Mean shrinkage improves the classification of ERP signals by exploiting additional label information

Johannes Höhne*, Benjamin Blankertz* Klaus-Robert Müller†‡, and Daniel Bartz†,
*Neurotechnology group, Berlin Institute of Technology, Berlin, Germany
†Dept. of Machine Learning, Berlin Institute of Technology, Berlin, Germany
‡Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Korea

***Abstract*—Linear discriminant analysis (LDA) is the most commonly used classification method for single trial data in a brain-computer interface (BCI) framework. The popularity of LDA arises from its robustness, simplicity and high accuracy. However, the standard LDA approach is not capable to exploit sublabel information (such as stimulus identity), which is accessible in data from event related potentials (ERPs): it assumes that the evoked potentials are independent of the stimulus identity and dependent only on the users' attentional state. We question this assumption and investigate several methods which extract subclass-specific features from ERP data. Moreover, we propose a novel classification approach which exploits subclass-specific features using mean shrinkage. Based on a reanalysis of two BCI data sets, we show that our novel approach outperforms the standard LDA approach, while being computationally highly efficient.**

I. Introduction

Brain-computer interfacing (BCI) is a highly interdisciplinary research area which aims to enable communication pathways that are independent from muscle activity [1], [2]. Generally, BCIs analyze brain signals of a user in real-time while advanced methods for data processing and classification allow to translate the users' intention into commands. Such BCI systems are developed for various applications, including communication, gaming, rehabilitation or mental state monitoring [3].

BCI paradigms which are based on event related potentials (ERPs) evaluate brain responses to a sequence of external stimuli. Within such paradigms, it is the objective to assess which stimulus the user is attending to. This yields to a binary classification task attended vs. unattended stimuli, also referred to as targets vs. nontargets.

Numerous studies have investigated classification techniques in order to optimally separate between evoked potentials of targets and nontargets [4], [5]. Most of these studies found LDA to be amongst the best performing methods. The standard LDA classifier however disregards the sublabel information: i.e. the stimulus identity.

Thus, for a given attentional state of the user, the same EEG response is assumed for each stimulus. This assumption conflicts with evidence for subclass-specific features in the ERP data found by several studies [6], [7], which arise from varying stimulus properties. In general, stimuli are designed to be unique and distinct in order to facilitate discrimination while they should also be highly standardized to prevent stimulus/subclass-specific EEG signatures.

The aim of this work is to question this assumption and to derive alternative classification approaches which enable to utilize such subclass-specific features.

978-1-4799-4149-0/14/$31.00

II. Methods

A. The standard approach: LDA with covariance shrinkage

In order to use a BCI based on event related potentials, a binary classification problem has to be solved. The task is to separate between brain responses to target and non-target stimuli. Linear discriminant analysis (LDA) is a simple and robust linear classification method which is frequently applied for ERP data. LDA assumes the data to follow a Gaussian distribution with all classes having the same covariance structure. LDA seeks a linear projection w such that within-class variance is minimized while the between-class variance is maximized. For the two-class scenario, it can be shown that the optimal projection $\boldsymbol{w}$ can be determined by

$$\boldsymbol{w} = \boldsymbol{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{1}$$

Thus, in order to compute the LDA classifier, the class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as well as the class-wise covariance $\boldsymbol{C}$ have to be estimated. However, the estimation of the covariance matrix might be distorted, as the features can be high dimensional and only a limited amount of data points are available. It is known, that this curse of dimensionality leads to sample estimates $\boldsymbol{C}^s$ of the unknown covariance $\boldsymbol{C}$ with a systematical distortion: directions with high variance are over-estimated, while low-variance directions are under estimated. For BCI data, this was discussed in [8]. In order to compensate for such distortions, one can introduce a regularization term when estimating the covariance

$$\boldsymbol{C}^{reg}(\lambda) = (1-\lambda)\boldsymbol{C}^s + \lambda\nu\boldsymbol{I}, \tag{2}$$

with λ and ν being regularization and scaling parameters. In the BCI framework, this regularization parameter λ is mostly determined with the shrinkage method [9]. Shrinkage seeks for an estimate of the covariance matrix, such that the expected mean squared error (EMSE) is minimized,

$$\begin{aligned}\lambda^* &= \underset{\lambda}{\operatorname{argmin}}\, \mathbb{E}\left[\sum_{i,j}\left(C_{ij}^{reg}(\lambda) - C_{ij}\right)^2\right] \\ &= \frac{\sum_{i,j}\left\{\operatorname{Var}\left(C_{ij}^s\right) - \operatorname{Cov}\left(C_{ij}^s, \nu I_{ij}\right)\right\}}{\sum_{i,j}\mathbb{E}\left[\left(C_{ij}^s - \nu I_{ij}\right)^2\right]}.\end{aligned} \tag{3}$$

Replacing the expectations with sample estimates yields an analytical formula for an estimator $\hat{\lambda}$, being highly favorable as model selection through cross validation is not required.

B. Naïve approaches to exploit subclass-specific information for LDA

When performing the classification task for ERP data, one has access to additional label information, which is not exploited by the standard approach: the stimulus identity is always known. The label which specifies the exact stimulus identity will be denoted *sublabel* g in the following. In the example of an auditory BCI with k different stimuli, the sublabel $g_i \in 1 \dots k$ specifies which auditory stimulus was presented. As stimuli may differ in pitch, direction or intensity, it is highly plausible that those differences lead to subclass-specific features in the ERPs. Those subclass-specific features may also impact the classification performance, such that the classifier accuracy could be improved if the sublabels are considered for classification.
The straight-forward way to extract such subclass-specific information is to split the training data and compute a classifier solely on the subclass-specific data. This approach, called "naïve mean-Cov subclass LDA (*naïve-mCsLDA*)" might however be suffering from the highly reduced number of data points. Estimates for the means and especially for the covariance might become inaccurate.
Under the assumption that the covariance of the data reflects the background noise in the EEG, it is reasonable to assume that the covariance $\boldsymbol{C}$ does not contain subclass-specific information. Thus, another straight-forward approach for a subclass-specific LDA classifier is to estimate a subclass-specific $\boldsymbol{\mu}$, while computing $\boldsymbol{C}$ pooled across all subclasses. The resulting classifier will be called "naïve mean subclass LDA (*naïve-msLDA*)".

C. Regularization of subclass mean towards global mean

In order to obtain a more robust estimator for the subclass mean one can regularize the sample estimator towards the mean of all other subclasses. Thus, one can define the regularized estimator for class i and subclass g by

$$\hat{\boldsymbol{\mu}}_{i,g}^{reg} = (1-\lambda)\hat{\boldsymbol{\mu}}_{i,g}^{s} + \lambda\hat{\boldsymbol{\mu}}_{i,\bar{g}}^{s} \tag{4}$$

with $\hat{\boldsymbol{\mu}}_{i,\bar{g}}^{s}$ denoting the sample mean of class i (e.g. targets), while excluding the datapoints from subclass g.

Figure 1 illustrates how the estimation of the mean directly impacts the LDA separation hyperplane. A binary classification task with four subclasses (marked by different symbols) is shown. Estimates for $\hat{\boldsymbol{\mu}}_{i,g}^{s}$ and $\hat{\boldsymbol{\mu}}_{i,\bar{g}}^{s}$ are marked in bold. Disregarding all subclass label information corresponds to a fixed value for λ defined by one minus the ratio of data points in subclass g and all other subclasses, i.e. $\lambda_0 = 1 - \frac{n_g}{n_{\bar{g}}}$. For the dataset depicted in Figure 1A, the global classwise sample mean ($\lambda_0 = 0.75$) is depicted with a star. Figure 1B depicts the LDA separation hyperplanes, when using $\hat{\boldsymbol{\mu}}_{i,g}^{s}$ (dashed bold line), $\hat{\boldsymbol{\mu}}_{i,\bar{g}}^{s}$ (solid bold line) or $\hat{\boldsymbol{\mu}}_{i}^{s}$ (narrow dashed line). The exact choice of the regularization parameter λ determines where the mean estimator is located on the line between $\hat{\boldsymbol{\mu}}_{i,g}^{s}$ and $\hat{\boldsymbol{\mu}}_{i,\bar{g}}^{s}$. In order to not downweight the impact of the subclass-specific data, it is reasonable to constrain the regularization parameter to be upper-bounded with $\lambda^* \leq \lambda_0$, as it is also visualized with the shaded areas in Figure 1B.

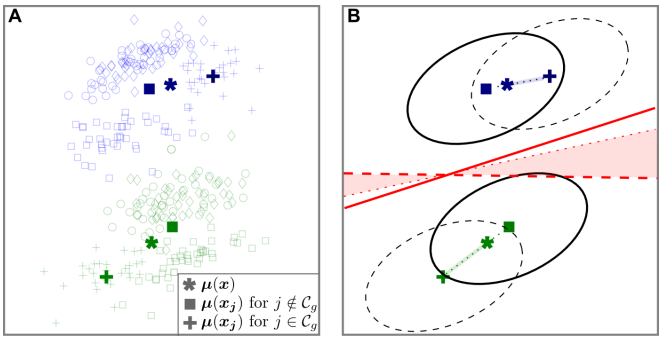

Fig. 1. Example for a binary classification task with subclasses. Plot A shows the distribution of datapoints with the color/symbol specifying the class/subclass respectively. The means are shown in bold. Plot B depicts the means and the covariance and the resulting LDA separation hyperplanes for the three mean estimates. The shaded area denotes the range of hyperplanes when regularizing between the subclass mean and the global mean.

The LDA classifier which is computed with a regularized mean estimator is called "regularized subclass mean LDA (*regsmLDA*)" in the following, while the parameters λ need to be estimated by cross-validation. This method is computationally highly inefficient, as each subclass mean might require its individual λ, resulting in $n_{subclasses} \times n_{classes}$ (e.g. 6×2) parameters which have to be estimated with cross validation. In order to reduce computational load, it is assumed that each subclass has the same parameter $\lambda \in \{0, 0.1, 0.2 ... 1\}$, which results in 2 parameters to be chosen.

A closely related approach was presented in [7]. Their approach also aims to extract subclass specific information in a BCI experiment, by artificially replicating the training data of the corresponding subclass. This can be formulated such that an optimized training data set $\mathcal{X}_g^{opt}$ is determined, which consists of the original data $\mathcal{X}^{orig}$ being artificially enriched with subclass-specific data $\mathcal{X}_g$,

$$\mathcal{X}_g^{opt} = \mathcal{X}^{orig} \underbrace{\cup \mathcal{X}_g \cup \cdots}_{\alpha - 1}. \tag{5}$$

This approach (called "alpha upweight subclass LDA (*ausLDA*)") is analog to the regularization approach as $\alpha = 1 \Rightarrow \lambda = \lambda_0$ and $\alpha \to \infty \Rightarrow \lambda \to 0$. It should be noted that *ausLDA* results in subclass-specific estimators for the covariance and also the mean. Moreover, the regularization strength for all classes i (targets/non-targets) is always equal. In order to reduce computational workload, this method was also implemented such that each subclass had the same $\alpha \in \{1..10\}$. Thus, there was only one parameter to be estimated by cross-validation.

D. Mean shrinkage

As for the covariance matrix, shrinkage allows for improved estimation of the mean with respect to expected mean squared error. James-Stein Shrinkage [10] yields an estimator

for the optimal shrinkage intensity in eq. (4),

$$\hat{\lambda}^{JS} = \frac{\sum_d \widehat{\mathrm{Var}}(\hat{\mu}_d^s)}{\sum_d \|\hat{\mu}_d^s - \hat{\mu}_d^t\|^2}, \tag{6}$$

with $\hat{\mu}_d^s$ being the sample mean in feature dimension d and $\hat{\mu}_d^t$ the corresponding shrinkage target.

The advantage of the shrinkage approach is that the optimal shrinkage strength can be calculated with very low computational cost. Thus, with the shrinkage approach there is no need to perform expensive cross-validation.

E. Subclass LDA with mean shrinkage

When computing subclass-specific LDA classifiers, the mean of all other subclasses resembles are reasonable shrinkage target for a subclass LDA classifier, as already described in eq. (4). High-variance directions tend to dominate the estimation of the shrinkage strength [11]. In order to downweight the impact of high-variance directions, data were whitened before applying shrinkage. An LDA classifier can then be computed with the shrinkage mean estimator, which resembles a weighted average between the sample subclass mean and the remaining subclasses. This is done for each subclass, resulting in $n_{subclasses} \times n_{classes}$ (e.g. $6 \times 2 = 12$) parameters. The resulting classifier is denoted "weighted shrinkage mean subclass LDA (*wsmsLDA*)".

F. Overview of classifiers in this study

All classifiers which were implemented for this study are listed in Table I.

TABLE I. LIST OF CLASSIFIERS AND THEIR SHORT DESCRIPTION.

stdLDA	LDA classifier with covariance shrinkage estimation (shrC); sublabels are disregarded.
naïve-mCsLDA	This classifier (shrC) is trained only on subclass specific data.
naïve-msLDA	The mean is computed on subclass specific data and shrC is done based on data from all subclasses.
regsmLDA	A weighted/regularized mean is computed for subclass specific data and shrC is done based on all subclasses. Regularization parameters λ are estimated with cross-validation, using data from remaining subclasses as regularization target.
ausLDA	This classifier is trained on manipulated data, in which the subclass specific data were artificially replicated. The weighting parameter α is chosen by cross validation.
wsmsLDA	A novel Classifier, which is based on a regularized mean for each subclass. Regularization parameters λ were estimated by mean-shrinkage using data from remaining subclasses as shrinkage target. The shrC is calculated based on data from all subclasses.

G. Evaluation data and preprocessing

To evaluate the novel classification approaches on real data, two ERP datasets were reanalyzed with the classification methods described above. For both data sets, the calibration data was analyzed only. Each dataset presented specific characteristics, as they were differing in the stimulus modality as well as in the number of trials and subjects – see Table II for details. For feature extraction, a widely used "subsampling approach" was taken [14], [15]: the EEG data were first epoched [-150 800] ms after stimulus onset and baselined between [-150 0]. EEG epochs containing eye artifacts were excluded by an heuristic, cf. [14]. Then, for each channel the mean amplitude value was computed in a fixed set of 10 intervals. Those intervals had a length of 40-60 ms and they were densely placed between 200 ms and 650 ms after stimulus onset. It should be noted that the global selection of such intervals circumvents any additional parameter selection, while the feature space becomes high-dimensional (e.g. 63 channels $\times$ 10 intervals = 630 dimensional feature space).

TABLE II. DETAILS OF THE TWO DATA SETS WHICH WERE REANALYZED TO EVALUATE THE CLASSIFIERS.

Dataset	AMUSE	HexoSpeller
Modality	auditory	visual
# subclasses	6	6
# subjects	21	13
# epochs	4320	2040
Reference	[12]	[13]

Based on those features, the classification accuracy was estimated with a 5-fold cross validation (with 2 repetitions) while the classifier weights and all additional parameters were solely estimated on the training data. For *regsmLDA* and *ausLDA*, a nested cross-validation was performed as additional regularization parameters needed to be selected with an inner 3-fold cross-validation.
Classification accuracy was assessed with $acc = 1 - AUC$, with AUC being the area under the ROC curve.

III. RESULTS

Figure 2 depicts the results of this study with scatter-plots. The estimated classification accuracy of each method is plotted against *stdLDA* as baseline on the x-axis. It can be seen that both naïve methods performed significantly worse than the *stdLDA* approach. The *regsmLDA* approach performed significantly worse than the standard approach for the HexoSpeller data while for the AMUSE data, several subjects benefit from the subclass-specific features. However, for the majority of subjects (62%), *regsmLDA* still underperformed *stdLDA*. The remaining two approaches which exploit subclass-specific features (*ausLDA* and *wsmsLDA*) could outperform the standard approach *stdLDA* significantly for the AMUSE data. For the HexoSpeller data, significance could not be found, which indicates that the sublabel information of the HexoSpeller data might not contain discriminant features.

IV. DISCUSSION

Linear discriminant analysis (LDA) is the most commonly used classification method for single trial ERP data in the BCI framework. The popularity of LDA arises from its robustness, simplicity and high accuracy. This work aims to improve binary linear classification approaches by exploiting subclass-specific features. Several novel methods were introduced and applied on two existing data sets.
It was found that one can improve upon the standard LDA approach by using regularized estimates of the subclass-specific mean. Regularization towards other subclasses was essential, as the sample estimates of the subclass means lead

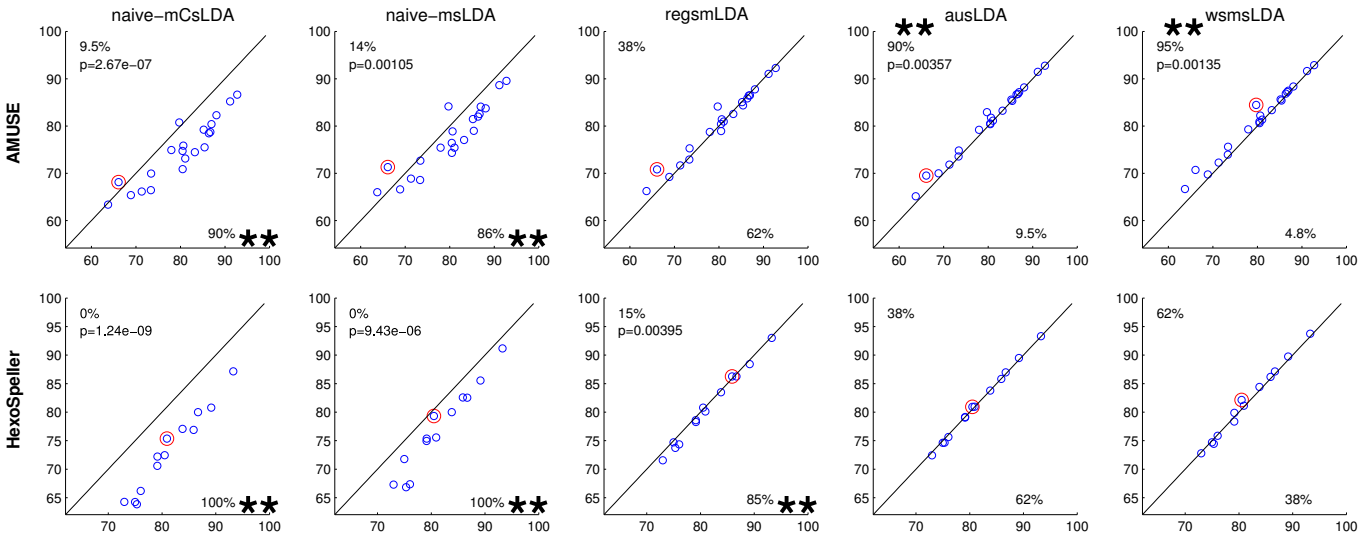

Fig. 2. Classification performances of subclass LDA methods. Each scatter plot shows the accuracies of the corresponding subclass-specific LDA method (y-axis) against *stdLDA* (x-axis). A circle corresponds to one subject. Significant differences ($p < 0.01$) are marked with **. The subject that is most benefiting from the corresponding subclass-specific LDA method is highlighted in red. Two datasets were analyzed: AMUSE (first row) and HexoSpeller (second row).

to a worsening of the classification performance - see Figure 2, *naïve-mCsLDA* & *naïve-msLDA*. However, determining suitable regularization parameters λ_i^g by cross-validation was computationally highly inefficient, as numerous parameters (≥ 10) had to be estimated. Therefore these approaches are either not applicable in practice, or simplifications have to be made. For this analysis, several simplifications were made and *regsmLDA* was trained on a sparse grid (10×10) of possible parameters. With those simplifications, *regsmLDA* underperforms *stdLDA*. Without such simplifications and under the assumptions of unlimited time and computing power, one can expect *regsmLDA* to perform at least as good as *wsmsLDA*.

In *wsmsLDA*, regularization parameters were estimated by shrinkage, which serves a computationally efficient, analytical expression for each λ_i^g. The *wsmsLDA* classifier outperformed the standard approach and all other methods that require model selection through cross-validation. While being computationally highly efficient, the results showed that *wsmsLDA* performed at least as good as *stdLDA* for each subject.

To conclude, novel classification algorithms were described which exploit subclass label information. The method *wsmsLDA* could improve the classification accuracy of ERP signals, using a shrinkage mean estimator. Even if subclass-specific features are not present in the data, *wsmsLDA* performs equal to the standard approach - cf. HexoSpeller data. This makes *wsmsLDA* a good candidate for ERP classifiers which can be applied for various BCI paradigms.

References

[1] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, Eds., *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[2] J. R. Wolpaw and E. W. Wolpaw, Eds., *Brain-computer interfaces : principles and practice*. Oxford University press, 2012.

[3] B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L. E. Ramsey, I. Sturm, G. Curio, and K.-R. Müller, "The Berlin Brain-Computer Interface: Non-medical uses of BCI technology," *Frontiers in Neuroscience*, vol. 4, p. 198, 2010, open Access.

[4] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, pp. R1–R13, Jun 2007.

[5] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, pp. 387–399, 2011.

[6] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "A novel 9-class auditory ERP paradigm driving a predictive text entry system," *Frontiers in Neuroscience*, vol. 5, p. 99, 2011.

[7] Y. Matsumoto, S. Makino, K. Mori, and T. M. Rutkowski, "Classifying P300 responses to vowel stimuli for auditory brain-computer interface," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, 2013, pp. 1–5.

[8] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial," *NeuroImage*, vol. 56, pp. 814–825, 2011.

[9] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, pp. 365–411, 2004.

[10] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 1961, 1961, pp. 361–379.

[11] D. Bartz and K.-R. Müller, "Generalizing analytic shrinkage for arbitrary covariance structures," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 1869–1877.

[12] M. Schreuder, T. Rost, and M. Tangermann, "Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI," *Frontiers in Neuroscience*, vol. 5, no. 112, 2011.

[13] M. S. Treder and B. Blankertz, "(C)overt attention and visual speller design in an ERP-based brain-computer interface," *Behavioral and Brain Functions*, vol. 6, p. 28, May 2010.

[14] J. Höhne, K. Krenzlin, S. Dähne, and M. Tangermann, "Natural stimuli improve auditory BCIs with respect to ergonomics and performance," *Journal of Neural Engineering*, vol. 9, no. 4, p. 045003, 2012.

[15] P.-J. Kindermans, D. Verstraeten, and B. Schrauwen, "A bayesian model for exploiting application constraints to enable unsupervised training of a p300-based bci," *PloS ONE*, vol. 7, no. 4, p. e33758, 2012.