Multi-Task Logistic Regression in Brain-Computer Interfaces

Karl-Heinz Fiebig*, Vinay Jayaram†, Jan Peters*† and Moritz Grosse-Wentrup†
* Department of Computer Science
Technische Universität Darmstadt, 64289 Darmstadt, Germany
Email: karl-heinz.fiebig@stud.tu-darmstadt.de, mail@jan-peters.net
† Department of Empirical Inference
Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
Email: {vjayaram, jan.peters, moritzgw}@tuebingen.mpg.de

***Abstract*—A Brain-Computer Interface (BCI) is used to enable communication between humans and machines by decoding elicited brain activity patterns. However, these patterns have been found to vary across subjects or even for the same subject across sessions. Such problems render the performance of a BCI highly specific to subjects, requiring expensive and time-consuming individual calibration sessions to adapt BCI systems to new subjects. This work tackles the aforementioned problem in a Bayesian multi-task learning (MTL) framework to transfer common knowledge across subjects and sessions for the adaptation of a BCI to new subjects. In particular, a recent framework that is able to exploit the structure of multi-channel Electroencephalography (EEG), is extended by a Bayesian hierarchical logistic regression decoder for probabilistic binary classification. The derived model is able to explicitly learn spatial and spectral features, therefore making it further applicable for identification, analysis and evaluation of paradigm characteristics without relying on expert knowledge. An offline experiment with the new decoder shows a significant improvement in performance on calibration-free decoding compared to previous MTL approaches for rule adaptation and uninformed models while also outperforming them as soon as subject-specific data becomes available. We further demonstrate the ability of the model to identify relevant topographies along with signal band-power features that agree with neurophysiological properties of a common sensorimotor rhythm paradigm.**

I. Introduction

Second only to brain signal acquisition, the decoding of subject intention is fundamental to the practice of practical brain-computer interfaces (BCIs). However, state-of-the-art BCIs suffer from high performance variations between subjects and even across sessions within the same subject [1], [2]. A calibration session with the subject prior to actual BCI usage is therefore necessary, which poses a major hindrance to out-of-the-box applications. On top of that, machine learning based decoders have to deal with rather few data points gained from the calibration phase, making them prone to overfitting and requiring very low dimensional feature spaces. Only a hand full of carefully selected features using expert knowledge of the paradigm are usually used to train decoders that generalize well to upcoming sessions [3]–[6].

Approaches to solve the problem of such performance variations in the past years have been dominated by domain adaptation techniques. These techniques use data acquired from different subjects and sessions to train decoders on invariant feature spaces, most commonly by preprocessing signals with common spatial patterns [7]–[9]. While domain adaptation is able to drastically reduced calibration time for new subjects and sessions with only slight performance losses, subject-specific variations are modeled exclusively by a fixed feature space. This makes it difficult to adapt BCIs to new subjects when calibration data becomes available. A more natural approach to this problem is given by rule adaptation techniques, which encode variations directly into the decision rule of decoding models.

Recent work in this area includes data fusion methods [10] and a Bayesian multi-task learning (MTL) framework first proposed by Alamgir et al. [11]. Jayaram et al. generalized the framework to a Bayesian hierarchy with a novel feature space to decompose EEG structure into a spatial and a frequency [12] or time domain component [13]. This approach did not only enable the authors to transfer knowledge between subject and sessions, but also further reduce the need for expert knowledge through simultaneous inference of topographic and broad-band features of the paradigm. The resulting framework produced reliable decoders (classification accuracies above 70%) for new subjects based on a sensorimotor rhythm (SMR) paradigm without or few calibration trials while outperforming models trained solely from pooled or subject-specific data when calibration data was available.

This work extends the MTL framework for transfer learning by a more suitable assumption on binary dependent variables with a probabilistic model for two-class classification. Exploiting EEG structure for dimensionality reduction yields a bilinear logistic regression model that is able to learn relevant topographic and band-specific features instead of relying on a small set of manually selected features.

The rest of this paper is structured as follows. Section II introduces the notation used throughout this work and derives a logistic regression model within a Bayesian MTL framework from previous work. The model is extended for bilinear feature decomposition (FD) to exploit multi-channel EEG structure and the final learning algorithm is presented. After describing the SMR based experimental setup used to

978-1-5090-1897-0/16/$31.00

evaluate the models, Section III shows that the derived model outperforms comparable models in calibration free decoding as well as in subject-adaptation. Section IV concludes this work with a short summary and future work.

II. Methods

A. Notation

Throughout this paper we denote scalars with lower case, vectors with bold lowercase, matrices with uppercase and sets with calligraphic uppercase letters. We regard the decoding problem for each subject or session as an individual task on binary dependent variables and denote the data set of m tasks with $\mathcal{T} = \left\{\mathcal{D}^{(t)}\right\}_{t=1}^{m}$. Each task data set $\mathcal{D}^{(t)} \in \mathcal{T}$ is formalized with

$$\mathcal{D}^{(t)} = \left\{\left(\boldsymbol{x}_i^{(t)}, y_i^{(t)}\right)\right\}_{i=1}^{n_t} \subset \mathbb{R}^d \times \{C_1, C_2\} \tag{1}$$

consisting of n_t data points with d-dimensional feature vectors extracted from EEG signals and corresponding class label C_1 or C_2 representing one of two brain conditions of interest. In case of FD, each feature vector $\boldsymbol{x} \in \mathbb{R}^d$ is replaced by the corresponding feature matrix $X \in \mathbb{R}^{k \times d}$ that organizes d band-power features from k channels. Matrix calculus follows denominator-layout notation.

B. Multi-task Logistic Regression

A popular method for simple binary classification using probabilistic predictions is to pass the linear model through the logistic sigmoid activation. The hypothesis model is then given by

$$h(\boldsymbol{x}; \boldsymbol{w}) = \left(1 + \exp\left(\boldsymbol{w}^T \boldsymbol{x}\right)\right)^{-1} \in]0, 1[\tag{2}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is an input feature and $\boldsymbol{w} \in \mathbb{R}^d$ is the parameter vector of the model family. The output of (2) can be interpreted as the probability $p\left(C_1 \mid \boldsymbol{x}\right) = h(\boldsymbol{x}; \boldsymbol{w})$ to observe condition C_1 in brain state $\boldsymbol{x}$. Likewise, the probability to observe C_2 in state $\boldsymbol{x}$ is given by the complementary event $p\left(C_2 \mid \boldsymbol{x}\right) = 1 - h(\boldsymbol{x})$.

Assume that we have gathered a set $\mathcal{T}$ as in (1). Following the MTL framework we model each task with an individual hypothesis from (2). Hence, we train a set of m weight vectors $\mathcal{W} = \left\{\boldsymbol{w}^{(t)}\right\}_{t=1}^{m}$ from the corresponding task data sets in $\mathcal{T}$. By representing the classes with $\{C_1, C_2\} = \{0, 1\}$ and assuming they follow a Bernoulli distribution parameterized with our hypothesis, we can define the *likelihood* of all our data (for iid feature samples) through

$$p\left(\mathcal{T} \mid \mathcal{W}\right) = \prod_{t=1}^{m} \prod_{i=1}^{n_t} \mathrm{Ber}\left(y_i^{(t)} \mid h\left(\boldsymbol{x}_i^{(t)}; \boldsymbol{w}^{(t)}\right)\right) \tag{3}$$

where $\mathrm{Ber}\left(y \mid h\left(\boldsymbol{x}, \boldsymbol{w}\right)\right) = h\left(\boldsymbol{x}, \boldsymbol{w}\right)^y \left(1 - h\left(\boldsymbol{x}, \boldsymbol{w}\right)\right)^{1-y}$. We can further state the *posterior* distribution over the weights using Bayes rule

$$p\left(\mathcal{W} \mid \mathcal{T}\right) = \frac{p\left(\mathcal{T} \mid \mathcal{W}\right) p(\mathcal{W})}{p\left(\mathcal{T}\right)} \tag{4}$$

based on the likelihood from (3), a *prior* distribution $p(\mathcal{W})$ and the *evidence* $p(\mathcal{T})$. This posterior is the entry point for the Bayesian MTL framework, where we assume that tasks have a common statistical distribution. In particular, we capture common structure between related tasks in a shared prior $p(\mathcal{W})$. Statistics of $p(\mathcal{W})$ can be used afterwards for immediate decoding or improving decoders by combining shared knowledge with subject-specific data.

We model the shared prior with a general multivariate Gaussian density function $p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right)$ parameterized by a mean $\boldsymbol{\mu}_{\boldsymbol{w}} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\boldsymbol{w}} \in \mathbb{R}^{d \times d}$. Hence, assuming that the task weights are iid, the prior reads

$$p(\mathcal{W}) = \prod_{t=1}^{m} p\left(\boldsymbol{w}^{(t)}\right) = \prod_{t=1}^{m} \mathcal{N}\left(\boldsymbol{w}^{(t)} \mid \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right). \tag{5}$$

Plugging (3) and (5) into (4) results in a parameterized posterior model,

$$\begin{aligned} p\left(\mathcal{W} \mid \mathcal{T}\right) \propto & \prod_{t=1}^{m} \prod_{i=1}^{n_t} \mathrm{Ber}\left(y_i^{(t)} \mid h\left(\boldsymbol{x}_i^{(t)}; \boldsymbol{w}^{(t)}\right)\right) \\ & \prod_{t=1}^{m} \mathcal{N}\left(\boldsymbol{w}^{(t)} \mid \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right). \end{aligned} \tag{6}$$

Our goal is to maximize $p\left(\mathcal{W} \mid \mathcal{T}\right)$ w.r.t. the weights in $\mathcal{W}$ and the prior parameters $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$. Using (6) and applying the negative logarithm yields an equivalent objective for a loss minimization of the form

$$\begin{aligned} L(\mathcal{W}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}) = & -\sum_{t=1}^{m} \sum_{i=1}^{n_t} E_{ce}\left(\boldsymbol{w}^{(t)}; \boldsymbol{x}_i^{(t)}, y_i^{(t)}\right) \\ & + \frac{1}{2} \sum_{t=1}^{m} \Omega\left(\boldsymbol{w}^{(t)}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right) \end{aligned} \tag{7}$$

where E_{ce} is the point-wise *cross-entropy error* function

$$E_{ce}\left(\boldsymbol{w}; \boldsymbol{x}, y\right) = y \log h\left(\boldsymbol{x}; \boldsymbol{w}\right) + (1 - y) \log\left(1 - h\left(\boldsymbol{x}; \boldsymbol{w}\right)\right)$$

derived from the likelihood of the data and Ω is a *regularization* term

$$\Omega\left(\boldsymbol{w}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right) = \left(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}}\right)^T \Sigma_{\boldsymbol{w}}^{-1} \left(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}}\right) + \log|\Sigma_{\boldsymbol{w}}|$$

arising from the prior distribution. Notice that Ω reduces to the same regularizer as for the Gaussian prior on the linear model presented in [12], and so we can interpret minimization of (7) in the same way: Ω penalizes weights that deviate too far from the prior mean while the covariance scaling acts as an implicit feature selector. However, the squared error in the loss objective of the linear model switched with a more suitable error measurement for binary classification, namely the cross-entropy loss [14].

In order to train a shared prior across tasks, we want to minimize (7) w.r.t. the prior parameters $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\Sigma_{\boldsymbol{w}}$. It turns out that L is minimized by standard Gaussian sample statistics from the optimal weights in $\mathcal{W}$, i.e. the mean is estimated with the average over all task weights

$$\mathrm{mean}(\mathcal{W}) = \frac{1}{m} \sum_{t=1}^{m} \boldsymbol{w}^{(t)} \tag{8}$$

and the covariates with the sample covariance matrix or some numerically more stable version like

$$\mathrm{cov}(\mathcal{W};\boldsymbol{\mu}) = \frac{\sum_{t=1}^{m}\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu}\right)\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu}\right)^{T}}{\mathrm{Tr}\left(\sum_{t=1}^{m}\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu}\right)\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu}\right)^{T}\right)} + \varepsilon I, \tag{9}$$

where an appropriate $\varepsilon > 0$ ensures practicable condition numbers on the estimate. Unfortunately, the maximum a-posteriori (MAP) estimates for the optimal weights has to minimize the cross-entropy term in L that has no closed form solution. However, L is differentiable w.r.t. each individual weight $\boldsymbol{w}^{(t)} \in \mathcal{W}$ yielding the gradient

$$\begin{aligned}\nabla L\left(\boldsymbol{w}^{(t)};\boldsymbol{\mu}_{\boldsymbol{w}},\Sigma_{\boldsymbol{w}}\right) = &\sum_{i=1}^{n_t}\left(h\left(\boldsymbol{x}_i^{(t)};\boldsymbol{w}^{(t)}\right)-y_i^{(t)}\right)\boldsymbol{x}_i^{(t)} \\ &+\Sigma_{\boldsymbol{w}}^{-1}\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu}_{\boldsymbol{w}}\right).\end{aligned} \tag{10}$$

This vector can be used in gradient based optimization procedures [15], [16] to obtain optimal weight estimates given the prior parameters. In fact, the learning procedure is a special case of the algorithm outlined for the FD case in Fig. 1, but without weight decomposition (see the next section).

C. Spatio-spectral Feature Decomposition

Many BCIs use expert knowledge of the paradigm to determine relevant band power features in d frequency bands recorded with k electrodes, making up a subset of the full kd-dimensional feature space applicable to smaller data sets obtained from calibration sessions. Jayaram et al. [12] proposed FD as a spatio-spectral feature space for EEG that significantly reduces the feature dimensionality from kd to $k+d$. In particular, the authors assumed that the spectral feature importance is independent from the spatial topography of each electrode. We can state a FD model in the same way by using a bilinear hyperplane as the decision boundary for logstic regression

$$h(X;\boldsymbol{w},\boldsymbol{a}) = \left(1+\exp\left(\boldsymbol{a}^{T}X\boldsymbol{w}\right)\right)^{-1} \in]0,1[\tag{11}$$

where $X \in \mathbb{R}^{k\times d}$ is an input feature, $\boldsymbol{w} \in \mathbb{R}^{d}$ is the parameter vector weighting the spectral features and $\boldsymbol{a} \in \mathbb{R}^{k}$ is the parameter vector weighting the spatial features.

The MTL derivation is analogous to the previously presented non-FD case, except that we have decomposed the original weights into a spectral and a spatial part. We therefore incorporate two Gaussian priors $\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{a}},\Sigma_{\boldsymbol{a}}\right)$ and $\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{w}},\Sigma_{\boldsymbol{w}}\right)$ multiplicatively into the posterior in (6) with spectral task weights $\mathcal{W} = \{\boldsymbol{w}_t\}_{t=1}^{m} \subset \mathbb{R}^{d}$ and spatial task weights $\mathcal{A} = \{\boldsymbol{a}_t\}_{t=1}^{m} \subset \mathbb{R}^{k}$. The loss objective for the FD case then becomes

$$\begin{aligned}&L(\mathcal{W},\mathcal{A},\boldsymbol{\mu}_{\boldsymbol{w}},\Sigma_{\boldsymbol{w}},\boldsymbol{\mu}_{\boldsymbol{a}},\Sigma_{\boldsymbol{a}}) = \\ &-\sum_{t=1}^{m}\sum_{i=1}^{n_t}E_{ce}\left(\boldsymbol{w}^{(t)};X_i^{(t)},y_i^{(t)}\right) \\ &+\frac{1}{2}\sum_{t=1}^{m}\left(\Omega\left(\boldsymbol{w}^{(t)},\boldsymbol{\mu}_{\boldsymbol{w}},\Sigma_{\boldsymbol{w}}\right)+\Omega\left(\boldsymbol{a}^{(t)},\boldsymbol{\mu}_{\boldsymbol{a}},\Sigma_{\boldsymbol{a}}\right)\right)\end{aligned} \tag{12}$$

Algorithm 1: FD Multi-task Logistic Regression

Data: Training sets $\mathcal{T}$ from m related tasks
Result: $\boldsymbol{\mu}_{\boldsymbol{w}}$, $\Sigma_{\boldsymbol{w}}$, $\boldsymbol{\mu}_{\boldsymbol{a}}$, $\Sigma_{\boldsymbol{a}}$

```
Initialize μ_w = 0 and Σ_w = I ;
Initialize μ_a = (1/√k)1 and Σ_a = I ;
Arbitrary initialize W = {w^(t)}_{t=1}^t and A = {a^(t)}_{t=1}^m ;
while μ_w, Σ_w, μ_a and Σ_a not converged do
    for w^(t) ∈ W and a^(t) ∈ A do
        while w^(t) and a^(t) not converged do
            Choose some learning rate η ∈ ]0, ∞[;
            Set w^(t) = w^(t) − η∇L(w^(t); a^(t), μ_w, Σ_w);
            Set a^(t) = a^(t) − η∇L(a^(t); w^(t), μ_a, Σ_a);
    Update μ_w = mean(W) using (8);
    Update Σ_w = cov(W; μ_w) using (9);
```

Fig. 1. Gradient based MTL logistic regression algorithm based on simultaneous MAP estimates of the weights followed by prior updates to capture common structure in FD space throughout each task.

where E_{ce} and Ω are defined accordingly as in (7). Minimizing (12) can be done again using gradient based numerical optimization. The spectral task weight gradient reads

$$\begin{aligned}&\nabla L\left(\boldsymbol{w}^{(t)};\boldsymbol{a}^{(t)},\boldsymbol{\mu}_{\boldsymbol{w}},\Sigma_{\boldsymbol{w}}\right) = \\ &\sum_{i=1}^{n_t}\left(h\left(X_i^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{a}^{(t)}\right)-y_i^{(t)}\right){X_i^{(t)}}^{T}\boldsymbol{a}^{(t)} \\ &+\Sigma_{\boldsymbol{w}}^{-1}\left(\boldsymbol{w}^{(t)}-\boldsymbol{\mu}_{\boldsymbol{w}}\right)\end{aligned} \tag{13}$$

and the spatial one is similarly given by

$$\begin{aligned}&\nabla L\left(\boldsymbol{a}^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{\mu}_{\boldsymbol{a}},\Sigma_{\boldsymbol{a}}\right) = \\ &\sum_{i=1}^{n_t}\left(h\left(X_i^{(t)};\boldsymbol{w}^{(t)},\boldsymbol{a}^{(t)}\right)-y_i^{(t)}\right)X_i^{(t)}\boldsymbol{w}^{(t)} \\ &+\Sigma_{\boldsymbol{a}}^{-1}\left(\boldsymbol{a}^{(t)}-\boldsymbol{\mu}_{\boldsymbol{a}}\right).\end{aligned} \tag{14}$$

In order to break the circular dependencies between spatial and spectral gradient, we have to alternatingly fix one set of weights to compute the MAP estimate of the others. A gradient descent based procedure to learn FD priors is depicted in Fig. 1.

D. Subject-specific Adaptation

Once we have trained the prior parameters using MTL we can immediately use the mean weights for prediction. When given a new feature $\boldsymbol{x}$ or X, we only need to compute $h\left(\boldsymbol{x};\boldsymbol{\mu}_{\boldsymbol{w}}\right)$ or $h(X;\boldsymbol{\mu}_{\boldsymbol{w}},\boldsymbol{\mu}_{\boldsymbol{a}})$, respectively, for an out-of-the-box prediction of the class probability. When decoding brain states for a new subject, we are faced with a new task that has subject-specific variations. The Bayesian framework naturally copes with this case; we have to just compute the MAP estimate with the trained prior of the adapted weights based on the new data set. However, we do not know how much

belief we should put into the prior to optimally trade-off between task-specific variations and shared task knowledge. This concept can be captured formally by introducing an additional regularization factor $\lambda \in [0, \infty]$ for subject-specific adaptation. Given we denote our new task data with $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^d \times \{C_1, C_2\}$ we obtain adapted weights in the non-FD case by minimizing

$$L_\lambda(\boldsymbol{w}) = -\sum_{i=1}^{n_t} E_{ce}\left(\boldsymbol{w}; \boldsymbol{x}_i, y_i\right) + \frac{\lambda}{2}\Omega\left(\boldsymbol{w}, \boldsymbol{\mu}_{\boldsymbol{w}}, \Sigma_{\boldsymbol{w}}\right) \quad (15)$$

w.r.t. to the weights $\boldsymbol{w}$. The gradient for numerical optimization is analytically given by

$$\nabla L_\lambda\left(\boldsymbol{w}\right) = \sum_{i=1}^{n_t} \left(h\left(\boldsymbol{x}_i; \boldsymbol{w}\right) - y_i\right)\boldsymbol{x}_i + \lambda \Sigma_{\boldsymbol{w}}^{-1}\left(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}}\right). \quad (16)$$

The FD case derives in the exact same way, but shares the regularizer for the spatial and spectral weights. An optimal regularization factor λ may be obtained using model selection techniques (e.g. by cross-validation)

E. Experimental Setup

We evaluated the model on real EEG signals recorded from ten healthy subjects (two female, eight male, 22-28 years old, nine subjects were naïve to BCIs and one participated twice in BCI experiments) using a two-class motor imagery paradigm of left or right hand movements.

Each subject sat in a comfortable chair in front of a screen and performed 300 trials in the experiment (150 per condition, stimuli were presented in pseudorandom order and no feedback on the performance was provided). Each trial consisted of an initial pause of three seconds, followed by an imagery phase lasting seven seconds in which a centrally displayed arrow pointing to the left or right informed the subject to perform haptic left or right hand motor imagery, respectively.

Brain activity during the experiment was recorded using EEG with 128 electrodes positioned according to the extended 10-20 system (referenced at Cz). The signals were sampled at 500Hz using BrainAmp amplifiers[1] and a temporal analog high-pass filter with 10 seconds time constant.

After the experiment was conducted, data preprocessing solely consisted of spatially filtering the signals with a surface Laplace [17] to keep the results unbiased for evaluation. FD feature matrices were generated by applying the discrete Fourier transform with a Hann window to the motor imagery phase of each trial in order to extract equidistant log-band power features of 2Hz width within the frequency range from 7Hz to 31Hz from all electrodes. Hence, the FD space was spanned by 128×12-dimensional features.

III. Experimental Results

A. Classification Performance

The performance of MTL logistic regression on the real-world BCI paradigm was evaluated by comparing classification accuracies when different amount of subject-specific data is available. In particular, one out of the ten subject data sets was taken out to be regarded as subject-specific calibration data. Three models where used for comparison: FD MTL Logistic Regression with Gaussian prior trained using the algorithm shown in Fig. 1, standard FD Logistic Regression with L2 regularization (i.e. uninformed prior) and finally FD MTL Linear Regression with a Gaussian prior and maximum-likelihood estimates for the variance hyperparameter [12]. Two out of the ten subjects performing near chance were taken out from prior training (i.e. priors were finally trained from seven task data sets).

After obtaining priors for the models, the 300 samples from the subject-specific data were randomly divided into a distinct training set (200 samples) and test set (100 samples). Each model was successively trained on an increasing subset of the training set using a step size of 50 trials and 5-fold cross-validation from $\{\exp(-10), \exp(-9), \ldots, \exp(9), \exp(10)\}$ for hyperparameter selection of the regularization value. Evaluation of the trained models occured on the test set to compute their accuracy. The whole procedure was conducted for each subject using 100 runs with random splits into training and test set. The mean accuracy development over all subjects and runs is shown in Fig. 2. Further, the development of the model deviation from the prior (cross-validated mean regularization factor) over the runs is visualized in Fig. 3.

The results reveal a decreasing performance gap between the models with increasing amount of calibration data to train on. However, MTL logistic regression with prior information outperforms the MTL model from the previous framework as well as standard logistic regression with uniformed prior and manages to reach the 70% mark (considered as the minimum requirement for reliable communication in BCIs) with much less training data. Development of the regularization factor shows a decreasing trend over an increasing amount of calibration data. This indicates that the decoder is deviating from the prior in order to learn more task-specific structure, which is in fact a plausible statistical behavior; we expect that with more data from a problem the underlying structures emerges stronger which can be captured by the model to improve on subject-specific variations.

B. Spatial and Spectral Prior

Trained models using FD features have weights associated to the topography and frequency distribution that indicate relevance of individual dimensions for prediction. In order to compare findings of the MTL algorithm with domain knowledge of the SMR paradigm, Gaussian prior parameters for the FD space were trained with the algorithm in Fig. 1 from eight subjects (leaving out two near chance performers). A visualization of the resulting priors is shown in Fig. 4.

The trained prior identifies spatial relevance on electrodes placed above the left and right sensorimotor cortex. Those features agree with domain knowledge of the neurophysiological characteristics for this paradigm and indicate that indeed neural activity is used to predict the corresponding brain

[1] BrainProducts GmbH, Gilching, Germany

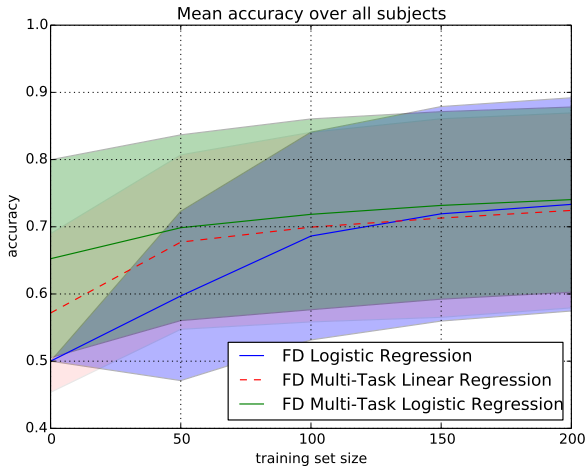

Fig. 2. Mean accuracy with shaded standard deviation on the test set over an increasing amount of subject-specific calibration data provided for training. The mean was taken over all ten subjects using 100 runs with random splits of 200 training and 100 test samples. All models improve with increasing amount of calibration data and eventually reach comparable performance levels. The accuracy development suggests that MTL logistic regression outperforms MTL linear regression and standard logistic regression and both MTL models outperform the uninformed model if less data is provided.

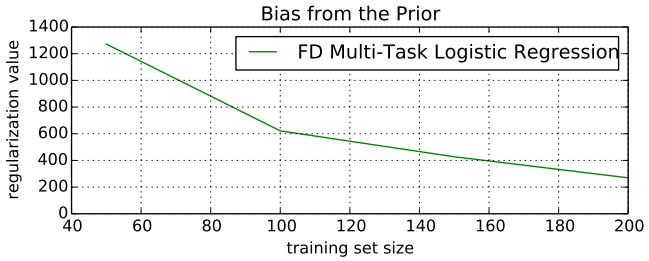

Fig. 3. Mean regularization value of MTL logistic regression (gathered through 5-fold cross-validation) for subject adaptation over all subjects and 100 runs (see Fig. 2). Initial high regularization towards the prior for few calibration trials demonstrates that the model is relying more strongly on the prior than the data. As more subject-specific data is added to the training set, regularization drops and incorporates more exposed structure specific to the new subject.

condition. Furthermore, the model puts highest priority on the frequency bin for 11-13Hz, which agrees with the μ rhythm band-power modulation characteristics of the paradigm, too. Implicit feature selection for subject adaptation is likewise consistent with high covariates between the α-rhythm (9-13Hz) and β-rhythm (19-23Hz and 27-29Hz).

C. Null Hypothesis Pairwise Permutation Test

In Section III-A we compared the mean accuracies over each subject with increasing amount of calibration data in 100 runs. Here, we perform a statistical test to examine if there is a significant difference in performance of the three tested models. In particular, a pairwise permutation test [18] between two models was conducted under the null hypothesis that their true mean performance is equal: The mean accuracy over the 100

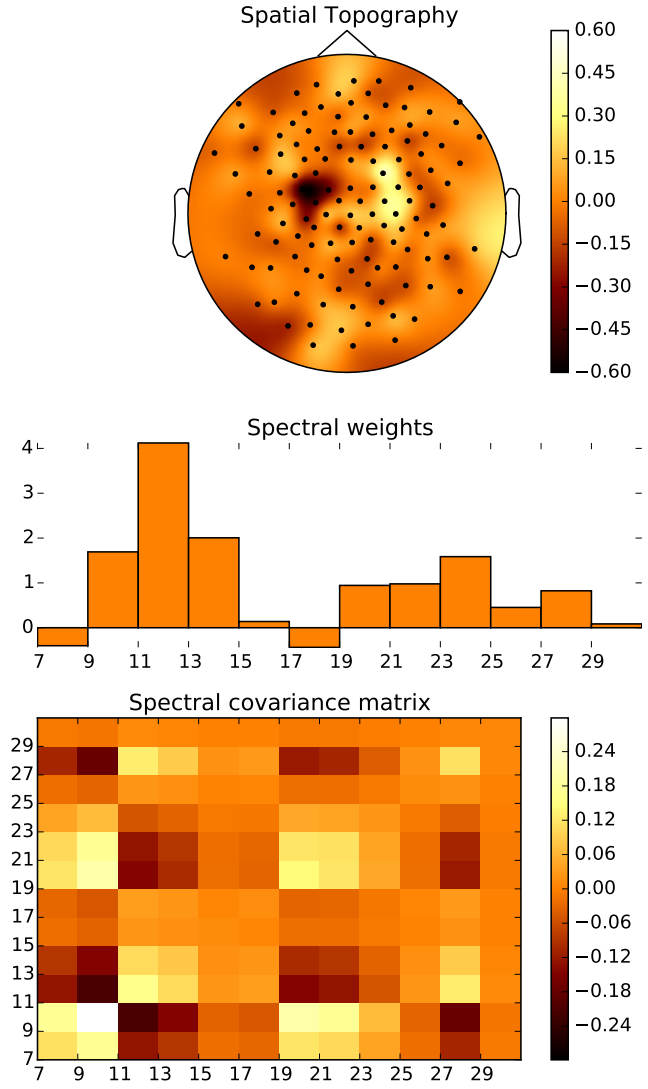

Fig. 4. The first plot shows the topography of the spatial weight prior trained by MTL logistic regression over eight subjects. The prior indicates relevant activity above the left and right sensorimotor cortex on electrodes C3 and C4. The second plot shows a bar chart of the trained spectral weight prior. They show high relevance on the frequency bin for 11-13Hz and its surrounding bins corresponding to the μ-rhythm, as well as moderate relevance within the β-range in 19-29Hz. The final third plot visualizes the spectral covariance prior trained by the algorithm. High positive covariates for spectral weights can be mainly found within the α- and β-frequencies while negative covariates show up between those bands.

runs of each of the ten subjects and three models were taken for one calibration set size, denoted by $\mathcal{P}_a = \{a_1, a_2, \ldots, a_{10}\}$ for the performance samples from FD MTL logistic regression, $\mathcal{P}_b = \{b_1, b_2, \ldots, b_{10}\}$ for the samples from FD mutli-task linear regression and $\mathcal{P}_c = \{c_1, c_2, \ldots, c_{10}\}$ for standard L2 FD logistic regression. Further, let μ_a, μ_b and μ_c denote the true mean performance from which $\mathcal{P}_a$, $\mathcal{P}_b$ and $\mathcal{P}_c$ were drawn, respectively. We tested two null hypotheses, the first was $H_0^* : \boldsymbol{\mu}_a = \boldsymbol{\mu}_b$ and the second $H_0^{**} : \boldsymbol{\mu}_a = \boldsymbol{\mu}_c$, i.e. MTL logistic regression is compared against both other models.

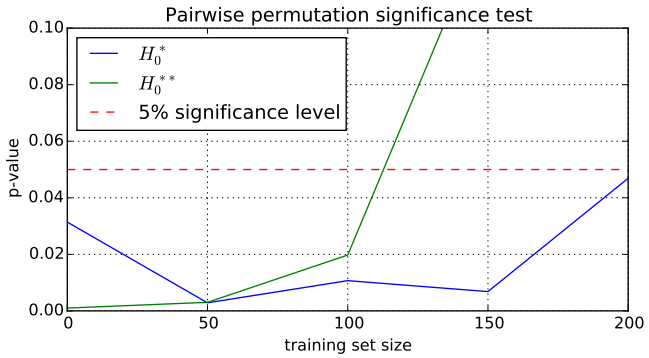

Fig. 5. p-value over an increasing amount of subject-specific calibration data using a pairwise permutation test for the null hypotheses H_0^* (MTL logistic and linear regression have same performance) and H_0^{**} (MTL logistic and standard logistic regression have same performance). At 5% significance level there is strong evidence against H_0^* on all amounts of trials (including out-of-the-box performance) and against H_0^{**} for up to 100 trials where we can reject the hypotheses.

Using the test statistic $T(\mathcal{P}_x, \mathcal{P}_y) = \text{mean}(\mathcal{P}_x) - \text{mean}(\mathcal{P}_y)$ where x and y are substitutes for a, b or c the p-value was computed by

$$p = \frac{1}{n} \sum_{i=1}^{n} \left[T\left(\mathcal{P}_x^{(i)}, \mathcal{P}_y^{(i)}\right) \geq T\left(\mathcal{P}_x, \mathcal{P}_y\right) \right]$$

where $\mathcal{P}_x^{(i)}$ and $\mathcal{P}_y^{(i)}$ are pseudorandomly generated pairwise permutations of $\mathcal{P}_x$ and $\mathcal{P}_y$. This means that each pair of samples (x_t, y_t) for the same subject t appears in $\mathcal{P}_x^{(i)}$ and $\mathcal{P}_y^{(i)}$ again, but with a chance of 50% that the positions have switched to (y_t, x_t). The results for the p-value using $n = 10^6$ permutations over different amounts of calibration data are shown in Fig. 5.

The results show that H_0^* is rejected at a 5% significance level on all up to 200 calibration data points. Hence, together with the classification results from Fig. 4, we can indeed observe a statistically significant improvement of the new MTL model over the MTL model from the previous framework. Against uninformed logistic regression we can reject H_0^{**} for moderate calibration sessions (up to 100 trials) only. However, it is notable that for short calibration phases (about 50 trials) we can reject both hypotheses even at a 1% significance level, indicating that the new model is able to make better use of prior knowledge.

IV. Discussion

This work extended a general framework from previous work by a logistic regression model with more suitable assumptions on the distribution of the dependent variable in case of binary classification. We demonstrated a significant improvement in classification accuracy of the new model over comparable models for calibration-free decoding and subject-specific adaptation ability. The new model was able to learn spatially important locations on the scalp as well es relevant spectral frequency bands, both consistent with expert knowledge of the paradigm. Further approaches by using different prior structures or loss functions, MTL derivations for other effective models used throughout BCI research as well as hybrid techniques of MTL together with advanced spatial filters are to be investigated and may further improve performance.

References

[1] M. Grosse-Wentrup and B. Schölkopf, *A Review of Performance Variations in SMR-Based Brain–Computer Interfaces (BCIs)*, ser. Springer-Briefs in Electrical and Computer Engineering. Springer, 2013, ch. 4, pp. 39–51.

[2] T. Dickhaus, C. Sannelli, K.-R. Müller, G. Curio, and B. Blankertz, "Predicting BCI performance to study BCI illiteracy," in *BMC Neuroscience 2009*, vol. 10, 2009, p. (Suppl 1):P84.

[3] U. Hoffmann, J. M. Vesin, T. Ebrahimi, and K. Diserens, "An efficient P300-based brain-computer interface for disabled subjects," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 115–125, 2008.

[4] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 4, pp. 317–326, 2008.

[5] M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones, "Brain-computer interfaces based on the steady-state visual-evoked response," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 211–214, 2000.

[6] T. Hinterberger, S. Schmidt, N. Neumann, J. Mellinger, B. Blankertz, G. Curio, and N. Birbaumer, "Brain-computer communication and slow cortical potentials," *IEEE Trans Biomed Eng*, vol. 51, no. 6, pp. 1011–1018, 2004.

[7] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multi-subject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, vol. 2011, no. 217987, pp. 1–9, Aug. 2011.

[8] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[9] W. Samek, F. C. Meinecke, and K. R. Müller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[10] S. Fazli, S. Dähne, W. Samek, F. Bießmann, and K. R. Müller, "Learning from more than one data source: Data fusion techniques for sensorimotor rhythm-based brain–computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 891–906, June 2015.

[11] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *JMLR Workshop and Conference Proceedings Volume 9: AISTATS 2010*, Max-Planck-Gesellschaft. Cambridge, MA, USA: JMLR, May 2010, pp. 17–24.

[12] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[13] V. Jayaram and M. Grosse-Wentrup, "A transfer learning approach for adaptive classification in P300 paradigms," in *Proceedings of the Sixth International BCI Meeting*, 2016.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[15] J. Nocedal and S. J. Wright, *Numerical Optimization*, ser. Springer Series in Operations Research and Financial Engineering. Berlin: Springer, 2006.

[16] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of research of the National Bureau of Standards*, vol. 49, pp. 409–436, 1952.

[17] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 3, pp. 386 – 394, 1997.

[18] W. J. Welch, "Construction of permutation tests," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 693–698, 1990.