feature selection procedure significantly reduces the number of support vectors and, therefore, it increases the confidence on *good* generalization to unseen cases.

Recall that the goal for applying feature selection is not to improve classification accuracy, but to find out which features are relevant for patient classification. Their location in the brain will help to understand the illness.

We can conclude that the feature selection procedure was very effective. This set is enough to accurately

Table 4: Number of features and actual sensors selected for the feature selection procedure. Last two columns show initial alignment of the dataset with all features and final alignment after feature selection.

Band	*#Feats.*	*#Sens,*	*I.Align.*	*F.Align.*
Full Band	10	18	0.105	0.762
δ	19	35	0.151	**0.838**
θ	7	13	0.114	0.667
α	15	26	0.088	0.670
β	11	20	0.044	0.602
lower γ	16	29	0.123	**0.861**
upper γ	22	36	0.128	**0.897**

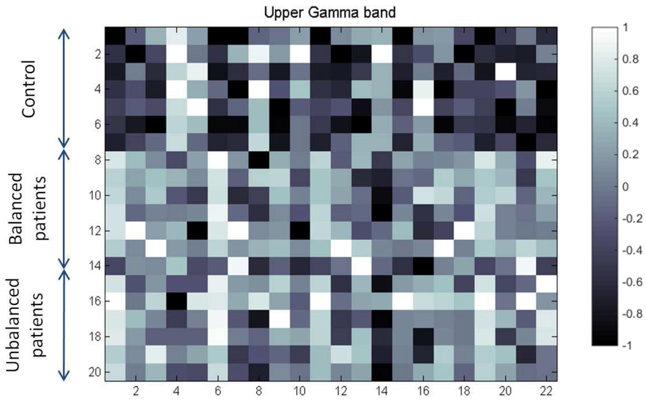

Figure 3: Matrix showing correlation values selected by the feature selection method on the upper γ band. Each column represents a feature selected. Each row represents a individual recording. Recordings are sorted so the first ones are control recordings and the later schizophrenic individual's recordings. Bar shows code for correlation values.

represent recordings with the goal of separating control individuals from those that do suffer schizophrenia. Notice that for the *upper* γ band only 22 of the features suffice to separate the data (table 4).

Figure 3 shows for each one of the 20 recordings (rows) the values of the correlations for each of the 22 features (columns). Each feature is the correlation between a pair of sensors, ranging in $[-1, 1]$. With a simple look at the table one may guess where the separation of individual's recordings is located.

Table 4 shows information about the relation between feature selection by alignment measure and possible overfitting. In the two last columns, alignment before and after feature selection is displayed. The bands with better final alignment show better results in the LOO procedure, suggesting that final alignment is highly correlated with accuracy. These results support the idea of doing feature selection by increasing the kernel alignment.

The second column of Table 4 indicates how many sensors are actually needed to compute the features describing the recordings. One would expect, being each feature the correlation of a pair of sensors, the number of sensors to be close to twice the number of features. For the *upper* γ band the number of sensors is 36, meaning that when we unfold the correlations in pairs of sensors, some sensors appear several times. We hypothesize that the number of times each sensor is needed to compute the relevant features is an index of its relevance.

Figure 4 shows, for the upper γ band, a map of the brain with the location of the sensors and how many times each one is used to compute the features relevant for classification. One sensor appears up to 8 times in the 22 feature-correlation definitions.

We infer that upper γ band may be the more relevant band for understanding schizophrenia. Still, lower γ and δ bands seem also to carry information about how to classify individuals.

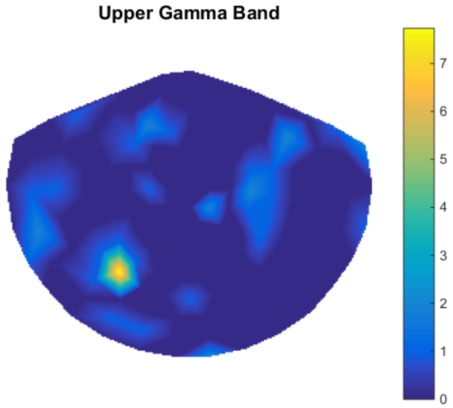

Figure 4: Brain map showing the number of times a sensor is needed to compute the set of selected features for the upper γ band. Nose is the north of the figure. Right and Left hemispheres are east and west sides respectively.

5 Validation step

Along all the analysis, we were concerned with overfitting, given the large number of features and the small number of examples. We have reduced parameter selection to the minimum. However, the C parameter for the SVM and the bands selected for the final results, remained. Most papers report only the E_{LOO} without realizing that those results could fail when extended to new examples. To test the validity of our results to new data, data for eight individuals randomly selected were hidden (before carrying the experiments) for a final validation step. The validation data set consists in 3 control individuals plus 3 balanced patients plus 2 unbalanced patients.

To test this methodology, we trained a SVM for the three selected bands: upper γ, lower γ and δ. We used a SVM with linear kernel, the C parameter set to 0.1 (value obtained in the LOO step of the research) and applied the feature selection procedure (see §4). This SVM trained with the filtered upper γ band of the initial 20 examples, resulted in only 1 error over the 8 validation examples. For the lower γ band we obtained 2 errors and, finally, on the δ band we obtained 3 errors.

These results should be considered orientative. There is not enough data to obtain a statistically significant result in terms of accuracy and confidence intervals. However, it is worth noting the tendency that bands with better results in LOO, also return better results in the validation step. So, validation results consistently suggest that γ bands actually carry useful information for separation of control from patients using the proposed feature selection procedure. Notice also that results in the validation set for each band are better for those bands with higher final alignment as shown in table 4. This fact supports the idea that alignment is a good indicator of generalization on unseen cases. So, overfitting can be avoided using our feature selection method.

5.1 Probabilistic analysis

However, we may still be concerned with overfitting. Note that we have 10,585 features describing a set of 20 observations. From this dataset, applying the feature selection method that relies on kernel alignment, a subset of 22 features is selected. The selected set shows the highest alignment found with ideal kernel. Still, two hypotheses could be the reason for such high alignment:

(1) Random effect: Among the huge number of features (10,585), it could exist a small set of features not actually related with the labels but that, by chance, present values able to separate the 20 examples in the two classes, and that is what the proposed method finds. This would be a clear overfitting case.

(2) Structural effect: The data present features related with labels that are captured by the method proposed. This would be the case of a correct modelization of the problem.

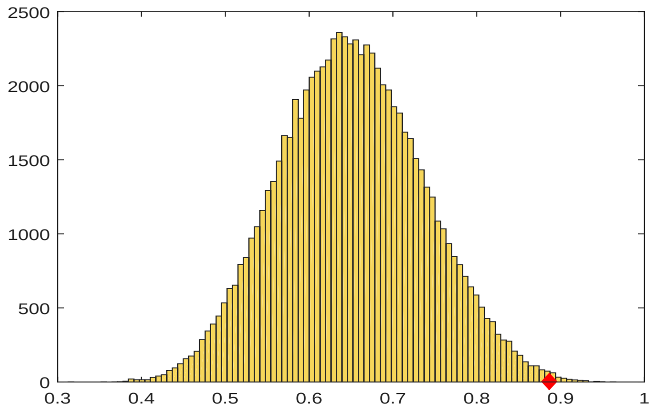

Figure 5: Distribution of alignment for all possible labelings of data with 7 positive examples and 13 negative examples. Mark shows true labeling alignment.

How to differentiate between the two cases? We hypothesize that if (1) holds, then this effect would appear with *any labeling* of the data. That is, for any labeling there will be a set of features with high alignment because of chance. From the large initial set of features, a small set would be enough to separate the 20 examples.

In order to test whether (1) or (2) holds, we did the following permutation test: We assigned an artificial labeling to data (following the original proportion of positive versus negative examples) and then we applied the method to find the set of features with maximum alignment. We repeated the experiment for all possible labeling with the same proportion of positive and negative examples (77,250 cases). Table 5 shows for each frequency band the *average* alignment obtained after feature selection for all possible labeling of data. For comparison, the same table shows alignment obtained using the original labels. For instance, Upper γ shows the expected best alignment with random labels of 0.66 but 0.90 with true labels. Figure 5 shows an histogram of the number of different labelings for each alignment value. Mark shows the final alignment using true labels. If we translate the position of the mark to *p-values*, we have a probability of 0.0022 to obtain by chance such high alignment using true labels (far from the 0.05 to accept standard statistical tests). So chance is not enough to explain the high alignment with the true labels and we must conclude that for the problem at hand (1) is not the case and that (2) holds.

Table 5: Number of features and sensors expected after FSiA when labeling randomly the dataset. Compare results with true labels in table 4. Last two columns show expected alignment of the dataset with random labels (*ERL Align*) and final alignment with true labels (*True L. Al*).

Band	*#Feats.*	*#Sens.*	*ERL Align.*	*True L. Al.*
Full Band	12	21	0.56	0.76
δ	11	19	0.51	0.84
θ	9	17	0.47	0.67
α	12	22	0.54	0.67
β	9	17	0.51	0.60
lower γ	17	30	0.67	0.86
upper γ	16	27	0.66	0.90

5.2 Comparison with Recursive Feature Elimination (RFE)

A popular and very effective method for feature selection using SVMs is RFE ([Guyon et al., 2002]). In order to compare the performance of our method with it, we designed and performed some experiments. However, there are two considerations to be made before applying RFE to our data.

The first one is that RFE needs a validation dataset in order to decide the number of features to keep. The method iteratively removes features until performance in the validation dataset decreases. In our dataset we have too few examples to extract a validation dataset for deciding when to stop the removal of features. Conversely, our method has a natural way to stop, that is, when alignment measure is not increased, so it does not need such validation dataset.

The second one is that RFE builds a SVM at each iteration in order to decide which features to remove. So, before applying RFE we have to decide the appropriate parameters for the SVM. Again this is solved using a cross-validation scheme where several combinations of parameters are tested against a validation dataset in order to find the best parameters *while* feature selection is done. In contrast, our feature selection method based on the alignment measure does not have any parameter to adjust. The finding of the best parameters of the SVM are done *after* selection of features has been done, with the according reduction in execution time.

So, a fair comparison of both methods cannot be done because in RFE we need a validation dataset that we cannot afford given the small size our dataset. However, we will do a comparison *assuming* that we know both the number of features needed to obtain a good accuracy and the parameters of the SVM: those will be the obtained in an *a priori* execution of the alignment method. The experiment will consist in a leave one out of the whole set of 40 examples and the goal will be to compare the accuracy and time execution of both methods.

Using the same number of features that the alignment method found, and parameters for the SVM described in section 3.2, RFE obtains a LOO of 95% of accuracy (2 failures out of 40), exactly the same than our method. Notice that in a fair comparison of RFE we should reduce the examples used for training because some of them should be used for determining when to stop removal of features, so we should expect a lower accuracy in a fairer comparison.

The running time of RFE with careful elimination (one feature at one iteration) is about 11 hours, while an aggressive approximation implementation that removes several features at each iteration reduces time spent up to 46 seconds. Notice that in a fair comparison we should add time needed for cross-validation to find best parameters of the SVM. In contrast, our method only spent 24 seconds.

So, in conclusion our method can take more profit of the number of examples, specially critical when the dataset is very small like on our case. The method is very fast and has an accuracy similar to the state-of-the-art algorithms like RFE.

6 Conclusions and Future Work

The main contribution of this paper is the methodology used to analyze MEG signals in order to obtain an objective criteria to discriminate schizophrenic individuals from control participants. Unlike other similar studies, for this one we have recruited a sample of patients through different stages of the illness (acute *vs* stable) that may be more representative of the heterogeneity of chronic psychotic patients.

From the experiments, we can confirm that MEG readings carry valuable information on schizophrenia condition, as it has been suggested in [Georgopoulos et al., 2007, Ince et al., 2008, Ince et al., 2009] and also in [Escudero et al., 2013]. We have shown that schizophrenia condition can be detected, while the subject is doing a simple task, by studying the correlation between sensors capturing MEG activity in γ frequency bands in different parts of the brain. Unlike other works, data was collected with a minimum of data preprocessing. Only band separation has been applied to better understand the schizophrenic condition. The experimental conclusions of this study shows that MEG signals can be a useful tool to successfully discriminate schizophrenic patients from healthy participants.

This analysis can be a complementary objective tool to help psychiatrists to obtain a reliable and objective diagnosis of schizophrenia. With the help of this tool, screening, preventive and early therapeutic strategies could be established in order to change the course of the illness. An objective diagnosis may favour patients' insight and better adherence to the treatment, which is the main prognostic factor. Research on the basis of a certain diagnostic label also would allow a better comprehension of the underlying mechanisms of the illness and lighten the future development of novel treatments.

Results have been validated after modeling the classifier. This contrast with most papers that only show results obtained with LOO for the best classifier obtained. It is known that the best LOO error after model selection is not a valid predictor of accuracy for unseen data because of overfitting, which is prone to appear in small datasets described with a lot of features as the one used in this paper. In addition, we have proposed suited simple procedure for feature selection for support vector machines applied to the discrimination task which shows:

• *State of the art performance* in the separation of control subjects from patients: Our LOO results are about 90% accuracy, which are better or equal that the ones shown in the previous literature (ranging from 77% to 94%). Consider that individual diagnosis of schizophrenia using the gold standard procedure (clinical interview by a specialist) can vary depending on the study but never overcomes a 90-95% ([Harvey et al., 2012], [Aboraya et al., 2006]).

• *Robustness to overfitting*: FSiA relies on the alignment measure, which shows the mathematical property of concentration. That means that with few data we can have good estimations in the feature selection process and, additionally, it will show good performance on unseen data (it does not overfit).

• *Efficiency in examples required for learning:* Most methods for feature selection in SVMs (for instance RFE [Guyon et al., 2002]) need a validation dataset to detect the stopping criteria for addition or removal of features in the feature selection procedure. FSiA shows a natural way to stop the addition of features: when the addition of one feature does not increase the kernel alignment. This saving in examples is specially important when dealing with small dataset, which is our case.

• *Efficiency in time:* The method has quadratic cost in the number of examples, linear with the number of features and linear with the number of features finally selected $O(n^2 f f_a)$, which is lower than the cost time of building a SVM. The selection of the final feature set using this new approach is in the order of seconds. So, it is much faster than RFE used in [Ince et al., 2008], which requires building several SVMs. Also, that time cost is much lower than other approaches that use genetic algorithms for feature selection ([Georgopoulos et al., 2007]), which requires several hours in order to deliver a feature set.

Finally, we have presented a visualization method that allows mapping, in a representation of the brain, which parts of the brain could be related to the illness. It shows the dissimilarities in the correlation of sensors from different regions of the brain that separate control subjects from schizophrenic patients.

We plan to extend this work in several directions. Firstly, we will complement and expand current data about schizophrenia patients with MEG registers in different experimental conditions, and study *how* diverse patients' states help to increase the accuracy in the diagnostic task. We will also use the collected results to study the neurophysiological relevance in the schizophrenic illness of brain regions detected by the feature selection mechanism. Also, we plan to extend this kind of analysis to other known mental disorders and to neurodegenerative diseases.

7 Acknowledgements

This work was partially supported by the Ministerio de Ciencia e Innovación. Fondo de Investigaciones Sociosanitarias (Grant Number: PI081008), also by TEC2014-56256-C2-2-P *Rehabilitación personalizada y adaptativa en tratamientos post-ictus: el* i-*Walker*, and also by the FI-DGR programme of AGAUR ECO/1551/2012. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[Aboraya et al., 2006] Aboraya, A., Rankin, E., France, C., El-Missiryd, A., and John, C. (2006). The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont)*, 3(1):41–50.

[Cawley and Talbot, 2001] Cawley, G. C. and Talbot, N. L. C. (2001). Manipulation of prior probabilities in support vector classification. In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, volume 4, pages 2433–2438 vol.4.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

[Cristianini et al., 2002] Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. (2002). On Kernel-Target Alignment. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*. MIT Press.

[Escudero et al., 2013] Escudero, J., Ifeachorl, E., Fernández, A., López-Ibor, J. J., and Hornero, R. (2013). Changes in the meg background activity in patients with positive symptoms of schizophrenia: spectral analysis and impact of age. *Physiological measurement*, 34(2):265.

[Georgopoulos et al., 2007] Georgopoulos, A. P., Karageorgiou, E., and et al. (2007). Synchronous neural interactions assessed by magnetoencephalography: a functional biomarker for brain disorders. *J. Neural Eng.*, 4:349–355.

[Guyon et al., 2002] Guyon, I., J Weston, a. S. B., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422.

[Harvey et al., 2012] Harvey, P. D., Heaton, R. K., Jr., W. T. C., Green, M. F., Gold, J. M., and Schoenbaum, M. (2012). Diagnosis of schizophrenia: Consistency across information sources and stability of the condition. *Schiz. Research*, 140(1–3):9 – 14.

[Hinkley et al., 2009] Hinkley, L. B. N., Owen, J. P., Fisher, M., Findlay, A. M., Vinogradov, S., and Nagarajan, S. S. (2009). Cognitive Impairments in Schizophrenia as Assessed Through Activation and Connectivity Measures of Magnetoencephalography (MEG) Data. *Front Hum Neurosci.*, 3(73).

[Ince et al., 2008] Ince, N., Goksu, F., Pellizzer, G., Tewfik, A., and Stephane, M. (2008). Selection of spectro-temporal patterns in multichannel meg with support vector machines for schizophrenia classification. In *30th Annual International Conference of the IEEE on Engineering in Medicine and Biology*, pages 3554–3557.

[Ince et al., 2007] Ince, N., Stephane, M., Tewfik, A., Pellizzer, G., and McClannahan, K. (2007). Schizophrenia classification using working memory meg erd/ers patterns. In *3rd International IEEE/EMBS Conference on Neural Engineering*, pages 457–460.

[Ince et al., 2009] Ince, N. F., Pellizzer, G., Tewfik, A. H., Nelson, K., Leuthold, A., McClannahan, K., and Stephane, M. (2009). Classification of schizophrenia with spectro-temporo-spatial MEG patterns in working memory. *Clinical Neurophysiology*, 120:1123–1134.

[Institute of Mental Health, 2013] Institute of Mental Health, N. (2013).

[Joachims, 1998] Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In *ECML '98: Proceedings of the 10tth European Conference on Machine Learning*, pages 137–142. Springer-Verlag.

[John et al., 1977] John, E. R., Karmel, B., Corning, W. C., Easton, P., Brown, D., Ahn, H., John, M., Harmony, T., Prichep, L., Toro, A., Gerson, I., Barllet, F., Thatcher, F., Kaye, H., and Valdes, P. (1977). Neurometrics. *Science*, 196(4297):1393–1410.

[Kandola et al., 2002] Kandola, J., Shawe-Taylor, J., and Cristianini, N. (2002). On the Extensions of Kernel Alignment. Technical Report NC-TR-02-120, NeuroCOLT.

[Lanckriet et al., 2004] Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72.

[Leahy et al., 1998] Leahy, R., Mosher, J., Spencer, M., Huang, M., and Lewine, J. (1998). A study of dipole localization accuracy for MEG and EEG using a human skull phantom. *Electroencephalogr Clin Neurophysiol.*, 107(2):159–73.

[Oldfield, 1971] Oldfield, R. (1971). The assessment and analysis of handedness: the Edinburgh Inventory. *Neuropsychologia*, 9:97–113.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

[Shenoy et al., 2006] Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N., and Müller, K.-R. (2006). Towards adaptive classification for bci. *Journal of Neural Engineering*, 3(1):R13.

[Srinivasan et al., 2007] Srinivasan, R., Winter, W. R., Ding, J., and Nunez, P. L. (2007). EEG and MEG coherence: Measures of functional connectivity at distinct spatial scales of neocortical dynamics. *Journal of Neuroscience Methods*, 166(1):41 – 52.

[Uhlhaas et al., 2011] Uhlhaas, P. J., Pipa, G., Neuenschwander, S., Wibral, M., and Singer, W. (2011). A new look at gamma? high-($>$ 60 hz) γ-band activity in cortical networks: function, mechanisms and impairment. *Progress in biophysics and molecular biology*, 105(1):14–28.

[Weston et al., 2000] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2000). Feature selection for SVMs. In *NIPS*, pages 668–674.