A supervised recalibration protocol for unbiased BCI

S. Perdikis[1], M. Tavella[1], R. Leeb[1], R. Chavarriaga[1], J. d. R. Millán[1]

[1]Chair in Non-Invasive Brain-Machine Interface, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

serafeim.perdikis@epfl.ch

Abstract

One important source of performance degradation in BCIs is bias towards one of the mental classes. Recent literature has focused on the general problem of classification accuracy drop, identifying non-stationarity as the generating factor, thus leading to several classifier adaptation approaches suggested as of today. In this work, we explicitly focus on bias elimination, demonstrating that the problem has two separate components, one related to non-stationarity and another one attributed to the nature of the feature distributions and the assumptions made by the classification methods. We propose a cued recalibration protocol including a supervised adaptation method and a novel framework for unbiased classification with a modified, unbiased Linear Discriminant Analysis classifier. Preliminary results show that our protocol can assist the subject to achieve quickly accurate and unbiased control of the BCI.

1 Introduction

Classification bias has proved to be a major problem in Brain-Computer Interfaces (BCIs), hindering user training and obstructing BCI operation, since one mental command can be heavily favored over the other, in which case the latter might often become unusable. However, bias elimination has received little attention per se so far, since in recent literature bias emergence has been only treated as part of the more general problem of accuracy degradation. Therefore, bias has been solely attributed to non-stationarity and thought to be largely eliminated by online adaptation of the classifier parameters [1, 2]. Alternatively, and although not reported in literature, the common code of practice in most labs in overcoming a biased classifier involves either a quick re-training session or "manual" adaptation of the classifier hyperplane.

Biasing effects become most prominent at the transition from the calibration (no feedback) to the online (BCI feedback) phase or in between consecutive online sessions, proving that non-stationarity accounts for a large component of the problem. Nevertheless, we discuss here other potential sources of bias and present a unified approach to tackle classification bias.

In this work, we present a novel method for supervised, adaptive estimation of Loss function parameters [3] leading to an unbiased Linear Discriminant Analysis (LDA) classifier. This supervised scheme will be applied in a cued recalibration protocol interleaved between the offline (calibration) phase and online operation of the BCI, thus achieving both classifier adaptation and explicit bias elimination for improved consecutive BCI experience.

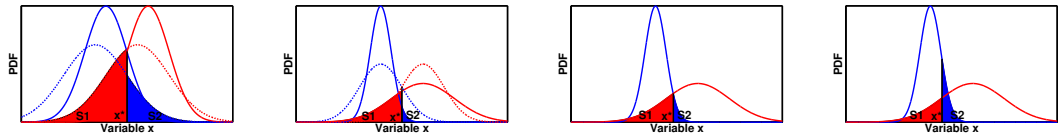
2 Methods

2.1 Motivation

Classification bias is evident from the confusion matrix of a BCI experiment where the per class accuracy may be significantly different, even for classifiers achieving high total accuracy. Such a biasing effect can occur for a variety of reasons. During online BCI operation bias can appear due to the violation of the stationarity assumption. In such case, the class distributions estimated on the training set and used to define the classifier's decision rule do not reflect any more the

class distributions currently generated by the subject (Figure 1(a)), thus introducing bias and general accuracy degradation. Non-stationarity is a known problem in BCI and several adaptive parameter estimation approaches have been proposed to eliminate its effects [1, 2], thus implicitly coping with this source of bias.

Nevertheless, classification bias can still persist, since there are additional reasons that may contribute to its appearance. These reasons include: (*i*) violation of the basic LDA assumption (when LDA is used) of identical covariance matrices for the two classes (Figure 1(b)), and (*ii*) the *inherent* bias emerging when classes are normally distributed with covariance matrices that are significantly different (Figure 1(c)). In the latter case, even a non-linear, quadratic classifier is not guaranteed to eliminate bias, since bias alleviation is not explicitly treated by Bayesian classifiers.



(a) Non-stationarity bias (b) LDA assumption bias (c) Inherent bias (d) Unbiased decision rule

Figure 1: Sources of bias for the case of a single feature and assuming equal prior probabilities without loss of generality. Dotted lines correspond to the sample distributions either estimated from the training set (a) or by assuming identical covariance matrices for the classes (b). Solid lines correspond to the actual sample distributions. Colored areas S_1, S_2 represent the error of the corresponding class. Potential bias is evident comparing the sizes of S_1, S_2.

The basic idea implemented in this work concerns a two-step algorithm, where in the first step a supervised LDA classifier adaptation technique is proposed to alleviate the non-stationarity-related bias and, in the second step, the LDA hyperplane constant term is further adjusted to eliminate the inherent- and assumption violation-related bias that might occur (Figure 1(d)).

2.2 Algorithm

Step 1 - Supervised adaptation and shrinkage: A supervised adaptation framework is employed, where for each incoming EEG sample $\mathbf{x}_t$ at time t the class mean vector $\boldsymbol{\mu}_i^t$ and covariance matrix $\boldsymbol{\Sigma}_i^t$ (where i the class that $\mathbf{x}_t$ belongs to) are iteratively estimated in the Maximum Likelihood (ML) approach as $\boldsymbol{\mu}_i^{t_i} = \boldsymbol{\mu}_i^{t_i-1} + \frac{1}{t_i+1}(\mathbf{x}^t - \boldsymbol{\mu}_i^{t_i-1})$ and $\boldsymbol{\Sigma}_i^{t_i} = \frac{t_i-1}{t_i}\boldsymbol{\Sigma}_i^{t_i-1} + \frac{1}{t_i}(\mathbf{x}^t - \boldsymbol{\mu}_i^{t_i-1})^T(\mathbf{x}^t - \boldsymbol{\mu}_i^{t_i-1})$ respectively, while the parameters of the other class retain their previous values. This straightforward supervised approach is possible, since data samples are acquired in a cued protocol where data labels are known. The "global" covariance matrix is also calculated at each step t as $\boldsymbol{\Sigma}^t = (t_1\boldsymbol{\Sigma}_1^t + t_2\boldsymbol{\Sigma}_2^t)/t$, $t = t_1 + t_2$, where iterators t_1, t_2 are only incremented when the sample at t belongs to the respective class i.

Parameters t_1, t_2 could generally be set to the respective sizes of the classes in the training set. However, we set this values to $t_i^0 = 1000$ in order to adapt faster to the distributions of the online session and recover quickly from any non-stationarity effect. Since for low values of t the ML estimates are known to be inaccurate and sensitive to outliers, we also employ an analytical covariance shrinkage method [4] to estimate the final covariance matrices $\widehat{\boldsymbol{\Sigma}_1^t}, \widehat{\boldsymbol{\Sigma}_2^t}, \widehat{\boldsymbol{\Sigma}}^t$. Shrinkage also avoids singularity problems. Then, we derive the conventional LDA hyperplane using a 0-1 Loss function [1] L_{0-1} at time t as $\mathbf{w_t}^T\mathbf{x} + b_{0-1,t} = 0$, where $\mathbf{w_t} = \widehat{\boldsymbol{\Sigma}^t}^{-1}(\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_2^t)$ and $b_{0-1,t} = -\frac{1}{2}(\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_2^t)^T\widehat{\boldsymbol{\Sigma}^t}^{-1}(\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_2^t) + ln(\frac{P(\omega_1)}{P(\omega_2)})$.

Step 2 - Hyperplane adjustment for unbiased classification: The main novelty of our approach lies on the introduction of a second adaptation step, aimed to alleviate the additional

[1] *0-1* subscript denotes that the quantity in question has been derived with a "0-1" Loss function, [3].

sources of bias based on the accurate estimation of the class distributions by the previous step. Figure 1(d) shows the basic idea consisting in finding a decision rule that "predicts" equal error rates for both classes, namely $P_e(c_1) = P_e(c_2)$.

By constraining our problem to linear decision rules $\mathbf{w_t'}^T\mathbf{x} + b_t = 0$ whose hyperplane is parallel to the one found by L_{0-1} LDA, $\mathbf{w_t'} = \mathbf{w_t}$, the problem reduces to a single degree of freedom independently of the dimension of the feature space. Intuitively, we wish to estimate the bias term b_t of the new linear rule that will lead to theoretically equal error rates, thus operating our LDA classifier in a different point on the ROC curve than that found by conventional LDA.

The geometrical interpretation of the above demand satisfies that the hyper-volumes $P_e(c_i) = \int\int\ldots\int_{D_i} N(\mathbf{x}, \mu_i, \mathbf{\Sigma_i})\, d\mathbf{x}$, $D_i : sgn(c_i)(\mathbf{w^T x} + b) > 0$, are equal. The solution b_{unbias} is the zero of the function $f(b) = P_e(c_1) - P_e(c_2)$. This complex equation can be significantly simplified by considering a rotation $\mathbf{R}$ of the n-dimensional coordinate system of the feature space, such that the first dimension of the rotated space is parallel to the normal vector $\mathbf{w_t}$, in which case it is easy to show that $f(b) = \frac{1}{2}\,\mathrm{er}(\frac{b-m_1}{\sqrt{2\sigma_1^2}}) + \frac{1}{2}\,\mathrm{er}(\frac{b-m_2}{\sqrt{2\sigma_2^2}})$, where $m_i = (\boldsymbol{R}\mu_i)_1$, $\sigma_i^2 = (\mathbf{R}^T\mathbf{\Sigma_i}\mathbf{R})_{11}$ and er is the error function (proof omitted due to lack of space). Solving the last non-linear equation is possible by means of the Taylor approximation of er and polynomial root identification through the companion matrix formation, so that finally the only real root is a very close approximation of the desired bias term b_{unbias}.

Formally, the obtained solution b_{unbias} defines a Loss function $L_{unbias} = \begin{vmatrix} 0 & e^{b_{0-1}-b_{unbias}} \\ 1 & 0 \end{vmatrix}$ allowing our linear classifier to operate on the point that ideally produces equal error rates for both classes when the normal distribution assumptions hold, and thus zero bias. It is also worth to note that all necessary operations are simple enough to allow online implementation of the algorithm even in MATLAB for a BCI working at 16 Hz.

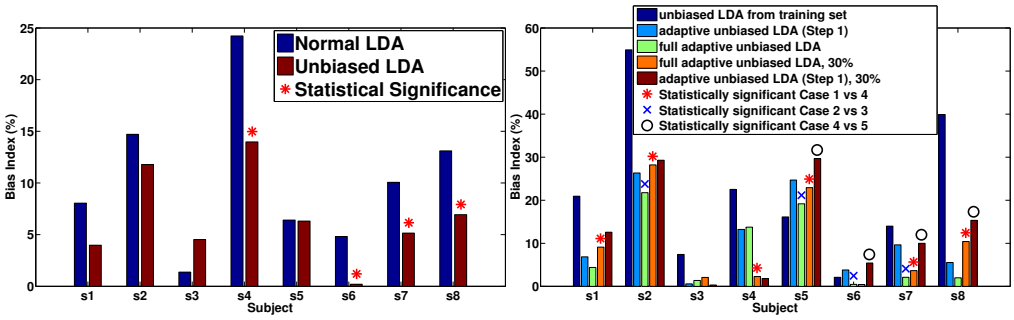
3 Results

In order to evaluate the effects of the extra bias factors we have identified and the effectiveness of the unbiased LDA framework in alleviating them, we compute the *Bias Index* $BI = \left|\frac{a_1}{k_1} - \frac{a_2}{k_2}\right|$, where a_i the number of correctly classified samples and k_i the total number of samples of class i. An ideal, unbiased classifier should have $BI = 0$.

The first comparison is on a calibration session dataset of 8 subjects with a cued Motor Imagery (MI) protocol consisting of at least 120, each 5 sec long trials for each subject. BI is is calculated on the training set, thus largely excluding non-stationarity related bias. In this case we compare our unbiased LDA approach to the normal LDA.

The results illustrated in Figure 2(a) show the existence of additional bias factors, as well as the ability of the proposed method to largely eliminate them. The fact that BI does not reach 0, as theoretically predicted, as well as the exception of subject s_3, are attributed to the fact that the assumption of normally distributed features does not absolutely hold. In the same experiment total accuracy was not affected, since the maximum reported difference was found to be less than 1% across all subjects and not statistically significant.

The ability of the proposed unbiased LDA framework to reduce bias is further demonstrated on extra (at least 120) MI trials executed the same day as the above calibration session for each subject. In this case, online feedback was driven by a Gaussian classifier and variable-length trials would end when a decision threshold on "accumulated" posterior probabilities was reached. The comparison on this dataset is done among the following variations of our unbiased LDA method: (*i*) unbiased LDA derived from the training set (calibration session) as above, (*ii*) adaptive unbiased LDA (Step 1 only) running over the whole dataset, (*iii*) full adaptive unbiased LDA (Steps 1 & 2) running over the whole dataset, (*iv*) full adaptive unbiased LDA running over the first 30% of the dataset and stopped afterwards, and (*v*) adaptive unbiased LDA (Step 1 only) running over the first 30% of the dataset and stopped afterwards. The last two cases are meant to evaluate the expected bias after the proposed re-calibration protocol has finished, where classifier (*iv*) would be the outcome of the proposed protocol. BI was calculated on the last 70% of the dataset.



(a) Calibration session MI data (b) Online feedback session MI data

Figure 2: Bias elimination through adaptive unbiased LDA on MI EEG data.

Figure 2(b) shows the overall potential of our framework, since for all subjects the proposed unbiased LDA (4^{th} bar) achieves a lower BI than the original unbiased LDA derived from the training set (1^{st} bar) for all subjects except s_5, who had a non-stationarity effect that made feature distributions change abruptly after adaptation was switched off. Differences are statistically significant for 5/8 subjects. Furthermore, although not shown in the figure due to space limitations, the unbiased LDA derived from the training set has a lower BI than the normal LDA for 7/8 subjects, 6 statistically significant. Statistical significance is computed at 99% confidence interval. Additionally, it is verified that the extra "unbiasing" procedure can assist in further reducing the total bias during the protocol execution (3^{rd} bar lower than the 2^{nd} for 6/8 cases, 4 statistically significant) as well as after protocol termination (4^{th} bar lower than the 5^{th} for 6/8 cases, 4 statistically significant).

4 Discussion

The overall trends support the utility of the proposed method, while the fact that non-favourable exceptions are not statistically significant proves that in the worst-case scenarios the recalibration protocol will not inflict further bias. Concerning the total accuracy, results (not shown due to lack of space) showed that accuracy can greatly improve when non-stationarity is intense between calibration and feedback trials, otherwise there is no significant improvement. It should also be mentioned that our method can equivalently be applied in other types of features or problems.

Ongoing and future work entails online experiments under this protocol, where it can be hoped that the mutual learning procedure when the protocol directly drives the feedback can further improve performance. We will also explore an unsupervised version of the unbiased framework.

Acknowledgements

This work is supported by the European ICT Programme Project FP7-224631 (TOBI).

References

[1] J. d. R. Millán. On the Need for On-Line Learning in Brain-Computer Interfaces. In *Proceedings of the International Joint Conference on Neural Networks*, 2004.

[2] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz. Machine-learning based co-adaptive calibration. *Neural Comput*, 23(3):791–816, 2011.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, November 2001.

[4] Y. Chen, A. Wiesel, Y.C. Eldar, and A.O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Trans Signal Process*, 58(10):5016–5029, 2010.