Proceedings of the
7th Graz Brain-Computer Interface Conference 2017 DOI: 10.3217/978-3-85125-533-1-68

Sparse Bayesian Learning for Multiclass Classification with application to SSVEP- BCI

V. P. Oikonomou[1], G. Liaros[1], S. Nikolopoulos[1], I. Kompatsiaris[1]

[1]Information Technologies Institute, Centre for Research and Technology Hellas, CERTH-ITI, 6th km Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece

E-mail: {viknmu,geoliaros,nikolopo,ikom}@iti.gr

ABSTRACT: Sparse Bayesian Learning (SBL) is a basic tool of machine learning. In this work, multiple linear regression models under the SBL framework (namely MultiLRM), are used for the problem of multiclass classification. As a case study we apply our method to the detection of Steady State Visual Evoked Potentials (SSVEP), a problem we encounter into the Brain Computer Interface (BCI) concept. The multiclass classification problem is decomposed into multiple regression problems. By solving these regression problems, a discriminant feature vector is learned for further processing. Furthermore by adopting the kernel trick the model is able to reduce its computational cost. To obtain the regression coefficients of each linear model, the Variational Bayesian framework is adopted. Extensive comparisons are carried out between the MultiLRM algorithm and several other competing methods. The experimental results demonstrate that the MultiLRM algorithm achieves better performance than the competing algorithms for SSVEP classification, especially when the number of EEG channels is small.

INTRODUCTION

Brain Computer Interface (BCI) is a communication system that allows a connection between the brain and the computer[1, 2, 3]. The basic goal of a BCI system is to help people, suffering from neuromuscular disorders, to establish a communication channel between their brain and external environment without using "traditional" pathways. The brain responses can be measured by adopting various acquisition modalities such as functional Magnetic Resonance Imaging (fMRI), functional Near-Infrared Spectroscopy (fNIRS) and electroencephalography (EEG). From the above acquisition modalities, the EEG signal is the most frequently used because of its noninvasiveness, its high temporal resolution, ease of acquisition, and cost effectiveness compared to other brain activity monitoring modalities. In the literature, there exists several BCI modalities which are characterized with respect to various brain responses such as sensorimotor responses, event-related potentials and visual-evoked potentials[4, 5, 6, 7, 8, 9, 10, 11, 12]. From the above modalities, SSVEP BCI systems have attracted special interest due to lower training requirements and higher information transfer rates (ITR)[12].

A SSVEP is the brain's response evoked in occipital and occipital - parietal areas of the brain by a visual stimulus flashing at a fixed frequency [10]. SSVEP responses normally include the fundamental frequency of the visual stimulus as well as its harmonics. SSVEP BCI systems detect the different frequency components corresponding to the visual stimuli and translate them into commands. The detection of SSVEP responses is achieved by using an EEG pattern recognition algorithm. Due to frequency characteristics of SSVEPs, power spectrum density analysis (PSDA)-based methods such as fast Fourier transform (FFT) were widely used for frequency detection. Also, Support Vector Machines (SVMs) and the Linear Discriminant Analysis (LDA) are used to detect SSVEPs. A comparison between the above approaches is presented in [13].

Others algorithms used for SSVEP detection are based on Canonical Correlation Analysis (CCA) methodology and its extensions [14]. The CCA-based approaches are multichannel techniques which consider a fixed set of ideal templates. However, in cases where the signal is of small duration the template is not able to represented it well. Furthermore, their performance is deteriorated when we have a small number of EEG channels. A situation which is present when new, low cost and wireless EEG acquisition devices are used such as Emotiv device[15]. To alleviate the above problems we can use the Multivariate Linear Regression (MLR) approach [16], since the MLR does not use templates. In addition, it is not strongly depended by the multichannel nature of the signal. However, the MLR approach is based on least squares problem formulation and hence lacks robustness to the outliers while it can not handle situations where the problem is ill - posed. On the other side, Sparse Bayesian Learning (SBL)[17] is a robust technique that can successfully solve the aforementioned problems of the MLR approach. Furthermore, SBL has been successfully applied to classify event-related potentials (ERP)[4].

In this work, we propose a method, named MultiLRM, for SSVEPs classification. The multiclass classification of SSVEPs is decomposed into multiple regression models. When using a regression model an important issue is how to determine its order. Estimating the proper order is very important since models of small order may lead

 DOI: 10.3217/978-3-85125-533-1-68

to underfitting, while large order values may become responsible for data overfitting. SBL framework provides an elegant solution to this problem due to the constraints that are imposed on the model through sparse priors. After learning the regression coefficients, the predictive distribution of each regression model is used to create new discriminant features helping the subsequent classification.

MATERIALS AND METHODS

Let $\mathcal{X}$ be a matrix of size $M \times P$ containing the samples from one EEG trial, where M is the number of channels and P the number of time samples. In our analysis we construct a feature vector from one EEG trial by concatenating the P temporal points from M channels into one vector $\mathbf{x}$. Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \Re^D$ be a set of EEG trials (feature vectors), where $D = M \times P$ the feature vector dimension and N is the number of training samples. It is worth noting that D is generally high compared to N in the context of BCI applications. The classes are represented by adopting the 1-of-K coding scheme, where K is the number of classes. More specifically, for a training sample $\mathbf{x}_i$ belonging to class m, its label is specified as:

$$\mathbf{y}_i = [y_1, y_2, \cdots, y_K], \text{ where } y_j = \begin{cases} 1, & \text{if } j = m \\ 0, & \text{otherwise} \end{cases}$$

The above formulation provides us with the indicator matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]^T, \in \Re^{N \times K}$. Assuming that each column of matrix $\mathbf{Y}$ can be expressed as a linear combination of feature vectors, we obtain the following K regression models:

$$\mathbf{y}_k = \mathbf{X}\mathbf{w}_k + \mathbf{e}_k, k = 1, \cdots, K \tag{1}$$

The above assumption leads us to K regression models where each regression model learns the labels of one class versus the rest. To obtain an estimate for the model parameters $\mathbf{w}_k$ we will resort to the framework of Sparse Bayesian Learning. But before that it is needed to provide relevant information related to Eq. (1). The vector $\mathbf{y}_k \in \Re^N$ contains 0's and 1's, with the n-th element being 1 if the n-th feature vector belongs to class k. The matrix $\mathbf{X} \in \Re^{N \times D}$ contains the EEG sampes (feature vectors) $\mathbf{x}_i, i = 1, \cdots, N$ and $\mathbf{e}_k$ denotes the noise of the model following a Gaussian distribution with zero mean and precision (inverse variance) β_k. Finally, the $\mathbf{w}_k \in \Re^D$ is a vector containing the model parameters.

Sparse Bayesian Learning: Since we make the assumption of independence between the K regression models, we can treat them independently. Our goal is to infer/learn the model parameters $\mathbf{w}_k$ and then use them to make predictions about the class labels of unseen EEG samples. For the remaining of this subsection we will omit the subscript k. In our study, we adopt a probabilistic view of model analysis, and more specifically a bayesian setting of the model through priors distributions. These types of models can be treated by using the bayesian evidence framework or the variational bayesian (VB) framework[17]. In our approach, we follow the VB framework since it provides us the ability to use prior (and hyperprior) distributions over all model parameters.

Sparsity is a very helpful property since processing is faster and simpler in a sparse representation where few coefficients reveal the information we are looking for. Hence, sparse priors help us to determine the model order in an automatic way and to reduce the complexity of the model. A natural choice for the prior distribution is the ARD prior [18, 19]. More specifically, the parameter vector $\mathbf{w}$ is treated as a random variable with Gaussian prior of zero mean and variance a_i^{-1} for each element in the vector $\mathbf{w}$:

$$p(\mathbf{w}|\mathbf{a}) = \prod_{i=1}^{D} N(0, a_i^{-1}), \tag{2}$$

where D is the length of the vector $\mathbf{w}$.

The overall precision (inverse variance) β of the noise follows a Gamma distribution: $p(\beta) = Gamma(\beta; b, c) = \frac{1}{\Gamma(c)} \frac{\beta^{(c-1)}}{b^c} \exp\left\{ -\frac{\beta}{b} \right\}$, where b and c are the scale and the shape of the Gamma distribution, respectively. We use the Gamma distribution for the noise components for two reasons: first, this distribution is conjugate to the Gaussian distribution, which helps us in the derivation of closed form solutions, and second it places the positivity restriction on the overall variance and the scaling parameters. Each parameter a_i, which controls the prior distribution of the parameters $\mathbf{w}$, follows a Gamma distribution, so the overall prior over all a_i is a product of Gamma distributions given by: $p(\mathbf{a}) = \prod_{i=1}^{D} Gamma(a_i; b_a, c_a)$. So, the overall prior over model parameters $\{\mathbf{w}, \mathbf{a}, \beta\}$ is given by: $p(\mathbf{w}, \mathbf{a}, \beta) = p(\mathbf{w}|\mathbf{a}) \prod_{i=1}^{D} p(a_i) p(\beta)$. The likelihood of the data is given by:

$$p(\mathbf{y}|\mathbf{w}, \beta) = \frac{\beta^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \cdot \exp\left\{ -\frac{\beta}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right\} \tag{3}$$

To apply the VB methodology[17] we need to define an approximate posterior based on one factorization over the parameters $\{\mathbf{w}, \mathbf{a}, \beta\}$. In our study we choose the following factorization: $q(\mathbf{w}, \mathbf{a}, \beta) = q(\mathbf{w}|\mathbf{a}) \prod_{i=1}^{D} q(a_i) q(\beta)$.

Applying the VB methodology, and taking into account the above factorization, the following posteriors are obtained:

$$q(\mathbf{w}) = N(\hat{\mathbf{w}}, \mathbf{C}_{\mathbf{w}}), \tag{4}$$

$$q(\beta) = Gamma(\beta; b', c'), \tag{5}$$

$$q(\mathbf{a}) = \prod_{i=1}^{D} Gamma(a_i; b'_{a_i}, c'_{a_i}), \tag{6}$$

 DOI: 10.3217/978-3-85125-533-1-68

where

$$\mathbf{C_w} = (\hat{\beta}\mathbf{X}^T\mathbf{X} + \hat{\mathbf{A}})^{-1}, \quad (7)$$

$$\hat{\mathbf{w}} = (\hat{\beta}\mathbf{X}^T\mathbf{X} + \hat{\mathbf{A}})^{-1}\hat{\beta}\mathbf{X}^T\mathbf{y}, \quad (8)$$

$$\frac{1}{b'_{a_i}} = \frac{1}{2}(\hat{w}_i^2 + \mathbf{C_w}(i,i)) + \frac{1}{b_a}, \quad (9)$$

$$c'_{a_i} = \frac{1}{2} + c_a, \quad (10)$$

$$\hat{a}_i = b'_{a_i}c'_{a_i}, \quad (11)$$

$$\frac{1}{b'_\beta} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + tr(\mathbf{X}^T\mathbf{X}\mathbf{C_w}) + \frac{1}{b}, \quad (12)$$

$$c'_\beta = \frac{N}{2} + c, \quad (13)$$

$$\hat{\beta} = b'_\beta c'_\beta, \quad (14)$$

In the above equations the matrix $\hat{\mathbf{A}}$ is a diagonal matrix with the mean of parameters a_i in its main diagonal. The Eqs. (7) - (14) are applied iteratively until convergence.

Given a feature vector $\mathbf{x}$, the full predictive distribution is given by: $p(y|\mathbf{x}) = \int\int p(y|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}, \beta)d\mathbf{w}d\beta$. However, the above integration over both $\mathbf{w}$ and β is intractable. But we can approximate the predictive distribution by $p(y|\mathbf{x}) = \int\int p(y|\mathbf{x}, \mathbf{w}, \hat{\beta})q(\mathbf{w})d\mathbf{w}$. The above integration results in a Gaussian distribution $p(y|\mathbf{x}) = \mathcal{N}(\mathbf{x}^T\hat{\mathbf{w}}, \hat{\beta}+\mathbf{x}^T C_\mathbf{w}\mathbf{x})$. In our analysis we use the predictive mean $\mathbf{x}^T\hat{\mathbf{w}}$ as a new feature. More specifically, when a new unseen feature vector $\mathbf{x}$ is provided, the K predictive means are calculated, constructing the new discriminant feature vector, and then the k-nearest-neighbour (k-NN) algorithm is applied to perform the classification.

Kernel approach: It is worth to note here that the regression models of Eq. (1) can be easily kernelized [20]. Instead of working on the original feature space described from the following equation $\mathbf{y}_k = \mathbf{X}\mathbf{w}_k + \mathbf{e}_k = \sum_{n=1}^{D} w_{kn}\mathbf{x}_n + \mathbf{e}_k$, we can work on kernel feature space by applying the kernel trick. In that case each regression model is described by $\mathbf{y}_k = \sum_{n=1}^{N} w'_{kn}k(\mathbf{x}, \mathbf{x}_n) + \mathbf{e}_k = \mathbf{X}'\mathbf{w}'_k + \mathbf{e}_k$ where the matrix $\mathbf{X}'$ is a $N \times N$ symmetric matrix with elements $X_{nm} = k(\mathbf{x}_n, \mathbf{x}_m)$, $k(\cdot)$ is the kernel function and $\mathbf{w}'_k \in \Re^N$ is the new vector of regression coefficients. Now in these regression models we can apply the same bayesian analysis procedure described in the previous subsection. It is worth to note here that the kernel method can be useful in high dimensional settings, even if we only use a linear kernel. More specifically, to compute the regression coefficients $\mathbf{w}_k$ into the original feature space (primal variables) the computational cost is $O(D^3)$, while in the kernel feature space is $O(N^3)$[20]. When $D >> N$, as it is the case for the SSVEP analysis, the computational cost of working into the original feature space is considerable compared to the computational cost of kernel feature space.

RESULTS

In order to validate the performance of the proposed classification algorithm for SSVEP classification, we use the EEG dataset described in [14]. In this dataset a 12-target visual stimuli were presented on a 27-inch LCD monitor. Ten healthy subjects with normal or corrected-to-normal vision participated in this study. EEG data were recorded with 8 electrodes covering the occipital area. For each subject, the experiment consisted of 15 blocks. In each block, subjects were asked to gaze at one of the visual stimuli indicated by the stimulus program in a random order for 4s, and complete 12 trials corresponding to all 12 targets. Data epochs, comprising eight-channel SSVEPs, were extracted according to event triggers generated by the stimulus program. All data epochs were down-sampled to 256Hz. The EEG data have been band-pass filtered from 6Hz to 80Hz with an infinite impulse response (IIR) filter using the filtfilt() function in MATLAB. As indicated in [14] a latency delay of 0.135ms in the visual system is considered. The experiments have been performed using the EEG processing toolbox[21].

The goal of a SSVEP pattern recognition algorithm is to to take as input one EEG trial, $\mathcal{X}$, and assign it into one of K(=12) classes where each class corresponds to a stimulation frequency $f_k, k = 1, \cdots, K$. CCA-based algorithms compare the EEG trial with reference signals in order to make the decision. The reference signals could be purely artificial such as sines and cosines or they could be constructed by using EEG trials. On the other side, methods, such as the MLR approach and the MultiLRM, do not need reference signals and are based on the linear regression model. In addition, for the MultiLRM approach we can use its kernelized version in order to reduce the computational cost.

In our study we compared the proposed algorithm with four algorithms reported in the literature. More specifically, the standard CCA, the individual template based CCA (itCCA), the combination method of standard CCA and itCCA (CombitCCA)[14], and the MLR approach [16] are used. In addition, a PCA-based preprocessing step was performed before using the MLR as described in [16]. For MultiLRM approach we use uniformative priors over a_i and β (i.e. $b_a = b = 10^6$,$c_a = c = 10^{-6}$) and the linear kernel. Also, for the MLR and the MultiLRM, the number of neighborhoods in k-NN classifier was set to five. Finally, for each method (except classical CCA), the performance of each classifier was evaluated using a leave-one-out cross-validation scheme.

The mean accuracy over all subjects for each method is provided in Fig. 1. At first we calculate the accuracy using all available channels of the occipital area (8 channels). The results are shown in Fig. 1(a). We can observe that when the duration of the trial is small enough ($\leq$0.5sec) the MultiLRM approach provides us with higher accuracy compared to others methods. Furthermore, McNemar's test analysis [22] has shown that the differences in classification accuracy are significant at 5% significance level (MultiLRM vs CompitCCA: $p = 4.8 \cdot 10^{-4}$, MultiLRM vs MLR: $p = 1 \cdot 10^{-3}$).

 DOI: 10.3217/978-3-85125-533-1-68

If the duration of the trial becomes larger($\geq$1sec) the CombitCCA approach presents the higher accuracy. This could be explained due to spatial filtering that it is performed inside this method. Furthermore, we can observe that MultiLRM and MLR approaches presents similar behaviour (with MultiLRM being slightly better) and clearly these two approaches achieve higher accuracy than itCCA and CCA when the duration of trial is small ($\leq$2secs), while the itCCA outperforms the above two approaches in larger trials duration ($>$2secs).

We have performed two additional analyses related to the number of channels. In the first experiment we have used 3 channels, the channel Oz and two other channels, which are based close to O1 and O2. In the second experiment we have used 2 channels where we have excluded the Oz from the previous 3 channels. The above settings correspond to devices such as the EPOC Emotiv [15] where very few channels in the occipital area are available. In both aforementioned experiments the MultiLRM approach presents the higher accuracy among all approaches. In addition we can observed in Figs. 1(c) and (e) that the performance of MultiLRM is considerably better when the trial duration is small ($\leq$2secs). Furthermore, we can observe that CombitCCA deteriorates significantly at these two experiments. This is expected since the spatial filters do not work sufficiently well when we have small number of channels. Finally, McNemar's test analysis, at 0.5sec, has shown that the differences in classification accuracy are significant at 5% significance level (MultiLRM vs CompitCCA: $p = 5 \cdot 10^{-6}$, MultiLRM vs MLR: $p = 3 \cdot 10^{-8}$ for 3 channels, MultiLRM vs CompitCCA: $p = 2 \cdot 10^{-4}$, MultiLRM vs MLR: $p = 1 \cdot 10^{-11}$ for 2 channels).

Furthermore in our study we compared the above methods by using the Information Transfer Rate (ITR)[10]. The ITR is a measure that takes into account, besides classification accuracy, the number of classes and the trial duration, which is needed, to achieve a particular classification. The results for the channel configuration (8, 3 and 2 channels) are reported in Fig. 1 (b),(d) and (f) for various values of trial duration. In the case of 8 channels, when the trial duration is 4 secs, we can observe that all methods present similar ITRs (around 1 bps). However, the interesting point is the behaviour of the methods when the trial duration is short ($\leq$ 1.5secs). We can observe that at 0.5sec the MultiLRM approach presents the best ITR values ($\sim$4bps) among all methods, all trials duration and all channels configuration. In addition by examining the results in the case of fewer channels (3 and 2 channels) the superiority of MultiLRM approach is terms of ITR measure is evident. To summarize, the MultiLRM approach presents the best performance in terms of ITR measure and among various channels configuration. Furthermore, when using accuracy as the comparison measure, we can see that the MultiLRM approach is superior to other methods when a small number of channels is used (2 or 3 channels).

CONCLUSION

In this work we propose a new method for SSVEP classification under the SBL framework. More specifically, our approach is able to handle multiclass classification problems by adopting multiple regression models and constructing a new discriminant vector of features. The MultiLRM approach has been used in order to study the detection of SSVEP responses in the field of BCI. The proposed method has shown superior performance, compared to other well - known methods of the SSVEP literature, in cases where the trial duration is small and we have few recordings channels. Furthermore, its kernelized version gives us a way to reduce the computational cost of the procedure when the method is applied in SSVEP-BCI problems. In future communications we intent to provide various versions of the MultiLRM by introducing dependencies between the linear models either by assuming a common covariance for the noise or by treating carefully the priors over the regression coefficients. Also, it would be useful to incorporate techniques for kernel learning.

ACKNOWLEDGEMENTS

This work is part of project MAMEM that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644780.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767 – 791, 2002.

[2] G. Pfurtscheller, R. Leeb, C. Keinrath, D. Friedman, C. Neuper, C. Guger, and M. Slater, "Walking from thought," *Brain Res.*, vol. 1071, no. 1, p. 145–152, 2006.

[3] N. Hill, T. Lal, M. Schroder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Scholkopf, A. Kubler, and N. Birbaumer, "Classifying eeg and ecog signals without subject training for fast bci implementation: Comparison of nonparalyzed and completely paralyzed subjects," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 14, p. 183–186, 2006.

[4] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, "Sparse bayesian classification of eeg for brain-computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–1, 2015.

[5] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral ap-

 DOI: 10.3217/978-3-85125-533-1-68

proaches to feature extraction for eeg-based motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, pp. 317–326, Aug 2008.

[6] A. Schlögl, F. Lee, H. Bischof, and G. Pfurtscheller, "Characterization of four-class motor imagery eeg data for the bci-competition 2005," *Journal of neural engineering*, vol. 2, no. 4, p. L14, 2005.

[7] C. Guan, M. Thulasida, and W. Jiankang, "High performance p300 speller for brain-computer interface," in *IEEE Int Workshop Biomed. Circuits Syst*, pp. 13–16, 2004.

[8] F. Piccione, F. Giorgi, P. Tonin, K. Priftis, S. Giove, S. Silvoni, G. Palmas, and F. Beverina, "P300-based brain computer interface: reliability and performance in healthy and paralysed participants," *Clinical neurophysiology*, vol. 117, no. 3, pp. 531–537, 2006.

[9] L. Citi, R. Poli, C. Cinel, and F. Sepulveda, "P300-based bci mouse with genetically-optimized analogue control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 1, pp. 51–61, 2008.

[10] S. Gao, Y. Wang, X. Gao, and B. Hong, "Visual and auditory brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 1436–1447, May 2014.

[11] G. Bin, X. Gao, Y. Wang, B. Hong, and S. Gao, "Vep-based brain-computer interfaces: time, frequency, and code modulations (research frontier)," *IEEE Computational Intelligence Magazine*, vol. 4, pp. 22–26, November 2009.

[12] M. Nakanishi, Y. Wang, Y. Wang, Y. Mitsukura, and T. Jung, "A high-speed brain speller using steady-state visual evoked potentials," *International Journal of Neural Systems*, vol. 24, no. 06, p. 1450019, 2014.

[13] V. Oikonomou, G. Liaros, K. Georgiadis, E. Chatzilari, K. Adam, S. Nikolopoulos, and I. Kompatsiaris, "Comparative evaluation of state-of-the-art algorithms for ssvep-based bcis." arXiv:1602.00904, February 2016.

[14] M. Nakanishi, Y. Wang, Y. Wang, and T. Jung, "A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials," *PLoS ONE*, p. e0140703, October 2015.

[15] "Emotiv." https://www.emotiv.com, 2016.

[16] H. Wang, Y. Zhang, N. R. Waytowich, D. J. Krusienski, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Discriminative feature extraction via multivariate linear regression for ssvep-based bci," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, pp. 532–541, May 2016.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, October 2007.

[18] D. J. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, pp. 415–447, 1992.

[19] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Mach. Learn. Research*, vol. 1, pp. 211–244, 2001.

[20] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[21] G. Liaros, V. Oikonomou, K. Georgiadis, E. Chatzilari, K. Adam, S. Nikolopoulos, and I. Kompatsiaris, "eeg-processing-toolbox." https://github.com/MAMEM/eeg-processing-toolbox, 2016.

[22] A. Agresti, *Categorical data analysis*. Wiley series in probability and statistics, Hoboken (N.J.): J. Wiley, 2002.

 DOI: 10.3217/978-3-85125-533-1-68

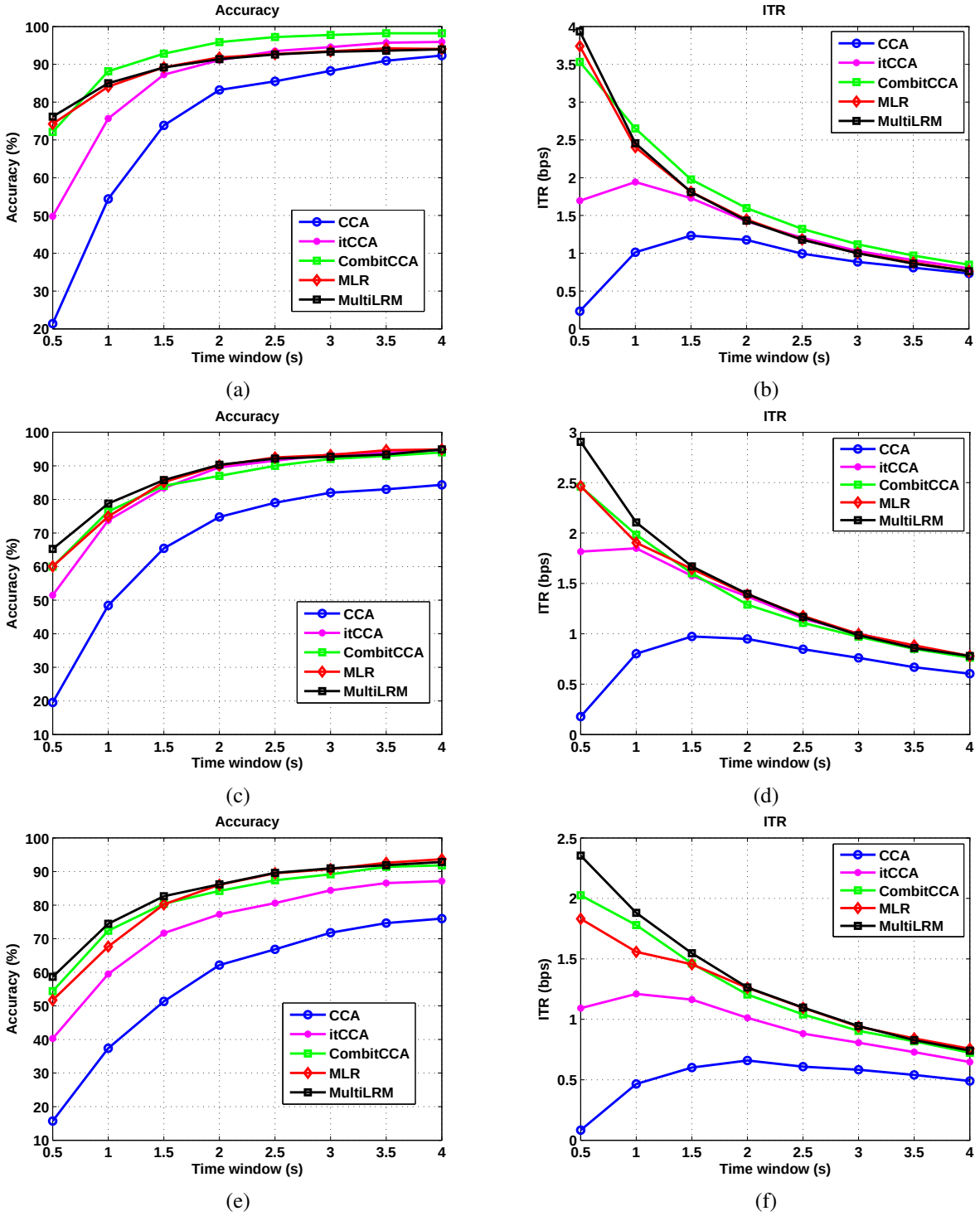

Figure 1: Mean Accuracy and Information Transfer Rate using 8 channels (a,b) using 3 channels (c,d) and using 2 channels (e,f).