Robust Modeling Based on Optimized EEG Bands for Functional Brain State Inference

Ilana Podlipski[1], Eti Ben Simon[1], Talma Hendler[1] and Nathan Intrator[2]

[1]Functional Brain Imaging Unit, Souraski Medical Center

[2]School of Computer Science, Tel Aviv University

Abstract

The need to infer brain states in a data driven approach is crucial for BCI applications as well as for neuroscience research. In this work we present a novel classification framework based on Regularized Linear Regression classifier constructed from time-frequency decomposition of an EEG (Electro-Encephalography) signal. The regression is then used to derive a model of frequency distributions that identifies brain states. The process of classifier construction, preprocessing and selection of optimal regularization parameter by means of cross-validation, is presented and discussed. The framework and the feature selection technique are demonstrated on EEG data recorded from 10 healthy subjects while requested to open and close their eyes every 30 seconds. This paradigm is well known in inducing alpha power modulations that differ from low power (during eyes opened) to high (during eyes closed). The classifier was trained to infer eyes opened or eyes closed states and achieved higher than 90% classification accuracy. Furthermore, our findings reveal interesting patterns of relations between experimental conditions, EEG frequencies, regularization parameters and classifier choice. This viable tool enables identification of the most contributing frequency bands to any given brain state and their optimal combination in inferring this state. These features allow for much greater detail than the standard Fourier Transform power analysis, making it an essential method for both BCI proposes and neuroimaging research.

Key words: Ridge Regression, Regularization, Classification, EEG, time-frequency.

1. Introduction

A central aim of functional brain imaging research is to reveal the mapping between brain signals and the mental states which elicit them. One of the common methods of investigating brain signals is EEG spectral analysis. Due to the rhythmic nature of many EEG activities, if several rhythms occur simultaneously, Fourier Transform enables separation of these rhythms and estimation of their frequencies independently of each other (Baar et al., 2001; Pfurtscheller and Lopes da Silva, 1999). In this method brain oscillations are divided to frequency bands that were found related to different brain states, functions or pathologies (Niedermeyer and Lopes da Silva, 2005). In the field of EEG-based Brain-Computer Interface (BCI) design, machine learning algorithms are used to identify 'patterns' of brain activity that identify a certain mental task (Anderson et al., 1998; Keirn and Aunon, 1990; McFarland et al., 2000). Typically these algorithms treat either raw EEG data or power of some predefined frequency band (such as motor-related α and β rhythms) as features. Those features are then fed into a classifier to produce the final classification. Most of these studies focus on classification performance, rather than examine the features that contribute to the classification. These features may reveal additional information about brain function. Importantly, classification tools enable us to quantify the relative contribution of each functional signal to task performance, while evaluating the predictive power of these signals (Cox and Savoy, 2003; Haynes and Rees, 2006; O'Toole et al., 2007) Hence, classification can be used as a functional analysis tool.

One of the techniques that have been used for this purpose, in the field of BCI, is regularized Fisher's Linear Discriminant analysis (LDA). The aim of LDA is to use hyper planes to separate the data representing the different classes (Duda et al., 2001). This classifier introduces a regularization parameter that can allow or penalize classification errors on the training set. The resulting classifier can accommodate outliers and obtain good generalization capabilities (Zhdanov et al., 2007). As outliers are common in EEG data, this regularized version of LDA may give better results for BCI than the non-regularized LDA.

Recently several machine learning approaches have been proposed for the study of neuronal imaging data (for reviews see Besserve et al., 2011; Lemm et al.,

2010) ranging from BOLD fmri signal (Pereira et al., 2009) to ERP studies (Tomioka and Müller, 2010). Tomioka et. Al., have proposed a method for classification using joint spatio-temporal information of the EEG while focusing on specific chosen frequency bands. This framework provides information of spatial distribution of classification weights in specific time points used for the classification model. This allows investigation of temporal evolution of spatial content, within a classification framework. The current study describes a framework for inferring the temporal evolution of frequency content within functional states, by derivation of linear combination of frequency bands most relevant to a given task. The inference problem is discrimination between two classes of signals time locked to experimental events. The framework is specifically adapted to the needs of EEG-based functional neuroimaging. It provides the means to investigate the frequency and spatial distribution of the EEG signal related to a given task. The framework utilizes a regularized linear classifier constructed from instantaneous signal values with cross validation. The behavior of regularization parameter and features contributing to the classification is investigated using EEG data taken from a combined EEG\fMRI study of an experiment that involves switching between eyes opened and closed states. For this change of state, it is known that the EEG power in the Alpha frequency band (8-12Hz) increases when eyes are closed, especially at rest, and decreases when eyes are opened (Berger, 1929). This phenomenon is commonly referred to as the "Berger effect" (Lopes da Silva et al., 1976; Niedermeyer and Lopes da Silva, 2005). However whether the Alpha band is a stand-alone frequency or other frequencies oscillate with the change in eyes state remains an open question.

Within our framework we chose the combination of EEG channel and regularization parameter that jointly optimize classification rate and constructed a frequency combination model from a family of models. This frequency combination identifies those frequencies that change power according to the change in eye state.

2. Methods

2.1. Problem formulation

In a standard block design experimental setup, EEG signals are continuously recorded from a number of electrodes located over the subject's scalp for the whole duration of the experiment. The signals can be transformed into a time frequency representation where each time interval is associated with a target label defined by the type of the corresponding experimental event, such as opened or closed eyes of the subject. This target label describes the brain state we are looking to infer. In this fashion, we obtain a set of labeled data samples for each channel. Each channel is represented by N_{freq}-by-N_{tps} signal matrix, where N_{freq} is the number of frequencies in the time frequency representation and N_{tps} is the number of time sampling points in the segmented interval.

2.2. Classification Method

The problem of inferring mental states can be treated as a classical high-dimensional pattern recognition problem where each frequency of the signal matrix is a separate feature. In order to make an inference about subject's states we derive a combination of features in the signal that exhibit a desired pattern. This simple linear regression approach provides a measure of importance of each frequency to a given task within a combination of all other frequencies of the signal, rather than the sole importance of each frequency individually.

Two types of classifiers were chosen for this inference:

1. Linear Ridge regression classifier, chosen for its robustness and for the fact it naturally provides the interpretation of the resulting weights.
2. Regularized Logistic Regression classifier, which fits statistically for dichotomous classification, and therefore was chosen as a good candidate for successful inference.

2.2.1. Ridge Regression

Consider

$$y = Xw + \varepsilon \tag{1}$$

Where $w \in R_m$ is the vector of regression coefficients to be determined using

observed data (X, y); ε represents noise in the response y; X_i, i = [1: n] are the explanatory variables or predictors.

In Ridge regression, the coefficients are obtained as:

$$\arg\min_w \{\|y - Xw\|^2 + \lambda\|w\|^2\} \quad (2)$$

$$w = (X'X + \lambda I_m)^{-1} X'y \quad (3)$$

Where $\lambda \geq 0$ is the ridge regularization parameter and I_m is the mxm identity matrix. This form of regularization is also known as Tikhonov regularization, and it was shown (Tikhonov, 1963) that for every $\lambda \geq 0$ there exists a unique w for which L_λ is minimal.

When X is ill-conditioned it has near zero singular values which usually correspond to noise components of X. When inverted, these components drastically amplify the contribution of noise to the solution and destroy the regression model. Therefore, a large enough λ reduces the contribution of noisy near-zero components and stabilizes the solution. However, if λ is too large, it eliminates the contribution from nearly all components of the data, producing a meaningless result with near-zero variance.

2.2.2. Logistic Regression

Given a binomial response variable y = 0, 1; it is useful to look at the probability of occurrence of y given a set of explanatory variables X.

Logistic regression predicts the probability that the log odds (also known as logit function) of an observation will have an indicator equal to 1. The odds of an event are defined as the ratio of the probability that an event occurs to the probability that it fails to occur. This is described by the logit function (eq**.** 3). The logit of the response is modeled with a linear term:

$$; \quad P(X) = P(y = 1|X) \ \log it\{P(X)\} = \log\left[\frac{P(X)}{1 - P(X)}\right] = Xw \quad (4)$$

From (3) it follows:

$$P(X) = \frac{\exp(Xw)}{1 + \exp(Xw)} \quad (5)$$

Maximum likelihood is the most popular method for estimating the logistic model. Let y_i, i = 1. . . m, be the response variable and x_{ij} , j = 1, . . . , n, the

predictor variables, for the above problem the log likelihood has the form:

$$l(w) = \sum_{i=1}^{m} \left[y_i \log\{P(x_i)\} + (1 - y_i) \log\{1 - P(x_i)\} \right] \quad (6)$$

For an ill-conditioned problem over-fitting is avoided by imposing a penalty on large fluctuations of the estimated parameter w and respectively on the fitted function. Here the strategy is similar to the one described in the previous section. The likelihood is penalized by a regularization term:

$$l^{\lambda}(w) = l(w) - \lambda \|w\|^2 \quad (7)$$

In our case minimization of this expression was done with a variation of the Newton-Raphson method, called Conjugate gradient ascent (Bishop, 2005; Komarek and Moore, 2003; Minka, 2003).

2.3. Selection of regularization parameter and performance estimation

The classifiers described so far depend on a regularization parameter λ. For the actual classification, one needs to select a value for this parameter and to assess classification accuracy. The standard approach to selecting the values of free parameters in pattern recognition is to search the parameter space for the values that minimize the error estimate (Duda et al., 2001; Picard and Cook, 1984; Tomioka and Müller, 2010). The error estimate may be calculated using cross-validation (CV) on all the available data.

Here a two stage (nested) m-k fold CV procedure was used (figure 1). In the first stage, the data is partitioned K times into two disjoint sets: a training set and a validation set. The training set is used for learning and fitting the parameters of the classifier. The validation set is used only to assess the performance of a fully-trained classifier.

In the training stage, a second, inner procedure of m-fold cross-validation is used to determine the optimal regularization parameter. Each training set k from the first stage is split into training and testing sets, M times. A regression model is determined using each training set m, for various lambdas within the range of interest (eq. 8). It is then tested by predicting m testing sets and calculating the average prediction MSE errors at each λ (eq. 8).

$$MSE_m = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - Y_i' \right]^2 \quad ; \quad \text{m} = 1 \ldots M \quad (8)$$

$$MSE = \frac{1}{m}\sum_{m=1}^{M} MSE_m \quad (9)$$

Where Y is the observed data and Y' is the fitted model. MSE_m is the error estimate of training CV iteration m. MSE is the average error estimate across all M training iterations.

The λ that minimizes the average MSE across M training iterations is chosen to be the optimal regularization parameter for the model (figure 2). Finally, performance accuracy is tested by prediction of the K validation sets from the first CV stage with the optimal model and calculation of average Error-rate estimate across validation sets. Error-rate is defined as the number of wrongly predicted samples divided by overall number of samples.

In this work, for each iteration of the CV a specific heuristic for dependant data was used (Burman et al., 1994). This CV scheme is known as hv-block CV, a cross validation method for dependent data proposed in (Racine, 2000). Briefly, for a given time period t, the validation sample is constructed using the v observations preceding and following t (2v +1 data points). The training sample is constructed using the observations from the beginning of the sample to the (t - h - v)-th observation and from the (t + h + v) +th observation to the end of the sample (n-2v- 2h-1 data points), the testing set is constructed from the 2h samples between the training and validation sets. The error estimate is then computed in the training sample and used to forecast the validation sample. In the test step the model is tested imputing the 2v + 1 removed observations with their expected value. The performance evaluation is then computed by averaging the prediction errors.

As a set of candidate λ values we used a set of 100 values uniformly sampled on the logarithmic scale (i.e. the ratio of the two successive samples is constant) from the interval [min (si) 10*max (si)], where si are the singular values of the data matrix X. The lower limit of the interval is chosen to be min (si) so that it yields a regularized matrix X with minimal filtering of data components. The upper limit of the range is selected to be larger than the value that is comparable in magnitude to the data components and thus it allows some contribution of the data to the solution. Therefore, as λ approaches 10*max (si) the regularization effect becomes strong enough to represent over-regularization where the solution becomes nearly constant.

Classifier accuracy was estimated with m=k=25 fold CV using all the available data for the given subject. In each iteration of the CV, 70% of the data was used as training set; each such set was divided into 70% training and 30% validation, the remaining 30% of the data was used for testing. The training and validation sets were comprised of continuous data blocks, with one classifier for every combination of training and validation sets and regularization parameter λ. For each subject, we used the labeled set of signals obtained after the preprocessing of one EEG channel at a time, as the input to the classifier problem. This yields N_{freq} = 50 and N_{tps} = 4500.

2.4. Interpretation of classifier weights

For Linear regression (eq. 1 and 3) where each frequency in the time-frequency decomposition of the signal is a feature of the classifier, the absolute value of w_i describes the relative importance of the i^{th} frequency contribution to the classification and the sign describes which of the two classification outcomes is supported by the positive value of the signal x_i. This differs from the classic electrophysiological data analysis, in which time frequency analysis is frequently used by inspection of power at each frequency under each experimental condition rather than relevance of each frequency to the discrimination between experimental conditions.

To define the most important frequencies that contributed to the classification, one can look at the significance of each frequency weight across cross-validations. In such a case the significance of the test must be corrected for multiple comparisons.

2.5. Spatial distribution of performance

The classification is applied to data recorded from a single EEG channel. To assess correctly classifier performance it is necessary to know which channel is most suitable for the inference. We chose not to enforce any preliminary knowledge about the selection of the most informative channel. The classifier was applied to each channel of the EEG separately and the channels with the lowest error estimate were chosen as the most informative for this inference. This way we obtained for each subject, a map of error estimates across channels. This map can be regarded as spatial distribution of most relevant information for inference of a mental state and can be used as a localization method.

2.6. Framework application

For the application of the proposed framework, we used EEG data from an experiment designed to examine the EEG Alpha wave correlates in fMRI. For this, a simultaneous EEG/fMRI study of the Alpha and Berger effect was performed. Only the EEG data of this experiment was used for the presented work. The combined EEG/fMRI results of the experiment are published in (Ben-Simon et al., 2008).

2.6.1. Participants & Study Design

14 healthy volunteers (6 men and 8 women), aged 19-35 (mean 24.8±3.7), signed an informed consent for this study, approved by the Helsinki committee. Subjects were equipped with earphones and were asked by means of audio instructions to open and close their eyes every 30 seconds for a total time of 3 minutes. Subjects were told to lie as still as possible and follow the instructions. Sponge cushions were used to minimize head movements.

The acquired data was examined for the presence of blinks following the instructions, an examination which led to the exclusion of 2 subjects. Two more subjects were excluded from the analysis due to movements in the scanner which induced large artifacts on the EEG. Thus our final analysis included 10 subjects.

2.6.2. EEG acquisition

Continuous EEG data was recorded simultaneously with fMRI acquisition. EEG was acquired using the MRI-compatible BrainAmp-MR EEG amplifier (Brain Products, Munich, Germany) and the BrainCap electrode cap with sintered Ag/AgCl ring electrodes providing 30 EEG channels, 1 ECG channel, and 1 EOG channel (Falk Minow Services, Herrsching-Breitbrunn, Germany). The electrodes were positioned according to the 10/20 system. The reference electrode was between Fz and Cz. Raw EEG was sampled at 5 kHz and recorded using the Brain Vision Recorder software (Brain Products).

2.6.3. EEG data preprocessing

EEG data underwent the following preprocessing stages, similarly to (Ben-Simon et al., 2008; Sadeh et al., 2008):

MR gradient artifacts removal. Artifacts related to the MR gradients were removed from all the EEG datasets using the FASTR algorithm (Iannetti et al., 2005; Niazy et al., 2005) implemented in FMRIB plug-in for EEGLAB (Delorme and Makeig, 2004), provided by the University of Oxford Centre for Functional MRI of the Brain (FMRIB).

Cardio-ballistic artifacts removal. EEG data recorded during an MRI scan is contaminated by Cardioballistic noises induced by the heart's electric and mechanic activity. Cardioballistic artifacts were also removed using the FMRIB plugin.

Down sampling to 250Hz followed by a visual inspection of the EOG data. In this inspection we validated the presence of blinks at the time of the instructions, in order to ensure that the subjects closed and opened their eyes at those times.

Eye movement artifacts removal. Artifacts caused by blinks and eye movements were removed from the EEG recording using Independent Component Analysis (ICA). Artifacts were distinguished from brain activities by inspection of the time course of the components and their projection to scalp sites. Eye blink components' time courses usually have brief large monopolar peaks. Components of Eye blinks should project most strongly to frontal sites on the scalp. Once a certain independent component with the above characteristics is identified, artifact removal can be achieved by simply subtracting the relevant independent component from the original EEG recording (Li and Principe, 2006).

Time-Frequency transformation of each channel signal was calculated using Stockwell (ST) Time Frequency Decomposition (Stockwell et al., 1996). The Stockwell Transform has good time and frequency resolution at low frequencies as well as high time resolution at high frequencies. It is an extension of the continuous wavelet transform (CWT) and is based on a moving and scalable Gaussian window. The transform frequency resolution was set to 1.5Hz with time resolution of 1/250sec.

3. Results

3.1. Comparison of Ridge and Logistic regression

Performance of Logistic Regression was similar to the performance of Linear Ridge Regression classifier and didn't prove to be different under a paired t-test (figure 3A). However prediction error was lower for Linear Ridge regression for all subjects, this was tested with a signed-rank test ($p<0.002$). Importantly, resulting weights for both methods are very similar and do not prove to be different under a paired t-test, $p<0.05$ FDR corrected (see supplementary figure s3).

Since the two regression methods produce highly similar results, from this point of the report, only results based on Linear Ridge regression are presented.

3.2. *Cross validation error estimate*

For a classification problem that uses regularization, one typically expects that the (estimated) classifier error, as function of regularization parameter, will exhibit a clear global minimum. In our case, when plotted against the regularization parameter, the classification error clearly revealed such minimum for all subjects. Error graphs of four subjects, with best classification performance, can be seen in figure 2 while the remaining subjects can be seen in supplementary figure s1.

Since minimizing the error over any free parameters biases the error estimate downwards (Vapnik, 1999), we compared the estimated error to the estimate obtained by applying exactly the same algorithm to the data with randomly scrambled class labels. Classification error of the data with scrambled labels was at chance level for all subjects. Furthermore, the difference between the mean error estimates of scrambled and unscrambled labels was significant for all subjects ($p<0.0001$, FDR corrected), estimated using a Student's t-test subjects (for details see supplementary figure s2).

3.3. Relation between classifier performance and regularization parameter

To inspect the influence of the regularization parameter on prediction error and resulting weight distribution, the results of the Ridge and Logistic classifier were examined with two additional regularization parameters:

1) Sub-optimal regularization parameter, which is 100 times lower than the

optimal regularization parameter chosen by cross-validation.

2) Above optimal regularization parameter which is 100 times greater than the optimal one.

These regularization values were taken in the range mentioned above because of the wide nature of the cross-validation minima as function of regularization parameter resulting in large error differences seen for lambdas 100 times far apart. Ridge regression cross-validation error graphs of four subjects with best classification rates can be seen in figure 2 while the corresponding graphs of the remaining subjects are shown in supplementary figure s2.

Prediction error with above and below optimal regularization was higher than resulting error with optimal regularization parameter for both Ridge and Logistic regression (figure 3B). Moreover, prediction error with sub-optimal regularization is highest with large variability as expected from an unconstrained model. Accordingly the error of the over regularized model is higher than optimal but has lower variability as expected from an over constrained model. Importantly, one should note the low error (under 10% for most subjects) was achieved with optimal regularization. Resulting frequency weights with optimal regularization parameter (chosen with cross validation), above and below optimal regularization parameters had significantly distinct distributions (paired t-test with optimal weights set, $p<0.05$ FDR corrected). Weights distributions for four subjects with best classification error are shown in figure 4 while the weights of the remaining subjects are presented in supplementary figure s5.

The over regularized model produced very similar models for all subjects, its weights distribution is similar to the classical FFT (Fast Fourier Transform) power distribution of the EEG signal in this experiment, showing mainly high power of the Alpha band frequency. The optimal regularization model revealed high contribution of the Alpha band to the prediction as well as lower but significant contribution of other frequencies such as Beta and Gamma. In addition the optimal model revealed a detailed division of the frequencies into bands that contribute positively and negatively to the prediction. This division is not seen by the classical FFT method or by the over constrained model. Lastly, similar to the regularization inspection reported above, the importance of data normalization was also explored using various normalization parameters (see details in supplementary text 3 and supplementary figure s4).

3.4. Comparison to Other Techniques

We compared performance of our regularized ridge regression framework to classical power analysis technique widely used in the neuroscience literature (Ben-Simon et al., 2008; Klimesch et al., 1998; Laufs et al., 2003; Lopes da Silva et al., 1976). This conventional technique uses the alpha power as a marker for eyes state. When alpha power is higher than baseline the subject is assumed to be with closed eyes. When Alpha power is below base line level the subject is assumed to be with opened eyes. Specifically, continuous alpha power from each subjects' time-frequency Stockwell decomposition, averaged across the relevant frequency band (8-12Hz) and across three occipital electrodes (O1, O2, Oz), was thresholded around 50th percentile. At any time point where Alpha power was above the threshold subject was assumed to be with eyes closed. Figure 3d shows that, for all subjects, this technique produced larger classification error than the ones obtained with our regression framework (t-test, $p<0.05$ FDR corrected).

3.5. Spatial Distribution of performance

Linear Ridge regression classifier was applied to data sets of each subject from each electrode separately. The classifier was trained to predict subjects' eyes state (opened or closed). Regularization parameter was chosen for each electrode separately and classifier performance was assessed by means of cross validation.

Good performance was achieved with data from only one electrode, however performance differed between electrodes as can be seen in the average scalp distribution of performance across subjects (figure 5). Distribution of performance for each subject is shown supplementary figure s6. For each subject only one electrode, with best classifier performance was chosen for further analysis (see chosen electrodes in Table 1). The occipital area was the main location of best prediction strength for all subjects. The locations of the electrodes on the scalp are consistent with known findings about the occipital origin of the Alpha band that modulates most evidently with eyes state (Ben-Simon et al., 2008) .

3.6. Feature selection – frequencies above 20Hz

In an effort to understand whether valuable or classifier relevant information may be found beyond 20Hz, the same Ridge Regression classification framework was applied for data with frequencies above 20Hz. The classifier (figure 3C) showed good prediction with this range of frequencies. This indicates that there is

important information for the classifier in higher than 20Hz frequency range. When looking at the resulting frequency weights of this classifier (figure 6 and supplementary figure s8), it can be seen that for subjects with good prediction strength, frequencies above 20Hz have high weights in both methods. This finding reinforces the importance of high frequencies for the classifier.

4. Discussion

We have presented a classification framework, based on Regularized Linear Regression classifier constructed from time-frequency decomposition of an EEG signal. This framework provides tools for constructing classifiers that can distinguish between different brain states and derive a model of frequency distributions that identifies each state. With this framework, we have presented a viable tool for extraction of the important EEG frequencies taking part in a range of mental states.

Application of the proposed framework to a real-world experimental EEG data set revealed interesting patterns of relations between experimental conditions, EEG frequencies, regularization parameters, and classifier choice.

Experimental results show that correct choice of regularization parameter significantly improves classification result and greatly affects the resulting frequency weights of the model. Therefore the choice of regularization parameter with cross-validation played an important role in a successful construction of the classifier. Importantly, it was evident that the over regularized model gives a frequency distribution result that is very similar to conventional state of the art method, i.e. the FFT power spectra of the EEG signal. This demonstrates that the classical FFT analysis approach is an inaccurate approximation of the frequency distribution identifying a brain state as shown with the classifier method proposed here. This emphasizes the advantages of using more sophisticated methods, such as machine learning, for identification of relevant EEG frequency bands. By using the classifier to examine frequency distribution of the Berger effect we are able to receive a more complicated in depth knowledge of the different frequencies involved even in such a simple task as eyes opening and closing.

Experimental results also show that the choice of a specific regression technique does not affect the result significantly. This could imply that the change of frequencies between the two explored brain states (eyes opened and closed) is robust enough to be detected equally well by the two regression methods.

It is widely accepted in the literature that Alpha band frequency from an occipital origin, changes power in transition from eyes closed to open under normal light condition (Ben-Simon et al., 2008; Gevins and Rémond, 1987; Klimesch, 1997; Laufs et al., 2003; Lopes da Silva et al., 1976; Niedermeyer and Lopes da Silva,

2005). However, change in oscillation of other frequencies with this change of state remained largely unexplored. The proposed method enabled us to identify frequencies that distinguish between two brain states (eyes opened or closed) and showed that in addition to Alpha other frequencies behave differently between the two states.

With this framework, we were able to achieve good (higher than 90%) classification performance. Hence, it can be used not only for classical BCI purposes but also for neuroimaging research as it provides an important tool for automatic identification of EEG frequencies that mostly contribute to a certain brain state. Moreover, our framework allows identification of specific frequency sub-bands for each subject and brain state. These findings cannot be achieved with conventional EEG analysis techniques such as Power Spectra analysis (Feige et al., 2005; Goldman et al., 2002; Mantini et al., 2007). In particular, this framework provides a measure for the importance of each frequency to a given task within a combination of all other frequencies of the signal, rather than the importance of each frequency individually. This makes the inference problem very high dimensional and emphasizes the importance of correct regularization due to the curse of dimensionality (Bellman, 1961).

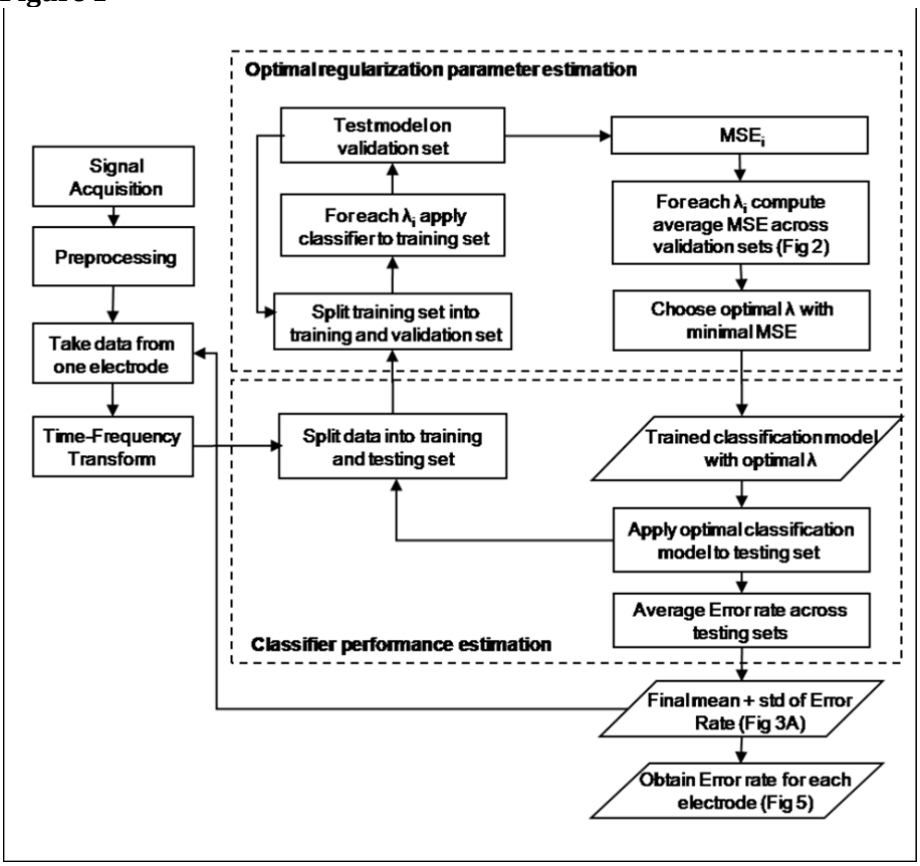
In the recent years a number of approaches for classification of EEG frequencies have been proposed (Besserve and Martinerie, 2011), some have proposed to include spatial information into the classification model as well (Lemm et al., 2010). The proposed framework does not use spatial information for the classification. The classifier is applied to data of each electrode separately. This allows a relatively simple, robust and computationally light tool, which can be used by users inexperienced in machine learning. The proposed method can also be used as a preliminary stage of the analysis, for example for identification of relevant frequency bands for further combined EEG-fMRI analysis (Ben-Simon et al., 2008; De Munck et al., 2009; Laufs et al., 2003), or further connectivity analysis between the frequency bands (Aftanas and Golocheikine, 2001). Inclusion of all electrodes into the model would require a very large amount of samples to avoid over-fitting, and will be computationally heavy. One can include a small number of electrodes into the classification model, for example, for investigation of frequency content of a hypothesized region of interest, or apply a prior spatial feature selection process to reduce problem dimensionality first

(Blankertz et al., 2008). The single electrode approach proposed here gives a classification rate result for each electrode separately; this can give some insight into the location of the most informative electrode with regard to the examined paradigm.

From a computational point of view, we have proposed a robust machine learning approach for prediction of brain states from the EEG signal. This approach uses regularization chosen by two levels of cross-validation. The right choice of regularization minimizes classification error and greatly improves classifier performance and stability. The framework described here was applied to a relatively well known effect in neuroscience (i.e. the Berger effect). Having established the most stable and robust framework in the current work, it can now be applied to more complex classification problems. For example a similar linear regression with nested cross validation approach has been successfully applied to investigation of temporal content of ERP (Hasson-Meir, 2011). Our framework can be used not only for BCI purposes but also for investigation of time-frequency decomposition of brain signals such as EEG or MEG and automatic identification of relevant frequency features of an explored brain state.

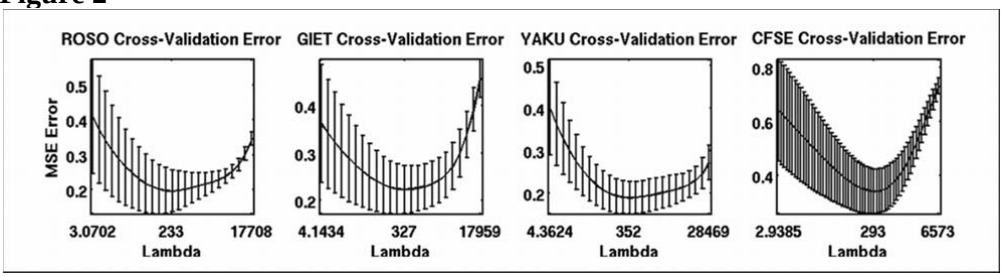
Figure Legends

Figure 1



Flow chart of the regression framework. Pre-processed data is introduced into a nested m-k fold cross-validation procedure. In each iteration of the CV, the data is partitioned into two disjoint sets: training and testing sets. The training sets are used for choosing the best model and the test sets are used to check the predictive accuracy of the model. For each training set an additional inner m-fold cross-validation procedure is applied for selecting the optimal regularization parameter on the training sets, where n is the number of averaged cross-validation iterations. Performance of the model with optimal parameters is evaluated on the testing set in the first level of CV. The average error of classifier performance across testing sets is the error rate of the model. The process is repeated for each electrode separately.

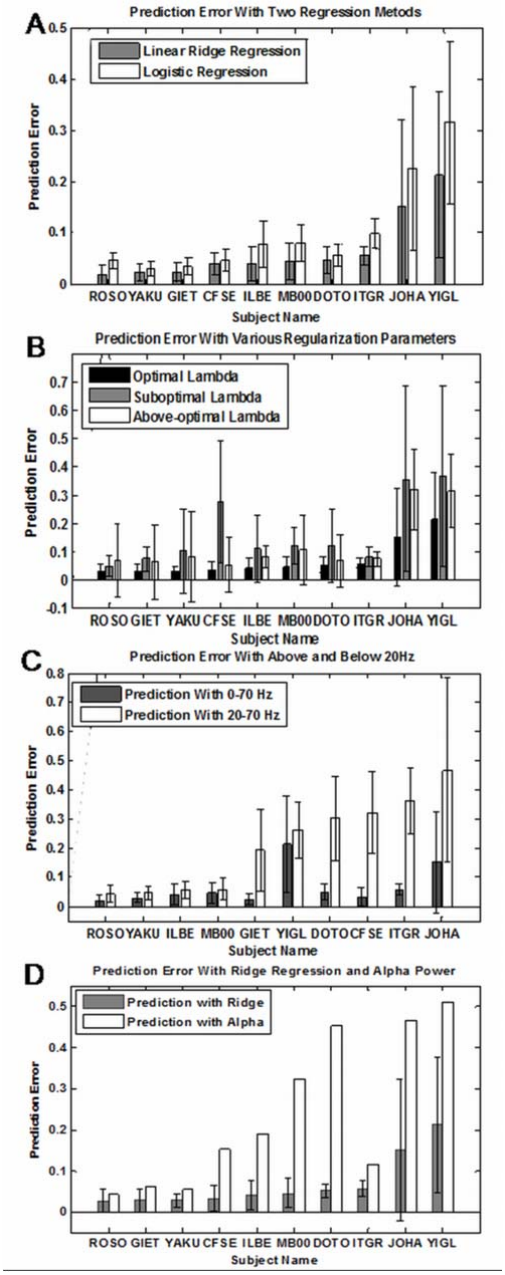
Figure 2



Cross validation error estimate obtained by 25-fold cross validation with Ridge regression. Error bars denote 1-std – wide margin around the error estimate. Only

a segment between sub-optimal and above-optimal lambda is shown. The middle value of the lambda axis is the optimal lambda that minimizes cross-validation error. First and last values of lambda are approximately 100 smaller and larger than the optimal lambda, respectively.

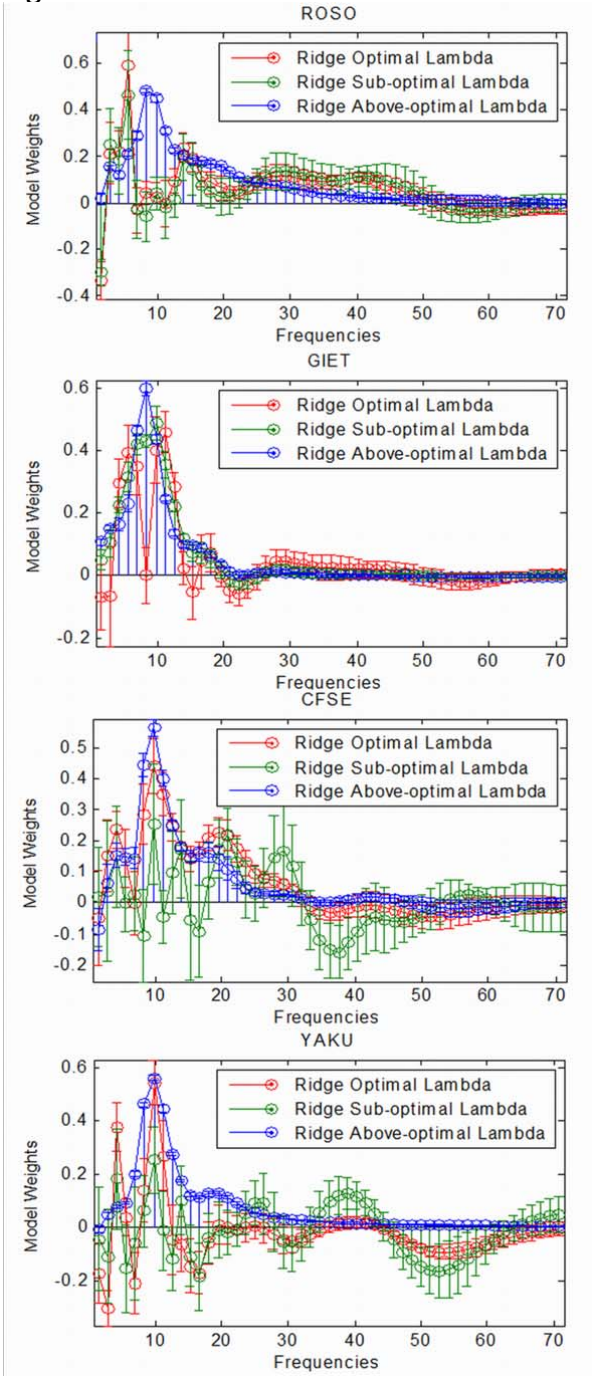
Figure 3



Comparison of various classification errors. Error bars denote 1-std margin across cross validation estimates. A. Comparison of Prediction Error achieved with Linear Ridge Regression (gray) and Regularized Logistic regression (white). B. Ridge Regression prediction error with optimal regularization parameter determined by cross validation (black), sub-optimal regularization which is 100 times smaller than optimal (gray) and above-optimal which is 100 time larger than optimal (white). C. Performance of Ridge regression with frequencies above and below 20Hz. Performance of linear ridge regression classifier with data from all frequencies (gray) and data with frequencies only above 20Hz included (white).

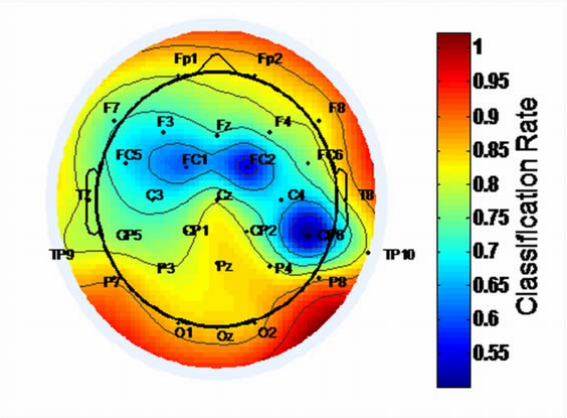
Performance without low frequencies is significantly better than chance for all subjects (t-test, p<0. 01, FDR corrected). D. Comparison of classification error between eyes opened and closed states in light conditions using regularized ridge regression framework (gray) and conventional Alpha power analysis (white).

Figure 4



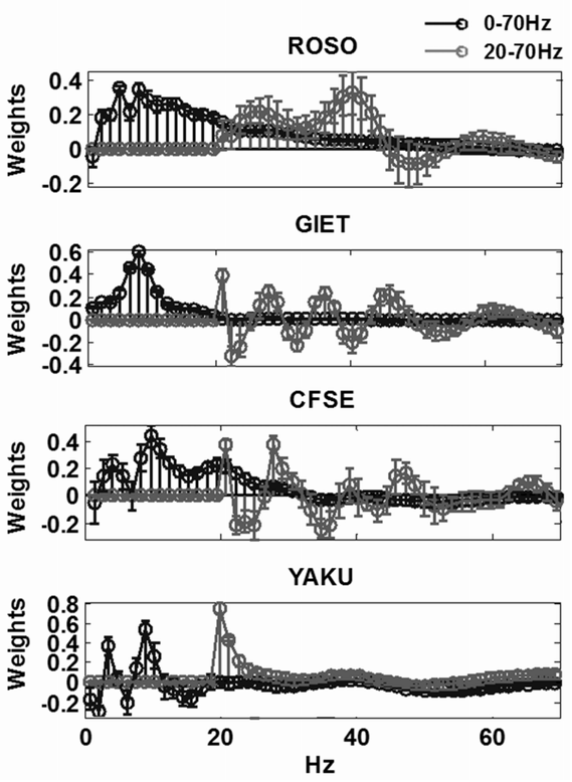
Resulting Ridge regression weight sets. In red: Result with optimal regularization parameter determined by cross validation. In green: result with regularization parameter 100 times smaller than the optimal one. In blue: result with regularization parameter 100 times larger than the optimal one. The two un-optimal weight sets differ significantly from the optimal one. Each weight is normalized by magnitude for comparison.

Figure 5



Spatial distribution of classifier performance in normal light condition. Linear Ridge regression Classifier was applied to each channel separately to identify channels with best prediction strength. This figure shows distribution of classifier performance across channels averaged for all subjects.

Figure 6



Comparison of weights resulting from classification using all frequencies and classification using only frequencies above 20Hz. Subjects are sorted by strength of prediction with frequencies above 20Hz. Note that for subjects with good prediction strength there are large weights for frequencies above 20 Hz for both methods, indicating there is significant information beyond the Alpha band.

References

Aftanas L, Golocheikine S. Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution EEG investigation of meditation. Neuroscience Letters, 2001; 310: 57-60.

Anderson CW, Stolz EA, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng, 1998; 45: 277-86.

Baar E, Baar-Eroglu C, Karaka S, Schürmann M. Gamma, alpha, delta, and theta oscillations govern cognitive processes. Int J Psychophysiol, 2001; 39: 241-8.

Bellman RE. Adaptive control processes : a guided tour. Princeton University Press: Princeton, 1961.

Ben-Simon E, Podlipsky I, Arieli A, Zhdanov A, Hendler T. Never resting brain: simultaneous representation of two alpha related processes in humans. PLoS One, 2008; 3: e3984.

Berger H. Über das elektrenkephalogramm des menschen. European Archives of Psychiatry and Clinical Neuroscience, 1929; 87: 527-70.

Besserve M, Martinerie J. Extraction of functional information from ongoing brain electrical activity. IRBM, 2011; 32: 27-34.

Besserve M, Martinerie J, Garnero L. Improving quantification of functional networks with EEG inverse problem: Evidence from a decoding point of view. Neuroimage, 2011.

Bishop C. Neural networks for pattern recognition. Oxford Univ Pr, 2005.

Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller KR. Optimizing spatial filters for robust EEG single-trial analysis. Signal Processing Magazine, IEEE, 2008; 25: 41-56.

Burman P, Chow E, Nolan D. A cross-validatory method for dependent data. Biometrika, 1994; 81: 351-8.

Cox D, Savoy R. Functional magnetic resonance imaging (fMRI) "brain reading" : detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage, 2003; 19: 261-70.

De Munck J, Gon alves S, Mammoliti R, Heethaar R, Lopes da Silva F. Interactions between different EEG frequency bands and their effect on alpha-fMRI correlations. Neuroimage, 2009; 47: 69-76.

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Meth, 2004; 134: 9-21.

Duda RO, Hart PE, Stork DG. Pattern classification, 2nd ed. Wiley: New York ; Chichester, 2001.

Feige B, Scheffler K, Esposito F, Di Salle F, Hennig J, Seifritz E. Cortical and subcortical correlates of electroencephalographic alpha rhythm modulation. Journal of neurophysiology, 2005; 93: 2864.

Gevins AS, Rémond A. Methods of analysis of brain electrical and magnetic signals. Elsevier: Amsterdam ; Oxford, 1987.

Goldman RI, Stern JM, Engel J, Jr., Cohen MS. Simultaneous EEG and fMRI of the alpha rhythm. Neuroreport, 2002; 13: 2487-92.

Hasson-Meir Y, Zhdanov, A., Hendler, T., Intrator, N. Inference Of Brain Mental States from Spatio-Temporal Analysis of EEG Single Trials. Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, 2011: 59-66.

Haynes J, Rees G. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience, 2006; 7: 523-34.

Iannetti G, Niazy R, Wise R, Jezzard P, Brooks J, Zambreanu L, Vennart W, Matthews P, Tracey I. Simultaneous recording of laser-evoked brain potentials and continuous, high-field functional magnetic resonance imaging in humans. Neuroimage, 2005; 28: 708-19.

Keirn ZA, Aunon JI. A new mode of communication between man and his surroundings. IEEE Trans Biomed Eng, 1990; 37: 1209-14.

Klimesch W. EEG-alpha rhythms and memory processes. Int J Psychophysiol, 1997; 26: 319-40.

Klimesch W, Doppelmayr M, Russegger H, Pachinger T, Schwaiger J. Induced alpha band power changes in the human EEG and attention. Neuroscience Letters, 1998; 244: 73-6.

Komarek P, Moore A. Fast robust logistic regression for large sparse datasets with binary outputs. Citeseer, 2003.

Laufs H, Kleinschmidt A, Beyerle A, Eger E, Salek-Haddadi A, Preibisch C, Krakow K. EEG-correlated fMRI of human alpha activity. Neuroimage, 2003; 19: 1463-76.

Lemm S, Blankertz B, Dickhaus T, Muller KR. Introduction to machine learning for brain imaging. Neuroimage, 2010.
Li R, Principe JC. Blinking artifact removal in cognitive EEG data using ICA. Conf Proc IEEE Eng Med Biol Soc, 2006; 1: 5273-6.
Lopes da Silva FH, van Rotterdam A, Barts P, van Heusden E, Burr W. Models of neuronal populations: the basic mechanisms of rhythmicity. Prog Brain Res, 1976; 45: 281-308.
Mantini D, Perrucci M, Del Gratta C, Romani G, Corbetta M. Electrophysiological signatures of resting state networks in the human brain. Proceedings of the National Academy of Sciences, 2007; 104: 13170.
McFarland DJ, Miner LA, Vaughan TM, Wolpaw JR. Mu and beta rhythm topographies during motor imagery and actual movements. Brain Topography, 2000; 12: 177-86.
Minka T. A comparison of numerical optimizers for logistic regression. Unpublished draft, 2003.
Niazy R, Beckmann C, Iannetti G, Brady J, Smith S. Removal of FMRI environment artifacts from EEG data using optimal basis sets. Neuroimage, 2005; 28: 720-37.
Niedermeyer E, Lopes da Silva FH. Electroencephalography : basic principles, clinical applications, and related fields, 5th ed. Lippincott Williams & Wilkins: Philadelphia ; London, 2005.
O'Toole A, Jiang F, Abdi H, Pénard N, Dunlop J, Parent M. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J Cognitive Neurosci, 2007; 19: 1735-52.
Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: A tutorial overview. Neuroimage, 2009; 45: S199-S209.
Pfurtscheller G, Lopes da Silva F. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clinical Neurophysiology, 1999; 110: 1842-57.
Picard R, Cook R. Cross-validation of regression models. J Am Stat Assoc, 1984; 79: 575-83.
Racine J. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. Journal of econometrics, 2000; 99: 39-61.
Sadeh B, Zhdanov A, Podlipsky I, Hendler T, Yovel G. The validity of the face-selective ERP N170 component during simultaneous recording with functional MRI. Neuroimage, 2008; 42: 778-86.
Stockwell R, Mansinha L, Lowe R. Localization of the complex spectrum: the S transform. IEEE transactions on signal processing, 1996; 44: 998-1001.
Tikhonov A. Solution of incorrectly formulated problems and the regularization method. 1963: 1035-8.
Tomioka R, Müller K. A regularized discriminative framework for EEG analysis with application to brain-computer interface. Neuroimage, 2010; 49: 415-32.
Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw, 1999; 10: 988-99.
Zhdanov A, Hendler T, Ungerleider L, Intrator N. Inferring functional brain States using temporal evolution of regularized classifiers. Comput Intell Neurosci, 2007: 52609.