ONLINE CLASSIFICATION ACCURACY IS A POOR METRIC TO STUDY MENTAL IMAGERY-BASED BCI USER LEARNING: AN EXPERIMENTAL DEMONSTRATION AND NEW METRICS

F. Lotte[1], C. Jeunet[1,2,3]

[1]Inria Bordeaux Sud-Ouest / LaBRI, Talence, France
[2]Inria Rennes Bretagne Atlantique / Irisa, Rennes, France
[3]EPFL, Geneva, Switzerland

E-mail: fabien.lotte@inria.fr

ABSTRACT: While promising for many applications, Electroencephalography (EEG)-based Brain-Computer Interfaces (BCIs) are still scarcely used outside laboratories, due to a poor reliability. It is thus necessary to study and fix this reliability issue. Doing so requires to use appropriate reliability metrics to quantify both signal processing and user learning performances. So far, Classification Accuracy (CA) is the typical metric used for both aspects. However, we argue in this paper that CA is a poor metric to study how well users are learning to use the BCI. Indeed CA is notably unspecific, discrete, training data and classifier dependent, and as such may not always reflect successful EEG pattern self-modulation by the user. We thus propose new performance metrics to specifically measure how distinct and stable the EEG patterns produced by the user are. By re-analyzing EEG data with these metrics, we indeed confirm that CA may hide some learning effects or hide the user inability to self-modulate a given EEG pattern.

INTRODUCTION

While they are very promising for numerous applications, such as assistive technology or gaming, Electroencephalography (EEG)-based Brain-Computer Interfaces (BCIs) are still scarcely used outside laboratories [1]. This is mostly due to their poor reliability, as they often recognize erroneous mental commands from the user. One of the main current challenges of the community is thus to improve BCI reliability [1]. This is currently addressed at different levels, such as trying to design more robust EEG signal processing algorithms, or trying to improve BCI user training approaches, which have been shown to be inappropriate and a major cause of poor performances, both in theory and in practice [1, 3, 8]. Improving these different aspects requires to measure this reliability and thus BCI performances. Indeed such performance metrics could identify what are the limitations of a given algorithm or training approach, which is a necessary first step towards fixing these limitations [1].

User performance metrics are particularly useful for studying and improving Mental Imagery (MI) BCI user training. Appropriate performance metrics could indeed help to understand what users have successfully learned or still need to improve, which can then be used to guide them, i.e., to provide them with appropriate training tasks and feedback. In EEG-based BCI, the most used metric is Classification Accuracy (CA), i.e., the percentage of mental commands that were correctly recognized by the BCI [11, 12]. CA, together with other machine learning evaluation metrics [11, 12], have been successfully used to quantify the decoding performance of the BCI, i.e., how well the BCI recognizes the users' commands. However, CA is also used to study BCI user learning, i.e., how well users can modulate/self-regulate their EEG signals to control the BCI, and how much they learn to do so. For instance, CA is typically used to study how different feedbacks impact BCI user training [5], or how different psychological factors impact BCI user learning and performances [4].

In this paper, we argue and demonstrate that CA alone, as used in online MI-BCI, is not enough to study BCI user learning and performances. Indeed, this metric is notably unspecific, discrete as well as classifier and training data dependent, among other. In order to fully understand BCI user skill acquisition, alternative or additional metrics are thus necessary. Therefore, in this paper, we also propose new, simple and computationally efficient metrics to quantify various aspects of MI-BCI skill acquisition and compare them with the classically used online CA. We show that using online (or simulated online) CA as metric may actually hide several relevant aspects of BCI skill acquisition. In particular, online CA may miss user learning effects or fail to identify that a mental task performed is actually no different than rest EEG. Our new metrics can overcome these limitations.

This paper is organized as follows: The next section presents the material and methods, and notably how online CA is measured, and what its limitations are, as well as the new metrics we propose. It also presents the data set on which these measures are compared. Then the Results section compares the performances estimated with all metrics, which are then discussed in the Discussion section. The last section concludes the paper.

MATERIALS AND METHODS

Classification accuracy and its limitations

As indicated before, CA is the most used metric to quantify BCI performances. Typically, the classifier is trained on the EEG signals from the trials of the first BCI runs (calibration runs) and applied to classify the users' EEG signals from the trials of the subsequent runs. CA is defined as the percentage of these EEG trials that were correctly classified [11]. From the classification results, we can also obtain a more detailed information on the performances from the Confusion Matrix (CM), which informs about how many trials from each class were estimated to be from each one of the possible classes. The CM is defined as follows for a two class problem:

Table 1: Confusion matrix for two classes

		Estimated class	
		Class 1	Class 2
Real Class	Class 1	a	b
	Class 2	c	d

Here, the number in row i, column j is the number of trials from class i that was classified as belonging to class j. Thus, a and d correspond to correct classifications (the real and estimated classes are the same), and c and b to erroneous classifications. CA (in %) can thus be computed as $\frac{a+d}{a+b+c+d} \times 100$. From there we can also estimate the CA of each class, e.g., $\frac{a}{a+b} \times 100$ is the percentage of trials from class 1 that were correctly classified.

This CA metric is very useful to quantify the decoding performance of a BCI [11]. However, when it comes to studying how well users can voluntarily modulate their EEG signals to control the BCI, we argue that such metric actually suffers from many limitations.

First, CA is unspecific: it only provides the global performance, but not what is (in)correctly classified, nor why it is so. Then, CA is a discrete measure: a trial is either correctly classified or not, there is no middle ground. As such, even if the user produces a stronger EEG modulation than before, but not strong enough to make the trial correctly classified, CA will not change.

CA is also strongly classifier and training data dependent. Changing the classifier type, its parameters, or the amount and/or quality of the training data will change the CA, irrespectively of how well users can modulate their EEG activity. Therefore variations of CA might not always reflect users' proficiency at BCI control. Classifiers are also sensitive to non-stationarities, and thus would lead to poor CA when applied on EEG data from a different distribution than that of the calibration runs. This is likely to happen if users are trying out various strategies or are learning. When based on a discriminative classifier such as Linear Discriminant Analysis (LDA), the most used classifiers for BCI [1], CA does not reflect how well a given mental command can be recognized but rather how distinct the mental commands are from each other. Therefore, if users are unable to modulate their EEG signals for one class (e.g., left hand MI), they may still obtain very high CA as long they can modulate their EEG for the other class (e.g., right hand MI), since the EEG signals from the two classes are distinct.

This leads to a last limitation: in BCI, CA usually considers the MI EEG signals only, but not the rest EEG signals. As illustrated just before, this prevents us from identifying whether the user's EEG patterns during MI are actually any different from rest EEG. For all these reasons, CA may not be able to reveal some important aspects of BCI user performance and BCI user learning, which thus calls for new metrics to quantify these aspects. This is what we propose in the following sections.

New Performance metrics

To address some of the limitations mentioned above, a possible approach (not new in itself but typically not used to study BCI user learning) would be to perform Run-Wise Cross-Validation (RWCV). The idea is to use CV to estimate offline the CA of each run. With RWCV, the trials from the current run are divided into K parts, K-1 parts being used for training the classifier, and the last part for testing it, the process being repeated K times, and the obtained CA averaged over the K testing parts. This thus also provides class-specific CV accuracies, as done with the standard CA. We will assess this approach in this paper. Since training and testing are performed on each run, and for different parts of each run, this makes RWCV CA much less sensitive to training data and to non-stationarities. This metric remains non-specific and discrete though, and still ignores the background EEG. It is also computationally expensive.

To further improve on the metrics mentioned above, we thus need metrics that are also specific, continuous, that consider rest EEG signals and that are computationally cheap. To go towards more specific metrics, we can first consider that the goal of BCI user training is typically defined as to enable users to produce stable and distinct EEG patterns, so that these patterns can then be recognized by the classifier. As such, it would make sense to design a set of metrics dedicated to estimating how stable and distinct the EEG patterns for each MI task actually are. A stable pattern would be a pattern that is not changing dramatically between trials, and thus with a small variance. A distinct pattern would be both 1) a pattern that is distinct from the rest EEG pattern, i.e., there is a specific signature to that pattern and 2) a pattern that is distinct from the EEG patterns of the other MI tasks.

Interestingly enough, metrics quantifying these various properties can be defined using distances in a Riemannian geometry framework. Indeed, Riemannian geometry offers an efficient and simple way to measure distances between covariance matrices, such matrices being increasingly used to represent EEG patterns [2, 14]. Given matrix $X_i \in \mathbb{R}^{N_c \times N_s}$ of EEG signals from trial i, with N_c the number of channels and N_s the number of samples per trial, the covariance matrix C_i of this trial is defined as $C_i = \frac{1}{N_s} X_i^T X_i$, with T being transpose. There-

fore, the diagonal elements of C_i represent the EEG band power for each channel, and the off-diagonal elements, their covariations. Such spatial covariance matrices are used - implicitly or explicitly - to represent EEG signals in numerous MI-BCI designs, notably those based on the Common Spatial Patterns (CSP) algorithm, and many others [9, 14]. The Riemannian distance $\delta_R(C_i, C_j)$ between covariance matrices C_i and C_j can be defined as:

$$\delta_R(C_i, C_j) = [\sum_{i=1}^{n} \log(\lambda_i)^2]^{1/2} \quad (1)$$

where the λ_i are the eigen values of $C_i^{-1}C_j$. This Riemannian distance is particularly interesting since it is affine invariant: it is invariant to linear transformations, i.e., to variations such as normalization or channel displacement [14]. As such, the Riemannian distance has been used successfully for robust EEG signal decoding, in various kinds of BCIs [14]. In this paper, we show that this distance can also be a very relevant tool to quantify how distinct and stable the EEG patterns produced by a BCI user are. In particular, how distinct the EEG patterns from two tasks are could be quantified using the Riemannian distance between the average covariance matrices for each task. Then, the stability of a given EEG pattern can be defined using the average distance between each trial covariance matrix and the average covariance matrix for this task, which is a form of Riemannian standard deviation [14]. More formally, let us first define the Riemannian mean $\bar{C}$ of a set of covariance matrices C_i [14] as:

$$\bar{C} = \text{argmin}_C \sum_{i=1}^{N} \delta_R^2(C_i, C) \quad (2)$$

and the standard deviation σ_C of a set of matrices C_i as:

$$\sigma_C = \frac{1}{N} \sum_{i=1}^{N} \delta_R(C_i, \bar{C}) \quad (3)$$

From there we propose to define the distinctiveness *classDis* of the EEG patterns from two classes A and B as:

$$classDis(A, B) = \frac{\delta(\bar{C}^A, \bar{C}^B)}{0.5 \times (\sigma_{C^A} + \sigma_{C^B})} \quad (4)$$

where $\bar{C}^K$ and σ_{C^K} are respectively the mean and standard deviation of the covariance matrices from class K. This equation can be seen as the extension of the t-statistic to covariance matrices. Similarly, we propose to define the distinctiveness *restDis* between the EEG patterns from one class and those from the rest state as:

$$restDis(A) = \frac{\delta(\bar{C}^A, \bar{C}^{rest})}{0.5 \times (\sigma_{C^A} + \sigma_{C^{rest}})} \quad (5)$$

where $\bar{C}^{rest}$ and $\sigma_{C^{rest}}$ are respectively the mean and standard deviation of the covariance matrices of the rest EEG. Finally, we can define the stability of the EEG patterns from one MI task as being inversely proportional to the standard deviation of the covariance matrices from that task:

$$classStab(A) = \frac{1}{1 + \sigma_{C^A}} \quad (6)$$

These are simple, intuitive and computationally efficient metrics to quantify some aspects of users skills at BCI control. They are also training data and classifier independent, as well as robust to some non-stationarities given the affine invariance of δ_R. In the following, we compare them offline with CA and RWCV CA.

Data set and evaluation

To compare the performance metrics, we used the motor imagery EEG data from the experiment described in [3]. This data set comprises the EEG signals of 20 BCI-naive participants, who had to learn to do 2 MI-tasks, namely imagining left- and right-hand movements. Participants first had to complete a calibration run, without feedback, followed by 4 feedback runs. Each run was composed of 20 trials for each of the two MI tasks. At the beginning of each trial, a fixation cross was displayed. Then, after 2s, a beep sound occurred. Then, at t = 3s, the instruction appeared as an arrow the direction of which indicates the MI task to be performed, i.e., an arrow pointing left indicated a left hand MI and an arrow pointing right a right hand MI. From t = 3.250s, a feedback was provided for 4s in the shape of a bar the direction of which indicating the mental task that has been recognized and the length of which representing the classifier output.

EEG data were filtered in 8-30 Hz, using a 5th order butterworth filter. For each trial, the MI EEG segment used was the 2s long segment starting 0.5s after the cue (left or right arrow), i.e., from second 3.5 to 5.5 of each trial. For the rest EEG signals, we used the 2s long segment immediately before the cue, i.e., from second 1 to 3 of each trial. For CA and RWCV, we used 3 pairs of Common Spatial Pattern (CSP) spatial filters and a LDA classifier, as in [3]. For the standard (here simulated online) CA, we trained the CSP and LDA on the EEG data from the calibration run and used it to classify the EEG data from the 4 subsequent runs, as in [3]. For the RWCV CA, we used 4-fold CV on each run. For *classDis*, *restDis* and *classStab*, the covariance matrices for each trial were estimated using automatic shrinkage using the method from [7].

RESULTS

Average results

Figure 1 shows the average measures of distinctiveness between classes (MI tasks), i.e., CA, RWCV CA and *classDist*, for each run. CA displays some oscillations in performance, but no global learning effect. On the other hand, both RWCV CA and *classDist* reveal a clear continuous increase in distinctiveness between classes over runs. Despite the high inter-subject variability, the 2-way ANOVA *Metric*Run* (*Metric*: CA, RWCV CA, *classDist*

- transformed to z-score to enable comparisons; *Run*: 2 to 5) for repeated measures showed a significant metric*run interaction [$F(1,19) = 4.432$; $p < 0.05$; $\eta^2 = 0.189$], see also Figure 2. Figure 3 shows the measures of distinctiveness per class (class-wise CA and *restDist*). Here as well, CA does not show any obvious learning, while RWCV CA shows some and *restDist* shows a continuous learning for one of the two classes. The 3-way ANOVA *Metric*Class*Run* (*Metric*: CA, RWCV CA, *classDist* (z-score); *Class*: left- vs. right-hand MI; *Run*: 2 to 5) for repeated measures showed a strong tendency towards a main effect of the metric [$F(1,19) = 3.826$; $p = 0.065$; $\eta^2 = 0.168$] but no metric*class*run interaction [$F(1,19) = 0.195$; $p = 0.664$; $\eta^2 = 0.010$]. Concerning the stability metric (*classStab*, Fig. 4), no clear learning is visible over a single session, at least on average.

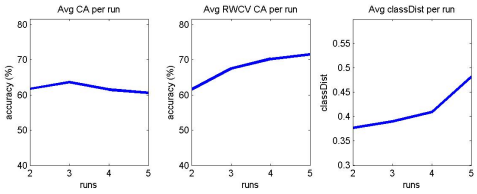

Figure 1: The average measures of distinctiveness between classes, across runs.

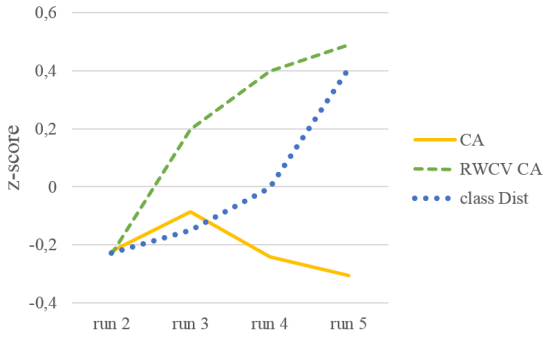

Figure 2: The z-score transformed distinctiveness measures, revealing learning with RWCV and *classDist* only.

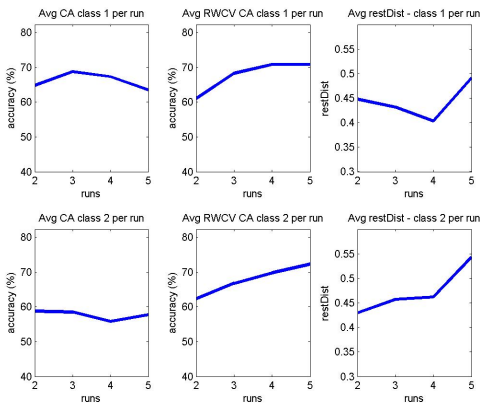

Figure 3: The average measures of class specific distinctiveness, across runs.

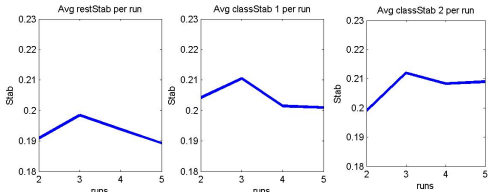

Figure 4: The average measures of stability.

Some subject specific results

As stated earlier, we observed a high inter-subject variability, therefore it is interesting to further investigate the different patterns observed in terms of metrics' evolution across the runs, for individual subjects. It will enable the analysis of the behavior of the different metrics and provide insights on their pros and cons.

For instance, all the distinctiveness measures for subject S5 could reveal a clear learning effect. However, the same metrics for subject S4 did not show any learning effect with the online CA, whereas both RWCV CA and *classDist* revealed clear learning over runs (see Fig. 5). Metrics for subject S9 (Fig. 6) revealed another interesting phenomenon. While both CA and RWCV CA did not show any learning, *classDist* did. However, *restDist* revealed that class 1 actually became increasingly more similar to rest EEG over the runs (*restDist* for class 1 sharply decreased), and thus that the increased *classDist* was probably due to the BCI discriminating rest vs right hand MI rather than left vs right MI. CA cannot identify such a phenomenon since it ignores rest EEG.

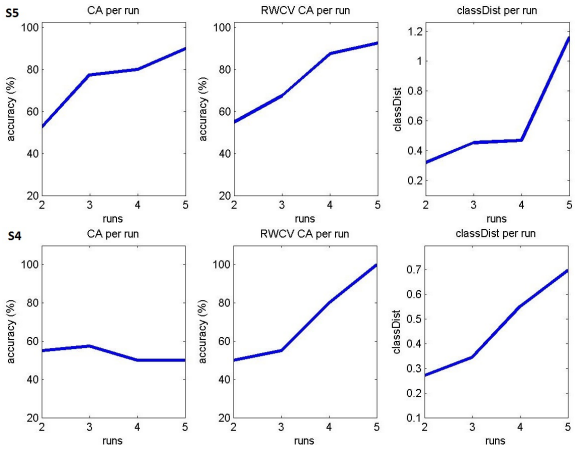

Figure 5: Examples of 2 subjects for which, either CA measured a learning effect like the other metrics (top), or did not whereas the other metrics did (bottom)

Finally, analyzes of Subject 19's data (Fig. 7) showed decreasing class discriminability with CA and *classDist*, however still revealed learning, with *restDist* continuously increasing over runs, for both classes. This could

mean this subject learned to modulate his EEG signals so that they differ from rest EEG, but may have more troubles generating consistently distinct patterns between the two MI tasks. Such phenomenon has also been observed with simultaneous EEG-fMRI in [15], in which some subjects showed modulations of brain activity during MI with respect to rest signals, but no lateralization of the patterns. The *restDist* metric could thus be a cheap and easy way to identify this phenomenon in EEG.

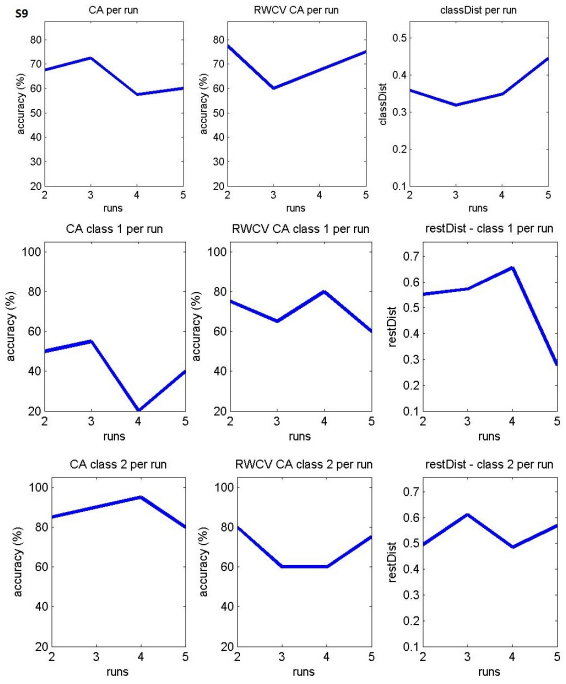

Figure 6: Subject 9, for which class 1 became like rest

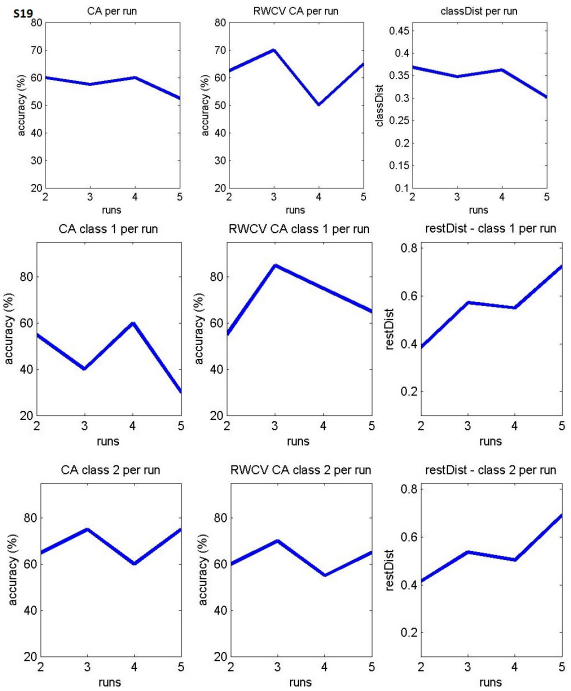

Figure 7: Subject 19 produced EEG patterns increasingly more different than rest, but not distinct from each other.

DISCUSSION

Globally, average results showed a significant metric*run interaction. This suggested that some metrics (here RWCV CA and *classDist*) revealed learning while another (CA) did not. This is all the more interesting given the fact that (1) we considered only one training session, so it is very likely that several subjects did not actually learn over such a short time and (2) the feedback was based on the CA metric. Indeed, participants were asked to make the blue bar feedback, that depended on the CA, as long as possible in the correct direction. Despite such feedback based on a possibly incomplete metric (as shown above), most of the participants showed that they were able to learn to modulate their EEG patterns, sometimes leading to metrics increase. This result is promising for the future as it suggests that with a better feedback, the ability of the participants to learn to modulate efficiently their EEG patterns, in order to improve their BCI control, could be enhanced. On the other hand, these results also suggested that different performance metrics can reveal different aspects of BCI user learning. Notably, they first showed that CA may not always reveal that users have learned to modulate their EEG patterns, whereas metrics such as RWCV CA and *classDist* can reveal such learning. They even revealed fast learning effects in several subjects, with continuous progress over runs, over a single day of training. This can have profound implications for the study of BCI user training. For instance, the present results may explain why in [6], it was concluded that most BCI studies - and notably those based on machine learning - do not actually involve human learning. Indeed, in most of the studies surveyed in [6], CA was used as the performance metric. As such, human learning might have occurred, but CA might not have been able to measure it. It thus seems necessary to re-analyse EEG data from previous studies with complementary performance metrics such as the ones proposed here, to assess whether or not human learning was actually involved. The fast learning over runs revealed by the alternative metrics also stresses the need for co-adaptive BCI systems, and explains the success of these approaches, see, e.g., [13]. Interestingly enough, these works also studied EEG changes due to BCI learning, in a univariate way at each channel, using the R^2 metric. The *restDist* metric also highlighted the need to consider rest EEG when evaluating the BCI user skills. Not doing so may prevent us from realizing that the user is not able to perform one of the MI tasks, and should thus probably be specifically trained to do so.

CONCLUSION

In this paper, we argued that CA, the most used metric to quantify BCI performance, should not be used alone to

study BCI user learning. We indeed identified many limitations of CA for this purpose and proposed new metrics, based on Riemannian distance, to do so. An evaluation of these metrics indeed confirmed that online CA may hide some learning effects and cannot identify how different an MI class is from rest EEG. We therefore conclude that, when studying user learning and performance, CA should be used with care, and complemented with metrics such as the ones proposed.

Naturally, this study needs to be extended by assessing these metrics on other data sets, as well as across several sessions, to measure long-term learning as well. Nonetheless, this study and metrics open many promising perspectives. In particular it would be interesting to re-analyze the relationship between users' profile, notably neurophysiological, personality and cognitive profile, and these new performance metrics (so far done by looking for correlation with CA only [4]), which could reveal new predictors of performance, and thus new ways of improving BCI user training. In the future, these metrics could also be used as the basis to design new feedbacks, and in particular explanatory feedbacks [10]. Indeed, these metrics being based on simple distance measures, they could be computed online, using incrementally estimated average covariance matrices. In contrast, the RWCV CA metric cannot be used online, notably due to its computational cost. The *classDist*, *restDist* and *classStab* metrics could thus be provided as online feedback, to tell users whether they should improve the distinctiveness with rest, with another class, or the stability of their patterns, for instance. These concepts being abstract and unusual for BCI users, a considerable work would be needed in terms of user-centered design and human-computer interaction to find out the best ways to provide such an explanatory feedback. These metrics revealing fast learning effects, they could also be used as a cheap, possibly online way (faster and more convenient that CV) to identify when to update and retrain classifiers. Finally, it would be relevant to further refine these metrics, for instance by defining sub-metrics, for subset of EEG channels, over specific brain areas, to study brain area specific learning processes. Overall, we are convinced that BCI user training should be further studied, and we hope these new metrics could be a new way to look at these aspects.

Acknowledgments: This work was supported by the French National Research Agency with the REBEL project and grant ANR-15-CE23-0013-01, as well as by by the EPFL/Inria International Lab.

References

[1] R. Chavarriaga, M. Fried-Oken, S. Kleih, F. Lotte, and R. Scherer. Heading for new shores! overcoming pitfalls in BCI design. *Brain-Computer Interfaces*, pages 1–14, 2016.

[2] M. Congedo, A. Barachant, and A. Andreev. A new generation of brain-computer interface based on riemannian geometry. *arXiv preprint arXiv:1310.8115*, 2013.

[3] C. Jeunet, E. Jahanpour, and F. Lotte. Why standard brain-computer interface (BCI) training protocols should be changed: An experimental study. *Journal of Neural Engineering*, 13(3):036024, 2016.

[4] C. Jeunet, B. N'Kaoua, and F. Lotte. Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates. *Progress in brain research*, 2016.

[5] T. Kaufmann, J. Williamson, E. Hammer, R. Murray-Smith, and A. Kübler. Visually multimodal vs. classic unimodal feedback approach for smr-bcis: a comparison study. *Int. J. Bioelectromagn*, 13:80–81, 2011.

[6] A. Kübler, D. Mattia, H. George, B. Doron, and C. Neuper. How much learning is involved in BCI-control? In *Int. BCI Meeting*, 2010.

[7] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[8] F. Lotte and C. Jeunet. Towards improved BCI based on human learning principles. In *Proc. Int BCI Winter Conf*, 2015.

[9] L. Roijendijk, S. Gielen, and J. Farquhar. Classifying regularized sensor covariance matrices: An alternative to CSP. *IEEE Trans Neur. Syst. Rehab.*, 24(8):893–900, 2016.

[10] J. Schumacher, C. Jeunet, and F. Lotte. Towards explanatory feedback for user training in brain-computer interfaces. In *Proc. IEEE SMC*, pages 3169–3174, 2015.

[11] E. Thomas, M. Dyson, and M. Clerc. An analysis of performance evaluation for motor-imagery based BCI. *J Neur Eng*, 10(3):031001, 2013.

[12] D. E. Thompson, L. R. Quitadamo, L. Mainardi, S. Gao, P.-J. Kindermans, J. D. Simeral, et al. Performance measurement for brain–computer or brain–machine interfaces: a tutorial. *Journal of neural engineering*, 11(3):035001, 2014.

[13] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz. Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural computation*, 23(3):791–816, 2011.

[14] F. Yger, M. Berar, and F. Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Trans Neur. Syst. Rehab.*, 2017.

[15] C. Zich, S. Debener, C. Kranczioch, M. G. Bleichner, I. Gutberlet, and M. De Vos. Real-time EEG feedback during simultaneous EEG–fMRI identifies the cortical signature of motor imagery. *Neuroimage*, 114:438–447, 2015.