Adaptive Multi-class classification for Brain Computer Interfaces [1].

A. Llera, V. Gómez, H. J. Kappen
Donders Institute for Brain, Cognition and Behaviour
Radboud University Nijmegen, the Netherlands.
a.llera@donders.ru.nl

Keywords: Brain computer interfaces, adaptive classification, unsupervised learning.

Abstract

We consider the problem of multi-class adaptive classification for brain computer interfaces and propose the use of multi-class pooled mean linear discriminant analysis (MPMLDA), a multi-class generalization of the adaptation rule introduced by (Vidaurre et al., 2010) for the binary class setting. Using publicly available EEG datasets and the tangent space mapping (Barachant et al., 2012) as feature extractor, we demonstrate that MPMLDA can significantly outperform state-of-the-art multi-class static and adaptive methods. Furthermore, efficient learning rates can be achieved using data from different subjects.

1 Introduction

Brain computer interfaces (BCI) (Vidal, 1973) aim to provide human subjects with control over devices or computer applications while bypassing the traditional muscular paths. In other words these interfaces might allow humans to control devices using only their measured brain activity as measured by e.g. electroencephalogram (EEG) (Haas, 2003) . The control of such devices could offer an additional channel of communication/action that could improve the quality of life for people with severe disabilities (Wolpaw et al., 2002).

The most common procedure for setting up a BCI is as follows: the user participates in a training (i.e. calibration) session in which the user is instructed to perform a specific mental task, while the brain activity generated is recorded. The recorded data (usually designated as 'training data') are used to extract discriminative features associated with various user intentions. The data are subsequently used to train a classifier that predicts the user's intention during the testing (feedback) session (van Gerven et al., 2009). To optimise the feature extraction procedure, a time window of interest should be selected, as well as the most discriminative frequency bands. This choice is typically experimental and subject dependent and it usually requires some level of prior physiological knowledge. Several algorithms are available for automatically optimizing these parameters (Blankertz et al., 2011; Fernandez-Vargas et al., 2013; Ang et al., 2012). In a

[1] Authors preprint

multi-class BCI problem, the most common feature extractors are based on supervised data projections, such as one-against-one common spatial patterns (CSP) (Blankertz et al., 2008; Dornhege et al., 2004) or multi-class CSP (MCSP) (Grosse-Wentrup et al., 2008). Another interesting approach that has been introduced recently is tangent space mapping (TSM) (Barachant et al., 2012), which is an unsupervised nonlinear projection of the data covariance matrices that can be used to optimise its classification.

Usually the binary class BCI classification problem is considered linear (Müller et al., 2003; Farquhar, 2009) and a linear discriminant analysis (LDA) classifier (Fisher, 1936) can be used effectively. When the dimension of the feature space is large in relation to the number of samples (training data), the classifier must be regularized (Bishop, 2007; Blankertz et al., 2011). In the multi-class case, the multi-class LDA (MLDA) has been shown to be the best choice (Yang et al., 2009; Felix et al. , 2005; Tang et al., 2008).

One common problem encountered by the classifiers in EEG-based BCIs involves changes in the feature statistics over time, due to the non-stationary character of the EEG data (Krauledat, 2008). These changes generally result in poor classifier generalization performance (Shenoy et al., 2006; Millán, 2004). To overcome this problem, several methods that adapt the feature space or the classifier parameters have been proposed (Tomioka et al., 2006; Hasan et al., 2009; Llera et al., 2012).

The most common (i.e. the safest) strategy is to update the model parameters that are class-independent, while keeping the rest fixed. In the case of binary LDA, this strategy can be used to update the global average covariance matrix (or its inverse) (Vidaurre et al., 2006) or the global mean of the data (Vidaurre et al., 2010). When the adaptation is performed in the feature space and not in the classifier parameters, the strategy can be used to reduce the non-stationary effect by transforming the CSP filters (Tomioka et al., 2006) or, more generally, the actual testing data (Arvaneh et al., 2013) in a linear fashion.

To adapt class-dependent classifier parameters, it is necessary to introduce uncertainty into the model, as in the unsupervised adaptive Gaussian mixture model classifier (Hasan et al., 2009). To the best of our knowledge, the unsupervised adaptation of class-dependent feature space extraction parameters has yet to deliver any applicable results.

Updating the global mean of the data allows the adaptation of the bias of the binary LDA discriminant function. This is usually a robust technique for BCI classifier adaptation and it is referred to as the pooled mean linear discriminant analysis (Pmean) (Vidaurre et al., 2010). It is able to adapt to shifts in the feature space that are commonly attributed to non-class-related non-stationarity in EEG-based imaginary movement binary BCI (Shenoy et al., 2006). Despite its simplicity, Pmean can achieve state-of-the-art binary unsupervised adaptive classification performance. Moreover, it has been shown to be a valuable tool for helping to increase the number of possible BCI users (Vidaurre et al., 2011). Similar binary performance has been reported using Data Space Adaptation (DSA), a feature-based adaptation method proposed recently by (Arvaneh et al., 2013).

In the multi-class setting, the adaptation process clearly becomes more complex and poses harder challenges. State-of-the-art unsupervised multi-class methods, such as enhanced Bayesian LDA (EBLDA) (Xu et al., 2011), perform unsupervised retraining of each of the pair-wise Bayesian LDA classifiers (BLDA) (MacKay, 1992). This adaptive approach uses a generative model for class-conditional distributions, and its performance is strongly dependent upon the quality of the initialization. For binary problems, however, evidence suggest that the performance of the Pmean update approaches that obtained using supervised updates Vidaurre et al. (2010), thus often outperforming adaptive unsupervised generative models (e.g. EBLDA).

In this paper we introduce a novel multi-class extension of the binary Pmean adaptation of the LDA classifier and demonstrate that this kind of adaptation is better suited for multi-class adaptation than are the previously mentioned state-of-the-art methods.

In Section 2, we present the proposed method, followed by a description of the three EEG datasets used in this work (Section 3). The results are presented in Section 4. The paper concludes with a discussion (Section 5).

2 Methods

In this section, we present the methods for feature extraction and classification that we consider in the rest of the work. In Sub-section 2.1 we specify the feature extraction procedure: tangent space mapping (TSM). We describe multi-class LDA in Sub-section 2.2 and present the proposed algorithm for multi-class adaptive classification, the MPMLDA, in Sub-section 2.3.

2.1 Tangent Space Mapping (TSM)

Tangent space mapping (TSM) as feature space was recently presented for BCIs in the context of the tangent space linear discriminant analysis (TSLDA) (Barachant et al., 2012). The referenced work highlights the potential of TSM as a feature extractor for the multi-class classification of covariance matrices in the context of BCI.

In light of the observation that covariance matrices belong to the Riemannian manifold of symmetric positive-definite matrices ($\mathcal{M}$) (Moakher, 2011), TSM performs a non-linear projection of the spatial covariance matrices of the data into the tangent space (do Carmo, 1976) of $\mathcal{M}$ at the Riemannian (or geometric) mean (Heath, 1981) of the spatial covariance matrices of the training data. The Riemannian mean ($\mathcal{C}_{\mathcal{R}}$) of a set of covariance matrices $\{C_1, \dots, C_n\} \in \mathcal{M}$ is defined as

$$\mathcal{C}_{\mathcal{R}} = \text{argmin}_{C \in \mathcal{M}} \sum_{k=1}^{n} d_{\mathcal{R}}(C_k, C)^2 \tag{1}$$

where $d_{\mathcal{R}} : (\mathcal{M} \times \mathcal{M}) \to \mathcal{R}^{\geq 0}$ denotes the Riemannian distance induced by the Riemann geometry on $\mathcal{M}$ and it can be computed as a generalized eigenvalue problem (Moakher, 2005). More precisely, for any two $C_{k_1}, C_{k_2} \in \mathcal{M}$

$$d_{\mathcal{R}}(C_{k_1}, C_{k_2}) = \left[\sum_{i=1}^{m} \log^2 \lambda_i \right]^{1/2} \tag{2}$$

where λ_i, $i \in \{1, \dots, m\}$, are the eigenvalues of $C_{k_1}^{-1} C_{k_2}$.

The geometric mean exists within the considered manifold, and it is unique (Karcher, 1977). Although there is no closed-form solution for its computation, it can be computed efficiently using iterative algorithms. In this work we consider the algorithm presented by (Fletcher et al., 2004).

Given a set of covariance matrices and denoting their geometric mean as $\mathcal{C}_\mathcal{R}$, the tangent space mapping at $\mathcal{C}_\mathcal{R}$ of a given covariance matrix C_k is (after several trivial simplifications) given by

$$\mathrm{TSM}_{\mathcal{C}_\mathcal{R}}(C_k) = \log({\mathcal{C}_\mathcal{R}}^{\frac{-1}{2}} C_k {\mathcal{C}_\mathcal{R}}^{\frac{-1}{2}}), \tag{3}$$

where the log is the logarithm of a matrix derived from its diagonalization (Barbaresco, 2008).

$\mathrm{TSM}_{\mathcal{C}_\mathcal{R}}(C_k)$ is a symmetric matrix that, in vectorized form (after eliminating redundant elements due to symmetry) can be used as features for classification in BCI problems (Barachant et al., 2012). In most cases and particularly in this work, one unique $\mathcal{C}_\mathcal{R}$ is computed using all of the covariances matrices in the training set, thus rendering TSM an unsupervised feature extractor for covariance classification.

2.2 Multi-class Linear Discriminant Analysis (MLDA)

The multi-class linear discriminant analysis (MLDA) classifier (Bishop, 2007) is defined by a discriminant function $\mathcal{D}_{i,j}(\mathbf{x})$ of the input feature vector $\mathbf{x} \in \mathbf{R}^n$ for each class pair (i, j). A majority vote or a probabilistic interpretation of the results can be used to produce unique output from each pair of binary classifiers (Tax et al., 2002).

Given a K-class classification problem and a set of m labeled data vectors $\{\mathbf{x}^1, \ldots, \mathbf{x}^m\}$, for each $k \in \{1, \ldots, K\}$ we can compute the class-wise means $\boldsymbol{\mu}_k \in \mathbf{R}^n$ and covariance matrices $C_k \in \mathbb{M}_{n\times n}$. Defining the per-class-pair average covariance matrices as $\Sigma_{i,j} = \frac{C_i + C_j}{2}$ for $j > i \in \{1, \ldots, K\}$, we define the discriminant function between classes i and j as

$$\mathcal{D}_{i,j}(\mathbf{x}) = \left[b_{i,j}, \mathbf{w}_{i,j}^\intercal\right] \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \tag{4}$$

$$\mathbf{w}_{i,j} = \Sigma_{i,j}^{-1} \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right) \tag{5}$$

$$b_{i,j} = -\mathbf{w}_{i,j}^\intercal \boldsymbol{\mu}_{i,j} \tag{6}$$

$$\boldsymbol{\mu}_{i,j} = \frac{1}{2}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right) \tag{7}$$

where each $\mathbf{w}_{i,j} \in \mathbb{R}^n$ describes the vector of weights and $b_{i,j} \in \mathbb{R}$ the bias term of the discriminant function between classes i and j. For each $\{(i, j) : j > i \in \{1, \ldots, K\}\}$, an input feature vector $\mathbf{x}$ is classified as class j if $\mathcal{D}_{i,j}(\mathbf{x}) > 0$, and as class i otherwise. The output of the discriminant function $\mathcal{D}_{i,j}(\mathbf{x}) \in \mathbb{R}$ can be interpreted probabilistically by assuming that the probabilities of classes i and j are given by a binomial distribution on the sigmoidal mapping of the discriminant function value (MacKay, 2003).

More concretely, given the discriminant function $\mathcal{D}_{i,j}$, $j > i$ between classes i and j, we define the probability of $\mathbf{x}$ belonging to class i as $Q_{i,j}(i|\mathbf{x}) := \sigma(\mathcal{D}_{i,j}(\mathbf{x}))$ and, consequently $Q_{i,j}(j|\mathbf{x}) := 1 - \sigma(\mathcal{D}_{i,j}(\mathbf{x}))$ with

$$\sigma(\mathcal{D}_{i,j}(\mathbf{x})) = \frac{1}{1 + \exp(-\mathcal{D}_{i,j}(\mathbf{x}))}. \tag{8}$$

Note that this probabilistic interpretation of the discriminant function of the LDA classifier makes a connection between the LDA and the logistic regression model.

The MLDA procedure produces probabilistic output from all of the pair-wise probabilities $Q_{i,j}$ in the following manner. For simplicity of notation, we define for $j < i$, $Q_{i,j}(k|x) = Q_{j,i}(k|x)$. We then define the probability vector $\mathbf{P}$ containing in the i-th coordinate the probability of class i given $\mathbf{x}$ as

$$\mathbf{P}_i(\mathbf{x}) = \frac{\sum_{j \neq i} Q_{i,j}(i|\mathbf{x})}{\sum_k \sum_{j \neq k} Q_{k,j}(k|\mathbf{x})}. \tag{9}$$

The final output of this MLDA classifier is assigned to the most probable class under this measure $\bar{k} = \max_i \mathbf{P}_i(\mathbf{x})$.

2.3 Multi-class pooled mean LDA (MPMLDA)

In the binary setting, under the assumption of balanced classes, the bias of the discriminant function can be adapted in an unsupervised manner using the global data mean (Vidaurre et al., 2010). See Appendix A for additional information.

In the multi-class setting, not every sample contributes to all discriminant functions. For this reason, not all discriminant functions must be updated with every sample. We extend the pooled-mean approach to the multi-class case using a probabilistic update for the pairwise class means $\boldsymbol{\mu}_{i,j}$ in equation (7) as

$$\boldsymbol{\mu}'_{i,j} = (1 - \gamma_{i,j}(\mathbf{x})\beta)\boldsymbol{\mu}_{i,j} + \gamma_{i,j}(\mathbf{x})\beta\mathbf{x} \tag{10}$$

where $\boldsymbol{\mu}'_{i,j}$ represents the updated $\boldsymbol{\mu}_{i,j}$, $\beta \in \mathbb{R}$ is the learning rate and $\gamma_{i,j}(\mathbf{x})$ is defined using the probability vector $\mathbf{P}(\mathbf{x})$ of equation (9) as:

$$\gamma_{i,j}(\mathbf{x}) := \mathbf{P}_i(\mathbf{x}) + \mathbf{P}_j(\mathbf{x}). \tag{11}$$

The update (10) allows for the adaptation of each bias $b_{i,j}$ through equation (6). The MPMLDA algorithm is summarized in algorithm 1.

In the binary setting, equations (10) and (11) reduce to the binary Pmean update when $\gamma_{1,2}(\mathbf{x}) = 1$. The multi-class MPMLDA is thus a natural extension of the original pooled-mean adaptation rule.

One particular feature of MPMLDA is that, given a new data point $\mathbf{x}$, the adaptation is automatically stronger for discriminant functions between pairs of classes that are more relevant, as demonstrated in equation (11). In a supervised scenario (i.e. $P_i = \delta_{i,k}$, with k being the true class label), it is easy to see that MPMLDA updates only the boundaries between the real label k and the other classes.

Algorithm 1 *Multi-class pooled mean LDA* (MPMLDA)

Require: $\mathcal{D}_{i,j}$ $j > i \in \{1, \ldots, K\}$ constructed from labeled training data using (4)-(7).
Learning rate β.
$\mathbf{x}$ new feature sample to classify.
1: $\mathbf{P}(\mathbf{x}) \leftarrow \mathcal{D}_{i,j}(\mathbf{x})$ using (8) and (9).
2: Compute the classifier output $\bar{k} = \max_i \mathbf{P}_i(\mathbf{x})$.
3: Update $\boldsymbol{\mu}_{i,j}$ using (11) and (10).
4: Update each bias $b_{i,j}$ of each $D_{i,j}$ using (6)
5: **return**
$\bar{k}$: Class membership of $\mathbf{x}$;
$\mathcal{D}_{i,j}$: updated discriminant functions.

3 Data sets and evaluation

We consider three different datasets, all containing multi-class imagery-movement tasks.

Physiobank: [2] (Goldberger et al. (2000); Schalk et al. (2004)). This dataset contains data from 109 subjects performing various combinations of real and imagined movements in one day of recordings. In this work we consider only the data related to imagery movement of the left hand, right hand and both feet. Each subject performed 22 trials of each class ($\approx$ 3 seconds per trial) and the EEG data were recorded using 64 electrodes. Due to the computational load required by the extended analysis presented in this work, we focus only on 20 EEG electrodes over the sensorimotor cortex. The large number of users makes this dataset convenient to evaluate the impact of feature extraction on a large scale.

BCI competition IV-2a: [3] (Brunner et al. (2008)). This dataset provides data from 9 subjects performing 4 different imagery movement tasks (right hand (RH), left hand (LH), both feet (F), tongue (T)) during 2 different days of recordings. Each day the subjects performed 72 trials of each task (3 seconds per trial) and the EEG data were recorded using 20 electrodes.

BSI-RIKEN: [4](Cichocki, A. & Zhao, Q. (2011)). This dataset provides data from several subjects performing binary or multi-class imagery movement tasks. In this work we consider only the two subjects performing multi-class problems (LH, RH, F) on different, well defined days (Subjects B and C). Subject B was recorded on 2 different days and Subject C in seven different days. Each day they performed ≈ 65 trials (3-4 seconds) of each task and the EEG data were recorded using 5 or 6 electrodes. In this work we always consider the same 5 EEG electrodes, 'C3', 'Cp3', 'C4', 'Cp4' and 'Cz'.

The dataset also contains three long additional sessions for Subject C (≈ 268 trials per session) spreaded across a different day. We denote these data as Subject

[2] `http://www.physionet.org/physiobank/database`

[3] `http://www.bbci.de/competition/iv`

[4] `http://www.bsp.brain.riken.jp/~qibin/homepage/Datasets.html`

C2.

In this work, all of the datasets were pre-processed in a similar way before the feature extraction step. In all cases, the data were divided into trials containing the imagery movement period. Channels and trials contaminated with artefacts were eliminated from the training set using an automatic variance based routine (Nolan et al., 2010). The contaminated channels were also removed from the testing set, after which the data were detrended and band-pass-filtered into the frequency bands 8-30 Hz. To reduce the presence of filter-induced artefacts in the data, we discarded the starting and ending half seconds from each trial.

We measure BCI performance using the Cohen's Kappa coefficient (Kraemer, 1982), which assigns a value of zero to random classification and a value of one to perfect classification. This measure is commonly used in multi-class BCI problems (Schlögl et al., 2007).

We use the Physiobank dataset only to compare the different feature extractors, and not to evaluate adaptive methods. Note that the reduced ammount of trials per class available per subject in this dataset does not allow the proper evaluation of any adaptive method. In this case, we report classification results based on leave one out cross-validation.

We use the BCI-IV-2a and BSI-RIKEN datasets to evaluate and compare the performance of the different classifiers in terms of adaptation. We perform training on the data from one day and testing on data from a subsequent day. Because these data were recorded on different days, they are more likely to present hard non-stationarities.

For subject C2, however, each model was trained and tested on the different sessions taking on the same day. Even though all recordings denoted as subject C2 were performed within the same day, we decided to include them in the analysis as each session provides many trials and the three sessions were distributed throughout the day, thereby allowing the presence of clear non stationary changes. This procedure yielded a total of 9 evaluations for the BCI-IV-2a dataset and 25 evaluations for the BSI-RIKEN dataset, which are decomposed to a value of 1 for subject B, 21 for subject C $(6 + 5 + ... + 1)$ and 3 for subject C2.

4 Results

This section describes the results. First, given that TSM has only recently been introduced and it is largely unknown in the BCI community, we present novel results that confirm the quality of TSM as an algorithm for multi-class feature extraction when using static classification. We then analyse and compare the performance of the proposed multi-class adaptive method. In Sub-section 4.3 we consider the method's dependence on the learning rate and, in Sub-section 4.4 we provide a qualitative analysis of the dynamic changes occurring in the feature space. We assess the significance of the performance differences between methods according to Wilcoxon statistical tests (Demšar., 2006). The observed differences are with either single or double asterisks to indicate p-values less than 0.05 and 0.01, respectively.

4.1 TSM as feature space for non-adaptive classification

In this subsection, we illustrate the ability of TSM to serve as multi-class feature extractor for BCI imagery movement problems on a large scale. We present results according to a static MLDA classifier (Section 2.2) and the three datasets considered in this work. Figure 1 provides a comparison of the Kappa values obtained using TSM and

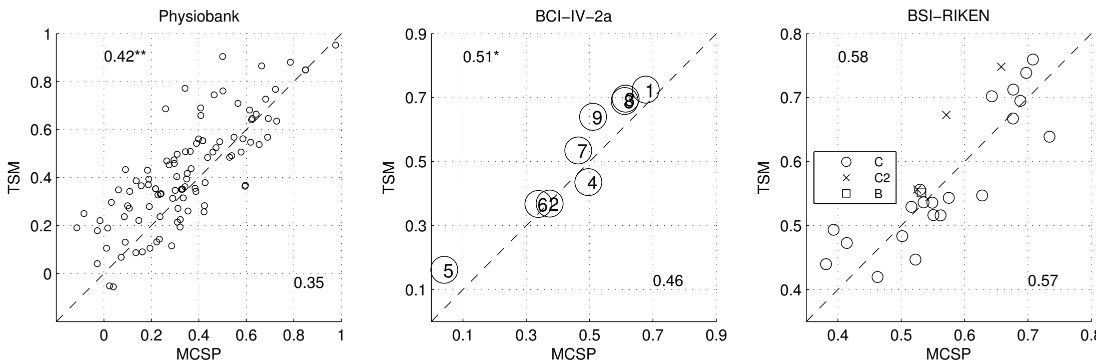

Figure 1: Kappa values obtained through MCSP (x axis) plotted against Kappa values obtained by TSM for the three datasets under consideration. The values printed in upper left and lower right corners of each figure represent the mean Kappa values obtained by each method. Wilcoxon statistical tests indicate the significance of the observed differences, noted by single or double asterisks representing p-values smaller than 0.05 and 0.01, respectively.

multi class CSP (MCSP) (Grosse-Wentrup et al., 2008), which is commonly used for multi-class BCI feature extraction. For the BCI-IV-2a dataset (middle panel), subject numbers are indicated inside the circles. For the BSI-RIKEN dataset (right panel), the circles represent Subject C, the square represents Subject B and the crosses represent the different possible test sessions of Subject C2 during one day. The mean Kappa values are displayed in the upper left (for TSM) and lower right (MCSP) corners of each panel.

For Physiobank and BCI-IV-2a datasets, TSM provides a significantly better feature space than MCSP does. These results confirm that TSM can be considered a state-of-the-art feature extractor for multi-class BCI problems. For the remainder of this paper, we consider TSM as the feature space.

4.2 Multi-class Adaptive Classification

In this subsection, we analyse the problem of inter-day classifier adaptation in the multi-class setting, using the BCI-IV-2a and BSI-RIKEN datasets and considering TSM as the feature extractor. We analyse the behaviour of the proposed method, MPMLDA, and we consider EBLDA and DSA for comparison (a brief description of the two methods is provided in Appendices B and C, respectively). For reference on the improvement achieved, we also present values computed for the performance of the initial (static) MLDA classifier

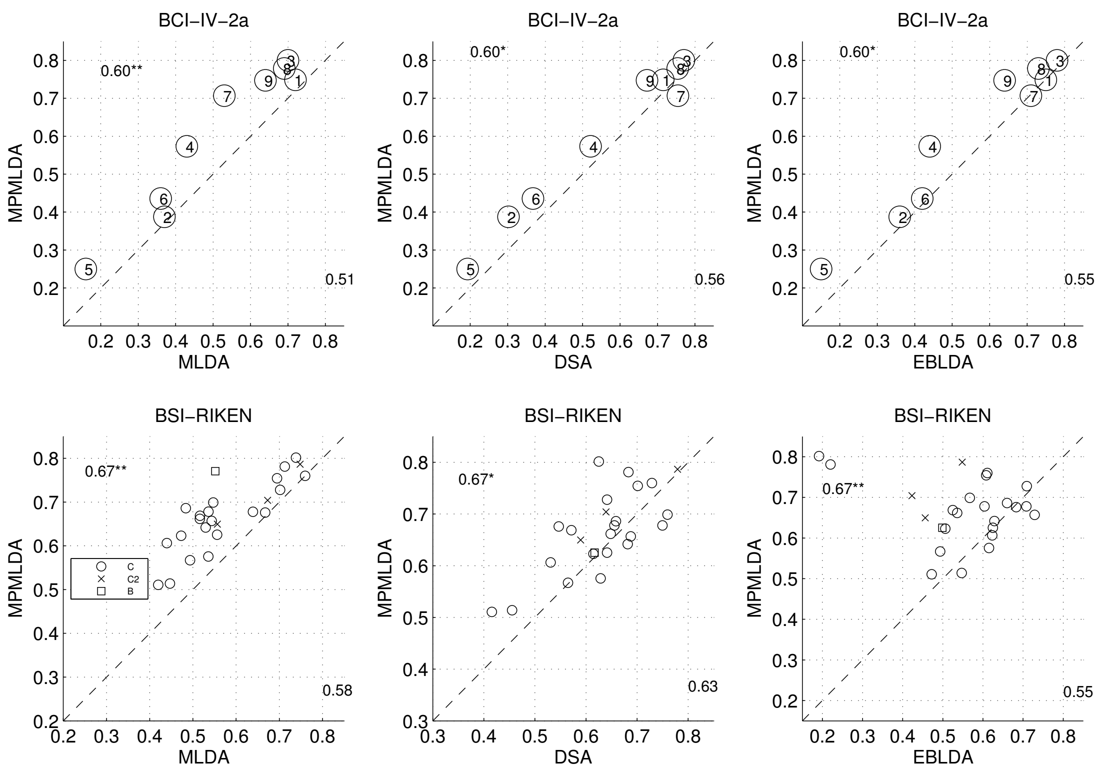

Figure 2: Comparison between MPMLDA and other multi-class methods. **All adaptive methods use individual optimal learning rates**. Each row represents a different data set, and each column compares it with a different model (MLDA, DSA and EBLDA). The values displayed in the upper left and lower right corners are the mean Kappa value averaged over all the subjects of the dataset. Wilcoxon statistical tests indicate the significance of the observed differences. Single and double asterisks indicate p-values smaller than 0.05 and 0.01, respectively.

In all cases, we first train a initial static MLDA (or BLDA [5]) classifier using data from a day of recordings, after which we selected a different (future) day/session of the same subject as testing data, in order to evaluate both static and adaptive classifiers. In this subsection, we optimise the learning rates for each combination of subject and adaptation method separetely. In the next sub-section we analyse the influence of the learning rate in more detail.

Figure 2 shows the comparison of MPMLDA in terms of Kappa values for the different methods (in columns) based on each of the datasets (in rows). Points above the discontinuous lines represent cases in which MPMLDA performance is superior. The values at the upper left and lower right corners indicate the mean Kappa values of the corresponding method averaged over all subjects.

First of all we note that on average any adaptation improves the mean performance of the static classifier, which has Kappa 0.51. Remarkably, MPMLDA significantly outperforms all of the other methods. Note that for the BSI-RIKEN dataset, EBLDA yields very poor performance in some isolated cases.

In contrast, although DSA provides excellent stable performance, its performance is still significantly worse than that achieved by MPMLDA. We conclude that for the datasets under consideration MPMLDA performs significantly better than DSA and EBLDA.

4.3 Influence of the Learning Rate

In Figure 3, we present results of our analysis of the influence of the learning rate. We plot the mean Kappa values averaged over all subjects, as a function of the learning rate for the BCI-IV-2a dataset (left panel) and the BSI-RIKEN dataset (right panel). Horizontal lines indicate the performance of the static MLDA method, and the adaptive EBLDA and DSA (the latter two adaptive methods use individually optimised learning rates).

Interestingly, the improvement of MPMLDA is notable with respect to the static MLDA for a wide range of learning rate values. The performance of this method is also superior or comparable to the other adaptive methods for both datasets.

Observe that the optimal learning rate differs between datasets. These differences can be explained by the dependence of the learning rate on the number of classes induced by equation (11).

We now set the MPMLDA learning rate to the corresponding optimal values shown above and compare the performance of MPMLDA against the other adaptive methods, for which the learning rates have been fully optimised. The results are displayed in Figure 4.

With this subject-independent but dataset-dependent learning rate, MPMLDA still generally outperforms the static classifier and also EBLDA. Although it does not yield any significant difference with respect to the optimal DSA (second column).

We conclude that effective subject independent learning rates can be learned for a fixed paradigm and a fixed EEG montage. Although they deviate from subject dependent optimal learning rates, this approach is able to significantly outperform static

[5] For EBLDA, a Bayesian LDA was used for training.

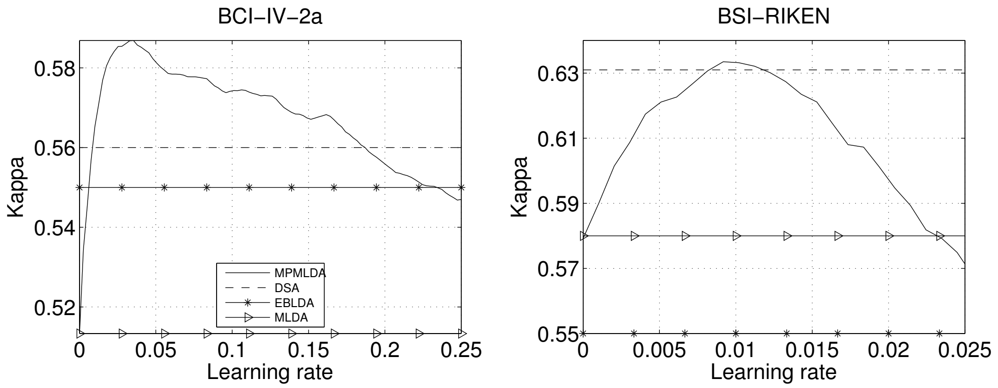

Figure 3: The continuous line represents the average Kappa value across subjects as a function of the MPMLDA learning rate using the datasets BCI-IV-2a and BSI-RIKEN datasets. Discontinuous and marked lines represent the average Kappa values obtained by MLDA, EBLDA, and DSA. Individual optimal learning rates are used EBLDA and DSA.

classification and optimal EBLDA.

4.4 Analysis of the feature space dynamics.

In order to understand why MPMLDA outperforms the other methods we project the class-wise training and testing feature distributions onto the first two Principal Components derived from the training data (Jolliffe, 2002). The results for subjects 4 and 8 (BCI-IV-2a) are displayed in Figure 5 left side.

Note the clear shift observed between the training and testing distributions occurring for Subject 8. In the right column of Figure 5, we represent the shift in the mean for each class. Observe that the shifts are class dependent in the case of Subject 4 and largely class-independent for subject 8. Adapting for class independent shifts it is obviously a much simpler task that can be achieved using DSA or a naive multi class Pmean update (i.e. assuming $\gamma_{i,j}(x) = 1 \forall i, j$ in equation 11). However, class dependent changes can definetly not be tracked by any of this methods. Further, note that the changes in the covariances are not too strong for any of both subjects. Such behaviour (i.e. strong mean shifts and small covariance changes) was observed in general in the BCI-IV-2a dataset. By construction MPMLDA can be able to adapt for class dependent shifts; this fact explains the superior performance of MPMLDA in the BCI-IV-2a dataset.

The PCA projections on different days for Subject C (BSI-RIKEN) are displayed in Figure 6. In the first row, we show the change between the first and the second day. Once more, strong means shifts represent properly the non stationary changes. However, the non-stationary character becomes stronger when comparing the projections of the first and the seventh day for the same subject (second row). Note that the shifts in the mean, is insufficient to represent the non-stationarity in this case; a strong class dependent covariance change is also present. Even though this is a difficult scenario for any adaptive model, it is clear that correcting for the bias is a necesary condition for any adaptation strategy to be succesfull.

Instead of using the principal components from the first day, we concentrate now on

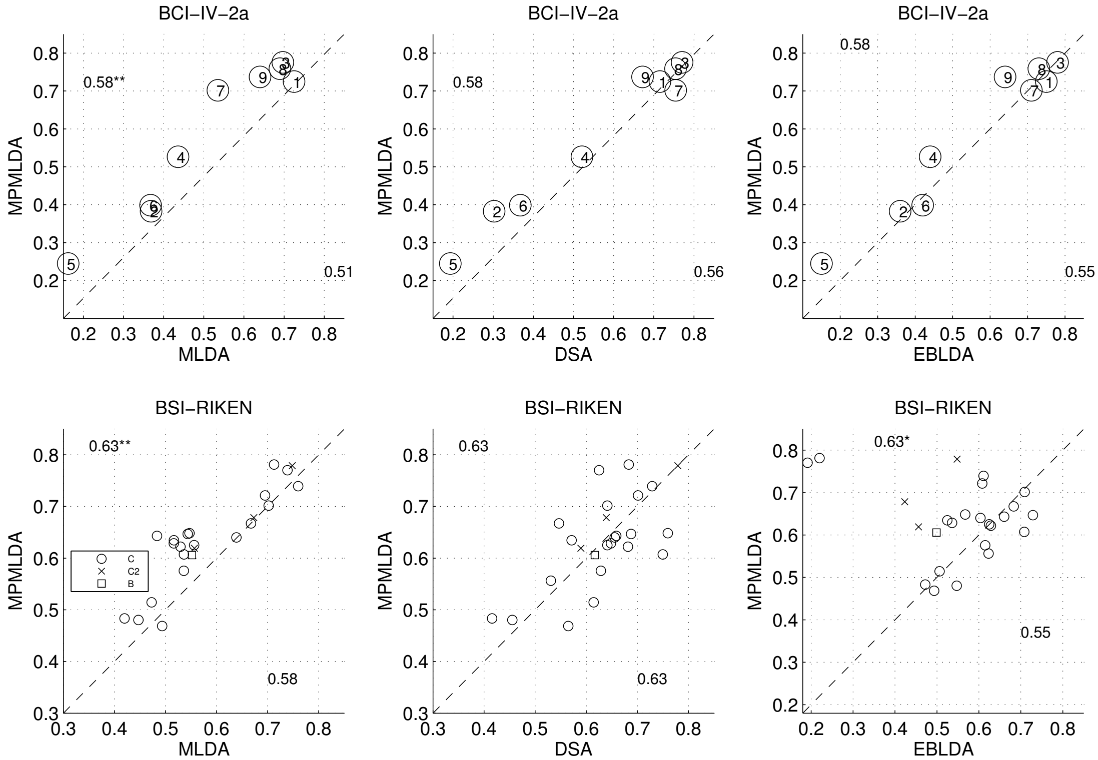

Figure 4: Comparison between MPMLDA and other multi-class methods. **The learning rates values of all the adaptive methods have been optimised for each subject independently, with the exception of MPMLDA, which corresponds to the optimised mean across subjects, (see Figure 3)**. The learning rate values used by MPMLDA are 0.03 and 0.01 for the BCI-IV-2a and BSI-RIKEN datasets respectively. Each row represents a different data set and each column compares it to a different model (MLDA, EBLDA and DSA). The values displayed in the upper left and lower right corners indicate the mean Kappa value averaged over all subjects in the dataset. Wilcoxon statistical tests indicate the significance of the observed differences. Single and double asterisks indicate p-values smaller than 0.05 and 0.01, respectively.

the projections after training and testing on the sixth and the seventh days, respectively, as shown in Figure 6 (bottom row). Remarkably, they are more concentrated than they were in the previous cases. One possible explanation for this reduced variance could that the subject learned to execute the task better (Curran, 2003; Barbero Jimenez, A. and Grosse-Wentrup, M. , 2010). However, also in this case class dependent changes are dominant (both mean shifts and covariance changes).

To conclude we note that in some cases, the translation vectors are grouped (right columns), suggesting that class-dependent bias updates add little improvement to a naive Pmean that would account for a class-independent shift, (i.e. assuming $\gamma_{i,j}(x) = 1 \forall i, j$ in equation 11). Using this approach, however, results in mean Kappa values of 0.541 (**) and 0.60 (**) for the BCI-IV-2a and BSI-RIKEN datasets respectively, which improve the static MLDA but are significantly worse than the performance obtained by MPMLDA. This result confirms that the use of class-dependent bias updates

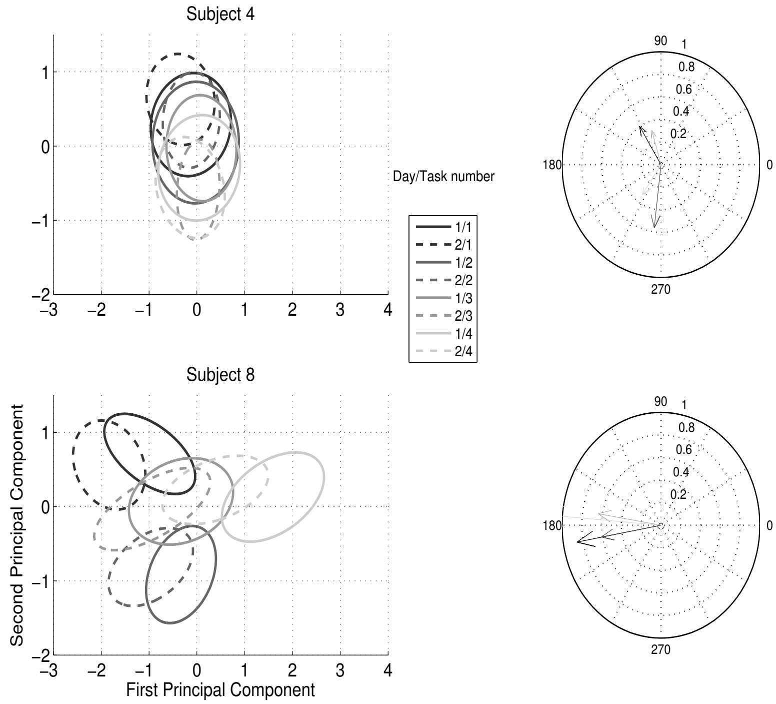

Figure 5: Left: Distribution of training (continuous ellipses) and testing (discontinuous ellipses) features of four different tasks (colour), projected onto the first two principal components for Subjects 4 and 8. Right: the shift in the mean for each class between day 1 and day 2.

is useful for tracking such multi-class class-dependent distribution shifts.

5 Discussion

In this work we propose a novel method for adaptive multi-class classification for BCI: MPMLDA. This method is a multi-class extension of the binary pooled mean LDA (Pmean) introduced in (Vidaurre et al., 2010). We demonstrate that the performance of MPMLDA is superior to that of state-of-the-art adaptive methods as EBLDA and DSA.

As feature space we use TSM as recently introduced in Barachant et al. (2012). Our results confirm previous findings; TSM based features yields better classification performance than MCSP features.

The most important feature of MPMLDA is that its parameter updates are class-dependent, thereby resulting in larger updates for discriminant functions between pairs of classes that are more suitable for explaining the current EEG pattern. Our results on different datasets suggest that such class-dependent updates are a key ingredient in explaining the improved performance of MPMLDA over the other methods.

One interesting observation is that MPMLDA can achieve higher classification than is possible with DSA. This is a remarkable result. By construction DSA has the potential to adapt for stronger non-stationarities than does MPMLDA. This is because, in principle, DSA can remove not only shifts but also rotations in feature space (if they

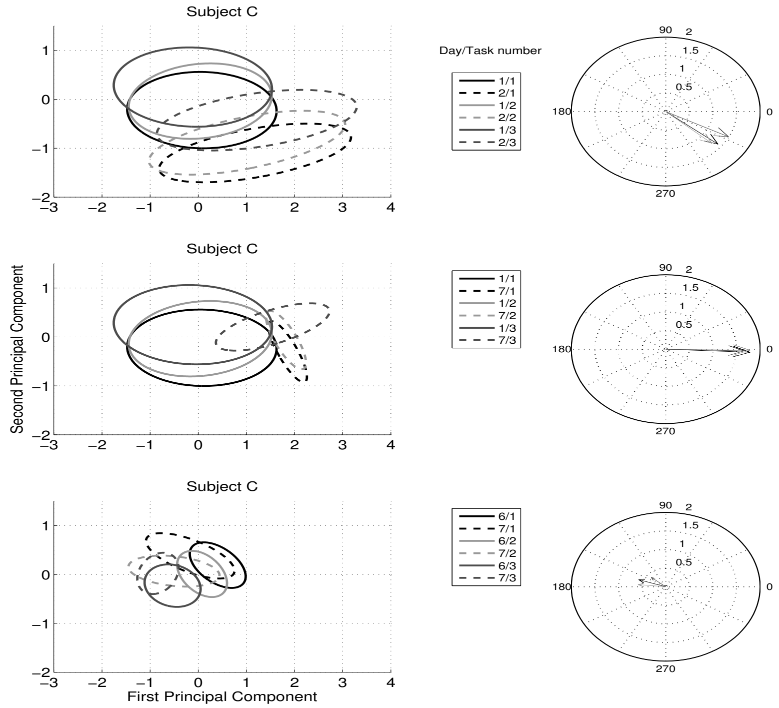

Figure 6: Left: Distribution of training (continuous ellipses) and testing (discontinuous ellipses) features of four different tasks (colour), projected onto the first two principal components for Subject C. Training and testing was performed for days (1,2), (1,7) and (6,7) respectively as indicated by the corresponding legend. Right: the shift in the mean for each class between day 1 and day 2..

are common to all classes). However, the presented results clearly show the presence of class dependent non stationarity components that can not be learned using DSA. As indicated by the empirical results, the MPMLDA is able to outperform also DSA and as such the bias adaptation has been proved to be a very powerful tool also in the multi-class setting.

Appendix

A - Binary linear discriminant analysis with pooled mean adaptation (*Pmean*)

Pmean (Vidaurre et al., 2010) is a binary unsupervised adaptive LDA algorithm which identifies the global mean of the data ($\boldsymbol{\mu}$) with $\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ represent the class wise means, under the assumption of balanced classes. Consequently $\boldsymbol{\mu}$ can be updated sequentially following

$$\boldsymbol{\mu}' = (1-\beta)\boldsymbol{\mu} + \beta\mathbf{x}, \tag{12}$$

where $\boldsymbol{\mu}'$ is the updated $\boldsymbol{\mu}$, $\mathbf{x}$ is the new observed feature vector and $\beta \in [0,1]$ is the learning rate controlling the adaptation. Because the bias of the LDA discriminant function is given by

$$b = -\boldsymbol{w}'\boldsymbol{\mu}, \tag{13}$$

with $\boldsymbol{w}$ representing its weights, the update of $\boldsymbol{\mu}$ (through equation 12) allows the unsupervised update of the bias.

Even though Pmean provides state-of-the-art binary unsupervised classification performance, it is important to note that no multi-class extension of Pmean has previously been considered in the literature.

B - Enhanced Bayesian linear discriminant analysis (EBLDA)

Bayesian Linear Discriminant Analysis (BLDA) is a Bayesian version of regularized LDA, in which regularization parameters are estimated with Bayesian regression (see (Xu et al., 2011) for more detailed information).

In order to improve the performance of the classifier EBLDA proposes training a new classifier by supplementing training sets with additional high probability test samples. In the binary case, the probability from the BLDA classifier for a test sample is computed. if this probability exceeds a threshold (e.g. 0.9), this test sample and its estimated label are added to the training set for classifier retraining. In this work we refer to the parameter value indicating this threshold as the learning rate.

In the multi-class setting, EBLDA uses combinations of binary BLDA classifiers as MLDA.

C - Unsupervised EEG Data Space Adaptation (*DSA*)

The DSA adaptation procedure (Arvaneh et al., 2013) performs an adaptive linear transformation of the testing data instead of adapting the classifier parameters. It provides a direct approximation of the discrepancy between the band-pass filtered training and testing data distributions by comparing the average distributions of the EEG data obtained regardless of the class labels (under Gaussian assumption). Let $N_1 = \mathcal{N}(0, \hat{C})$ be the average distribution of the training data, where $\hat{C}$ is obtained by averaging the covariance matrices over all available EEG training trials. Denote the average distribution of the testing data after a linear transformation as $N_2 = \mathcal{N}(0, W^t C W)$, with W as a linear transformation matrix and C as the average covariance matrix of the testing data. The DSA method optimises the matrix W by minimizing the KL divergence between N_2 and N_1. Interestingly, W can be written in closed form as

$$W = (\hat{C}^{-1}C)^{\frac{-1}{2}} = C^{\frac{-1}{2}}\hat{C}^{\frac{1}{2}}.$$

In DSA, the linear transformation W is recomputed sequentially after a certain number of trials, and the new testing pattern is projected onto W before applying the pre-trained feature extraction and classification algorithms. The number of trials used for re-computing C and W determines the length scale of the adaptation. We therefore refer to this parameter as the learning rate for the DSA method.

The original unsupervised DSA method considering common spatial patterns (CSP) as feature extractor is very similar to the unsupervised adaptation of CSP as proposed in (Tomioka et al., 2006). The latter updates the CSP filters using W (i.e. $CSP \rightarrow WCSP$) while DSA filters the testing data (X) using W (i.e. $X \rightarrow W'X$). Consequently in both cases the features are extracted from the linear transformation $X \rightarrow CSP'W'X$.

In this work, we consider a multi-class variant of DSA with TSM as feature space. In both senses, this unsupervised strategy differs from the proposals in (Tomioka et al., 2006; Arvaneh et al., 2013).

Acknowledgments

The authors gratefully acknowledge the support of the Brain-Gain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

References

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., Zhang, H. (2012). Filter Bank Common Spatial Pattern algorithm on BCI Competition IV Datasets 2a and 2b. *Frontiers in Neuroscience*, *6, 39*.

Arvaneh, M., Guan, C., Ang, K. K., Quek, C. (2013). EEG Data Space Adaptation to Reduce Intersession Nonstationarity in Brain-Computer Interface. *Neural Computation*,*8-25*, 2146–2171.

Barachant, A., Bonnet, S., Congedo, M., Jutten, C. (2012). Multiclass Brain-Computer Interface Classification by Riemannian Geometry. *IEEE Transactions on Biomedical Engineering*, *59-4*, 920–928.

Barbaresco, F. (2008). Innovative tools for radar signal processing based on Cartans geometry of SPD matrices and information geometry. *IEEE Radar Conference*.

Barbero Jimenez, A. and Grosse-Wentrup, M. (2010). Biased Feedback in Brain-Computer Interfaces. *Journal of NeuroEngineering and Rehabilitation*, *7,34*, 1–4.

Bishop, C. M. (2007). Pattern Recognition and Machine Learning. *Springer*.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller K-R. (2008). Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine*, *56-2*, 41–56.

Blankertz, B., Lemm S., Treder M., Haufe S., Müller K-R. (2011). Single-trial analysis and classification of {ERP} components. A tutorial. *NeuroImage*, *56-2*, 814–825.

Brunner, C., Leeb R., Müller-Putz, G.R, Schlögl, A., Pfurtscheller, G. (2008). BCI competition 2008, Graz dataset 2A

do Carmo, M. P. (1976). Differential Geometry of Curves and Surfaces. *Prentice-Hall*, Englewood Cliffs, NJ.

Cichocki, A. & Zhao, Q. (2011). EEG motor imagery dataset. *Technical report*, Laboratory for Advanced Brain Signal Processing, BSI, RIKEN, Saitama, Japan.

Curran, E. (2003). Learning to control brain activity: A review of the production and control of EEG components for driving brain computer interface (BCI) systems *Brain and Cognition*, *51-3*, 326–336.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dornhege G., Blankertz B., Curio G., Müller K.R. (2004). Increase information transfer rates in BCI by CSP extension to multi-class *Advances in Neural Information Processing Systems*, 733–740, MIT Press.

Farquhar, J. (2009). A linear feature space for simultaneous learning of spatio-spectral features in BCI *Neural Networks*, *22-9*, 1278–1285.

Felix, L., Scherer R., Leeb R., Neuper C., Bischof, H., Pfurtscheller G.(2005) A Comparative Analysis of Multi-Class EEG Classification for Brain Computer Interface. *Proceedings of the 10th Computer Vision Winter Workshop CVWW 2005*.

Fernandez-Vargas, J., Pfaff, H. U., Rodríguez, F. B., Varona, P. (2013) Assisted closed-loop optimization of SSVEP-BCI efficiency. *Frontiers in Neural Circuits*, *27-7*.

Fletcher, P. T., Joshi, S. (2004). Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. *ECCV Workshops CVAMIA and MMBIA*, 87–94, Springer-Verlag.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems *Annals of Eugenics*, *2-2*. 179–188.

Gerven van, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., Haselager, P., Vuurpijl, L., Gielen, S., Desain, P. (2009). The brain computer interface cycle. *Journal of Neural Engineering*, *6-4*.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, *100-23*, 215–220.

Grosse-Wentrup, M., Buss, M. (2008). Multi-class Common Spatial Patterns and Information Theoretic Feature Extraction. *IEEE Transactions on Biomedical Engineering*, *55-8*,

Haas, L. F. (2003). Hans Berger (1873 - 1941), Richard Caton (1842 - 1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, *74, 1*.

Hasan, B.A.S. & Gan, J.Q. (2009). Sequential EM for unsupervised adaptive Gaussian mixture model based classifier. *Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, Springer Berlin / Heidelberg*, vol. 5632, 96–106.

Heath, Sir T. (1981). A History of Greek Mathematics. Vol. 1. From Thales to Euclid. *Dover Publications.*.

Jolliffe, I. T. (2002). Principal Component Analysis. Springer.

Kachenoura, A., Albera, L., Senhadji, L., Comon, P. (2008). ICA: a potential tool for BCI systems. *IEEE Signal Processing Magazine*, *25, 1*, 509–541.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Wiley Subscription Services, Inc., A Wiley Company*, *30,5*, 57–68.

Kraemer, H. C. (1982). Encyclopedia of Statistical Sciences.New York: Wiley.

Krauledat M. (2008). Ph.d thesis *Technische Universität Berlin, Fakultät IV – Elektrotechnik und Informatik*.

Llera, A., Gómez, V., Kappen, H. J. (2012). Adaptive classification on Brain Computer Interfaces using reinforcement signals. *Neural Computation*, *4:11*, 2900–2923.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *3,4*, 415–447.

MacKay, D. J. C. (2003). Information Theory, Inference and Learning Algorithms. *Cambridge University Press*.

Millán, J. del R. (2004). On the Need for On-Line Learning in Brain-Computer Interfaces. *IEEE Proc. of the Int. Joint Conf. on Neural Networks*, vol.4, 2877 - 2882.

Moakher M.(2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl*, *26*.

Moakher M., Zéraï M.(2011). The Riemannian Geometry of the Space of Positive-Definite Matrices and Its Application to the Regularization of Positive-Definite Matrix-Valued Data. *J. of Mathematical Imaging and Vision*, *40-2*, 171–187.

Müller, K. R., Anderson, C.W., Birch, G.E. (2003) Linear and nonlinear methods for brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11,2*, 165–169.

Meng Jincheng, Yang Wanlin (2005). Nearest neighbor classifier based on Riemannian metric in radar target recognition *IEEE International Radar Conference*, 851–853.

Nolan, H., Whelan, R., Reilly, R. B.(2010). FASTER: Fully Automated Statistical Thresholding for {EEG} artifact Rejection. *Journal of Neuroscience Methods*, *192,1*, 152–162.

Schalk, G., Mcfarl, D. J., Hinterberger, T., Birbaumer, N., Wolpaw, J. R.(2004). BCI2000: A general-purpose brain-computer interface (BCI) system *IEEE Transactions on Biomedical Engineering*, *51*.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. & Müller, K. R. (2006). Towards adaptive classification for BCI. *Journal of neural engineering*, *3*, R13-R23.

Schlögl, A., Kronegg, J., Huggins, J. E., Mason, S. G.(2007). Evaluation Criteria for BCI Research. *Toward Brain-computer Interfacing*, *19*, 327–342, MIT Press.

Tang, Y. and Tang, J. and Gong, A. (2008). Multi-Class EEG Classification for Brain Computer Interface based on CSP. *Proceedings of the International Conference on BioMedical Engineering and Informatics*, *2*, 463–472. IEEE Computer Society, Washington, DC, USA.

Tax D. M. J., Duin, R. P. W. (2002). Using two-class classifiers for multi-class classification. *Proceedingof the 16th IEEE International Conference on Pattern Recognition*, *2*, 124–127.

Tomioka, R., Hill, J.N., Blankertz, B. & Aihara, K. (2006). Adapting Spatial Filter Methods for Nonstationary BCIs. *Proceedings of 2006 Workshop on Information-Based Induction Sciences (IBIS 2006)*, 65–70.

Tuzel O., Porikli F., Meer P. (2007). Human detection via classification on riemannian manifolds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Vidal, J. J. (1973). Toward Direct Brain-Computer Communication. *Annual Review of Biophysics and Bioengineering*, *2*, 157–180.

Vidaurre, C., Schlöogl, A., Cabeza, R., Scherer, R., Pfurtscheller, G. (2006). A fully on-line adaptive BCI. *IEEE Transactions on Biomedical Engineering*, *6*, *53*, 1214–1219.

Vidaurre, C., Kawanabe, M., von Bünau, P., Blankertz, B. & Müller, K. R. (2010). Toward an unsupervised adaptation of LDA for Brain-Computer Interfaces. *IEEE Trans. Biomed. Eng*, vol.58, no.3, 587 – 597.

Vidaurre, C., Sanelli, C., Müller, K. R., & Blankertz, B. (2011). Machine-Learning Based Co-adaptive Calibration. *Neural computation*, vol.23, no.3, 791 – 816.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical neurophysiology*, *6-113*, 767–791.

Xu, P., Yang, P., Lei, X., Yao, D. (2011). An enhanced probabilistic LDA for multi-class brain computer interface *PloS One*, *6-1*: e14634. doi:10.1371/journal.pone.0014634.

Yang, P., Xu, L., Liu, T., Xu, P., Yao, D-Z. (2009). Probabilistic Methods in Multi-Class Brain Computer Interfaces *Journal of Electronic Science and Technology of China*, *7,1*, 12–16. International Society for Bioelectromagnetism.