Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Mind Reading with Regularized Multinomial Logistic Regression

Heikki Huttunen · Tapio Manninen · Jukka-Pekka Kauppi · Jussi Tohka

Received: date / Accepted: date

Abstract In this paper, we consider the problem of multinomial classification of magnetoencephalography (MEG) data. The proposed method participated in the MEG mind reading competition of ICANN'11 conference, where the goal was to train a classifier for predicting the movie the test person was shown. Our approach was the best among 10 submissions, reaching accuracy of 68 % of correct classifications in this five category problem. The method is based on a regularized logistic regression model, whose efficient feature selection is critical for cases with more measurements than samples. Moreover, a special attention is paid to the estimation of the generalization error in order to avoid overfitting to the training data. Here, in addition to describing our competition entry in detail, we report selected additional experiments, which question the usefulness of complex feature extraction procedures and the basic frequency decomposition of MEG signal for this application.

H. Huttunen
Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland
Tel.: +358-40-849-0799
E-mail: heikki.huttunen@tut.fi

T. Manninen
Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland

J-P. Kauppi
University of Helsinki, Department of Computer Science and HIIT, P.O. Box 68, FI-00014 Helsinki, Finland.

J. Tohka
Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland

Keywords logistic regression · elastic net regularization · classification · decoding · magnetoencephalography · natural stimulus

1 Introduction

Functional neuroimaging relies on statistical inference for explaining relations between measured brain activity and an experimental paradigm. During the recent years, supervised classification has become increasingly important methodology in analyzing functional neuroimaging data [49] within functional magnetic resonance imaging (fMRI) [41] as well as in electroencephalography (EEG) and magnetoencephalography (MEG) (for reviews, see, e.g., [32,37]) with earliest papers tracing back to early 90's [24,27,31]. The significance of the topic is witnessed, for example, by special issues dedicated to brain decoding in technical [49] and more applied journals [16]. In brain research, pattern classifiers can be used to answer the questions *is there* information about a variable of interest (pattern discrimination), *where* is the information (pattern localization) and *how* is that information encoded (pattern characterization) as explained in more detail in [37] and [35]. A major benefit of this pattern classification approach over the more traditional analysis methods is that it, in principle, allows the identification of the set of data patterns, which are diagnostic for engagement of a particular task [39]. Important results regarding, e.g., face processing [15] and early visual processing [17,22] in human brain have been obtained using the technique.

The uses of pattern classifiers on EEG and MEG data have concentrated around applications in the brain computer interfaces (BCIs) for which there is a large body of literature [4,3]. The majority of studies in BCIs

have focused on EEG with relatively few channels. For example, Van Gerven et al. [50] used regularized logistic regression to classify imagined movements of right or left hand based on EEG data from 16 channels. Perhaps more relevant to the present work is [48], where regularized logistic regression was applied for two different problems with 64-channel EEG data: 1) 2-category self-paced finger tapping task from the BCI Competition 2003 and 2) a P300 speller system task from the BCI competition III, which is a 36-category classification problem. In the BCI-IV competition[1], one task was the classification of the direction of wrist movements based on 10-channel MEG data [46]. Surprisingly, only one of four entries to the competition clearly exceeded the chance level in classification accuracy. Other studies focusing on the decoding of MEG data include Zhdanov et al. [53], who applied regularized linear discriminant analysis to MEG signals recorded while subjects were presented with images from two different categories (faces and houses). Chan et al. [6] applied a support vector machine (SVM) classifier to decode the data that was recorded using simultaneous scalp EEG and MEG while the subjects were performing auditory and visual versions of a language task. Rieger et al. [43] applied an SVM to test whether it is possible to predict the recognition of briefly presented natural scenes from single trial MEG-recordings of brain activity and to investigate the properties of the brain activity that is predictive of later recognition. Besserve et al. [2] applied an SVM to classify between MEG data recorded during a visuomotor task and resting condition.

In this paper, we propose a method for supervised classification of MEG data. Our method is based on multinomial logistic regression with elastic net penalty [9,54] and it was the most accurate in the "Mind Reading from MEG" challenge organized in conjunction with the International Conference on Artificial Neural Networks (ICANN 2011) in June 2011 [19]. The task in the competition was to train a classifier for predicting the type of a movie-clip being shown to the test subject based on MEG recordings. In more detail, the subject was viewing video stimuli from five different categories (football match, 2 different feature films, recording of natural scenery, and artificial stimulus) while MEG signal was recorded. The MEG signal recorded from 204 channels was cut into non-overlapping one-second epochs by the competition organizers. These epochs along with the corresponding category labels were released to the participants as training samples, and the task was to build a classifier that can predict categories of unseen samples. The classification performance was assessed by the competition organizers based on the independent test set, which was hidden from participants during the competition. Our method achieved 68 % accuracy on the test samples and was a clear winner among 10 methods participating to the competition. In addition to ICANN MEG data, we highlight the good performance of our method with MEG data from the BCI-IV competition where the experimental paradigm is much simpler than with the ICANN MEG data.

There are certain key differences between typical BCI and our "Mind reading from MEG" decoding applications as laid out by Zhdanov et al. [53]. Perhaps most importantly, the dimension of input data is much higher in MEG than that of EEG typically used in BCI applications, the number of samples is much smaller, and the behavioral paradigm is much more complex here. Moreover, all the above cited uses of supervised classifiers in MEG [53,6,43,2] differ from the prediction task in the ICANN competition in that they were based on strictly controlled behavioral paradigm and the knowledge about the paradigm was often applied in the feature extraction. Moreover, all except Chan et al. [6] considered only a binary (two-category) classification problem.

The elastic net has been used in neuroimaging with fMRI data sets in the context of classification [10,41] and regression [5] problems, but not with MEG data and, in the classification setting, not in a naturalistic behavioral paradigm like movie watching studied in this paper. As our main contribution, we propose using a linear classifier for classification of MEG data and illustrate its efficiency using simple features such as the mean and standard deviation of the measurement signals. Despite the simplicity of our approach, the experimental results confirm an excellent performance in cases with complex behavioral paradigms (which movie is shown) as well as in simple setups (which direction the hand is moving).

The rest of the paper is organized as follows. After introducing the data and the acquision setup, we will describe the details of the proposed method in Section 2. In Section 3 we present results of applying the method for the ICANN dataset with basic set of mean and standard deviation features (Section 3.1), with a larger set of statistical quantities as features (Section 3.2), and with frequency band energy features (Section 3.3). Moreover, we consider the classification performance for a modified version of the ICANN challenge problem in Section 3.4, and experiment with data from an earlier BCI-IV MEG decoding challenge [46] in Section 3.5. Finally, Section 4 discusses the results and concludes the paper.

[1] `http://www.bbci.de/competition/iv/index.html`

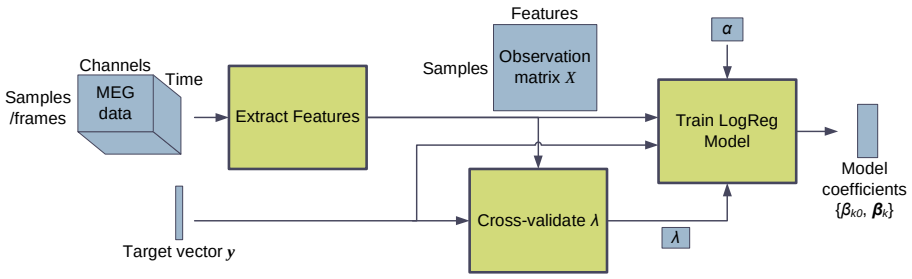


Fig. 1: Block diagram of the proposed method.

2 Methods

2.1 Material

In the results section we study the efficiency of our method using the data released in the mind reading competition[2]. The data set consists of within-subject MEG signals recorded from a single test person while watching five different video stimuli without audio:

1. **Artificial:** Animated shapes or text
2. **Nature:** Nature documentary clips
3. **Football:** Soccer match clips
4. **Bean:** Part from the comedy series "Mr. Bean"
5. **Chaplin:** Part from a Chaplin movie

The provided measurements consist of 204 gradiometer channels, and the length of each individual epoch is one second and the sampling rate is 200 Hz. Moreover, the five band-pass filtered versions of the signal are also included in the measurement data, with bands centered on the frequencies of 2 Hz, 5 Hz, 10 Hz, 20 Hz, and 35 Hz.[3]

The MEG measurements were recorded on two separate days such that the same set of video stimuli was shown to a test person on both days. Stimuli labeled as either Artificial, Nature, or Football (short clips) were presented as randomly ordered sequences of length 6 – 26 s with a 5 s rest period between the clips, while Bean and Chaplin (movies) were presented in two consecutive clips of approximately 10 minutes. In the competition data, the measurements are cut into one-second epochs that are further divided into training and testing such that the training data with known class labels contains 677 epochs of first day data and 50 epochs of second day data while the secret test data contains 653 epochs of second day data only. Note that the ground truth class labels for the test recordings have been released after the end of the competition.

The data is provided in a randomized order, and the complete signal cannot be reconstructed based on the individual signal epochs. During the competition, the competitors were told that the secret test data comes from the second day measurements only and that—similar to the training data—it is approximately class-balanced. The division between training and test data was elaborate. In particular, 33 % of the test samples consist of recording during stimuli not seen in the training phase in order to test the ability of the classifiers to generalize to new stimuli. A more detailed description of the data can be found in the challenge report by Klami et al. [25].

2.2 Overview

Our method follows the strategy of feature extraction followed by a multinomial linear logistic regression classifier. We use elastic net penalization in estimating the model parameters, which results in a sparse model and works as an embedded feature selector. The structure of our approach is illustrated in the block diagram of Figure 1.

More specifically, the training and error estimation procedures consist of two nested cross-validation (CV) loops as illustrated in Algorithm 1. The outer loop is used for estimating the performance for the unlabeled test data using, e.g., $N = 200$ splits of the training data, while the inner (e.g., $M = 5$ fold) CV loop is used for selection of classifier parameter λ (see Section 2.4).

The high computational complexity of simultaneous error estimation and parameter selection can be clearly seen from the pseudo code. In order to speed up the development, our method uses parallel validation spread over numerous processors as also described in Section 2.5. A Matlab implementation of our method is available for download[4].

[2] The data can be downloaded from `http://www.cis.hut.fi/icann2011/meg/measurements.html`

[3] Note, that the challenge report [25] erroneously states the frequency features to be *the envelopes* of the frequency bands. However, the data consists of the plain frequency bands; see the erratum at `http://www.cis.hut.fi/icann2011/meg/megicann_erratum.pdf`.

[4] `http://www.cs.tut.fi/~hehu/mindreading.html`

Algorithm 1 Error estimation and parameter selection using nested cross validation.

Initialize α to a fixed value, e.g., $\alpha = 0.8$.
// Outer CV loop:
for $n = 1 \rightarrow N$ **do**
  Divide the training data into training and validation sets as described in Section 2.5
  // Select the best λ using M-fold CV:
  for $\lambda = \lambda_{\min} \rightarrow \lambda_{\max}$ **do**
    // Inner M-fold CV loop:
    for $j = 1 \rightarrow M$ **do**
      Train with all training data except the j^{th} fold.
      Estimate the error e_j by classifying the j^{th} fold.
    end for
    The error estimate $e_{\lambda,n}$ is the mean of e_j.
  end for
end for
The error estimate for each λ is the mean of $e_{\lambda,n}$ ovar all $n = 1, \ldots, N$.
Classify the test data using the λ with smallest error.

2.3 Feature Extraction

There were 204 (channels) $\times$ 200 (time points, N) = 40800 measurements per one exemplar and, thus, the possible number of features is much larger than the number of training samples. We approach this problem by first deriving a pool of simple summary features, and then feed these to the joint feature selection and classification. The full set of features consists of 11 simple statistical quantities listed in Table 1. Some of the features are proposed earlier in the literature (e.g., the mean [29]), but most were chosen due to their simplicity and widespread use in statistics.

The ICANN MEG challenge data includes a nonzero DC component, and thus many epochs exhibit either increasing or decreasing linear trend. As the significance of these random fluctuations for predicting the brain activity is unclear, we decided to calculate the features using also a detrended version of the time series in addition to the raw measurement. Moreover, this is in coherence with the usual preprocessing of several MEG studies, which use a bandpass filter to remove low frequency components, e.g., below 5 Hz. However, detrending does not have boundary problems and is thus favorable especially with short segments. For example, a frequency selective finite impulse response (FIR) filter with N taps requires $N - 1$ past samples in the delay line, which are not available in the beginning of the signal.

Detrending simply fits a slope into the time series and calculates the residual. More specifically, denote one epoch of the MEG signal from the i^{th} channel by $\mathbf{s}_i = [s_i(1), \ldots, s_i(N)]^T$, for $i = 1, ..., 204$. Then the linearly detrended signal $\tilde{s}_i(n)$ is defined by

$$\tilde{s}_i(n) = s_i(n) - \hat{a}_i(n - 100.5) - \hat{b}_i, \quad (1)$$

where the slope $\hat{a}_i$ and intercept $\hat{b}_i$ are obtained by a least squares fit minimizing $\sum_{n=1}^{N} \tilde{s}_i^2(n)$. Note that the value 100.5 subtracted from n is selected as the midpoint of time indices $1, 2, \ldots, 200$. With this particular value the intercept $\hat{b}_i$ becomes equal to the sample mean.

With the above notation, we can define our pool of features consisting of the following set of statistics $\{x^{(1)}, x^{(2)}, \ldots, x^{(11)}\}$, where each element $x^{(j)} = \{x_i^{(j)} \mid i = 1, 2, \ldots, 204\}$ contains the specific feature values calculated from the measurement signals of all the channels. The extracted features are listed in Table 1.

With this selection, the total number of features extracted from the ICANN MEG challenge data becomes $p = 11 \times 204 = 2244$ if using only the raw measurements; or $p = 6 \times 11 \times 204 = 13464$ if using also the five bandpass filtered channels. With only a few hundred training samples, the problem is clearly ill-posed with a large set of highly correlated predictors. Thus, a natural direction is to seek for an efficient feature selection method, which we will consider next.

Table 1: The pool of features extracted from the MEG signals. Notation "(d)" is used for indicating the detrended features.

Intercept	$x_i^{(1)} = \hat{b}_i$
Slope	$x_i^{(2)} = \hat{a}_i$
Variance (d)	$x_i^{(3)} = \frac{1}{N}\sum_{n=1}^{N} \tilde{s}_i^2(n)$
Std. dev. (d)	$x_i^{(4)} = \sqrt{x_i^{(3)}}$
Skewness (d)	$x_i^{(5)} = \frac{1}{N}(x_i^{(4)})^{-3}\sum_{n=1}^{N} \tilde{s}_i^3(n)$
Kurtosis (d)	$x_i^{(6)} = \frac{1}{N}(x_i^{(4)})^{-4}\sum_{n=1}^{N} \tilde{s}_i^4(n)$
Variance	$x_i^{(7)} = \frac{1}{N}\sum_{n=1}^{N}(s_i(n) - \hat{b}_i)^2$
Std. dev.	$x_i^{(8)} = \sqrt{x_i^{(7)}}$
Skewness	$x_i^{(9)} = \frac{1}{N}(x_i^{(8)})^{-3}\sum_{n=1}^{N}(s_i(n) - \hat{b}_i)^3$
Kurtosis	$x_i^{(10)} = \frac{1}{N}(x_i^{(8)})^{-4}\sum_{n=1}^{N}(s_i(n) - \hat{b}_i)^4$
Fluctuation	$x_i^{(11)} = \frac{1}{N-1}\left\lvert\sum_{n=2}^{N} \operatorname{sgn}(s_i(n) - s_i(n-1))\right\rvert$

2.4 Logistic Regression with Elastic Net Penalty

After extracting multiple candidate features for each of the 204 channels, we still have a large number of features compared to the number of training samples making the prediction problem ill-posed. Further, we are not sure, which features work the best or even turn out useful in our case. To cope with the ambiguity, we use a *logistic regression* model (also known as the *logit model*) with *elastic net* regularization [54]. In addition

to designing a classifier, the elastic net includes a sparsity enforcing regularization term and thus works as an embedded feature selector that automatically selects the set of relevant features and channels from the pool of candidates.

More specifically, the symmetric multinomial logistic regression models the conditional probability of class $k = 1, 2, \ldots, K$ given the p-dimensional feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ as

$$p_k(\mathbf{x}) = \frac{\exp(\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x})}{\sum_{j=1}^{K} \exp(\beta_{j0} + \boldsymbol{\beta}_j^T \mathbf{x})}, \quad (2)$$

where β_{k0} and $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})^T$ are the coefficients of the model [14]. For this model to be valid we have to assume mixture or $\mathbf{x}$-conditional sampling [1] or—in a more relaxed form—that the class frequencies are (approximately) the same in the training and test data. Despite of the apparent nonlinearity of Equation (2), the resulting classifier is linear and the class k^* of a test sample $\mathbf{x}$ is selected as $k^* = \arg\max_k \left\{\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}\right\}$.

In the elastic net framework, the training of the logistic regression model consists of estimating the unknown parameters $\{\beta_{k0}, \boldsymbol{\beta}_k\}_1^K$ by maximizing the penalized log-likelihood

$$\sum_{i=1}^{M} \log p_{k_i}(\mathbf{x}_i) - \lambda \sum_{k=1}^{K} \left(\alpha ||\boldsymbol{\beta}_k||_1 + (1-\alpha)||\boldsymbol{\beta}_k||_2^2\right), \quad (3)$$

where $k_i \in \{1, 2, \ldots, K\}$ denotes the true class of the i^{th} training sample $\mathbf{x}_i$ $(i = 1, 2, \ldots, M)$. The regularization term is a combination of the ℓ_1 and ℓ_2 norms of the coefficient vectors $\boldsymbol{\beta}_k$, and the weights for both types of norms are determined by the mixing parameter $\alpha \in [0, 1]$. The extent of regularization is controlled by the second regularization parameter $\lambda \geq 0$.

The role of parameter α is to determine the type of regularization. When $\alpha = 0$, the ℓ_1 norm vanishes and the purely ℓ_2 regularized result can be expected to work well in cases, where there are several noisy and mutually correlating features. This is because penalizing the ℓ_2 norm brings the coefficients of the correlating features closer to each other resulting in noise reduction in form of averaging. On the other hand, when $\alpha = 1$, the ℓ_2 norm disappears, which produces a generalized version of the *Least Absolute Shrinkage and Selection Operator (LASSO)* [47]. The LASSO is widely used in regression problems and known for its ability to produce sparse solutions where only a few of the coefficients are nonzero, and this property carries over to the elastic net (except for the case $\alpha = 0$). Thus, both the LASSO and the elastic net can be efficiently used as an implicit feature selectors.

The role of the parameter λ is to control the strength of the regularization effect: The larger the value of λ, the heavier the regularization. For small values of λ, the solution is close to the maximum likelihood solution, while large values of λ allow only restricted solutions and push the coefficients towards zero. In practice, the values of both regularization parameters α and λ are determined by CV, i.e., all combinations over a fixed grid are tested and the CV errors are compared. Note that this CV round is separate from that of Section 2.5, nested inside the error estimation loop of the entire solution.

In the case of Equation (2), the logit model is symmetric unlike the traditional multinomial logistic regression model, where one category is selected as the base category against which all the other categories are compared (see, e.g., Kleinbaum and Klein [26]). Note that while the coefficients of the model (2) are not identifiable without constraints, the penalty term in Equation (3) solves the ambiguity in a natural way [9]. This symmetry is useful here because regression coefficients are easier to interpret in the classification context. The linear discriminant functions of the classifier have the same parametric form $g_k(\mathbf{x}) = \beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}$ for every class $k = 1, \ldots, K$. Thus, the larger the $|\beta_{kj}|$, the more important the feature j is for the discrimination assuming the features are normalized to an equal variance (note that this is different from normalizing the original data). The traditional asymmetric model would lead to discriminant functions $g_k(\mathbf{x}) = \beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}$ for $k = 1, \ldots, K-1$ and $g_K(\mathbf{x}) = 0$, i.e., parametric form of the discriminant functions would differ between the base class K and the other classes [52, Page 161].

Elastic net regularized generalized linear models including logistic regression models can be efficiently fit using a coordinate descent algorithm proposed by Friedman et al. [9]. There is also a MATLAB implementation available[5].

2.5 Performance Assessment

An important aspect for the classifier design is the error assessment. This was challenging in the mind reading competition, because only a small amount (50 samples) of the test dataset was released with the ground truth. There is reason to believe that the characteristics of the data from the two days can be different. Additionally, we obviously wanted to exploit the second day data also for training the model. Since we wanted to maximize our accuracy on the secret test data, we concentrated

[5] `http://www-stat.stanford.edu/~tibs/glmnet-matlab`

our error estimation to the 50 second day samples with annotation.

A natural cross-validation error estimation technique would be the leave-one-out error estimator for the second day data. More specifically, we would train with all the first day data and 49 samples of the second day data and test with the remaining second day sample. This way there would be 50 test cases whose mean would be the leave-one-out error estimate. However, we were concerned about the small number of test cases and the high variance of the CV error estimator (see, e.g., [8] and references thereof), and decided to consider alternative divisions of the second day data to training and testing.

As a result, we randomly divided the 50 test day samples into two parts of 25 samples. The first set of 25 samples was used for training, and the other for performance assessment[6]. Since the division can be done in $\binom{50}{25} > 10^{14}$ ways, we have more than enough test cases for estimating the error distribution. This approach gives slightly too pessimistic error estimates because only half of the second day data is used for training (as opposed to 98 % with leave-one-out), but has smaller variance due to larger number of test cases. Moreover, the pessimistic bias is not a problem, because we are primarily interested in comparing feature sets during method development rather than actually assessing the absolute prediction error.

The imbalance in the number of samples between the first and second day data is quite significant, because with the above division the training set contains more than 25 times more first day data than second day data. Since we wanted to emphasize the role of the second day data, we assigned a higher cost to their misclassification. After experimentation, the misclassification cost for all second day samples was set three-fold the cost of first day samples.

The remaining problem in estimating the error distribution is the computational load. One run of training the classifier with CV of the parameters takes typically $10 - 30$ minutes. If, for example, we want to test with 100 test set splits, we would be finished after a day or two. For method development and for testing different features this is definitely too slow. However, the error estimation can be easily parallelized; simply by testing each division of the test data on a different processor. For example, in our case we had access to a grid computing environment with approximately 1000 processors, and we were able to obtain an accurate error estimate in a matter of minutes instead of hours or days.

[6] In the subsequent sections we refer to the first day data as *training data*, the 25 training samples from the second day as *validation data* and the remaining 25 samples from the second day as *test data*. The 653 originally unlabeled test samples from the second day are called *secret test data*.

Table 2: Choice of parameter α for selected experiments.

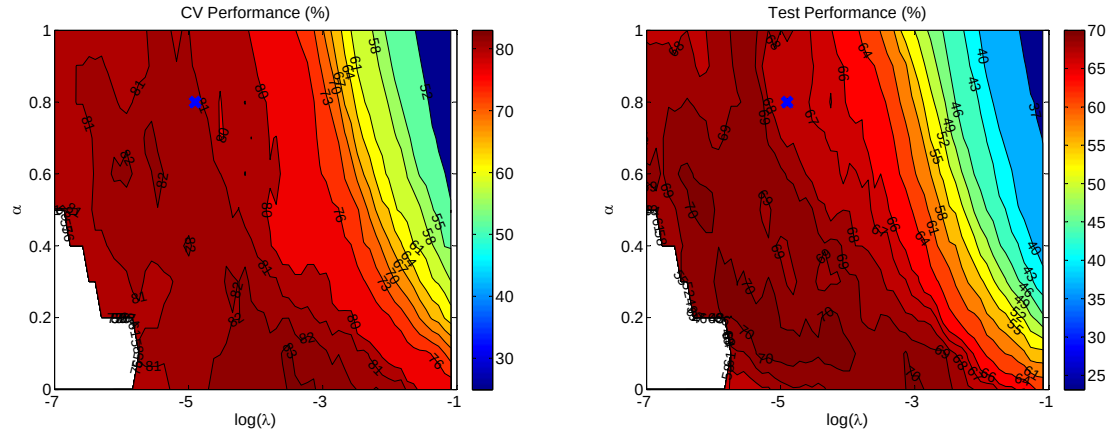
Experiment	α *selected by CV*
Section 3.1	0.8
Section 3.3 (1020 features)	0.1
Section 3.3 (510 features)	0.9
Section 3.5 (10 features; S1)	0.6
Section 3.5 (10 features; S2)	1.0
Section 3.5 (220 features; S1)	1.0
Section 3.5 (220 features; S2)	0.0

2.6 Parameter Selection

The elastic net classifier has two parameters affecting the solution: α and λ of Eq. (3), and they can be selected using cross-validation. However, it turns out that the obvious method of selecting both α and λ together in a 2-dimensional grid (i.e., inside the outer CV loop of Algorithm 1) results in a worse validation and test performance than Algorithm 1, where the parameter α is selected outside both CV loops. More specifically, different values are tested for the regularization parameter α while λ is automatically selected by 5-fold CV. The performance surface for the training data (using cross-validation) and the secret test data is shown in Figures 2a and 2b, respectively.

Figures 2a and 2b also describe the robustness of the model to the choice of parameters. It can be seen that the plots are very similar in shape, and the top is flat for a wide range of α and λ. In particular, the algorithm is insensitive to the selection of α: For any choice of α, there exists a value of λ, which is within 1.7 percentage points from the absolute optimum for the secret test data. Thus, the performance is insensitive to slightly erroneous parameters settings.

In Section 3 we cross-validated also the parameter α over the grid $\alpha = 0, 0.1, \ldots, 1.0$ for the cases listed in Table 2. In all, the optimal values tend to be in the upper end (close to ℓ_1 penalty) and the value used in our original submission ($\alpha = 0.8$) coincides with the median of Table 2 and can be used as a good compromise. Different CV rounds give slightly different error surfaces, and improper selection of α is often masked by random variation of the CV folds. Thus, there is a lot of variation on the optimal value of α suggested by the CV, and the full optimization may not be worth the extra computation. A separate selection of α parameter can also be interpreted as a separate selection of model



(a) (b)

Fig. 2: (a) The cross-validated performance for training data with different values for the parameters λ and α. (b) The true performance for the secret test data with different λ and α. The blue cross denotes the parameters of our ICANN submission.

family (i.e., ℓ_1 penalized LR, ℓ_2 penalized LR, etc...), which is typically done less frequently than the model parameter optimization.

3 Results

3.1 Submission to the ICANN MEG Challenge

In our submission to the ICANN MEG challenge, we achieved a classification performance of 68.0 % on the secret second day test data. This was the winning result with a clear margin to other submissions. The accuracies of all participants are listed in Table 3. More detailed analysis can be found from the challenge proceedings [25].

Before the submission we experimented with various combinations among the feature sets $\{x^{(1)}, x^{(2)}, \ldots, x^{(11)}\}$ and soon found out that increasing the number of feature sets beyond two only degraded the CV performance estimate. This led us to conclude that despite the feature selector embedded in the elastic net regularized model, too many or too complicated features were causing over-learning. Moreover, the limitations of the feature selection start to become more evident as the search space grows exponentially. Although the experiments in our recent paper [20] indicate that the elastic net framework is superior to simpler feature selection strategies, it is still sub-optimal, and manual expert design of feature sets should always complement the automatic selection.

Eventually, we ended up using only the feature sets $\{x_i^{(1)}, x_i^{(4)}\}_1^{204}$, i.e., mean and detrended standard deviation of the signal. This results in $2 \times 204 = 408$ features per each one second sample. Thus, the number of model parameters for the five-class case is $408 \times 5 = 2040$ (plus 5 bias terms). This is the starting point for the elastic net to do further selection to compensate the limited number of training samples.

Table 3: Results of the ICANN MEG Mind Reading Challenge.

Team	*Accuracy*
Huttunen *et al.*	68.0 %
Santana *et al.*	63.2 %
Jylänki *et al.*	62.8 %
Tu *et al.* (1)	62.2 %
Lievonen *et al.*	56.5 %
Tu *et al.* (2)	54.2 %
Olivetti *et al.*	53.9 %
van Gerven *et al.*	47.2 %
Grozea *et al.*	44.3 %
Nicolaou *et al.*	24.2 %

Figure 3 illustrates the results of 200 CV trials with random splitting of the second day data. Our estimate of the prediction error can be seen in the error distribution for the test data shown in Figure 3c. In this case the error rate is 0.397, or 60.3 % correct classification. The error is slightly higher than the error for the secret test data, which is due to using only half of the public second day data for training. The final submission used all public class labeled data for training and achieved an error rate of 0.32. This biasedness, however, was not a problem for our benchmarking, where the goal was only to find the best possible prediction model.

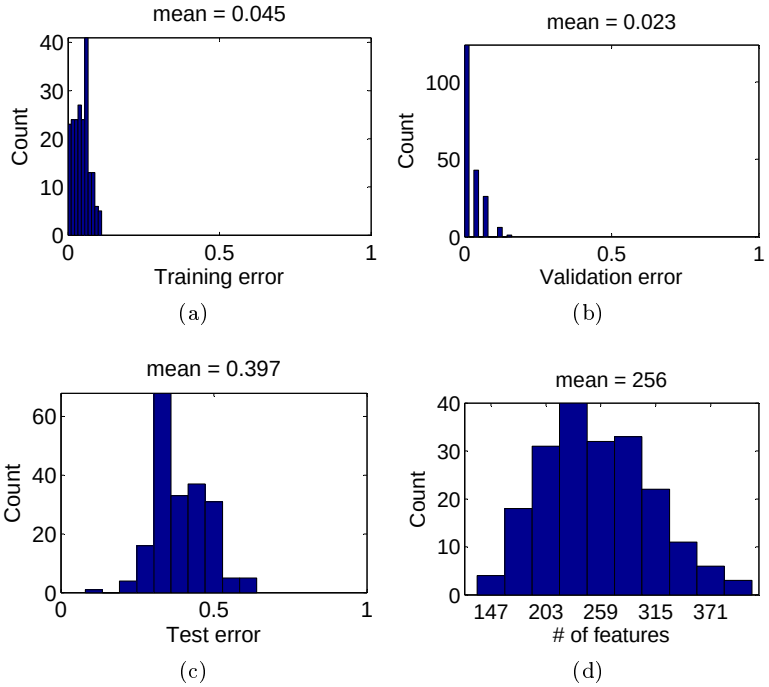


Fig. 3: The validation results of the final prediction model. Figures (a), (b), and (c) illustrate the error distributions with 200 trials for training data (complete first day data), validation data (25 second day samples used in training), and test data (25 second day samples left for testing), respectively. Figure (d) shows the distribution of the number of non-zero model coefficients determined by elastic net regularization.

Figure 3d shows the histogram of the number of features chosen by the elastic net after cross-validating the regularization parameter λ. On the average, only 256 of the 2040 model coefficients are non-zero, which indicates that there is a high number of redundant or noninformative features. The elastic net removes these, because with parameter $\alpha = 0.8$ the ℓ_1 penalty dominates over ℓ_2.

Since the model coefficients are directly related to spatial locations of the sensors, the locations of the nonzero coefficients are of interest. The MEG measurements were recorded in 102 nodes, each of which generated four input features for our model (2 gradiometer channels per node and 2 features per channel). Since the features were normalized in the prediction model, their coefficient values directly indicate the relative significance of particular MEG channels from the viewpoint of the classification.

Figure 4 illustrates the relative importance of different areas for prediction. In the figure, the locations of the 102 gradiometer sensors are marked by black dot, and the color indicates the sum of absolute values of all coefficients for all four features in each sensor location. More specifically, each class k in our logistic regression model has its own coefficient vector $\boldsymbol{\beta}_k$, and Figures 4a – 4e show the sum of coefficient magnitudes for each class separately, while Figure 4f shows the average over coefficients for all classes. Note that due to the sparsity of the classifier, the darkest areas in the topoplots correspond to coefficient values exactly equal to zero, and are thus not used by the classifier at all.

Because our approach is data-driven, care must be taken for not to over-interpret the results shown in topographic plots. Moreover, it was recently shown that the visualization and interpretation of a model depends on the regularization although the predictive performance may seem stable over a range of regularization parameter values [41]. In the following, however, we provide some possible interpretations of the plots. The spatial distribution of the regression coefficients with high magnitudes seems to be reasonable because most of the coefficients are located around visual and left temporal areas, which are known to be responsible of processing visual and linguistic information.[7] In addition, several studies have suggested that temporal lobes have a central role in narrative comprehension [30].

[7] Note, that this is relevant although the stimuli were presented without audio: language processing is not limited to the processing of spoken language [33].

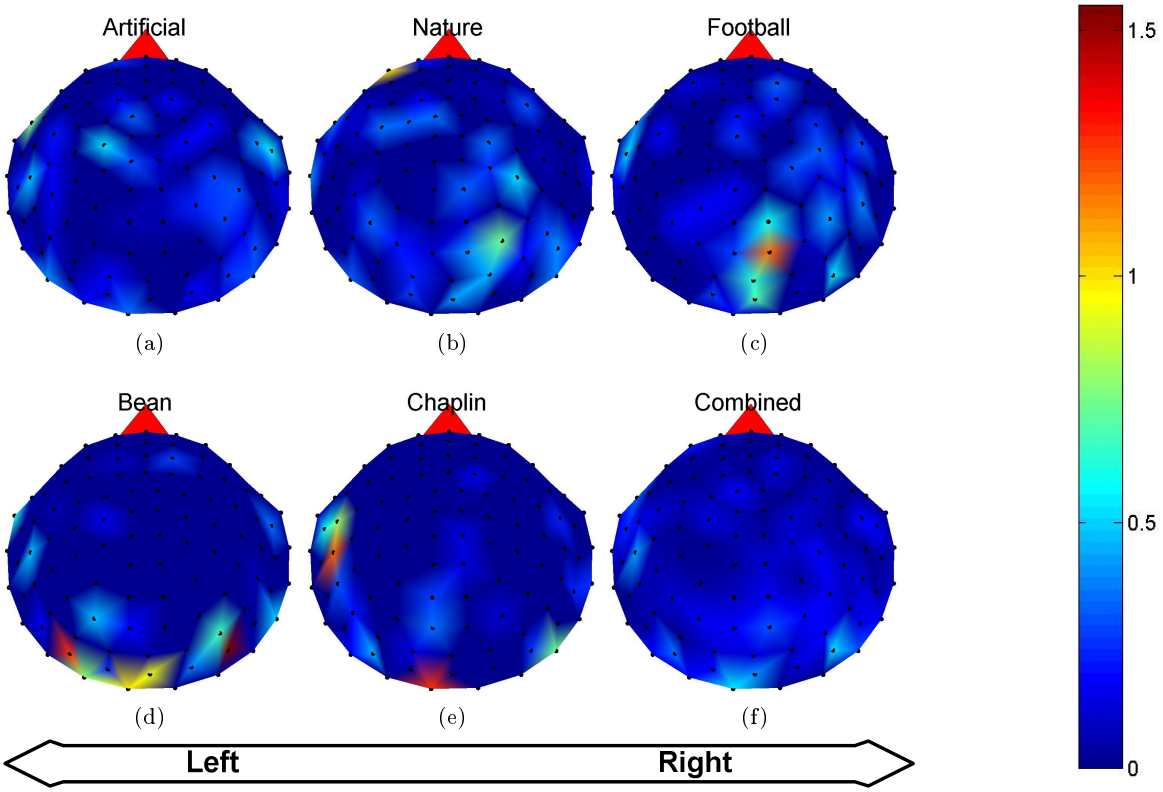


Fig. 4: Illustration of the sum of the regression coefficient magnitudes for each class separately (a — e) and for all classes combined (f). The features have been normalized in order for the coefficient values to be comparable.

High magnitudes of the coefficients on the right side of the occipital lobe might reflect active information processing in the visual motion area (V5) during the stimulus presentation. The Chaplin category has a coefficient with very high magnitude on the left side of the temporal lobe, possibly indicating that this category was (partially) discriminated from the other categories based on the neural activity related to the processing of linguistic information. Interestingly, the football category contains coefficients with high magnitudes in the parietal lobe; the area which is known to be active during the imagery of complex motor skills [34]. The nature category contains a high-magnitude coefficient in the anterior part of the prefrontal cortex. It is possible that the signal measured from the corresponding channel contains an artefact component related to the eye-movements. However, even though classification result would be partially affected by eye-movements, the overall distribution of spatial weights supports the fact that the classification is mostly based on brain-related neuronal activity.

3.2 Using a Larger Set of Features

Due to the limited preparation time, the feature set used in our ICANN MEG challenge submission was not a result of systematic comparison. Instead, the two sets of features $\{x_i^{(1)}, x_i^{(4)}\}_1^{204}$ were selected based on manual experimentation. Regardless of the success of the selection, the question of their optimality still remained open. Thus, we investigated if a more thorough initial feature selection would have further improved the performance. Since the use of more than two feature sets had always degraded the performance, we decided to compare only combinations of one or two feature sets from our list of 11. There are altogether 66 such feature set combinations, which were all tested.

Figure 5 illustrates the classification performance of the 66 feature set combinations estimated using the public data. The figure shows the error estimates and standard errors (of the mean) for all 66 combinations. It can be seen that there are several combinations with very good performance in the upper left corner. The ten best ones are enumerated in the close-up picture at the bottom. The results show that the feature set $\{x_i^{(1)}, x_i^{(4)}\}_1^{204}$ that we used in our final contest submis-

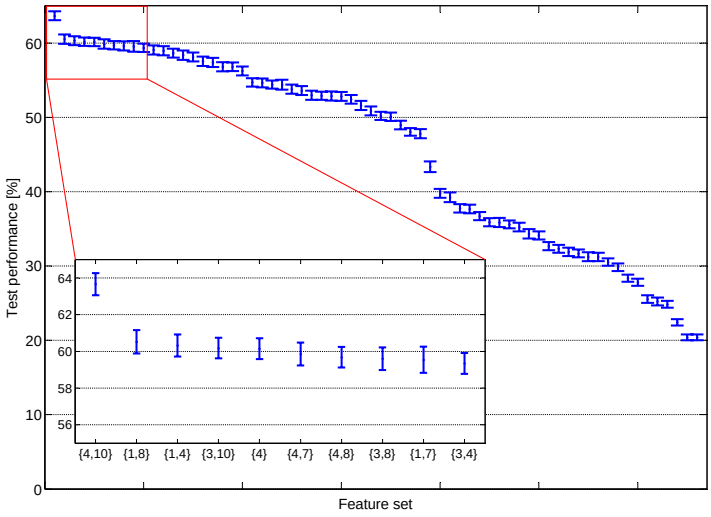


Fig. 5: Mean test performance and the standard error of the mean for different feature sets after 200 test trials. In the small figure, the labels on the x-axis indicate the top ten feature sets.

sion is the third best among all 66 sets. The differences between the second to the tenth best feature sets are not significant, however. Interestingly, clearly the best feature set is $\{x_i^{(4)}, x_i^{(10)}\}_1^{204}$, i.e., the detrended standard deviation together with the kurtosis, which yields a performance estimate of 63.7 %.

The performance of the feature set $\{x_i^{(4)}, x_i^{(10)}\}_1^{204}$ can be analyzed more carefully using Figure 6, which shows also the training and validation errors in Figures 6a and 6b, respectively. Compared to the result of the submitted classifier (Figure 3), the training and validation errors are about twice as big, while the test error shows significant improvement. Thus, had we experimented with this feature set combination, it would have been a strong candidate for our choice for the challenge submission.

However, when testing the performance with the secret test data disclosed after the challenge, it turns out that the classification performance is only 57 % as opposed to the performance of the original submission of 68 %. One explanation to this is that the feature sets $\{x_i^{(4)}, x_i^{(10)}\}_1^{204}$—the detrended standard deviation and the kurtosis—are already too complicated and prone to overlearning. This may be due to inherent Gaussianity of the measurements, which makes the higher order statistics fragile, and one should be careful when adopting them as classifier inputs. In fact, after detrending, 84.8 % of the first day samples and 83.7 % of the second day samples pass the Lilliefors test for Gaussianity with 95 % confidence level [28], thus making Gaussianity a dominant characteristic of the data. Another possible explanation for the poor generalization of the feature set $\{x_i^{(4)}, x_i^{(10)}\}_1^{204}$ is the high variance of the CV error estimator with small sample sizes [8].

Our final note is that there are eight feature sets among the ten best that include either the standard deviation (8) or the detrended standard deviation (4). Moreover, the detrended standard deviation alone is the fifth best feature set of all. This suggests that the standard deviation of the MEG signal has quite a significant predictive power in the MEG decoding task. This is not surprising, since it is known that a change in stimulus can trigger either a decrease or increase of power in the measured MEG activity [38].

3.3 The Effect of Frequency Bands

When preparing the submission to the MEG challenge, we were unable to gain any benefit from using the frequency bands provided by the organizers. However, ongoing spontaneous brain activity is characterized by the presence of more or less regular oscillations in various frequency bands [45] and therefore it might be expected that using the frequency band information might be useful for classification. Moreover, there are several different feature extraction methods in the literature, which often are stimulus type specific, but still many of them always include a frequency band decomposition [2,43,13]. Therefore, we tested in a systematic manner if this information would be useful in our setting. Figure 7 shows the validation results of a classifier using the variances of the band pass filtered signals as

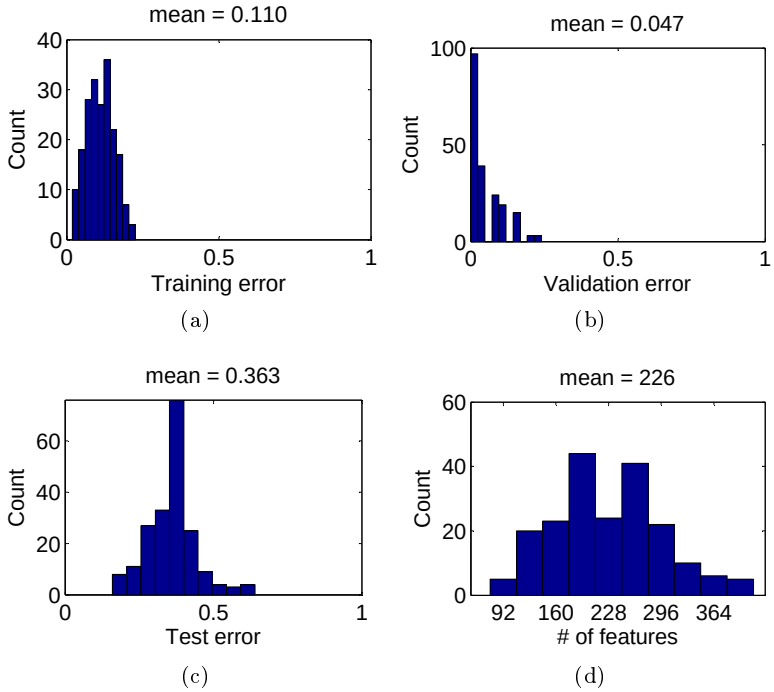


Fig. 6: The validation results using detrended standard deviation and kurtosis as features. Error distributions with 200 trials are shown for training data (complete first day data), validation data (25 second day samples used in training), and test data (25 second day samples left for testing) in Figures (a), (b), and (c), respectively. Figure (d) shows the distribution of the number of non-zero model coefficients determined by elastic net regularization.

features. The results are compared against the submitted result and the CV was repeated 200 times.

Figure 7 shows the training, validation, and test performances for three feature sets: The one used in our ICANN submission with (total 408 features), a feature set with variances of all five frequency bands from each channel (total $204 \times 5 = 1020$ features), and a feature set with variances of all five frequency bands from each sensor (total $102 \times 5 = 510$ features). The latter feature set was generated by averaging the band energies from the two gradiometer channels of each sensor, and was included in the test as it has a comparable cardinality to that of our ICANN submission.

As seen in Figure 7, the full frequency feature set of 1020 features ("band en.") has a worse performance compared to the two other methods. This is most likely due to the limitations of the feature selection procedure as the search space grows exponentially. When the feature space is subsampled by averaging two gradiometer channels, the performance becomes comparable to that of our ICANN submission. In addition to these CV tests, we tested the feature sets against the secret second day test data, with classification performance of 59 % for the 1020 frequency features, 62 % for the subsampled frequency features and 68 % for the baseband features. This suggests that the frequency bands do contain the information required to separate the classes efficiently, but are more prone to overlearning than the baseband features used in the original ICANN submission. This is in line with the results of Section 3.2, which show that the initial feature space can be constructed in various ways with minor effect on the CV performance. However, it seems that the more complicated feature sets have weaker performance on the secret test data, probably due to overlearning.

3.4 Movies vs. Short Clips

In the ICANN challenge proceedings [25], the organizers assessed the performance of the submitted multinomial classifiers also in the binary problem of discriminating between movies with a storyline (classes 4–Bean and 5–Chaplin) and short video clips with no storyline (classes 1–Artificial, 2–Nature, and 3–Football). This was done by combining classifier outputs 1, 2, and 3 as one class and 4 and 5 as another class.

Although our method clearly outperformed the other methods in the multinomial problem, the performance was inferior to the median performer in the binary problem. A possible explanation for this proposed by Klami

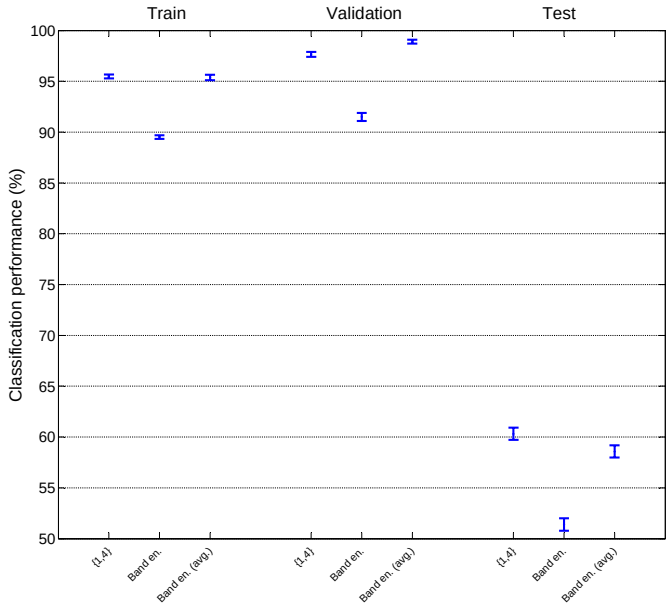


Fig. 7: The results for the prediction model that uses the variance computed separately from five distinct frequency sub-bands using all 408 channels ("band en.") and using the mean of the two gradiometer channels of each sensor ("band en. [avg]") as features. Average prediction performances along with standard error bars after 200 trials are shown for training data (complete first day data), validation data (25 second day samples used in training), and test data (25 second day samples left for testing). For comparison, labels "$\{1, 4\}$" indicate the corresponding performances by using mean and detrended standard deviation of the unfiltered signal.

et al. [25] is that the other methods have overlearned the easier binary problem, which has ultimately made the multinomial performance poorer. Our interpretation of this result is instead that our 5-class classifier is not regularized enough to achieve the best possible generalization performance for the 2-class case. In a more detail, a 5-class linear classifier applied to a 2-class problem in the above manner is not a linear classifier, but a classifier with a piece-wise linear decision surface.

To study the point further, we tested how well our classifier can make a distinction between movies and short clips if we train a binary classifier particularly for this 2-class task. Instead of the symmetric logistic regression model in Equation 2, we use a traditional logistic regression model:

$$p(\mathbf{x}) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}, \tag{4}$$

where $p(\mathbf{x})$ estimates the probability of class 1 given $\mathbf{x}$ while $1 - p(\mathbf{x})$ is the probability of class 2 given $\mathbf{x}$. Parameters β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ are estimated similarly to the multinomial case by maximizing the elastic net penalized log-likelihood

$$\sum_{i \in C_1} \log p(\mathbf{x}_i) + \sum_{i \in C_2} \log(1 - p(\mathbf{x}_i)) - \lambda \left(\alpha ||\boldsymbol{\beta}||_1 + (1 - \alpha) ||\boldsymbol{\beta}||_2^2 \right), \tag{5}$$

where $C_1 = \{i \mid k_i = 1\}$ and $C_2 = \{i \mid k_i = 2\}$ denote the index sets of samples from the first and second class, respectively.

As a result, the classification performance of the specifically designed binary classifier is 96.5 % on the secret test data while our earlier performance with the multinomial classifier was only 89.7 %. This indicates that here a simple (linear) classifier is preferred over a more complex classifier (piecewise linear). While the result is greatly improved, it is still slightly outperformed by two methods by Tu and Sun, which were the top results in the binary classification test in the ICANN challenge report having performances of 97.1 % and 96.6 %. For simplicity, we used the same initial feature set of mean and detrended standard deviation and the elastic net mixing parameter $\alpha = 0.8$ as in the multinomial case. Hence, slight improvement could be expected by optimizing these for the binary problem.

In Figure 8, we illustrate the absolute values of the model coefficients after feature standardization. As in Figure 4, the black nodes mark the locations of the sensors, which are a source of four model coefficients each. Due to the traditional logistic regression model (and not the symmetric one) we only have one set of coefficients $\{\beta_0, \boldsymbol{\beta}\}$ instead of one set per class. Using the symmetric model in the 2-class case would result in

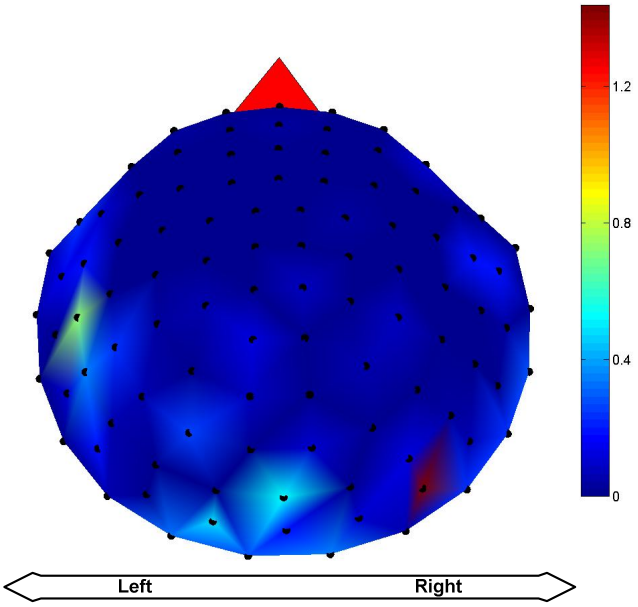


Fig. 8: The regression coefficient magnitudes of a binary classifier trained to distinct movies with a storyline from short clips.

two identical plots because the coefficients between the classes only differ by their signs. The single discriminant function is now of form $g(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}$. The figure shows the areas that the classifier uses to discriminate between the two classes. Interestingly, comparison of Figure 8 to Figure 4f indicates that almost the same sensors were important for the binary classification task as for the 5-class classification.

The coefficients with highest magnitudes are located around the right side of the occipital lobe and the left side of the temporal lobe. The result is reasonable from the neuroscientific point of view as it suggests that the clips with the storyline were discriminated from the clips with no storyline based on the brain activation related to processing of visual motion and linguistic information. However, we acknowledge that these explanations are speculative and cannot be fully verified based on this study. Especially, different experimental setups for short clips and movies[8] complicate the neuroscientific interpretation of these results. However, we re-iterate that this complication does not influence our interpretation of the result from the classification point of view.

[8] While short term clips from movie categories 1, 2 and 3 (see Section 2.1) were shown by the organizers in an intermingled fashion, the "storyline" movies (categories 4 and 5), have been presented in one continuous block, each at the end of the experiment [25]. Therefore, the acquired signals in categories 1,2, and 3 might be different to the signals in categories 4 and 5 purely for 'chronological' reasons, e.g., decreasing vigilance.

Table 4: Results on the BCI competition IV MEG dataset. The table shows the amount of correct classifications of the test set for two subjects (S1 and S2) using the four submissions to the BCI competition (ID-1,...,ID-4) and the proposed ℓ_1-regularized logistic regression method (LR (orig) and LR (mod)); see text.

Method	*S1*	*S2*	*Overall*	*p-value*
ID-1	**59.5 %**	34.3 %	**46.9 %**	2.65e-9
ID-2	31.1 %	19.2 %	25.1 %	0.44
ID-3	16.2 %	31.5 %	23.9 %	0.59
ID-4	23.0 %	17.8 %	20.4 %	0.88
LR (orig)	35.1 %	30.1 %	32.6 %	0.014
LR (mod)	41.9 %	**38.4 %**	40.1 %	1.87e-5

Table 5: The confusion matrix of classifying the BCI MEG data. The class labels are hand movement directions: L = left, R = right, F = forward, B = backward.

		True class				
		R	**L**	**F**	**B**	**Σ**
Predicted	**R**	7.5 %	6.1 %	2.7 %	4.8 %	21.1 %
	L	4.1 %	13.6 %	4.8 %	2.7 %	25.2 %
	F	4.8 %	3.4 %	8.2 %	8.2 %	24.5 %
	B	6.1 %	9.5 %	2.7 %	10.9 %	29.3 %
	Σ	22.4 %	32.7 %	18.4 %	26.5 %	100 %

3.5 Experiments with Other Data

The significance of our framework depends naturally on its generalization to other MEG datasets. Since MEG data can be captured in various settings, a classification method should be able to learn what is essential for a particular scenario. Since the strength of our logistic regression based framework is in the ability to select the most significant features, we believe in efficient generalization ability.

There exists numerous publicly available benchmark EEG datasets, but only a few benchmarks with MEG data. Among the few, probably the best known is the Brain-Computer Interface (BCI) competition IV (task 3) [51,46]. The BCI dataset contains MEG signals recorded while subjects performed hand movements in four directions. The motivation for the study is to facilitate the use of decoded MEG activity in the rehabilitation of, e.g., spinal injury or stroke patients, who have lost their natural hand movement.

The amount of data in the BCI dataset is significantly smaller than in the ICANN MEG dataset: there are only ten sensors (channels). Training data contains 160 samples (40 per class) and the secret test set contains 73 or 74 samples for two test subjects, respec-

tively. Since the ground truth for the secret test set is now available, we are not in the same position as the competitors. Thus, our goal is not to exceed their result, but to illustrate that our framework is able to reach comparable accuracy without extensive tuning.

The results of the four submissions to the BCI competition IV and our method (LR (orig) and LR (mod)) are shown in Table 4. The best submission—ID-1—used a set of statistical, frequency and wavelet features together with a genetic algorithm for feature selection and a fusion of SVM and LDA classifiers [46]. In fact, it is the only one exceeding chance level, which can be seen as follows.

The number of samples, k, correctly classified by a random classifier is given by the binomial distribution $\mathrm{Bin}(k; n, q)$, where n denotes the number of test samples and q is the probability of correct classification for a random classifier; i.e., our null hypothesis H_0 is that $q = 0.25$ for a 4-class case. Now, the probability of a random classifier correctly predicting the label of at least k samples is given by

$$P(\text{at least } k \text{ correct}|H_0) = 1 - \sum_{j=0}^{k-1} \mathrm{Bin}(k; n, q).$$

The above formula gives the p-values of observing the classification performances by chance under H_0, and these are listed in Table 4. Using 5 % significance level, one can say that only ID-1 and our method exceeds the chance level. Note that this analytical test is valid and to be preferred over permutation tests if the test examples can be assumed to be independent trials [36].

The second row from the bottom shows the result of using our ICANN submission algorithm without modifications. It can be seen that the accuracy is clearly above chance level, but slightly inferior to the winning submission ID-1 (except for subject S2). However, the used feature is simple and designed for data with significantly higher number of channels (204 instead of 10). The low dimensionality of our simple features calculated from the BCI dataset probably does not contain enough information for successful classification and may not enable linear separability of the classes. Thus, additional features are required.

We did another experiment, where the number of features was increased 10-fold by splitting each MEG measurement into 10 nonoverlapping blocks and calculating the mean and standard deviation from each. In total, this produces 220 features (20 from each block and 20 from the entire signal). This set of features enables rudimentary assessment of frequency content, which is a key component in all original submissions. The result of using these aggregate features is shown on the bottom row of Table 4. It is clearly seen that the additional features increase the performance significantly, although not to the level of the best submission.

The review of the BCI competition [46] reports a 62 % overall accuracy on the test dataset using Regularized Linear Discriminant Analysis [51]. Although the high accuracy is probably partly due to the availability of larger number of test subjects during development, we believe that inclusion of frequency domain and other derived features would render our logistic regression framework comparable to [46].

However, we decide to skip this additional feature engineering step since that would only adjust the model to the particular properties of this relatively small set of secret test data that is now available. Moreover, the dataset is not very suitable for a discriminative classifier, because the training data has an even class distribution but the test data does not. The effect can be seen from Table 5, which shows the confusion matrix of our classifier. The class L is overrepresented (32.7 %) and the class F is underrepresented (18.4 %) in the test data, although the predictor learns an equal proportion of all classes from the training data.

4 Discussion

We have proposed a method for multinomial classification of magnetoencephalography (MEG) data. The method is based on multinomial logistic regression with elastic net penalty [9,54], which combines feature selection and classification into a single penalized maximum likelihood optimization problem. The method achieves a classification performance of 68.0 % on the ICANN MEG test data set [25], which was the winning result in the ICANN challenge with a clear margin to the second best submission with 63.2 % accuracy. The method was also tested with another set of data, and despite the simplicity of the features, the results were comparable to the submissions of the BCI competition IV [51,46]. The BCI dataset revealed that the proposed method is not optimal with low-dimensional data, probably because the feature selection is useless. However, it was shown that the performance increases with additional features describing the frequency content.

One of the key components of our approach was to apply model regularization or feature selection in order to cope with the large number of possibly correlated measurements. The efficiency of the feature selection allowed us to experiment with a large pool of input features without careful manual selection process. A traditional approach to regularize classification models is explicit feature selection, where only a subset of all available features is used in the classification model.

We tested various iterative feature selection methods including the *Floating Stepwise Selection* [40] and *Simulated Annealing Feature Selection* [7], but their results were inferior to those of the regularized logistic regression model (for details, see [20,23]).

We have presented experimental results concerning feature extraction / selection for MEG decoding analysis. Our results indicate that simple features based on full-band data performs best for this application. We found slightly surprisingly that using the frequency band information was not as beneficial as one could expect. Namely, this is somewhat in contrast with works, which aim at the classification of MEG signals in a more tightly controlled experimental paradigm. For example, Rieger et al. [43] found that the theta band around 5 Hz was clearly most important for predicting the recognition success of natural scene photographs and Besserve et al. [2] found that the beta band from 15 to 30 Hz was the most important for classification between a visuomotor task and a resting condition. Perhaps the reason for this result is that in a naturalistic setups—as in here—no single frequency band alone can provide the most discriminative features combined with the increased difficulty of the feature selection with the increased number of features. We have also pointed out that based on a CV on a limited number of samples, a single best feature set can be challenging to identify, and there are several well-performing feature sets. However, as recognized by many, a simple solution should usually be preferred over a complicated one [12,18].

The proposed classifier has linear decision boundaries. In Section 3.5 we discovered that the classes of the BCI-IV data, in fact, seem to be linearly separable, since nonlinear classifiers do not improve the accuracy. Moreover, the classes of the ICANN MEG dataset seem to be linearly separable, as well: Our recent paper [20] compares the proposed method with the SVM using linear and RBF kernels, and finds no improvement from the nonlinearity. Note, however, that in some cases a simple solution may not have enough descriptive power to represent nonlinear structures in the data: For example, Rasmussen *et al.* [42] describe an fMRI classification task, where a nonlinear classifier clearly outperforms a linear one. When applicable, the benefits of the linear model are in its simplicity, which makes it robust against overlearning, in particular when the sample size is small.

An interesting topic for future work is possible filtering operations of the data. The current algorithm does not apply any preprocessing steps besides the usual calibration and sample rate related operations. The performance could be improved with denoising either in the time domain (within-channel smoothing) or in the spatial domain (between-channel smoothing).

Finally, it is interesting to contrast our approach to the other top performers in the ICANN competition. Santana et al. [44] applied an ensemble of different classifiers (including elastic net penalized logistic regression) preceded by a complex feature extraction step and reached a classification accuracy of 63.2 %. Jylänki et al. [21] used a feature selection approach based on binary classifiers succeeded by a multinomial Gaussian process classifier and reached a classification accuracy of 62.8 %. There are two key differences between our approach and these competing approaches. First, both of these solutions used frequency band decomposition to construct features (in Jylänki's approach the features were almost the same as those experimented with in Section 3.3, Santana's feature extraction approach was more complicated). Second, both Santana et al. [44] and Jylänki et al. [21] used a filter approach[9] to pre-select the features used for the classification. Related to these key differences, we speculate that our good results in the competition support the hypothesis that the wrapper/embedded feature selection methods are preferable over more simple filter methods and even for the embedded feature selection methods the pool of original features should not be too large, i.e., simplicity is to be preferred.

Acknowledgements The research was funded by the Academy of Finland grant no 130275. We also want to thank professor R. Hari (Brain Research Unit, Low Temperature Laboratory, Aalto University School of Science, Finland) for her valuable remarks concerning our study.

References

1. Anderson, J., Blair, V.: Penalized maximum likelihood estimation in logistic regression and discrimination. Biometrika **69**, 123–136 (1982)
2. Besserve, M., Jerbi, K., Laurent, F., Baillet, S., Martinerie, J., Garnero, L.: Classification methods for ongoing EEG and MEG signals. Biol Res **40**(4), 415–437 (2007)
3. Blankertz, B., Müller, K.R., Krusienski, D.J., Schalk, G., Wolpaw, J.R., Schlögl, A., Pfurtscheller, G., del R Millán, J., Schröder, M., Birbaumer, N.: The BCI competition III: Validating alternative approaches to actual BCI problems. IEEE Trans Neural Syst Rehabil Eng **14**(2), 153–159 (2006)
4. Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L., Sturm, I., Curio, G., Müller, K.R.: The Berlin Brain-Computer Interface: Non-medical uses of BCI technology. Front Neurosci **4**, 198 (2010)

[9] The term filter (see Guyon and Elisseeff [11]) here refers to the application of a feature selection method that is independent of the classifier.

5. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. Neuroimage **44**(1), 112–122 (2009)
6. Chan, A.M., Halgren, E., Marinkovic, K., Cash, S.S.: Decoding word and category-specific spatiotemporal representations from MEG and EEG. Neuroimage **54**(4), 3028–3039 (2011)
7. Debuse, J.C., Rayward-Smith, V.J.: Feature subset selection within a simulated annealing data mining algorithm. Journal of Intelligent Information Systems **9**, 57–81 (1997)
8. Dougherty, E.R., Sima, C., Hua, J., Hanczar, B., Braga-Neto, U.M.: Performance of error estimators for classification. Current Bioinformatics **5**(1), 53–67 (2010)
9. Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1–22 (2010)
10. Grosenick, L., Greer, S., Knutson, B.: Interpretable classifiers for FMRI improve prediction of purchases. IEEE Trans Neural Syst Rehabil Eng **16**(6), 539–548 (2008)
11. Guyon, I., Elisseeff, A.: An introduction to variable and feature seletion. Journal of Machine Learning Research **3**, 1157 – 1182 (2003)
12. Hand, D.J.: Classifier technology and the illusion of progress. Statistical Science **21**(1), 1–14 (2006). URL `http://www.jstor.org/stable/27645729`
13. Hanke, M., Halchenko, Y.O., Sederberg, P.B., Olivetti, E., Fründ, I., Rieger, J.W., Herrmann, C.S., Haxby, J.V., Hanson, S.J., Pollmann, S.: PyMVPA: A unifying approach to the analysis of neuroscientific data. Front Neuroinform **3**, 3 (2009)
14. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction, Second edn. Springer Series in Statistics. Springer (2009)
15. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science **293**(5539), 2425–2430 (2001)
16. Haynes, J.D.: Multivariate decoding and brain reading: Introduction to the special issue. NeuroImage **56**(2), 385 – 386 (2011)
17. Haynes, J.D., Rees, G.: Predicting the orientation of invisible stimuli from activity inhuman primary visual cortex. Nat Neurosci **8**(5), 686–691 (2005)
18. Holte, R.C.: Elaboration on two points raised in "classifier technology and the illusion of progress". Statistical Science **21**(1), 24–26 (2006). URL `http://www.jstor.org/stable/27645732`
19. Huttunen, H., Kauppi, J.P., Tohka, J.: Regularized logistic regression for mind reading with parallel validation. In: Proceedings of ICANN/PASCAL2 Challenge: MEG Mind-Reading, pp. 20–24 (2011). URL `http://www.cis.hut.fi/icann2011/meg/megicann_proceedings.pdf`
20. Huttunen, H., Manninen, T., Tohka, J.: MEG mind reading: Strategies for feature selection. In: Proceedings of the Federated Computer Science Event 2012, pp. 42–49 (2012). URL `http://www.cs.helsinki.fi/u/starkoma/ytp/YTP-Proceedings-2012.pdf`
21. Jylänki, P., Riihimäki, J., Vehtari, A.: Multi-class Gaussian process classification of single trial MEG based on frequency specific latent features extracted with binary linear classifiers. In: Proceedings of ICANN/PASCAL2 Challenge: MEG Mind-Reading, pp. 31–34 (2011). URL `http://www.cis.hut.fi/icann2011/meg/megicann_proceedings.pdf`
22. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. Nat Neurosci **8**(5), 679–685 (2005)
23. Kauppi, J.P., Huttunen, H., Korkala, H., Jääskeläinen, I.P., Sams, M., Tohka, J.: Face prediction from fMRI data during movie stimulus: Strategies for feature selection. In: Proceedings of ICANN 2011, *Lecture Notes in Computer Science*, vol. 6792, pp. 189–196. Springer (2011)
24. Kippenhan, J.S., Barker, W.W., Pascal, S., Nagel, J., Duara, R.: Evaluation of a neural-network classifier for pet scans of normal and alzheimer's disease subjects. J Nucl Med **33**(8), 1459–1467 (1992)
25. Klami, A., Ramkumar, P., Virtanen, S., Parkkonen, L., Hari, R., Kaski, S.: ICANN/PASCAL2 Challenge: MEG Mind-Reading — Overview and Results (2011). URL `http://www.cis.hut.fi/icann2011/meg/megicann_proceedings.pdf`
26. Kleinbaum, D., Klein, M.: Logistic Regression. Statistics for Biology and Health. Springer (2010)
27. Lautrup, B., Hansen, L., Law, I., Mørch, N., Svarer, C., Strother, S.: Massive weight sharing: a cure for extremely ill-posed problems. Supercomputing in Brain Research: From Tomography to Neural Networks. pp. 137–148 (1994)
28. Lilliefors, H.W.: On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association **62**(318), 399–402 (1967)
29. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain–computer interfaces. Journal of Neural Engineering **4**(2), R1 (2007)
30. Mar, R.: The neuropsychology of narrative: story comprehension, story production and their interrelation. Neuropsychologia **42**(10), 1414–1434 (2004)
31. Mørch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O.B.: Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover (1997)
32. Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L.: Encoding and decoding in fMRI. NeuroImage **56**(2), 400 – 410 (2011)
33. Nickels, L.: The hypothesis testing approach to the assesment of language. In: B. Stremmer, H. Whitaker (eds.) The handbook of neuroscience of language. Academic press (2008)
34. Olsson, C.J., Jonsson, B., Larsson, A., Nyberg, L.: Motor representations and practice affect brain systems underlying imagery: an fMRI study of internal imagery in novices and active high jumpers. Open Neuroimaging Journal **2**, 5–13 (2008)
35. O'Toole, A.J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J.P., Parent, M.A.: Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. Journal of Cognitive Neuroscience **19**(11), 1735–1752 (2007)
36. Pereira, F., Botvinick, M.: Information mapping with pattern classifiers: a comparative study. Neuroimage **56**(2), 476–496 (2011). DOI 10.1016/j.neuroimage.2010.05.026. URL `http://dx.doi.org/10.1016/j.neuroimage.2010.05.026`
37. Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: a tutorial overview. NeuroImage **45**(Suppl 1), S199 – S209 (2009)
38. Pfurtscheller, G., Lopes da Silva, F.H.: Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol **110**(11), 1842–1857 (1999)

39. Poldrack, R.A., Halchenko, Y.O., Hanson, S.J.: Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol Sci **20**(11), 1364–1372 (2009)
40. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recogn. Lett. **15**(11), 1119–1125 (1994)
41. Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C.: Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition **45**(6), 2085 – 2100 (2012)
42. Rasmussen, P.M., Madsen, K.H., Lund, T.E., Hansen, L.K.: Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage **55**(3), 1120–1131 (2011)
43. Rieger, J.W., Reichert, C., Gegenfurtner, K.R., Noesselt, T., Braun, C., Heinze, H.J., Kruse, R., Hinrichs, H.: Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. Neuroimage **42**(3), 1056–1068 (2008)
44. Santana, R., Bielza, C., Larranaga, P.: An ensemble of classifiers approach with multiple sources of information. In: Proceedings of ICANN/PASCAL2 Challenge: MEG Mind-Reading, pp. 25–30 (2011). URL `http://www.cis.hut.fi/icann2011/meg/megicann_proceedings.pdf`
45. Stam, C.: Use of magnetoencephalography (MEG) to study functional brain networks in neurodegenerative disorders. Journal of the Neurological Sciences **289**(1-2), 128 – 134 (2010)
46. Tangermann, M., Müller, K.R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K.J., Mueller-Putz, G., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S., Blankertz, B.: Review of the BCI competition IV. Frontiers in Neuroscience **6**(00055) (2012)
47. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1994)
48. Tomioka, R., Müller, K.R.: A regularized discriminative framework for EEG analysis with application to brain-computer interface. NeuroImage **49**(1), 415–432 (2010)
49. van De Ville, D., Lee, S.W.: Brain decoding: Opportunities and challenges for pattern recognition. Pattern Recognition, Special Issue on Brain Decoding **45**(6), 2033 – 2034 (2012)
50. van Gerven, M., Hesse, C., Jensen, O., Heskes, T.: Interpreting single trial data using groupwise regularisation. Neuroimage **46**, 665 – 676 (2009)
51. Waldert, S., Preissl, H., Demandt, E., Braun, C., Birbaumer, N., Aertsen, A., Mehring, C.: Hand movement direction decoded from MEG and EEG. Journal of neuroscience **28**(4), 1000–1008 (2008)
52. Webb, A.: Statistical Pattern Recognition, Second edn. John Wiley and Sons Ltd. (2002)
53. Zhdanov, A., Hendler, T., Ungerleider, L., Intrator, N.: Inferring functional brain states using temporal evolution of regularized classifiers. Comput Intell Neurosci p. 52609 (2007)
54. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2), 301–320 (2005)