ℓ_1-penalized Linear Mixed-Effects Models for high dimensional data with application to BCI

Siamac Fazli[a,b,*], Márton Danóczy[a], Jürg Schelldorfer[c],
Klaus-Robert Müller[a,b,d]

[a]*Berlin Institute of Technology, Franklinstr. 28/29, 10587 Berlin, Germany*
[b]*Bernstein Focus: Neurotechnology Berlin (BFNT-B), 10587 Berlin, Germany*
[c]*ETH Zürich,Rämistrasse 101, 8092 Zürich, Switzerland*
[d]*Institute for Pure and Applied Mathematics, UCLA, Los Angeles, CA 90095-7121*

Abstract

Recently, a novel statistical model has been proposed to estimate population effects and individual variability between subgroups simultaneously, by extending Lasso methods. We will for the first time apply this so-called ℓ_1-penalized linear regression mixed-effects model for a large scale real world problem: we study a large set of brain computer interface data and through the novel estimator are able to obtain a subject-independent classifier that compares favorably with prior zero-training algorithms. This unifying model inherently compensates shifts in the input space attributed to the individuality of a subject. In particular we are now for the first time able to differentiate *within-subject* and *between-subject variability*. Thus a deeper understanding both of the underlying statistical and physiological structure of the data is gained.

Keywords: mixed-effects model, sparsity, BCI, subject-independent

[*]corresponding author
email: fazli@cs.tu-berlin.de
tel: +49 30 314 28680
fax: +49 30 314 78622

1. Introduction

When measuring experimental data we typically encounter a certain inbuilt heterogeneity: data may stem from distinct sources that are all additionally exposed to varying measuring conditions. Such so-called group, respectively individual effects need to be modeled separately within a global statistical model. Note that here the data are not independent: a part of the variance may come from the individual experiment, while another may be attributed to a *fixed* effect. Such mixed-effects models (Pinheiro and Bates, 2000) are known to be useful whenever there is a grouping structure among the observations, e.g. the clusters are independent but within a cluster the data may have a dependency structure. Note also that mixed-effects models are notoriously hard to estimate in high dimensions, particularly, if only few data points are available.

In this paper we will for the first time use a recent ℓ_1-penalized estimation procedure (Schelldorfer et al., 2010) for high-dimensional linear mixed-effects models in order to estimate the mixed effects that are persistent in experimental data from neuroscience. This novel method builds upon Lasso-type procedures (Tibshirani, 1996; Meier et al., 2008; Yuan and Lin, 2006), assuming that the number of potential fixed effects is large and that the underlying true fixed-effects vector is sparse. The ℓ_1-penalization on the fixed effects is used to achieve sparsity. The idea of ℓ_1-penalized likelihood approaches in linear mixed-effects models is not novel. The work of (Bondell et al., 2010) and (Ibrahim et al., 2010) present ℓ_1-penalized methods for linear mixed effects models. While the latter (Bondell et al., 2010; Ibrahim et al., 2010) only studied the low-dimensional setting, only (Schelldorfer et al., 2010) have succeeded in investigating the high-dimensional case (i.e. $n \ll p$).

We will study Brain Computer Interfacing (Dornhege et al., 2007), where we encounter high variability both between subjects and within repetitions of an experiment for the same subject. The novel approach splits up the overall inherent variance into a within-group and a between-group variance and therefore allows us to model the unknown dependencies in a meaningful manner. While this is a conceptual contribution to adapt the mixed effects model for BCI, our paper also contributes practically. Due to the more precise modeling of the dependency structure we cannot only quantify both sources of variance but also provide an improved ensemble model that is able to serve as a one-size-fits-all BCI classifier – the central ingredient of a so-called zero-training BCI (Krauledat et al., 2008; Fazli et al., 2009b; Alamgir et al., 2010). In other words we can minimize the usually required calibration time for a novel subject – where the learning machine adapts to the new brain (e.g. (Blankertz et al., 2002, 2007)) – to practically zero.

The following section will introduce the novel statistical model, section 3 introduces the BCI setup and data basis and in section 4 we discuss the experimental results. Section 5 concludes the work and in the appendix we show the algorithm as well as a flowchart of when mixed-effects models should be considered.

2. Statistical Model

Since its early precursors in the 70's (Vidal, 1973) the main goals of the BCI community has been to reduce setup cost on one hand, as well as increasing Information Transfer Rates (ITR) on the other. Setup cost being the actual setup of EEG-related hardware, as well as acquiring training data to estimate efficient subject-dependent classifiers. Modern machine learning techniques have enabled the BCI-user to operate a high-speed BCI system with no training on the user side and short calibration sessions for training data generation (Cheng et al., 2002; Wang et al., 2006; Blankertz et al., 2008; Parra et al., 2008; Thomas et al., 2009), as compared to classical operant conditioning, which still required users to adapt to the system at hand. Very recently a novel approach has been suggested to completely overcome the need of calibration sessions (Fazli et al., 2009b,a) with the help of ℓ_1-regularized regression.

In this work we will work with a so called linear mixed-effects model (Pinheiro and Bates, 2000), due to the dependence structure inherent to the two sources of variability: within-subject (dependence) and between-subject (independence). The classical linear mixed-effects framework has two limiting issues: (1) it cannot deal with high-dimensional data (i.e. the total number of observations is smaller than the number of explanatory variables) and (2) fixed-effects variable selection gets computationally intractable if the number of fixed-effects covariates is very large. By using a Lasso-type concept (Tibshirani, 1996) these limits can be overcome in the present method (Schelldorfer et al., 2010), thus allowing application in the real world as we will see in the next sections.

2.1. Model Setup

Let $i = 1, \ldots, N$ be the number of subjects, $j = 1, \ldots, n_i$ the number of observations per subject and $N_T = \sum n_i$ the total number of observations. For each subject we observe an n_i-dimensional response vector y_i. Moreover, let X_i and Z_i be $n_i \times p$ and $n_i \times q$ covariate matrices, where X_i contains the fixed-effects covariates and Z_i the corresponding random-effects covariates. Denote by $\beta \in \mathbb{R}^p$ the p-dimensional fixed-effects vector and by $b_i, i = 1, \ldots, N$ the q-dimensional random-effects vectors. Then the linear mixed-effects model can be written as ((Pinheiro and Bates, 2000))

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \qquad i = 1, \ldots, N \quad , \tag{1}$$

where we assume that *i*) $b_i \sim \mathcal{N}_q(0, \tau^2 I_q)$, *ii*) $\varepsilon_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$ and *iii*) that the errors ε_i are mutually independent of the random effects b_i.

From (1) we conclude that

$$y_i \sim \mathcal{N}_{n_i}(X_i\beta, \Lambda_i(\sigma^2, \tau^2)) \quad \text{with} \quad \Lambda_i(\sigma^2, \tau^2) = \sigma^2 I_{n_i} + \tau^2 Z_i Z_i^{\mathsf{T}}. \tag{2}$$

It is important to point out that assumption *i*) is very restrictive. Nevertheless, it is straightforward to relax this assumption and assume that $b_i \sim \mathcal{N}_q(0, \Psi)$ for a general (or possible structured) covariance matrix Ψ. For the data described in the next section, assumption *i*) seems to hold.

To give the reader an intuition of the method, we generated a simple toy example that demonstrates why estimating mixed-effects can help in finding a superior solution that takes possible shifts in the input-space of multiple-subject data into account: The data is generated with the model given in Equation (1) and by setting $Z_i = \mathbf{1}_{n_i}$ and $b_i \in \mathbb{R}$ we assume a random-intercept model or one bias per group. The top left panel of Figure 1 shows the five groups of input data we generated, each consisting of 40 trials with the following parameters: $\beta_{\text{ORIG}} = 0.5$, $\mathbf{b}_{\text{ORIG}} = [-2; -1; 0; 1; 2]$ and a noise level of $\epsilon_{\text{ORIG}} \sim \mathcal{N}(0, 0.2)$. While least-square regression (LSR) estimates $\beta_{\text{LSR}} = 0.048$ and $b_{\text{LSR}} = 0.075$, the proposed mixed-effects model is far more accurate and estimates $\beta_{\text{LMM}} = 0.504$ and the individual biases to be $b_{\text{LMM}} = [-1.96; -1.015; -0.014; 0.973; 2.013]$, as can be seen in the lower part of Figure 1.

2.2. ℓ_1-penalized Maximum Likelihood Estimator

Since we have to deal a large number of covariates, it is computationally not feasible to employ the standard mixed-effects model variable selection

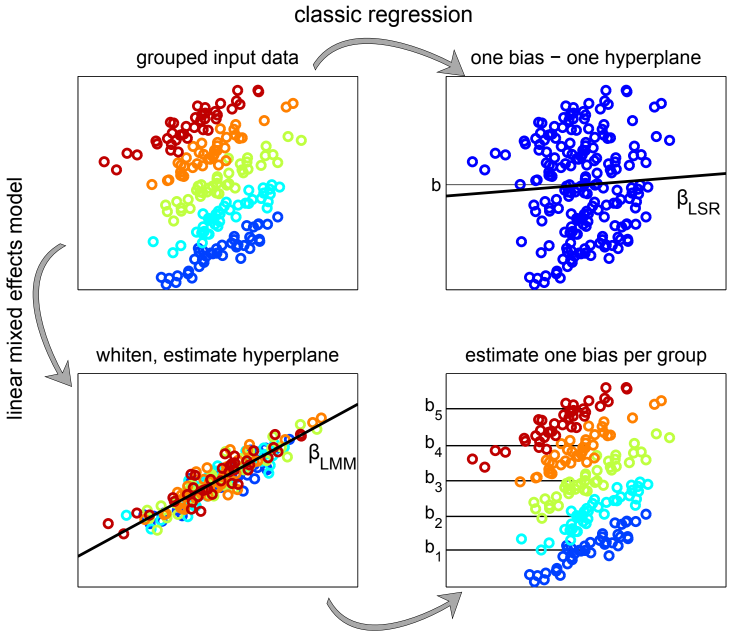

Figure 1: Illustration of the fitting procedure for a linear mixed-effects model with $Z = \mathbf{1}_{n_i}$, i.e. a random intercept model: groups have the same slope but different intercepts. The colors distinguish groups. If fitted with a classical regression, the fixed-effect is not recovered correctly. By applying Algorithm 1, the data are first whitened with Λ_i and then the fixed-effect is estimated from the whitened data by linear regression. In as a second step, the random effects are recovered.

strategies. To remedy this problem, in (Schelldorfer et al., 2010) a Lasso-type approach is proposed by adding an ℓ_1-penalty for the fixed-effects parameter β. This idea induces sparsity in β in the sense that many coefficients $\beta_j, j = 1, \ldots, p$ are estimated exactly zero and we can perform simultaneously parameter estimation and variable selection. Consequently, from (2) we derive the following objective function

$$S_\lambda(\beta, \sigma^2, \tau^2) := -\frac{1}{2} \sum_{i=1}^{N} \left\{ \log |\Lambda_i| + (y_i - X_i\beta)^\intercal \Lambda_i^{-1} (y_i - X_i\beta) \right\} - \lambda \sum_{k=1}^{p} |\beta_k| \quad , \tag{3}$$

where λ is a nonnegative regularization parameter.
Hence, estimating the parameters β, σ^2 and τ^2 is carried out by maximizing $S_\lambda(\beta, \sigma^2, \tau^2)$:

$$\hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2 = \underset{\beta, \sigma^2, \tau^2}{\operatorname{argmax}} \, S_\lambda(\beta, \sigma^2, \tau^2) \quad . \tag{4}$$

It is worth noting that $S_\lambda(\beta, \sigma^2, \tau^2)$ is a non-concave function, which implies that we can not apply a convex solver to maximize (3).

2.3. Prediction of the random-effects

The prediction of the random-effects coefficients $b_i, i = 1, \ldots, N$ is done by the maximum a posteriori (MAP) principle. Given the parameters β, σ^2 and τ^2, it follows by straightforward calculations that the MAP estimator for $b_i, i = 1, \ldots, N$ is given by $b_i = [Z_i^\intercal Z_i + \sigma^2/\tau^2 I_q]^{-1} Z_i^\intercal (y_i - X_i\beta)$. Since the true parameters β, σ^2 and τ^2 are not known, we plug in the estimates from (4). Hence the random-effects coefficients are estimated by

$$\hat{b}_i = [Z_i^\intercal Z_i + \hat{\sigma}^2/\hat{\tau}^2 I_q]^{-1} Z_i^\intercal (y_i - X_i\hat{\beta}). \tag{5}$$

2.4. Model Selection

The optimization problem in (4) is applied to a fixed tuning parameter λ. In practice, the solution of (4) is calculated on a grid of λ values. The choice of the optimal λ-value is then achieved by minimizing a criterion, i.e. a k-fold cross-validation score or an information criteria. We propose to use the Bayesian Information Criterion (BIC) defined as

$$-2\ell(\hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2) + \log N_T \cdot \hat{df}_\lambda \quad , \tag{6}$$

where $\hat{df}_\lambda = |\{1 \leq j \leq p; \hat{\beta}_j \neq 0\}|$ denotes the number of nonzero fixed regression coefficients and $\ell(\hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2)$ denotes the likelihood function following from the model assumptions in (1). The BIC works well in the simulation examples presented in (Schelldorfer et al., 2010) and is computationally fast.

2.5. Computational Implementation

With τ and σ fixed, the cost function (3) is equivalent to an ℓ_1-penalized linear regression after whitening by the covariances Λ_i:

$$\hat{\beta} = \underset{\beta|\tau,\sigma}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \Lambda_i^{-1/2} \left(X_i\beta - y_i \right) \right\|_2^2 + 2\lambda \sum_{k=2}^{p} |\beta_k| \tag{7}$$

We solve the resulting convex optimization problem for b with fixed σ and τ using the orthant-wise limited memory quasi-Newton algorithm (Andrew and Gao, 2007). As suggested in (Schelldorfer et al., 2010), the optimization is performed over a grid of (σ^2, τ^2) to find the optimum of the considered parameters.

Preliminary analysis indicates that a so called random-intercept (i.e. one bias per group) is appropriate for our data, i.e., $Z_i = \mathbf{1}_{n_i}$ and $b_i \in \mathbb{R}$. Then, in the context of (1), σ^2 corresponds to the *within-subject variability* and τ^2 to the *between-subject variability.* By estimating σ^2 and τ^2 we are able to allocate the variability in the data to these to sources.

3. Available Data and Experiments

We use two different datasets of BCI data to show different aspects of the validity of our approach. The first consists of 83 BCI experiments (sessions) from 83 individual subjects and each session consists of 150 trials (Blankertz et al., 2010b). Our second dataset consists of 90 sessions from only 44 subjects. The number of trials of a single session varies from 60 trials to 600 trials (Fazli et al., 2009b). In other words, our first dataset can be considered to be *balanced* in the number of *trials per subjects* and *sessions per subject.* Our second dataset is *unbalanced* in this sense. As one may expect, the balanced data is more suitable for building a zero-training classifier and enables us to obtain a 'clean' model. However, the unbalanced dataset enables us to examine how individual sessions of the same subject affect the estimation of our model and leads to a more thorough understanding of the underlying processes.

Each trial consists of one of two predefined movement imaginations, being left and right hand, i.e. data was chosen such that it relies only on these 2 classes, although originally three classes were cued during the calibration session, being left hand (L), right hand (R) and foot (F). 45 EEG channels, which are in accordance with the 10-20 system, were identified to be common in all sessions considered. The data were recorded while subjects were immobile, seated on a comfortable chair with arm rests. The cues for performing a movement imagination were given by visual stimuli, and occurred every 4.5-6 seconds in random order. Each trial was referenced by a 3 second long time-window starting at 500 msec after the presentation of the cue. Individual experiments consisted of three different training paradigms. The first two training paradigms consisted of visual cues in form of a letter or an arrow, respectively. In the third training paradigm the subject was instructed to follow a moving target on the screen. Within this target the edges lit up to indicate the type of movement imagination required. The experimental procedure was designed to closely follow (Blankertz et al., 2006a). Electromyogram (EMG) on both forearms and the foot were recorded as well as electrooculogram (EOG) to ensure there were no real movements of the arms and that the movements of the eyes were not correlated to the required mental tasks.

3.1. Generation of the Ensemble

The generation of the ensemble is a requirement for our aim of classifying the single trials of movement imagination of a novel subject. By exploiting our large database of previously recorded subjects, we generate a large set basis functions, each of which consist of subject-dependent temporal and spatial filters as well as their matching linear classifiers (LDA). This procedure is visualized in the upper panel of Figure 2, which is adopted from (Fazli et al., 2009b). For a new subject, each trial is now processed by this set of basis functions and the resulting outputs are combined with a weighted sum to predict its class label. Finding an appropriate weighting for the classifier outputs of these basis functions is of paramount importance for the accurate prediction. This weighting is found by linearly regressing each classifier's output onto the known targets.

The design matrix X and targets y for the regression are generated as follows: Each trial of each subject is first processed by 18 predefined bandpass filters, CSPs and then linearly classified. Since we have 83 subjects with 18 classifiers each, the total number of features is $18 \cdot 83 = 1494 \Rightarrow \beta \in \mathbb{R}^{1494}$. Each of the 83 subjects performed 150 trials, therefore we have $150 \cdot 83 = 12450$ data points. The data matrix X and the targets y have thus the dimensionalities $X \in \mathbb{R}^{12450 \times 1494}$ and $y \in \mathbb{R}^{12450}$. Note that contrary to the usual use case of ℓ_1-regularization, our regression problem is not ill-posed, i.e., in our case, $n > p$. We employed different forms of regression,

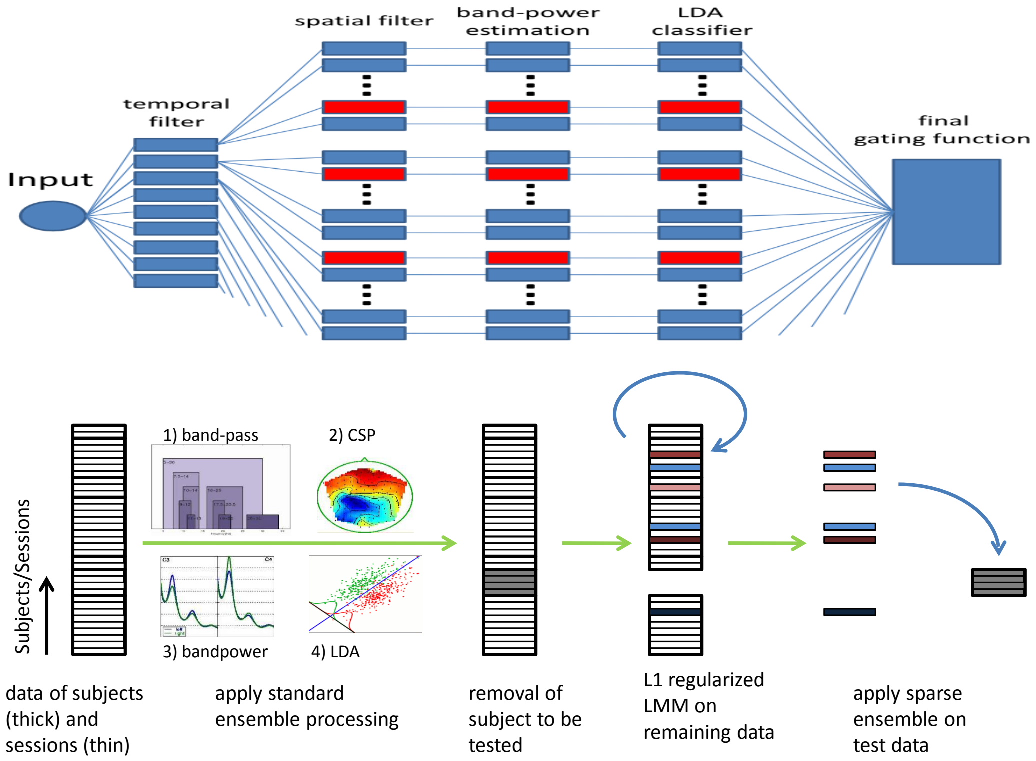
spatial filter
band-power estimation
LDA classifier
temporal filter
Input
final gating function
Subjects/Sessions
1) band-pass
2) CSP
3) bandpower
4) LDA
data of subjects (thick) and sessions (thin)
apply standard ensemble processing
removal of subject to be tested
L1 regularized LMM on remaining data
apply sparse ensemble on test data

Figure 2: Two flowcharts of the ensemble method. The red patches in the top panel illustrate the inactive nodes of the ensemble after sparsification.

namely the previously described linear mixed-effects model, as well as the more classical ℓ_1-regularized regression in order to find an optimal and sparse weighting for predicting the movement imagination data of unseen subjects. Since we have more data points than features we penalize with ℓ_1 not just for regularization, but rather to obtain a sparse and therefore easily interpretable solution. While this problem could also be solved with classification methods, previous results have shown, that no substantial benefit would be gained by doing so (Fazli et al., 2009a). Instead of taking the ensemble members' outputs, passing them through a sigmoid function and learning the weighting (for the summation of ensemble outputs) via logistic regression, we simply perform linear regression on the classifiers' linear outputs directly.

The validation was done by leave-one-subject-out cross-validation, i.e. the session of a particular subject was removed, the algorithm trained on the remaining trials (of the other subjects) and then applied to this subject's data (see lower panel of Figure 2).

3.1.1. Temporal Filters

The μ-rhythm (9-14 Hz) and synchronized components in the β-band (16-22 Hz) are macroscopic idle rhythms that prevail over the postcentral somatosensory cortex and precentral motor cortex, when a given subject is at rest. Imaginations of movements as well as actual movements are known to suppress these idle rhythms contralaterally. However, there are not only subject-specific differences of the most discriminative frequency range of the mentioned idle-rhythms, but also session differences thereof.

We identified 18 neurophysiologically relevant temporal filters, of which 12 lie within the μ-band, 3 in the β-band, two in between μ- and β-band and one broadband $7-30$Hz. In all following performance related tables we used the percentage of misclassified trials, or 0-1 loss.

3.1.2. Spatial Filters and Classifiers

Common spatial patterns (CSP) is a popular algorithm for calculating spatial filters, used for detecting event-related (de-) synchronization (ERD/ERS), and is in connection with the appropriate classifier considered to be the gold-standard of ERD-based BCI systems (Koles and Soong, 1998; Ramoser et al., 2000; Blankertz et al., 2002; Dornhege et al., 2006; Parra et al., 2005; Blankertz et al., 2008; Tomioka and Müller, 2010). The CSP algorithm maximizes the variance of right hand trials, while simultaneously minimizing the variance for left hand trials. Given the two covariance matrices Σ_1 and Σ_2, of size *channels* × *concatenated timepoints*, the CSP algorithm returns the matrices W and D. W is a matrix of projections, where the i-th row has a relative variance of d_i for trials of class 1 and a relative variance of $1 - d_i$ for trials of class 2. D is a diagonal matrix with entries $d_i \in [0, 1]$, with length n, the number of channels:

$$W\Sigma_1 W^{\mathsf{T}} = D \qquad \text{and} \qquad W\Sigma_2 W^{\mathsf{T}} = I - D \tag{8}$$

Best discrimination is provided by filters with very high (emphasizing one class) or very low eigenvalues (emphasizing the other class), we therefore chose to only include projections with the highest 2 and corresponding lowest 2 eigenvalues for our analysis. We use Linear Discriminant Analysis (LDA) (Blankertz et al., 2003), each time filtered session corresponds to a CSP set and to a matched LDA. Note that in principle also nonlinear filter-classifier combinations could be employed as 'basis functions' of the ensemble (see also (Müller et al., 2001; Müller et al., 2003; Blankertz et al., 2006b, 2010a)).

3.2. Validation

The subject-specific CSP-based classification methods with automatically, subject-dependent tuned temporal filters (termed reference methods) are validated by an 8-fold cross-validation, splitting the data chronologically. The chronological splitting for cross-validation is a common practice in EEG classification, since the non-stationarity of the data is thus preserved (Dornhege et al., 2007).

To validate the quality of the ensemble learning we employed a leave-one-subject out cross-validation (LOSO-CV) procedure, i.e. for predicting the labels of a particular subject we only use data from other subjects.

4. Results

4.1. Subject-to-Subject Transfer

As explained in Section 3, we use our first balanced dataset to find a zero-training subject-independent classifier. The left part of Figure 3 shows the results of fitting an ℓ_1-regularized least-squares regression model to fit a) a linear model with one bias and b) a mixed-effects model with one bias per subject. We are able to improve classification accuracy by use of the mixed-effects model. As can be seen in Figure 4 (left panel) the LMM method needs less features per subject ($N_{\text{LMM}} \approx 310$) as compared to estimating only one

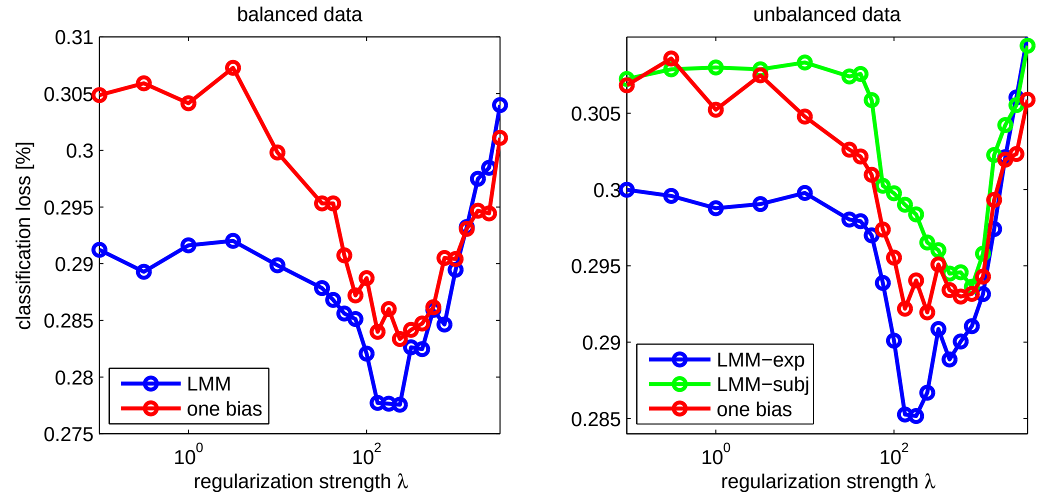

Figure 3: The figures show the mean classification loss over subjects for the *balanced* dataset (left) and the *unbalanced* dataset (right) as a function of the regularization strength λ. The LMM approach is compared to classical L1 regularized least squares regression (one bias). LMM-subj estimates one bias per subject and LMM-exp one bias per experiment (session).

bias ($N_{\ell_1} \approx 500$).

Besides from selecting less features in total, the LMM chose a higher fraction of features with low self-prediction errors, where 'self-prediction error' denotes the average cross-validation error when using the training data of a subject to predict his test data, i.e., performing conventional, subject-dependent BCI. This is shown in the middle panel, where we display the cumulative sum of features, sorted by increasing self-prediction accuracy.

To visualize differences between weight vectors resulting from the LOSO-CV procedure, the right panel displays these vectors, projected to two dimensions. The matrix of Euclidean distances between all pairs of weights was embedded into a 2×83-dimensional space and projected onto the resulting point cloud's first two principal axes for visualization. The mixed effects model absorbs more of the variability into its bias terms and thus results in more consistent weight vector estimates. The sparsity of our results becomes apparent from Figure 5, where we display the magnitude of weights for each run of the LOSO-CV. For LMM on average 28.9% of all features are active, while for 'one-bias' 33.5% of all features are non-zero. Note that for both methods most of the active features lie within a vertical line, indicating that

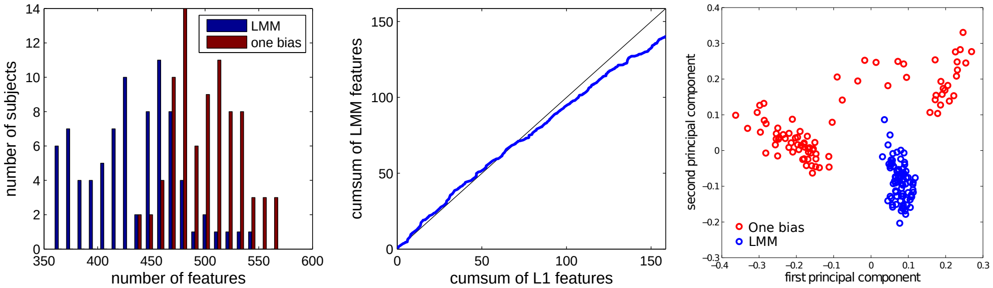

Figure 4: Left: histogram of the number of selected features for all subjects. Middle: cumulative sum of features, sorted by 'self prediction'. LMM rather chooses features, that had a good 'self prediction', and needs less features in total. Right: Variability between classifier weights b of the two models for each of the $N = 2 \times 83$ LOSO-training runs using the best regularization strength.

the feature is also active for most other subjects and can thus be considered particularly stable. In Figure 6 we compare the performance of our method on the basis of individual subjects with other methods and perform t-tests to examine their statistical significance. The p-values are included within the figure. As the most simple baseline method we used 'Laplace features' by calculating the difference of two motor related channels (namely 'C3' and 'C4') within a time interval of $750 - 3500$ ms, after broadband ($7 - 30$ Hz) temporal and Laplacian spatial filtering of the individual channels. This method scored an average loss of 33.9%. As can be seen on the left side of Figure 6 our novel method performs very favorably. LMM improves classification performance for 89.2% of the subjects considered with high significance and leads to an average loss of 27.6%. Furthermore, we compare with a recently proposed zero-training procedure (Fazli et al., 2009a), which is very similar to the LMM method described here, except that it performs ℓ_1-regularized regression for combining the outputs of the individual classifiers (average loss 28.3%). Also here we achieve a significant improvement. Finally, we compare our method to the subject-dependent, cross-validated classifier loss, derived from the data themselves (average loss 27.5%). A per se unfair comparison. Given that the subject-dependent classifier is not significantly better ($p = 0.93$), we may state that we are on par.

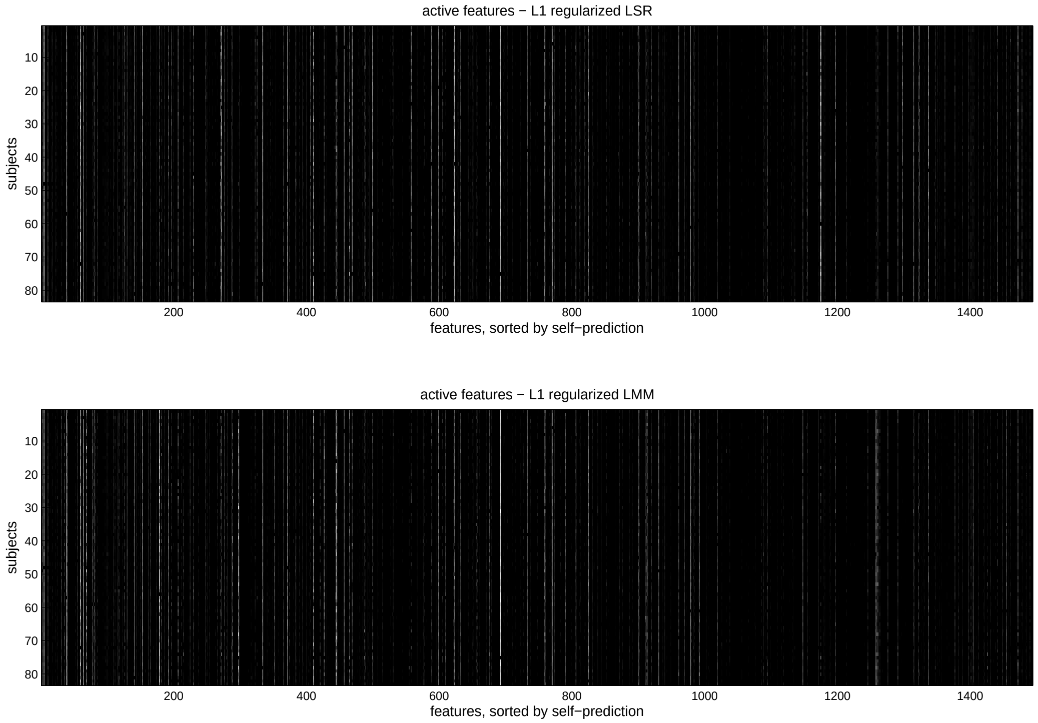

Figure 5: Both plots show the selected features in white, while inactive features are black. The x-axis represents all possible features, sorted by their cross-validated self-prediction. The y-axis represents each subjects resulting weight vector. 'LSR' stands for least-squares regression.

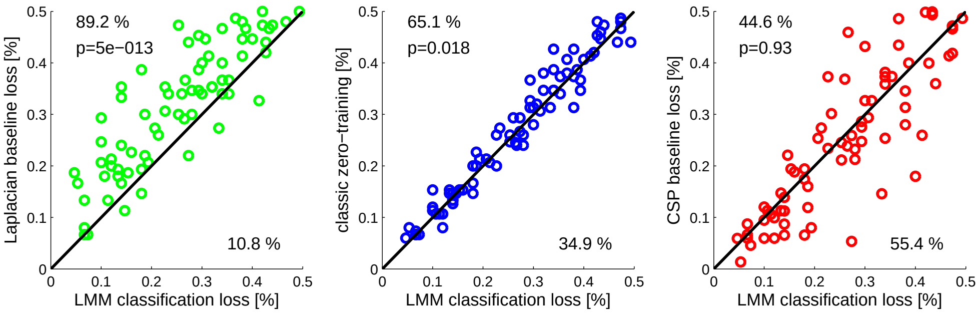

Figure 6: Scatter plot, comparing the proposed method with various baselines on a subject specific level.

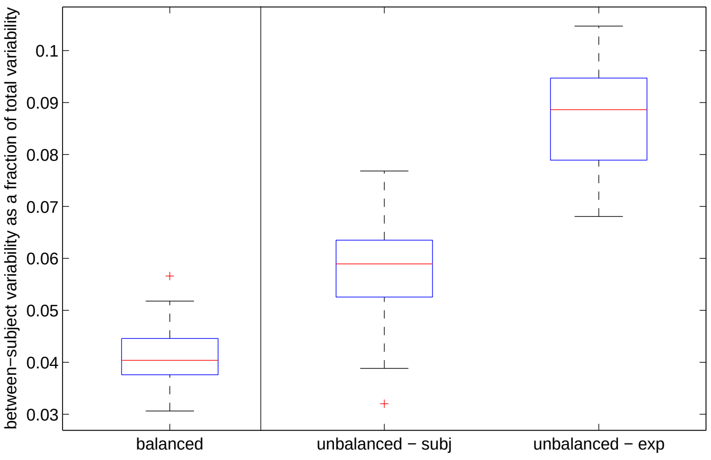

Figure 7: The figure show the magnitude of between-subject variability as a fraction of total variability. On the left: Results for the first *balanced* dataset. One the right: Results for the *unbalanced* dataset. *subj* stands for estimating one bias per subject and *exp* for estimating one bias per experiment.

4.2. Session-to-Session Transfer

To investigate how the results of the method can be understood in terms of individual subjects and their (possibly multiple) sessions, we validated the method in two ways. First we allow each experiment to have an individual bias. In the second approach, we allow only one bias per subject, i.e. multiple experiments/sessions from the same subject will be grouped. The results are shown in the right panel of Figure 7. They indicate a substantially higher between-group-variability if we allow biases for each experiment. This does not only confirm knowledge from previous publications, that the transfer of classifiers from sessions to sessions required a bias correction (Krauledat et al., 2008; Shenoy et al., 2006b), but also underlines the validity of our approach in the sense that we are able to capture a meaningful part of the variability which would otherwise be ignored as noise. As can be seen in Figure 7, a substantial fraction of the variability can be attributed to within-subject differences.

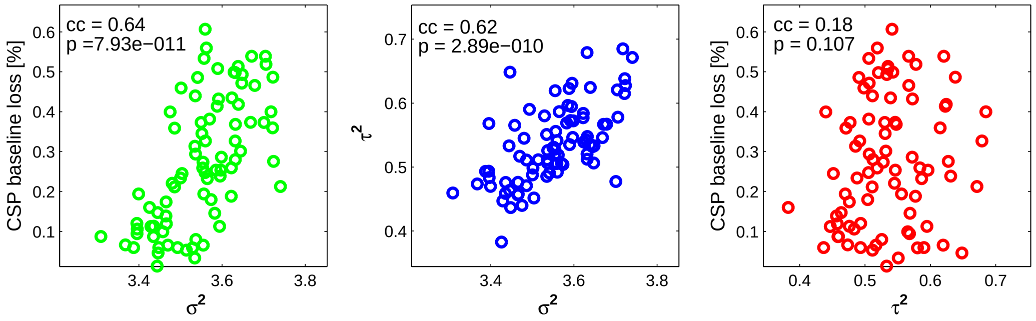

Figure 8: The three scatterplots show relations between *within-subject variability* $\sigma.^2$, *between-subject variability* τ^2 and the baseline cross-validation misclassification for every subject. *cc* stands for correlation coefficient and *p* stands for paired t-test significance.

4.3. Relation of baseline misclassification to σ^2 and τ^2

Using standard methods for ERD-related BCI decoding (Blankertz et al., 2008), we obtain a mean classification loss for each subject within our *balanced* dataset, based on the cross-validation of band-pass and spatially filtered features. In Figure 8 we examine the relationship between this *baseline loss* and the *within-subject variability* $\sigma.^2$ and *between-subject variability* $\tau.^2$. The *baseline loss* and σ^2 have a strong positive correlation, with high significance. This makes intuitive sense: a dataset that is well classifiable should also exhibit low variance of its residuals. We furthermore examine the relation of τ^2 and σ^2 and find a strong positive relation.
Interestingly we do not find a significant relation between the baseline loss and τ^2. In other words it is not possible to draw conclusions about the quality of a subject's data by the variance of its assigned biases.

4.4. Effective spatial filters and distances thereof

To estimate the similarity of effective spatial filters, we use a transfer function as described in (Fazli et al., 2009b): By injecting a sinusoid into a given channel and processing it by the spatial filter, estimating the bandpower and applying the classifier, we obtain a response for one particular channel. Repeating this procedure for each channel results in a response matrix that can be easily visualized. We define a distance measure for each individual subject between her original CSP filter and those estimated via 'LMM' and 'one bias' methods. The measure we use is the angle between their vectorized response matrices (see (Krauledat et al., 2008)).

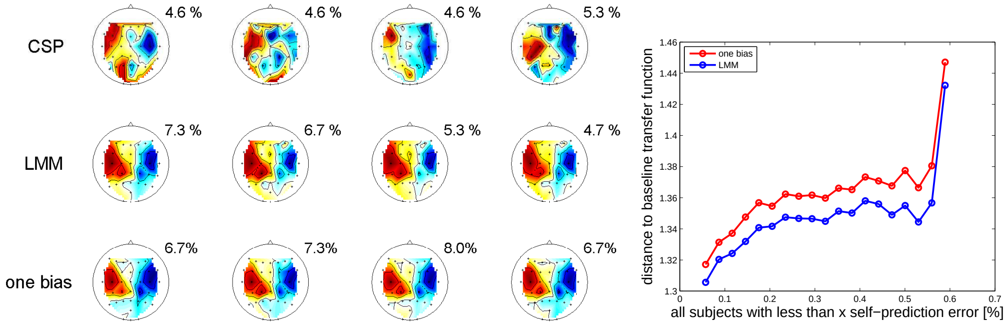

Figure 9: Left part: Response matrices of the four best subjects for 'original CSP', 'LMM' and 'one bias'. Classification loss is given as percentage numbers. Right part: Response distances of 'LMM' and 'one bias' versus self-prediction error [%].

For four subjects the resulting response matrices, based on the original CSP pattern, are shown on the top row of the left part of Figure 9. To obtain a response matrix for the ensemble approaches, we calculate the weighted sum of responses, determined by β (see middle and lower parts of Figure 9).
In the right part of Figure 9 the resulting distances between 'LMM' or 'one bias' and the original CSP based response function are plotted against all subjects with self-prediction loss of less than x. As one would expect both distances increase on average, as more subjects with higher self-prediction loss are added to the analysis. It shows that the linear mixed-effects model is consistently closer, irrespective of the subject's self-prediction error.

5. Discussion and Conclusions

When analyzing experimental data, it is of generic importance to quantify variation both across the ensemble of acquired data and within repetitions of measurements. Distinguishing and modeling such mixed effects is of high interest e.g. in medicine, biology, physics and the neurosciences. In this paper we have applied a recent sparse modeling approach from statistics (Schelldorfer et al., 2010) based on a so-called ℓ_1-penalized linear mixed-effects model and proposed its first time use for a large BCI data set: leading to a novel BCI zero-training model (see also (Krauledat et al., 2008; Fazli et al., 2009b)). In this manner we could efficiently model the different dependencies and variabilities between and within subjects. Note that the novel statistical model not only gave rise to a better overall prediction – in other words to an *improved zero-training model* – but it furthermore allowed to quantify the differences in variation more transparently and also interpretably. By attributing some of the total variability, in other methods considered as noise, to differences between subjects, we are now able to obtain a solution that is sparser and at the same time superior in prediction accuracy. Not only features with high prediction performance are preferably chosen, but also responses of the novel ensemble are more similar to it's original counterpart. Furthermore, we would like to note that while more complex random effects would in principle be conceivable, our random intercepts model was not just chosen by intuition but from our experience with BCI: When performing an experiment with the same subject on two subsequent days, on the second day the classifier can often be reused without much retraining, only the bias needs to be adjusted (Krauledat et al., 2008; Shenoy et al., 2006a). We have developed a statistical framework that can be applied to a large number of scientific experiments from a large number of domains, where inter-dependencies of input space exist and have shown that our approach leads to more robust feature selection and is superior in its classification accuracy. Future research will study online adaptation of penalized linear mixed-effects models in the context of medical diagnosis.

Appendix A. Algorithm

```
foreach $(\sigma^2, \tau^2, \lambda)$ do
    foreach $i$ do
        (Whiten data and labels)
        $\Lambda_i = \sigma^2 I_{n_i} + \tau^2 Z_i Z_i^\mathsf{T}$
        $\bar{X}_i = \Lambda_i^{-1/2} X_i, \quad \bar{y}_i = \Lambda_i^{-1/2} y_i$
    end
    (Fit $\ell_1$-penalized least-squares to concatenated data)
    $\hat{\beta} = \operatorname{argmin} \left\| \bar{X}\beta - \bar{y} \right\|_2^2 + 2\lambda \sum_{k=2}^{p} |\beta_k|$
    foreach $i$ do
        (Find random effects)
        $\hat{b}_i = [Z_i^\mathsf{T} Z_i + \sigma^2/\tau^2 I_q]^{-1} Z_i^\mathsf{T} (y_i - X_i \hat{\beta})$
    end
end
```

Algorithm 1: algorithm for fitting the mixed effects model

Appendix B. Flowchart and Visualization

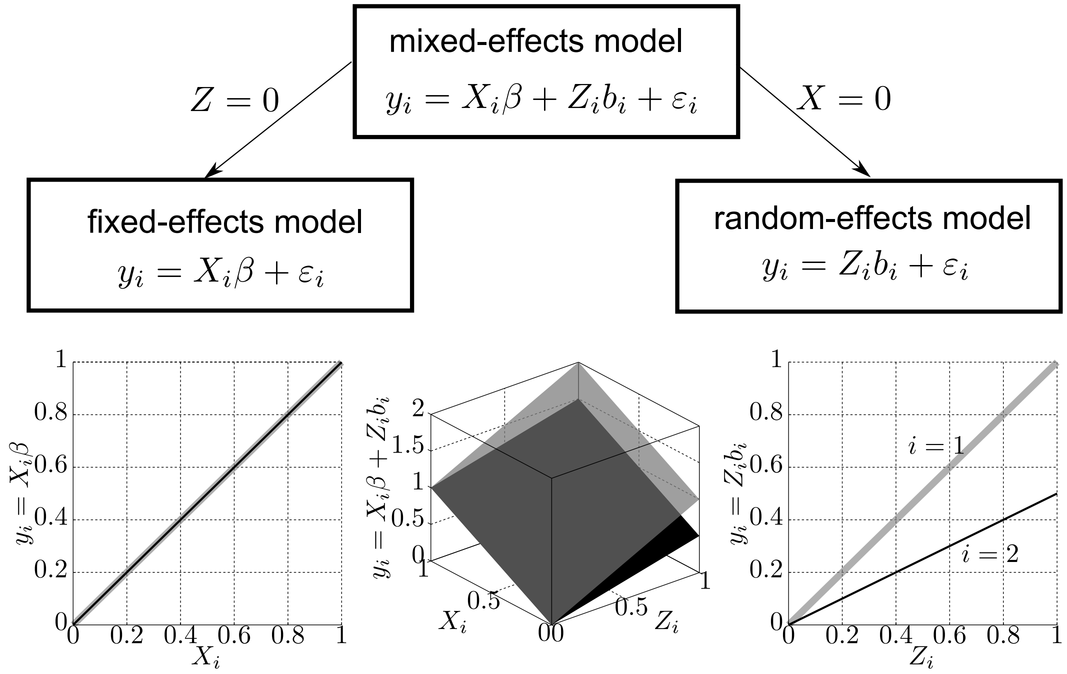

Figure B.10: Upper part: The flowchart gives an overview of the mixed-effects, random-effects and fixed-effects models. Lower part: Plot of the mixed-effects model $y = X_i\beta + Z_i b_i$ without noise, with $i = \{1, 2\}$, $\beta = 1$, $b_1 = 1$, $b_2 = 1/2$. Grey: group 1, black: group 2. The figure in the middle shows the plot for the general case. In the left plot, where $Z_1 = Z_2 = 0$, i.e. with only fixed effects present, the model reduces to a linear function in X_i: the two curves coincide. In the right plot, where $X_i = 0$, i.e. only random effects are present, the groups are decoupled and form two independent linear functions.

References

Alamgir, M., Grosse-Wentrup, M., Altun, Y., 2010. Multitask learning for brain-computer interfaces, in: Teh, Y.W., Titterington, M. (Eds.), Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010), pp. 17–24.

Andrew, G., Gao, J., 2007. Scalable training of L_1-regularized log-linear models, in: Proceedings of the 24th international conference on Machine learning, ACM, Corvalis, Oregon. pp. 33–40.

Blankertz, B., Curio, G., Müller, K., 2002. Classifying single trial EEG: Towards brain computer interfacing, in: Diettrich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Inf. Proc. Systems (NIPS 01), pp. 157–164.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K., Curio, G., 2007. The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. NeuroImage 37, 539–550.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K., Kunzmann, V., Losch, F., Curio, G., 2006a. The Berlin Brain-Computer Interface: EEG-based communication without subject training. IEEE Trans Neural Syst Rehabil Eng 14, 147–152.

Blankertz, B., Dornhege, G., Lemm, S., Krauledat, M., Curio, G., Müller, K., 2006b. The Berlin Brain-Computer Interface: Machine learning based detection of user specific brain states. Journal of Universal Computer Science 12, 2006.

Blankertz, B., Dornhege, G., Schäfer, C., Krepki, R., Kohlmorgen, J., Müller, K., Kunzmann, V., Losch, F., Curio, G., 2003. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. IEEE Trans. Neural Sys. Rehab. Eng. 11, 127–131.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K., 2010a. Single-trial analysis and classification of ERP components – a tutorial. Neuroimage .

Blankertz, B., Sannelli, C., Halder, S., Hammer, E.M., Kubler, A., Müller, K., Curio, G., Dickhaus, T., 2010b. Neurophysiological predictor of SMR-based BCI performance. Neuroimage 51, 1303–1309.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K., 2008. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Proc Magazine 25, 41–56.

Bondell, H.D., Krishna, A., Ghosh, S.K., 2010. Joint variable selection of fixed and random effects in linear mixed-effects models. Biometrics In Press.

Cheng, M., Gao, X., Gao, S., Xu, D., 2002. Design and implementation of a brain-computer interface with high transfer rates. Biomedical Engineering, IEEE Transactions on 49, 1181 –1186.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., Müller, K., 2006. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. IEEE Trans Biomed Eng 53, 2274–2281.

Dornhege, G., Millán, J.R., Hinterberger, T., McFarland, D., Müller, K. (Eds.), 2007. Toward Brain-Computer Interfacing. Cambridge, MA: MIT Press.

Fazli, S., Grozea, C., Danoczy, M., Blankertz, B., Popescu, F., Müller, K., 2009a. Subject independent EEG-based BCI decoding, in: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. MIT Press, pp. 513–521.

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.R., Grozea, C., 2009b. Subject-independent mental state classification in single trials. Neural Netw 22, 1305–1312.

Ibrahim, J.G., Zhu, H., Garcia, R.I., Guo, R., 2010. Fixed and random effects selection in mixed effects models. Biometrics In Press.

Koles, Z., Soong, A.C.K., 1998. EEG source localization: implementing the spatio-temporal decomposition approach. Electroencephalogr. Clin Neurophysiol 107, 343–352.

Krauledat, M., Tangermann, M., Blankertz, B., Müller, K., 2008. Towards zero training for brain-computer interfacing. PLoS ONE 3, e2967.

Meier, L., van de Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. Journal of the Royal Statistical Society 70, 53–71.

Müller, K., Anderson, C.W., Birch, G.E., 2003. Linear and nonlinear methods for brain-computer interfaces. Neural Systems and Rehabilitation Engineering, IEEE Transactions on 11, 165–169.

Müller, K., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Neural Networks 12, 181–201.

Parra, L., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., Sajda, P., 2008. Spatiotemporal Linear Decoding of Brain State. IEEE Signal Proc Magazine 28, 107–115.

Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. Neuroimage 28, 326–341.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-Plus. Springer, New York.

Ramoser, H., Müller-Gerkin, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehab. Eng 8, 441–446.

Schelldorfer, J., Bühlmann, P., van de Geer, S., 2010. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. arXiv preprint 1002.3784 .

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R., Müller, K., 2006a. Towards adaptive classification for BCI. Journal of Neural Engineering 3, R13–R23.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R.P.N., Müller, K., 2006b. Towards adaptive classification for BCI. J Neural Eng 3, 13–23.

Thomas, K.P., Guan, C., Lau, C.T., Vinod, A., Ang, K.K., 2009. A new discriminative common spatial pattern method for motor imagery brain-computer interfaces. Biomedical Engineering, IEEE Transactions on 56, 2730 –2733.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tomioka, R., Müller, K., 2010. A regularized discriminative framework for EEG analysis with application to brain-computer interface. Neuroimage 49, 415–432.

Vidal, J.J., 1973. Toward direct brain-computer communication. Annual Review of Biophysics and Bioengineering 2, 157–180.

Wang, Y., Wang, R., Gao, X., Hong, B., Gao, S., 2006. A practical vep-based brain-computer interface. Neural Systems and Rehabilitation Engineering, IEEE Transactions on 14, 234 –240.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B 68, 49–67.