BCI Competition III. Data Set V: Algorithm Description

Ferran Galán fgalan@ub.edu
Department of Behavioural Sciences Methods, Faculty of Psychology, University of Barcelona
Francesc Oliva foliva@ub.edu
Department of Statistics, Faculty of Biology, University of Barcelona
Joan Guàrdia jguardia@ub.edu
Department of Behavioural Sciences Methods, Faculty of Psychology, University of Barcelona

Algorithm

Our statistical discriminator works with the precomputed samples. It has been trained off-line with data of first three sessions and it is structured in three stages: preprocessing and feature extraction statistical discrimination, and online discrimination improvement.

Preprocessing and Feature Extraction

First of all, before any analysis data are transformed by means normalizing of each PSD sample. Each spectral component of channel *i* from sample *t* is normalized dividing by the energy of $PSD_t(i)$

$$\mathbf{PSD}_{nmh_t}(i) = \frac{\mathbf{PSD}_{h_t}(i)}{\sum_{h=1}^{n} \mathbf{PSD}_{h_t}(i)} \quad (1)$$

being $\mathbf{PSD}_{h_t}(i)$ the h^{th} power spectral component.
With data normalized, the feature extraction process is guided by canonical variates transform [1], a generalization of Fisher's linear discriminant function to more than two groups. This transformation permits the projection of a *p*-dimensional dataset **X** to be classified into *c* classes in a *(c-1)*-dimensional feature space where classes separation is maximized. This is achieved by finding vectors **a** that maximize the quotient

$$\gamma = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} \quad (2)$$

where **B** and **W** are dispersion matrix between and within classes respectively:

$$\mathbf{B} = \sum_{l=1}^{c} n_l(\mathbf{x}_l - \bar{\mathbf{x}})(\mathbf{x}_l - \bar{\mathbf{x}})' \quad (3)$$

$$\mathbf{W} = \sum_{l=1}^{c} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)' \quad (4)$$

being $\bar{\mathbf{x}} = \frac{1}{n}\sum_{l=1}^{c} n_l \bar{\mathbf{x}}_l$. Consequently, vectors **a** are the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ with eigenvalues larger than zero. In our case, we find the eigenvector's matrix $\mathbf{A}_{(97\times 2)}$ with significant eigenvalues larger than zero from $\mathbf{PSD}_{nmh}$ matrix of samples from the first three sessions. The new feature space **Y** is defined by the projection of $\mathbf{PSD}_{nmh}$ samples in **A**:

$$\mathbf{Y} = \mathbf{PSD}_{nmh}\mathbf{A} \quad (5)$$

Statistical Discrimination

After normalization and canonical variates transform, we discriminate amongst three mental tasks produced by subjects in the fourth session with the distance based DB discriminator [2] working with an Euclidean metric. As we already said, we use the samples of first three sessions as training set.
Given *c* subpopulations or classes $C_1, ..., C_c$ from population Ω, accomplishing $\bigcup_{l=1}^{c} G_l = \Omega$, with $C_l \bigcap C_m = \emptyset$ for $l \neq m$, the *p*-dimensional random vector **y** can be expressed as $\mathbf{y} = \sum_{l=1}^{c} \mathbf{y}_{C_l} I_{C_l}$ where $\mathbf{y}_{C_l}$ represents the random vector and I_{C_l} the C_l indicator. Defined a distance function d_l for class C_l, the proximity measurement for pattern $w_0 \in \Omega$ with vector $\mathbf{y}_0 = \mathbf{y}(w_0)$, is defined as

$$\phi_l(\mathbf{y}_0) = V_l(\mathbf{y}_{G_l}|\mathbf{y}_0) - V_l(\mathbf{y}_{G_l}) \quad (6)$$

where

$$V_l(\mathbf{y}_{G_l}|\mathbf{y}_0) = E_{G_l}[d_l^2(\mathbf{y}_0, \mathbf{y}_{G_l})] \quad (7)$$

$$V_l(\mathbf{y}_{G_l}) = \frac{1}{2} E_{G_l G_l}[d_l^2(\mathbf{y}_{G_l}, \mathbf{y}_{G_l})] \quad (8)$$

In this way, DB discriminator assigns w_0 to C_l, if

$$\phi_l(\mathbf{y}_0) = min_h[\phi_h(\mathbf{y}_0)] \quad (9)$$

Taking a random sample $\mathbf{y}_1, ..., \mathbf{y}_n$ of **y**, it is possible to estimate geometric variability and relative geometric variability to pattern $\mathbf{y}_0$

$$\hat{V}_d(\mathbf{y}) = \frac{1}{2n^2}\sum_{i,j=1}^{n} d^2(\mathbf{y}_i,\mathbf{y}_j) \quad (10)$$

$$\hat{V}_d(\mathbf{y}|\mathbf{y}_0) = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{y}_0,\mathbf{y}_i) \quad (11)$$

obtaining as a proximity function estimation

$$\hat{\phi}(\mathbf{y}_0) = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{y}_0,\mathbf{y}_i) - \frac{1}{2n^2}\sum_{i,j=1}^{n} d^2(\mathbf{y}_i,\mathbf{y}_j). \quad (12)$$

Then, the DB discriminator, which is trained by a sample of *n* patterns of Ω originating from *c* classes $C_1,...,C_c$ where n_l patterns are included in C_l class, operates by assigning w_0 to C_l if

$$\phi_l(\mathbf{y}_0) = min_h[\hat{\phi}_h(\mathbf{y}_0)] \quad (13)$$

being

$$\hat{\phi}_l(\mathbf{y}_0) = \frac{1}{n_l}\sum_{j=1}^{n_l} d^2(\mathbf{y}_0,\mathbf{y}_{lj}) - \frac{1}{2n^2}\sum_{j,j'=1}^{n_l} d^2(\mathbf{y}_{lj},\mathbf{y}_{lj'}). \quad (14)$$

In our case, the final assignment of a projected sample $\mathbf{y}_t$ incoming from fourth session to C_l is produced if

$$\overline{\psi}_l(\mathbf{y}_t) = min_h[\overline{\psi}_h(\mathbf{y}_t)] \quad (15)$$

where

$$\overline{\psi}_l(\mathbf{y}_t) = \frac{1}{N_{av}}\sum_{i=1}^{N_{av}} \psi_l(\mathbf{y}_{t-i+1}) \quad (16)$$

is an average proximity over the preceeding $N_{av} = 8$ consecutive samples, and $\psi_l(\mathbf{y}_t) = \frac{\hat{\phi}_l(\mathbf{y}_t)}{\sum_{l=1}^{c}\hat{\phi}_l(\mathbf{y}_t)}$ is the relative proximity to C_l at time *t*.

Online Discrimination Improvement

One of the biggest problems of the statistical discrimination task comes from the subject inconsistency in mental tasks production process, resulting in mislabeled samples by human or automatic operator which makes more difficult posterior class assignations. In this context, there is a need to explore some controller process that overcomes this limitation and permits to maintain an acceptable classification accuracy level. With this in mind, we have designed a parallel discriminant process guided by a mental task transition detector.

For each new incoming sample, after normalization and canonical variates projection, the algorithm works as follows:

1. Calculate an index to detect transition. It is easy to detect a mental task transition at time *t* with the index
$$I(\mathbf{PSD}_{nmh_t}) = \Phi(\mathbf{PSD}_{nmh_{t-1}}, \mathbf{PSD}_{nmh_t}) - \Phi(\mathbf{PSD}_{nmh_{t-2}}, \mathbf{PSD}_{nmh_{t-1}}) \quad (17)$$
if $|I(\mathbf{PSD}_{nmh_{t-1}}), I(\mathbf{PSD}_{nmh_t})| > \theta$, where θ is a fixed threshold and $\Phi(\cdot)$ a proximity function.
2. Classify with DB discriminator.
3. If $|I(\mathbf{PSD}_{nmh_{t-1}}), I(\mathbf{PSD}_{nmh_t})| > \theta$, calculate class proportions $p(C_l)$ given by DB discriminator in the gap limited by two last transitions or by first sample and first transition. Else, do nothing.
4. If $max_h[p(C_h)] > \xi$, being ξ a fixed threshold, until next transition remove from training samples those labeled as $max_h[p(C_h)]$ and reclassify once again with DB discriminator into resting classes. Else, do nothing (maintain classification from step 2).

Note that this algorithm utilizes the existence of transitions to discard the class that can be assumed as predominant in the anterior mental activity gap, improving chance classification of posterior samples of our dataset from .33 to .50. Is important to say that this benefit depends on ξ. If the threshold value is too high (too restrictive), reclassification doesn't work and first DB classification is maintained, given that the anterior gap is composed with a more heterogenic collection of samples we want to assume. Contrary, if the threshold level is minimum (minimum recommendable is .5), it is easy that reclassification process works and classification accuracy improves. We have to note that this is only true assuming that $max_h[p(C_h)]$ is representative of the mental task produced in the anterior gap. The label assigned by the operator or communicated by the subject may not correspond with $max_h[p(C_h)]$, producing a misclassification process of next gap. To avoid this problem, overall in subjects without previous experience with BCI's and working under operation modes without feedback, it is recommendable to fix a conservative threshold. Meanwhile, this algorithm seems to be useful improving classification accuracy in situations of scarcely consistent subjects, that is subjects that are not very consistent, case where reclassification achieves small improvements, are sufficiently consistent to accept the assumption that class with $max_h[p(C_h)]$ is representative of labeling carried out by subject or operator, case where reclassification can produce a great classification accuracy improvement.

Results and Discussion

The algorithm has been tested with *sessions 2 and 3*. To do this, the algorithm has been trained with the

first session in the case being tested with the second session, and with the two first sessions individually and jointly in the case being tested with the third session. The transition detector threshold has been fixed with $\theta = ,2$, and the probability threshold with $\xi = ,55$.
Table I shows the algorithm performance over the three subjects relative to test conditions mentioned above.For each subject the first row corresponds to performance of the algorithm without online improvement and the second corresponds to complete algorithm.

Table I

Performance over the three subjects relative to test condition.

Subject	Test Condition			
	$1 \to 2$	$1 \to 3$	$2 \to 3$	$1+2 \to 3$
1	0,6774	0,6973	0,7287	0,7486
	0,7284	0,7209	0,7564	0,7646
2	0,5194	0,5999	0,5864	0,6241
	0,5958	0,6803	0,6486	0,6872
3	0,5231	0,4297	0,3869	0,4273
	0,6168	0,4645	0,3910	0,4349
Average	0,5733	0,5756	0,5673	0,6000
	0,6470	0,6219	0,5987	0,6289

Performance is measured in proportion of correct classifications. For each subject and average, the first row corresponds to performance of the algorithm without online improvement and the second corresponds to complete algorithm.

These results show two aspects worth highlighting. By one hand, the online improvement systematically improve classification performance. By the other hand, in the case of testing algorithm with third session is convenient to join two previous sessions to train. This action produces the best performance over the two first subjects and similar results to the best performance over the third subject, achieved this when only the first session is used to train. However, despite these benefits, results achieved over third subject remain still far from acceptable performance.
For each subject, *figures 1, 2 and 3* show in graphs positioned in the first row labeling along the third session ($1 + 2 \to 3$ condition) produced by the algorithm without online improvement, with improvement, and produced by subjects. In the second row there are plotted the same data projected into canonical variates space. These figures help understanding the obtained results and online improvement functioning. Note that original labels (*2, 3 and 7*) have been changed into *2, 3 and 1* respectively.
First of all, we have to point out that temporal window graphs give intrasessions info while projections into canonical variates space give intersessions info (note that we are projecting samples of third session into a space constructed from two first sessions). In this sense, it is possible to say that relative good results of the algorithm over subject 1 are due to the consistent dissimilarity of different mental task patterns produced in a session, and the ability to maintain this in the same way along different sessions. In this sense, it is observable in any temporal window graph, without online improvement, that in most of gaps where subject say to be doing a mental task there are few samples classified in another class, fitting subject's labeling. By the other hand, similarity between projections into canonical variates of samples labeled by algorithm respect labeled by subject show that feature extraction from two first sessions constructs a valid space to represent and discriminate samples of different mental activities from *session 3*. Otherwise, plots of the algorithm results over *subject 3* show a very different behaviour. Firstly, this subject produces patterns too similar for each mental task, this is clearly shown in any temporal gap of third session where subject says to be doing a mental task and there are a similar proportion of samples classified into three possible classes (see without online improvement graph). Simultaneously, this produces that mental tasks can not be represented as consistent patterns along different sessions. This is dramatically represented in canonical variates plots where labeling carried out by subject totally mismatch with algorithm classification, showing the inappropriateness of canonical variates obtained from *sessions 1 and 2* to represent samples from *session 3*. Finally, results over *subject 2* show a middle point between the other two subjects and this is easily viewed in *figure 2*.
In reference to differential functioning of online improvement over the three subjects, we can give a particular explanation for each one. Starting with *subject 3*, we can say that online performance doesn't improve performance substantially because there are only two gaps (specifically the first and the fifth) where $max_h[p(C_h)] > \xi$ (see without online improvement graph). By another way, although *subject 1* shows a little improvement too the reason is totally different. There is only one gap where the expression $max_h[p(C_h)] > \xi$ is not true (eighth) but the margin to improve is minimal due to initial good performance, which is nearest to the maximum possible, when being used a good representation space. Otherwise, algorithm performance over *subject 2* shows the biggest improvement. This is caused by the existence of several gaps where expression $max_h[p(C_h)] > \xi$ is true, jointly with existence of a representation space that, being far from the appropriate, maintains a basic structure in common with samples from third session giving a broad margin to be modified by online improvement. In this case, we can say that this subject is a scarcely consistent subject, the kind of subject where online improvement has maximum performance.

Finishing, *figures 4, 5 and 6* show the same kind of graphs as discussed above plotting data of the fourth session whose estimated labels we have to provide to participate in the competition. We can see that without online improvement temporal window graphs seem clearer than obtained in $1 + 2 \rightarrow 3$ condition, overall over *subjects 2 and 3*. This is explained by the increase of number of mental activity gaps where $max_h[p(C_h)] > \xi$ is true (fixed $\xi =, 5$), which is the effect of a hypothetical subject learning process. This is especially outstanding over *subject 3*, wich is the subject that after four sessions starts showing a scarcely consistent behaviour. This is the reason for the apparent improvement introduced by 'online improvement' (see with online improvement temporal window graphs) from which we could get the unknown subject labeling. Of course, this improvement will only be real if the assumption of representativeness explained in algorithm section agrees. Under this assumption, we estimate that our algorithm achieves an average accuracy near .71 (.79 .70 and .64 for each subject).

References

[1] Rao C. R. *Advanced Statistical Methods in Biometric Research*, Wiley, New York, 1952.

[2] Cuadras C. M, Fortiana J. and Oliva F. The Proximity of an Individual to a Population with Applications in Discriminant Analysis. *Journal of Classification*, **14**(1), 117-136, 1997.

Subject 1

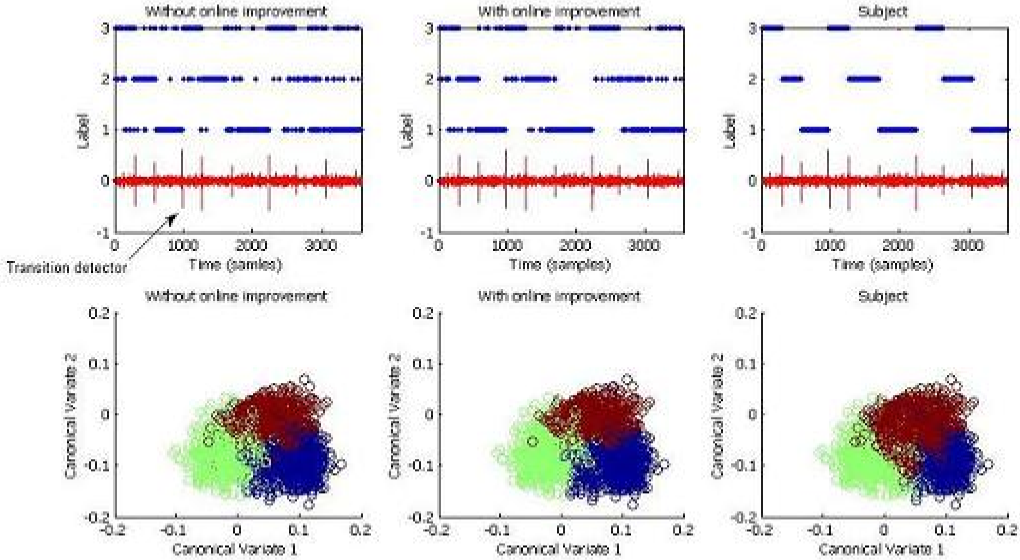

Figure 1. Temporal window graphs of $1 + 2 \rightarrow 3$ condition with labeling obtained from algorithm without online improvement, algorithm with online improvement, and subject 1. Each one shows labeling along time and index value of transition detector. Below, corresponding projections into canonical variates space .

Subject 2

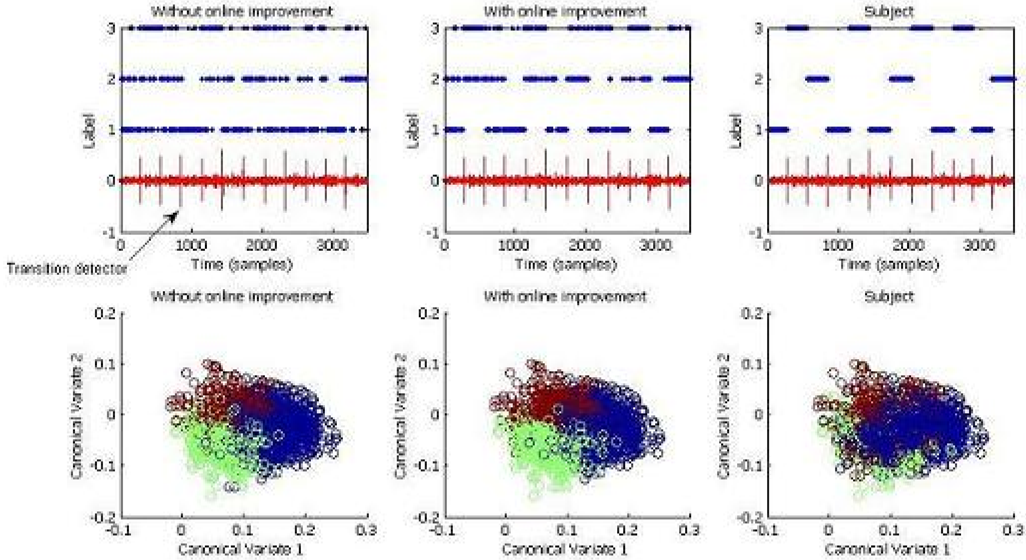

Figure 2. Temporal window graphs of $1 + 2 \rightarrow 3$ condition with labeling obtained from algorithm without online improvement, algorithm with online improvement, and subject 2. Each one shows labeling along time and index value of transition detector. Below, corresponding projections into canonical variates space .

Subject 3

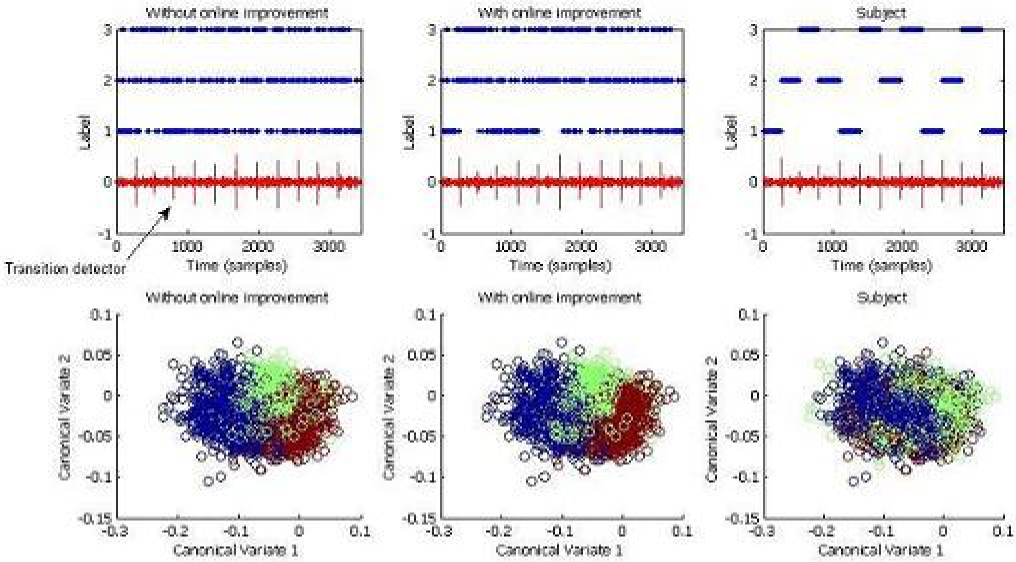

Figure 3. Temporal window graphs of $1 + 2 \rightarrow 3$ condition with labeling obtained from algorithm without online improvement, algorithm with online improvement, and subject 3. Each one shows labeling along time and index value of transition detector. Below, corresponding projections into canonical variates space .

Subject 1

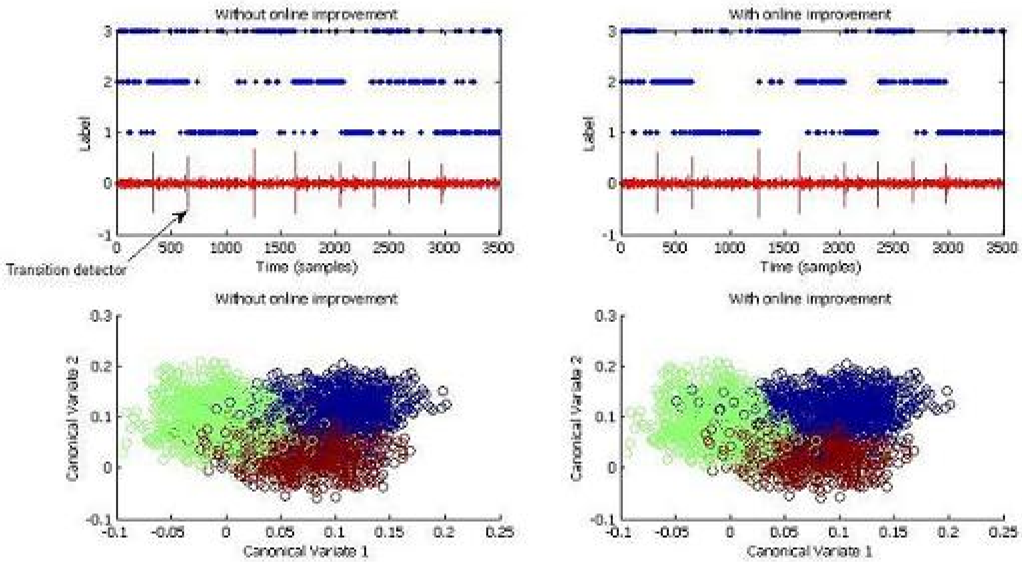

Figure 4. Temporal window graphs of $1 + 2 + 3 \rightarrow 4$ condition (test condition) with labeling obtained from algorithm without online improvement and algorithm with online improvement over subject 1. Each one shows labeling along time and index value of transition detector. Below, corresponding projections into canonical variates space .

Subject 2

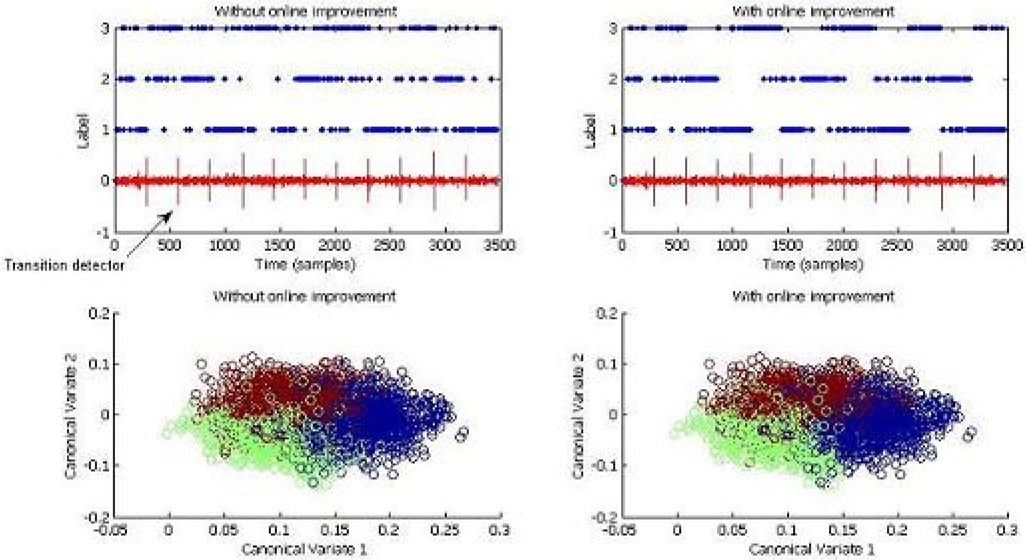

Figure 5. Temporal window graphs of $1 + 2 + 3 \rightarrow 4$ condition (test condition) with labeling obtained from algorithm without online improvement and algorithm with online improvement over subject 2. Each one shows labeling along time and index value of transition detector. Below, corresponding projections into canonical variates space .

Subject 3

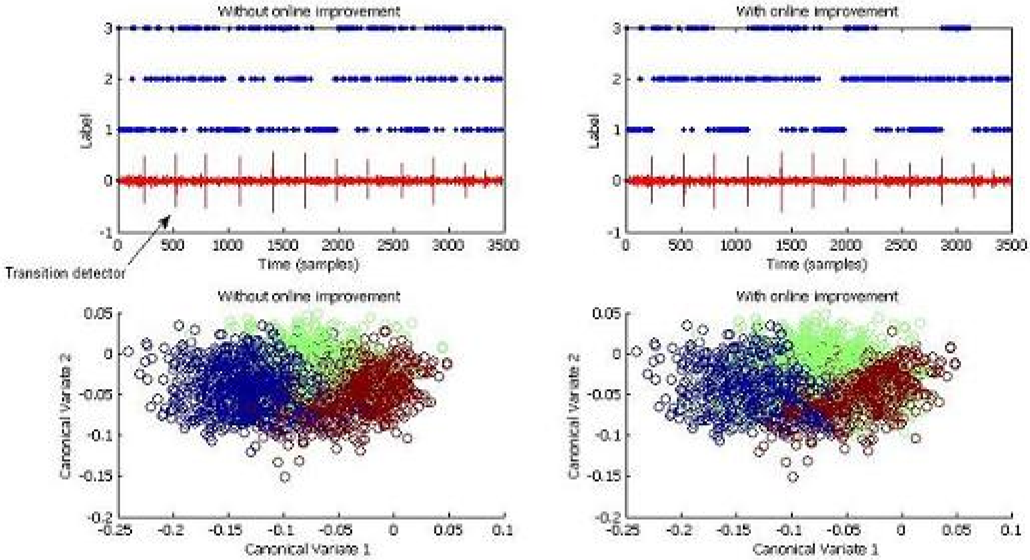

Figure 6. Temporal window graphs of $1 + 2 + 3 \rightarrow 4$ condition (test condition) with labeling obtained from algorithm without online improvement and algorithm with online improvement over subject 3. Each one shows labeling along time and index value of transition detector. Below, corresponding projections into canonical variates space .