The Seventh International BCI Meeting: "BCIs: Not Getting Lost in Translation", Asilomar May 21 – 25, 2018, California, USA, May 21 – 25, 2018

Effects of data sample dependence on the evaluation of BCI performance

Serafeim Perdikis[1,2], Fabien Bourban[1], Vincent Rouanne[1], José del R. Millán[2] and Robert Leeb[1]
[1]MindMaze SA, Chemin de Roseneck 5, 1006, Lausanne, Switzerland
[2]Chair in Brain-Machine Interface, Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne (EPFL), Chemin des Mines 9, 1202, Genève, Switzerland

Introduction:
Pattern Recognition (PR) methodology is nowadays the prevalent design model in Brain-Computer Interface (BCI). Consequently, classification accuracy is the most widely employed metric for BCI performance evaluation [1], so that best practices like cross-validation or training/testing set split to assess generalization capabilities are vital and popular in BCI research.

Nevertheless, since BCIs are natively also biosignal time-series processing machines, they are likely to violate common assumptions that otherwise render such PR practices optimal like the independence of training and testing folds. Abstaining from random data shuffling, which is normally a standard step of cross-validation, is shown to be enough to avoid accuracy overestimation arising from autocorrelations of the brain signals and its extracted features. This problem becomes more intense when analysis is done in overlapping sliding windows, as it is often the case [2]. We hereby intend to draw attention on this issue, as well as study the magnitude of accuracy overestimation that can occur with real BCI data when this caveat is overlooked.

Material, Methods and Results:
We analyzed 64-channel EEG data of 20 able-bodied participants performing a single session of open-loop training with two cue-based BCI paradigms, a 2-class (right and left hand) motor imagery (MI) protocol and a 6-class Steady-State Visually Evoked Potential (SSVEP) protocol (stimuli flickering at 7.5, 9, 10, 12, 15 and 20 Hz). The session consisted of 4 SSVEP and 3 MI runs. In total 45 MI and 24 SSVEP trials, each 5-second long, were collected per task. Data were recorded with a Biosemi ActiveTwo amplifier (BioSemi B.V., Amsterdam, Netherlands) at 2048 Hz sampling rate. Raw data were pre-processed with linear detrending and DC removal, spatially filtered with a cross Laplacian derivation and high-pass filtered above 1 Hz with a Butterworth filter. For both paradigms, Power Spectral Density (PSD) features are extracted for each channel (8-30 with 2 Hz resolution for MI and 1-30 Hz with 0.5 Hz resolution for SSVEP) in 1-sec long sliding overlapping windows. We repeat the feature extraction increasing the window overlap $\boldsymbol{O}$ from 0 to 875 ms with a step width of 125 ms to control the dependency of consecutive PSD samples.

Classification accuracy is then computed by means of linear discriminant analysis (LDA) classifiers using the $\boldsymbol{N}$ best (according to r^2 discriminant power) features and 10-fold cross-validation in two conditions: Prior to splitting the dataset in folds, data samples are either randomly shuffled as per regular convention (*SampleShuffle*), or only trials are shuffled forcing all within-trial samples to be in the same fold (*TrialShuffle*), thus respecting fold independence given that trials are separated enough in time. The total number of used features $\boldsymbol{N}$ is increased from 10 to 100 with a step of 10. Classification accuracy $\boldsymbol{A}$ is extracted for each paradigm, subject, $\boldsymbol{O}$ and $\boldsymbol{N}$, by averaging the testing fold accuracy across cross-validation repetitions and then across subjects. Fig. 1a-b illustrate the accuracy difference

$A_{SampleShuffle}$- $A_{TrialShuffle}$, reflecting the overestimation bias. Fig. 1c demonstrates the effect of ***N*** on ***A*** for both conditions and all ***O*** values using SSVEP data, an ordinary procedure to determine the optimal ***N***.

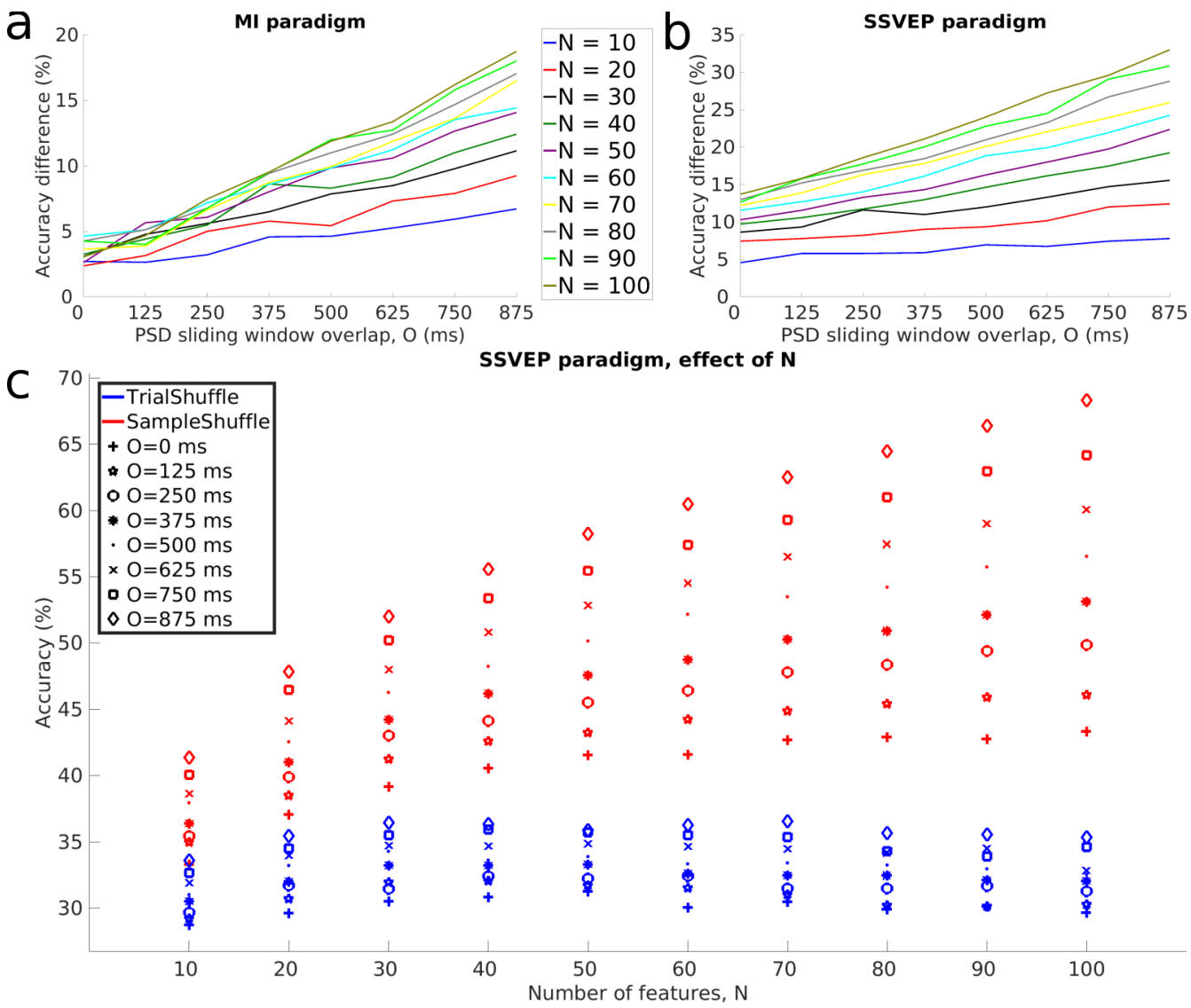

Figure 1. Difference $A_{SampleShuffle}$-$A_{TrialShuffle}$ of average (across 20 subjects) 10-fold cross-validation classification accuracy ***A*** between conditions *SampleShuffle and TrialShuffle*, as a function of the PSD sliding window overlap ***O***, for different numbers of features used ***N*** (as color-coded in the legend) and two BCI paradigms: (a) MI and (b) SSVEP. (c) Accuracy ***A*** of the SSVEP paradigm as a function of ***N***, for conditions *SampleShuffle* (red) *and TrialShuffle* (blue) and different values of ***O***, as shown in the legend.

Discussion:

Fig. 1a (MI) and Fig. 1b (SSVEP) verify deteriorating bias trends as the amount of data dependence, assessed by ***O***, increases. Interestingly, the bias is considerable even without overlapping (***O***=0), especially for the multi-class SSVEP paradigm, as consecutive EEG segments are still bound to correlate. Of note, this consequently also leads to underestimation of chance-level accuracy computed with random permutation tests, so that the significance of results will be falsely overstated. Fig. 1c shows that, unlike with *TrialShuffle* (blue), for *SampleShuffle* (red) ***A*** does not asymptote soundly as ***N***

increases, potentially misleading the experimenter to overestimate the number of informative features. This results in greater bias, since the latter is shown to (besides ***O)*** also increase with ***N*** (Fig.1a-b). These effects extend to train/test set split scenarios and all BCI paradigms.

Significance:
Sound offline performance evaluation is crucial for establishing optimal methods and parametrization before closed-loop application. Neglecting to address the common, but not highlighted enough, issue of data dependence harms the reliability and online replicability of BCI studies.

References:
1. DE Thompson, LR Quitadamo, L Mainardi, KU Rehman Laghari, S Gao, P-J Kindermans, JD Simeral, R Fazel-Rezai, M Matteucci and TH Falk. Performance measurement for brain–computer or brain–machine interfaces: a tutorial, Journal of Neural Engineering 11(3), 035001, 2014.
2. R Leeb, S Perdikis, L Tonin, A Biasiucci and M Tavella et al. Transferring brain-computer interfaces beyond the laboratory: Successful application control for motor-disabled users, in Artificial Intelligence in Medicine, 59 (2), 121-132, 2013.