Multitask Learning for Brain-Computer Interfaces

Morteza Alamgir **Moritz Grosse-Wentrup** **Yasemin Altun**
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany
{morteza, yasemin.altun}@tuebingen.mpg.de, moritzgw@ieee.org

Abstract

Brain-computer interfaces (BCIs) are limited in their applicability in everyday settings by the current necessity to record subject-specific calibration data prior to actual use of the BCI for communication. In this paper, we utilize the framework of multitask learning to construct a BCI that can be used without any subject-specific calibration process. We discuss how this out-of-the-box BCI can be further improved in a computationally efficient manner as subject-specific data becomes available. The feasibility of the approach is demonstrated on two sets of experimental EEG data recorded during a standard two-class motor imagery paradigm from a total of 19 healthy subjects. Specifically, we show that satisfactory classification results can be achieved with zero training data, and combining prior recordings with subject-specific calibration data substantially outperforms using subject-specific data only. Our results further show that transfer between recordings under slightly different experimental setups is feasible.

1 Introduction

In recent years, machine learning methods have been applied with great success to non-invasive brain-computer interfaces (BCIs), replacing the need for intensive subject training (Wolpaw et al., 1991, Wolpaw and McFarland, 2004, Birbaumer et al., 1999) by a comparatively brief calibration time (Blankertz et al., 2007). In spite of this progress, setting up a non-invasive BCI experiment still involves a time-consuming process. Besides the arduous electrode placement procedure, the main obstacle for out-of-the-box usable BCIs is the current necessity to acquire subject-specific training data for learning suitable spatial filters and classifiers.

While it is well known that inter-subject variability of informative spatial- and spectral features is substantial, the principle feature characteristics remain invariant across subjects. For example, haptic motor imagery, currently the most popular paradigm in research on non-invasive BCIs (Mason et al., 2007), typically induces a decrease in power, termed event-related desynchronization (ERD), of the μ- (roughly $10-14$ Hz) and β-rhythms (roughly $20-30$ Hz) over contralateral sensorimotor areas (Pfurtscheller and Neuper, 1997). Accordingly, one approach to reducing calibration time is to incorporate this prior knowledge into the learning process. This can either be done manually, e.g., by designing spatial filters that explicitly focus on sensorimotor areas (Grosse-Wentrup et al., 2008), or automatically by using previously recorded data in order to learn feature characteristics that are consistent across subjects.

In terms of the latter approach, the problem of session-to-session transfer, i.e., repeated BCI experiments with the same subject, has been addressed in (Krauledat et al., 2008). In that work, the method of Common Spatial Patterns (CSP) (Koles, 1991, Ramoser et al., 2000) is used to first compute session-specific spatial filters and classifiers. Then, a clustering procedure is employed to select prototypical spatial filters and classifiers, which are in turn applied to newly recorded data. Using this approach, the authors demonstrate that calibration time can be greatly reduced with only a slight loss in classification accuracy. While (Krauledat et al., 2008) only deals with session-to-session transfer, the problem of inter-subject transfer is addressed in (Fazli et al., 2009). Also building upon CSP for spatial filtering, the authors utilize a large database of pairs of spatial filters and classifiers from 45 subjects

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

to learn a sparse subset of these pairs that are predictive across subjects. Using a leave-one-subject-out cross-validation procedure, the authors then demonstrate that the sparse subset of spatial filters and classifiers can be applied to new subjects with only a moderate performance loss in comparison to subject-specific calibration.

While these are substantial advances, a framework capable of utilizing prior information while at the same time being able to adjust to subject-specific variations has not been presented yet in the context of non-invasive BCIs. Indeed, it would be desirable to construct a learning procedure which can be used out-of-the-box to instantaneously provide subjects with feedback at the beginning of the experiment, yet that can also be further improved as more subject-specific data becomes available. Furthermore, previous work has only shown that calibration time can be reduced with only a moderate *loss* in classification accuracy. However, combining prior information from other subjects with subject-specific data should ultimately results in an *increase* in classification accuracy in comparison to using subject-specific data only.

In this work, we address these questions by applying the framework of multitask learning (Evgeniou et al., 2005, Yu et al., 2005) to the domain of non-invasive BCIs. Multi-task learning methods, a subfield of machine learning, investigate the challenge of combining information from several related tasks in order to overcome the data scarcity problem. The goal is to discover important shared characteristics of the related task predictors via shared regularization (Evgeniou et al., 2005) or via shared prior (Yu et al., 2005) on the predictor functions. Although BCI has been an active research field, there has been very few real life problems that has shown the success of multi-task learning (Daumé III, 2007).

In this paper, we provide a successful real world application for multi-task learning, namely BCI problems. We treat each subject in a BCI experiment as one task and employ a parametric probabilistic approach that uses shared priors as proposed in (Yu et al., 2005). In an iterative manner, our model infers the model parameters and the shared prior parameters. We define an out-of-the-box BCI with respect to these shared prior parameters, and adapt the BCI to novel subjects by inferring the subject parameters with respect to the shared priors efficiently in an online fashion. We evaluate this system on experimental EEG data recorded from 19 subjects during a motor imagery paradigm. The experiments show that our system does indeed outperform the baseline model that trains a predictor for individual subjects with or without feature selection.

Multi-task learning is a well-studied framework in machine learning and has been applied to many fields ranging from Natural Language Processing to Computational Biology. According to our knowledge, the application of multitask learning, which is the main goal of this paper, is novel. This application provides a natural and principled approach to adapt a subject-independent BCI to individual novel subjects. The implicit feature selection property of our approach improves over feature selection approaches based on individual subjects as it enables inter- as well as intra-subject information transfer and removes the individual subject noise. The online adaptation can be performed very efficiently and removes the necessity to keep the whole dataset for future updates.

The remainder of this paper is structured as follows. In Section 2, we introduce the multitask learning framework and describe its application in the context of BCIs. In Section 3, we review some of existing approaches in multitask learning, and in Section 4 experimental results are shown, demonstrating the utility of the multitask framework for non-invasive BCIs. The paper concludes with a discussion of the results in Section 5.

2 Multitask Learning

2.1 Training off-line tasks

In multitask learning, we are interested in K related inference tasks given training data of i.i.d input vectors $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^n$ with corresponding outputs $\mathbf{y}_t = \{y_i^t\}_{i=1}^n$ for each task t, where $\mathbf{x}_i^t \in \mathcal{R}^d$ and $y_i \in \mathcal{R}$ [1]. Our goal is to infer K linear functions $f_t(\mathbf{x}; \mathbf{w}_t) = \langle \mathbf{w}_t, \mathbf{x} \rangle$ associated to each task such that $y_i^t = f_t(\mathbf{x}_i^t; \mathbf{w}_t) + \epsilon_t$, where ϵ_t is an additive noise normally distributed with zero mean and σ^2 variance. In this model, the conditional distribution of the output for task t given the input and the weights is given by

$$p(y|\mathbf{x}, \mathbf{w}_t, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}(\langle \mathbf{w}_t, \mathbf{x} \rangle - y)^2\right).$$

The Bayesian framework characterizes the uncertainty in parameters $\mathbf{W} = \{\mathbf{w}_t\}_{t=1}^K$ by defining a probability distribution $p(\mathbf{W})$. We specify the prior as the Gaussian distribution, where each $\mathbf{w}_t$ is normally distributed with $\boldsymbol{\mu}$ mean and $\boldsymbol{\Sigma}$ covariance. The posterior distribution for $\mathbf{W}$ is then given by the Bayes rule as

$$p(\mathbf{W}; D, \sigma^2) \propto \prod_t p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{w}_t) p(\mathbf{w}_t),$$

[1] More formally, for a fixed task, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and $\mathbf{y} = [y_1, \dots, y_n]^T$

where $D = \{d_1, \ldots, d_K\}$ with $d_t = \{\mathbf{X}_t, \mathbf{y}_t\}$.

In standard prediction problems, the prior is defined as a Gaussian distribution with $\mathbf{0}$ mean and unit covariance, $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Using this prior in the multitask setting corresponds to training each task independently. Our goal is to infer $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from all the tasks jointly with $\mathbf{W}$ in order to infer the common structure shared across all tasks. This can be achieved by maximizing the posterior or equivalently minimizing the negative log-posterior

$$\begin{aligned} L(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}; D, \sigma^2) &= \frac{1}{\sigma^2} \sum_t \|\mathbf{X}_t \mathbf{w}_t - \mathbf{y}_t\|^2 \\ &+ \frac{1}{2} \sum_t (\mathbf{w}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w}_t - \boldsymbol{\mu}) \\ &+ \frac{K}{2} \log \det(\boldsymbol{\Sigma}). \end{aligned} \tag{1}$$

By minimizing negative log-posterior we penalize elements of $\widehat{\mathbf{w}_t} = \mathbf{w}_t - \boldsymbol{\mu}$ by $\boldsymbol{\Sigma}^{-1}$. To see how exactly this penalization works, we expand one of them:

$$\widehat{\mathbf{w}_t}^T \boldsymbol{\Sigma}^{-1} \widehat{\mathbf{w}_t} = \sum_i \sum_j \boldsymbol{\Sigma}^{-1}_{i,j} \widehat{\mathbf{w}_{t,i}} \widehat{\mathbf{w}_{t,j}}. \tag{2}$$

This means that $\boldsymbol{\Sigma}^{-1}_{i,j}$ penalizes the relation between dimension i and dimension j, so $\boldsymbol{\Sigma}^{-1}$ works as an implicit feature selector.

We minimize (1) with respect to $\mathbf{W}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iteratively by holding $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{W}$ constant, respectively. For fixed $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we get the $\mathbf{w}_t$ updates by taking the derivative with respect to $\mathbf{w}_t$ for all t and equating to 0,

$$\mathbf{w}_t = \left(\frac{1}{\sigma^2} \mathbf{X}_t^T \mathbf{X}_t + \boldsymbol{\Sigma}^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} \mathbf{X}_t^T \mathbf{y}_t + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right). \tag{3}$$

In order to avoid inverting $\boldsymbol{\Sigma}$, which is a $O(d^3)$ operation, we perform the equivalent update

$$\mathbf{w}_t = \left(\frac{1}{\sigma^2} \boldsymbol{\Sigma} \mathbf{X}_t^T \mathbf{X}_t + \mathbf{I} \right)^{-1} \left(\frac{1}{\sigma^2} \boldsymbol{\Sigma} \mathbf{X}_t^T \mathbf{y}_t + \boldsymbol{\mu} \right). \tag{4}$$

For fixed $\mathbf{W}$, the derivation of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ update is more involved.

$$\begin{aligned} dL(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{2} \operatorname{tr} \Big[-\boldsymbol{\Sigma}^{-1} (d\boldsymbol{\Sigma}) \boldsymbol{\Sigma}^{-1} \\ & \quad \sum_t (\mathbf{w}_t - \boldsymbol{\mu})(\mathbf{w}_t - \boldsymbol{\mu})^T \\ &- \quad 2 \boldsymbol{\Sigma}^{-1} \sum_t (\mathbf{w}_t - \boldsymbol{\mu})(d\boldsymbol{\mu})^T \Big] \\ &+ \frac{K}{2} \operatorname{tr} \left[\boldsymbol{\Sigma}^{-1} (d\boldsymbol{\Sigma}) \right] \end{aligned}$$

This is equivalent to 0 when the terms multiplying $d\boldsymbol{\mu}$ and $d\boldsymbol{\Sigma}$ are 0. This gives the standard mean update for $\boldsymbol{\mu}$,

$$\boldsymbol{\mu} = \frac{1}{K} \sum_t \mathbf{w}_t. \tag{5}$$

Combining terms involving $d\boldsymbol{\Sigma}$, we get

$$\tfrac{1}{2} \operatorname{tr} \left(\boldsymbol{\Sigma}^{-1} (d\boldsymbol{\Sigma}) \left[K\mathbf{I} - \boldsymbol{\Sigma}^{-1} \textstyle\sum_t (\mathbf{w}_t - \boldsymbol{\mu})(\mathbf{w}_t - \boldsymbol{\mu})^T \right] \right).$$

This yields the $\boldsymbol{\Sigma}$ update given by

$$\boldsymbol{\Sigma} = \frac{1}{K} \sum_t (\mathbf{w}_t - \boldsymbol{\mu})(\mathbf{w}_t - \boldsymbol{\mu})^T. \tag{6}$$

Experimentally, we observe that $\boldsymbol{\Sigma}$ update is more stable when

$$\boldsymbol{\Sigma} = \frac{\sum_t (\mathbf{w}_t - \boldsymbol{\mu})(\mathbf{w}_t - \boldsymbol{\mu})^T}{\operatorname{tr} \left(\sum_t (\mathbf{w}_t - \boldsymbol{\mu})(\mathbf{w}_t - \boldsymbol{\mu})^T \right)} + \epsilon \mathbf{I}. \tag{7}$$

Note the changes are in the additional ϵI term which ensure $\boldsymbol{\Sigma}$ be invertible and in the scaling of $\boldsymbol{\Sigma}$ with the trace of the original update. While this scaling does not change the optimization problem overall, it ensures that $\boldsymbol{\Sigma}$ will have trace 1. This can be advantageous as it renders the method insensitive to hyper-parameters, such as σ^2.

Finally, σ^2 can be obtained by cross-validation. The learning procedure is outlined in Algorithm 1.

Algorithm 1 Multi-task optimization

```
Input: $D$, $\sigma^2$
Set $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \mathbf{I})$.
repeat
    Update $\mathbf{w}_t$ for all $t = 1, \ldots, k$ using (4)
    Update $\boldsymbol{\mu}$ using (5)
    Update $\boldsymbol{\Sigma}$ using (7)
until convergence
Output: $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
```

2.2 Novel task adaptation

In Section 2.1, we outlined a simple yet effective multitask learning approach. We now discuss an important issue, namely how to use this method when data from a novel task is available. Let $d = (\mathbf{X}, \mathbf{y})$ be the dataset from a task. Adaptation to this task using the system trained from the previous K tasks and the new data d can be achieved easily by

$$\mathbf{w} = \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} + \boldsymbol{\Sigma}^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{y} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right), \tag{8}$$

where $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the shared priors obtained from the previous K tasks via Algorithm 1. An important criteria for this adaptation procedure is efficiency in order to avoid causing delay between trials.

In (8), the computation involving the data of the novel task consists of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$. In some cases, the data for the novel task is not provided in batch, but rather is streamlined. For this online setting, we denote the cumulative data at time i by $(\mathbf{X}_i, \mathbf{y}_i)$ and the data point for the data at time $i+1$ by (x_{i+1}, y_{i+1}). Then the covariance term of the data at the $i+1$th time point is given by

$$\mathbf{X}_{i+1}^T\mathbf{X}_{i+1} = \mathbf{X}_i^T\mathbf{X}_i + \mathbf{x}_{i+1}\mathbf{x}_{i+1}^T.$$

Computing the variance in this iterative manner reduces the adaptation to the new task significantly. Furthermore, it removes the necessity to store the data gathered at previous time points. Similarly, the mean term can be computed in an iterative manner by

$$\mathbf{X}_{i+1}^T\mathbf{y}_{i+1} = X_i^T\mathbf{y}_i + y_{i+1}\mathbf{x}_{i+1}.$$

Once these two terms are updated, the discriminator at time $i+1$ can be obtained efficiently by solving a system of d linear equation, where the number of variables is the number of dimensions, d.

2.3 Multitask learning for BCIs

In the setting of BCIs, a task corresponds to an individual subject. Accordingly, $\mathbf{x}_i^t$ refers to the features derived from the recorded brain signals of subject t during trial i, and y_i^t denotes the subject's corresponding intention. The parameters $\mathbf{w}_t$ constitute the weights assigned to the individual features used to predict the subject t's intention, and the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ encode the information on relevant feature characteristics shared across subjects. As such, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, learned from previously recorded subject data, define an out-of-the-box BCI that can be used to classify data recorded from a novel subject without any subject-specific calibration process. As subject-specific data from a novel subject becomes available, this data can then be used to adapt the out-of-the-box BCI to subject-specific variations as described in Section 2.2.

3 Related Work

There is a large number of methods proposed in the literature that investigate multitask learning.

In (Obozinski et al., 2006), a regularization is proposed based on sparsity assumptions: only a subset of variables is relevant for prediction. Moreover and more importantly, the sparsity is assumed to be shared across tasks: the same variables are relevant to all the tasks. In order to enforce this assumption, the authors propose the use of the regularizer $\|\mathbf{W}\|_{2,1}$, consisting of building a vector with the ℓ_2-norm of each row in $\mathbf{W}$, and then the ℓ_1-norm of that vector.

A generalization of this idea is proposed in (Argyriou et al., 2008). Again, sparsity is assumed, so that the optimal predictors for every task belong to the same subspace. Thus, the same regularizer $\|\mathbf{W}\|_{2,1}$ is assumed to be still valid after a full rank transformation performed on the data. The cost function is given by $\sum_t \sum_i L(\mathbf{w}_t^T(\boldsymbol{\Sigma}^{-1})^T\mathbf{x}_i, y_i) + \lambda\|\mathbf{W}\|_{2,1}^2$. Because the joint optimization over $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{W}$ is a non-convex problem, an alternative and equivalent formulation is proposed. This is very similar to the approach presented in this paper but the optimization yields different updates.

In the context of semi-supervised learning, it has been proposed in (Ando and Zhang, T., 2005) to decompose the predictors into a task-specific part and part given by a shared low-dimensional representation across tasks. Thus, if we refer to the projection on this low dimensional subspace as $\boldsymbol{\Theta}$, each predictor is defined as $\mathbf{w}_t = \mathbf{u}_t + \boldsymbol{\Theta}^T\,\mathbf{v}_t$ where $\mathbf{u}_t$ and $\mathbf{v}_t$ are vectors in the feature space and the lower dimensional space, respectively. Only the task-specific component is regularized, $\|\,\mathbf{u}_t\,\|^2 = \|\,\mathbf{w}_t - \boldsymbol{\Theta}^T\,\mathbf{v}_t\,\|^2$. The joint optimization over $\{\mathbf{u}_t\}$, $\{\mathbf{v}_t\}$ and $\boldsymbol{\Theta}$ is again non-convex, therefore an iterative algorithm for alternatively obtaining $\boldsymbol{\Theta}$ and $\{\mathbf{u}_t, \mathbf{v}_t\}$ is proposed. Eventually, both the works proposed in (Ando and Zhang, T., 2005) and (Argyriou et al., 2008) can be interpreted as a kind of principal component analysis on the set of predictors obtained for the different tasks.

Finally, (Yu et al., 2005) introduced the multitask learning framework outlined here. The main difference is that whereas (Yu et al., 2005) take a hierarchical Bayesian approach, we perform a maximum likelihood optimization where the hyperparameters are selected by cross validation.

4 Experimental Results

4.1 Experimental paradigm & data

To experimentally evaluate the utility of the proposed multitask framework for BCIs, we recorded EEG data during a standard motor-imagery paradigm. Specifically, subjects were placed in front of a screen with a centrally displayed fixation cross. Each trial started with a pause of $3s$. A centrally displayed arrow then instructed subjects to initiate haptic motor imagery of either the left or right hand, as indicated by the arrow's direction. After further seven seconds the arrow was removed from the screen, marking the end of the trial and instructing subjects to cease motor imagery.

Two data sets were recorded in different laboratories. For data set A, ten healthy subjects participated in the study (one subject with previous BCI experience). EEG data was recorded from 128 channels, placed according to the extended 10-20 system, with electrode Cz as reference and sampled at 500 Hz. BrainAmp amplifiers (BrainProducts, Munich) with a temporal analog high-pass filter with a time constant of 10 s were used for this purpose. A total of 150 trials per class (left/right hand motor imagery) and subject were recorded, with no feedback provided to the subjects during the experiment.

For dataset B the same experimental paradigm and recording procedure was employed. However, a different amplifier (QuickAmp, BrainProducts, Munich) and a different EEG cap with only 124 channels was employed. Furthermore, electrode Fz was used as reference. Accordingly, electrode locations and setup slightly differed between datasets A and B. Nine healthy subjects with no prior BCI experience participated in the study for dataset B, and 45 trials per class and subject were recorded without any feedback provided to the subjects.

4.2 Feature computation

For feature extraction, recorded EEG data was first spatially filtered using a surface Laplacian setup (McFarland et al., 1997). We did not employ more sophisticated methods for spatial filtering, such as Common Spatial Patterns (Koles, 1991, Ramoser et al., 2000) or Beamforming (Grosse-Wentrup et al., 2008), in order to keep the spatial filtering setup data independent. Only data from electrodes C3 and C4, situated over left and right sensorimotor cortex, were used for further processing. For each subject, trial and electrode, frequency bands of 2 Hz width, ranging from $1-41$ Hz, were then extracted using a sixth-order Butterworth filter. Log-bandpower within the last seven seconds of each trial in each frequency band then formed the 40-dimensional feature vector.

4.3 Classification procedure

We evaluate performance of our algorithm by running experiments in different scenarios. At each round we select one subject as an online subject and consider data from the other subjects as our offline data. We discard noisy subjects from offline subjects, where we define a noisy subject as one that performs near chance level. To identify noisy subjects, we learn a simple classifier for each subject independently using standard ridge regression. Note that this corresponds to training our model with fixed mean and covariance $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \mathbf{I})$. This procedure identifies subjects 1,7 in group A and subjects 5,7 in group B as noisy subjects with the performance near chance (between 55-60%). After selecting qualified subjects, we apply our multitask learning method to infer priors from the offline data using Algorithm 1. In general, the algorithm converges in few iteration steps. We set the upper bound on the number of iterations to 5 iterations. We select regularization parameter σ^2 by 5-fold cross validation.

4.3.1 Regularization and Cross validation

Here, we discuss setting regularization parameter σ^2 in learning for a new subject. For a small n ($n < 10$), performing cross-validation to find optimal hyperparameters is not feasible due to high noise level. The number of trials per classes is between 0 and 100 and dimensionality is 40, so cross-validation will not be stable. Also for the streaming case we can not perform cross validation, so we simply set $\sigma^2 = 0.5$ in all of experiments. It worth mentioning that the method is not sensitive to σ^2 and by changing it in small ranges, performance does not change significantly. Thus, hyperparameter selection is not problematic for this method. We conjecture that with a large number of trials for the non-streaming case, cross-validation can further improve our results [2]. We compared the scenario with and without cross validation: the average performance of the two differs only by 0.18%. Also, by setting σ^2 to values ranging between $[0.01, 1]$ the average performance does not change more than 0.2%.

4.3.2 Learning for current subject

After learning priors from other subjects, we randomly permute trials from the current subject and observe them one by one. By observing a new trial, we update the classifier and record its performance on held-out test data. We repeat this procedure 100 times, to ensure that our results are not affected by a special arrangement of trials.

4.4 Classification results

4.4.1 Learning multitask inside own group

In the first experiment, a subject from group A is considered as the new subject using BCI trained with respect to off-line subjects from group A. For each sub-

[2] When we have few training points, using large regularization values renders the classifier stable, particularly when the training data are noisy. Hence, it is conceivable to update the regularization constant during trials. In our experiments, we did not investigate this issue and set σ^2 to a constant.

ject we have 150 trial per class. We randomly hold out 50 of trials from each class as test data.

The results are shown in Figure 1. In this figure, we also include the results from single task ridge regression, single task L1-regularized logistic regression (Koh et al., 2007) and single task ridge regression with pooled data. Note that L1-regularized logistic regression performs feature selection due to the L1 regularizer. In pooling data, we collect all data from qualified offline subjects and training data from the current subject and train the model using ridge regression. Hence, this method provides a baseline in one extreme where all the tasks are assumed to have the same model. On the other hand, standard single task ridge regression implements the assumption that there is no relation between tasks. For single task ridge regression and L1-regularized logistic regression we only run experiments for 10, 20, ..., 100 trials per class. For less than 10 trials per class, the classification in 40 dimensional space produces results near chance level.

Analysing Figure 1, we observe that the multitask approach significantly outperforms the other methods. Its good performance on the initial trials shows that the adaptation of an out-of-the-box BCI is very successful, whereas the other methods need 10-70 trials in order to reach the performance of the multitask method. For some subjects, the performance level is not approached even after 100 trials.

4.4.2 Effects of implicit feature selection

Investigating classifier vectors for different classes in group A in Figure 2, we observe that the classifiers of all of good performing subjects have bigger coefficients (absolute value) in the 11-14 Hz frequency range. The classifiers of Subjects 1 and 7 who have low performance they have small coefficients in this range. Note that small coefficients means that the corresponding features do not have informative values in this range. We can also observe a smaller peak in 25-30 Hz in many subjects.
Here we use simple feature computation and don't use prior knowledge on frequencies in feature computation, but multitask approach helps us to use appropriate frequencies in classification.

If we take a look at covariance matrix learned as prior in group A, we can easily see that it encourages classifiers to consider data in 11-15 Hz and also in 25-30 Hz with more importance. This acknowledges the importance of μ- and β-bands (Pfurtscheller and Neuper, 1997).

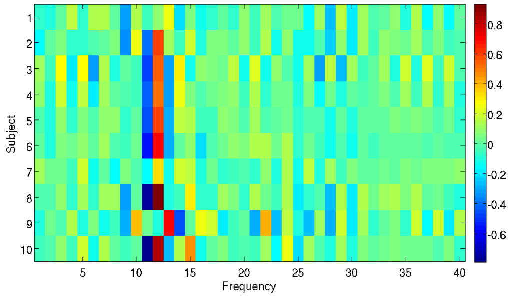

Figure 2: W matrix for group A shows weights on different frequency bands for different subjects.

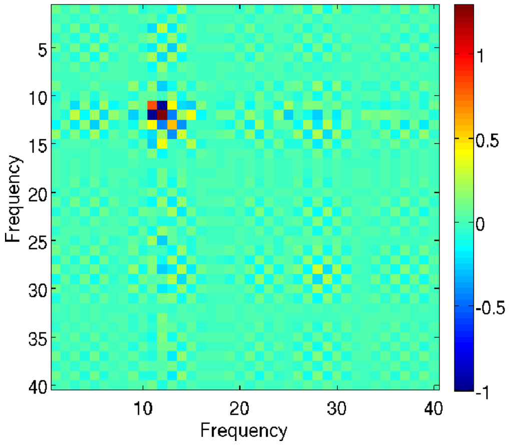

Figure 3: Covariance $\mathbf{\Sigma}$ matrix learned from group A

4.4.3 Transferring information between two groups

In the second experiment, a subject from group B is considered as the new subject using BCI trained with respect to off-line subjects from group A. Note that group A and B are evaluated with different amplifiers and caps. We hence compare between-group to within group multitask learning. The average performance for different number of trials is shown in Figure 4. We also report the performance of single task ridge regression as a baseline.
We observe a rather surprising result, namely that transfer from parameters trained on group A yields better results than learning priors from group B even though the experimental setup in two groups have some differences. We conjecture that this behaviour is due to the fact that the number of trials in the first group is higher than the second group. Hence, one can argue that the priors $(\boldsymbol{\mu}, \mathbf{\Sigma})$ learned from group A are of higher quality. The prior learned from the second group is based on 45 trials per subject, so it is of lower quality in comparison to the other prior.

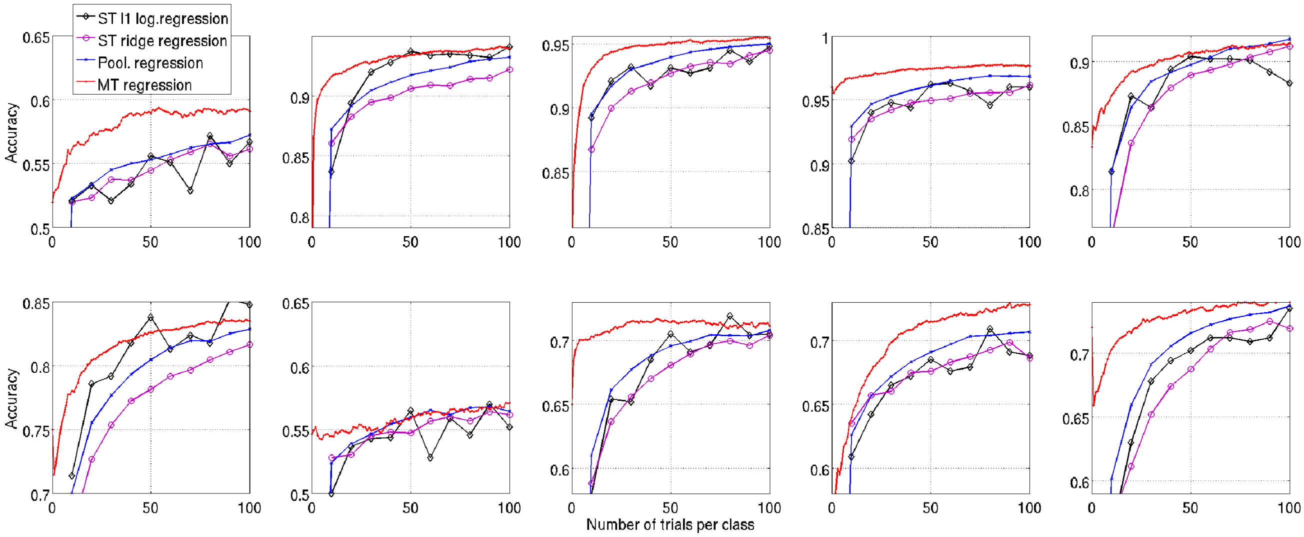

Figure 1: Accuracy for subjects in group A using 0 to 100 trials per class. ST stands for single task and MT for multitask

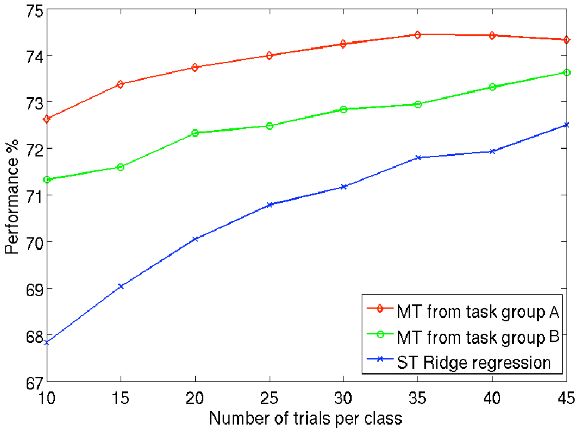

Figure 4: Learning for a subject in group B, with priors learned from group A or group B. ST stands for single task and MT for multitask

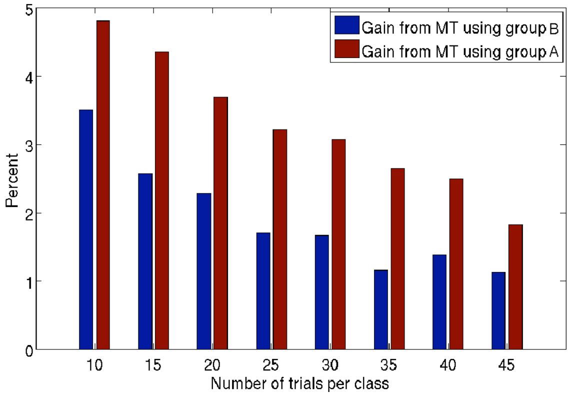

Figure 5: Gain achieved by transferring information from group A, comparing to gain achieved by using offline data from group B for an online subject in group B.

The main improvement achieved by multitask is in the cases with few learning data. We can see the average accuracy between tasks gained by transferring knowledge in learning procedure in Figure 5. As the number of training samples increases, the training data can better describes the classifier and the gain from multitasking decreases.

5 Discussion

In this paper, we presented a general framework for learning feature characteristics that are consistent across subjects for use in non-invasive BCIs. We demonstrated how this knowledge can be used to construct out-of-the-box BCIs, which can be used instantaneously by novel subjects with a satisfactory classification performance. Furthermore, our results show that combining the out-of-the-box BCI with subject-specific calibration data leads to a substantial increase in prediction accuracy compared to using subject-specific data only. Importantly, updating the out-of-the-box BCI with subject-specific data was shown to be computationally simple, rendering this approach feasible for an online setting. Finally, we demonstrated that multitask learning is also beneficial when transferring knowledge between datasets recorded with slightly different experimental setups. Since in the multitask framework all information on consistent feature characteristics is encoded in the mean vector and covariance matrix output of Algorithm 1, this should facilitate data and knowledge transfer between laboratories with access to large and very small subject databases, respectively.

While performance improvements reported in this study are already substantial, it should be noted that the multitask learning framework was only applied to the spectral domain. An even more prominent increase in prediction accuracy can be expected if the approach presented here is also extended to the spatial domain, i.e., if instead of only utilizing electrodes over sensorimotor cortices we learn priors on relevant recording locations as well. Finally, only data-independent spatial filtering was employed in this study. It remains to be established how efficient multitask learning is in settings where spatial filtering is done in a data-dependent fashion, as it is the case for CSP and beamforming.

References

Rie K. Ando and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6:1817–1853, November 2005.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73 (3):243–272, 2008.

N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kuebler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

B. Blankertz, G. Dornhege, M. Krauledat, K.R. Müller, and G. Curio. The non-invasive berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 27(2): 539–550, 2007.

Hal Daumé III. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, 2007.

Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6: 615–637, 2005.

Siamac Fazli, Florin Popescu, Mrton Danczy, Benjamin Blankertz, Klaus-Robert Müller, and Cristian Grozea. Subject-independent mental state classification in single trials. *Neural Networks*, 22(9):1305 – 1312, 2009. Brain-Machine Interface.

M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss. Beamforming in non-invasive brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 56(4):1209–1219, 2008.

K. Koh, S.-J. Kim, and S. Boyd. An interior point method for largescale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8: 1519–1555, 2007.

Z.J. Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalography and Clinical Neurophysiology*, 79:440–447, 1991.

M. Krauledat, M. Tangermann, B. Blankertz, and K.R. Müller. Towards zero training for brain-computer interfacing. *PLoS One*, 3(8):1–12, 2008.

S.G. Mason, A. Bashashati, M. Fatourechi, K.F. Navarro, and G.E. Birch. A comprehensive survey of brain interface technology designs. *Annals of Biomedical Engineering*, 35(2):137–169, 2007.

D.J. McFarland, L.M. McCane, S.V. David, and J.R. Wolpaw. Spatial filter selection for EEG-based communication. *Electroencephalography and Clinical Neurophysiology*, 103:386–394, 1997.

Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. Technical report, Department of Statistics, University of California, Berkeley, 2006.

G. Pfurtscheller and C. Neuper. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2-3):65–68, 1997.

H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

J.R. Wolpaw and D.J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Sciences*, 101:17849–17854, 2004.

J.R. Wolpaw, D.J. McFarland, G.W. Neat, and C.A. Forneris. An EEG-based brain-computer interface for cursor control. *Electroencephalography and clinical Neurophysiology*, 78:252–259, 1991.

Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML 2005: Proceedings of the 22nd international conference on Machine learning*, pages 1012–1019, New York, NY, USA, 2005. ACM.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 221680.