Sparse Bayesian Learning for Subject Independent Classification with Application to SSVEP- BCI

Vangelis P. Oikonomou*,[1], Anastasios Maronidis[1], George Liaros[1],
Spiros Nikolopoulos[1] and Ioannis Kompatsiaris[1]

***Abstract*— Sparse Bayesian Learning (SBL) is a widely used framework which helps us to deal with two basic problems of machine learning, to avoid overfitting of the model and to incorporate prior knowledge into it. In this work, multiple linear regression models under the SBL framework are used for the problem of multiclass classification when multiple subjects are available. As a case study, we apply our method to the detection of Steady State Visual Evoked Potentials (SSVEP), a problem that arises frequently into the Brain Computer Interface (BCI) paradigm. The multiclass classification problem is decomposed into multiple regression problems. By solving these regression problems, a discriminant vector is learned for further processing. In addition the adoption of the kernel trick and the special treatment of produced similarity matrix provides us with the ability to use a Leave-One-Subject-Out training procedure resulting in a classification system suitable for subject independent classification. Extensive comparisons are carried out between the proposed algorithm, the SVM classifier and the CCA based methodology. The experimental results demonstrate that the proposed algorithm outperforms the competing approaches, in terms of classification accuracy and Information Transfer Rate (ITR), when the number of utilized EEG channels is small.**

I. Introduction

Brain Computer Interface (BCI) is a communication system that allows a connection between the brain and the computer [1]–[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. In the literature, there exist several BCI systems which are characterized with respect to various brain responses such as sensorimotor responses, event-related potentials and visual-evoked potentials [4]. From the above modalities special interest has been placed on SSVEP-based BCI systems due to lower training requirements and higher ITR [5]. It is worth to mention here that the brain responses can be measured by adopting various acquisition modalities such as fMRI, fNIRS and EEG. From the above acquisition modalities, the EEG signal is the most frequently used because of its noninvasiveness, its high time resolution, ease of acquisition, and cost effectiveness compared to other brain activity monitoring modalities.

*Corresponding author, email: viknmu@iti.gr

[1]V. P. OIkonomou, A. Maronidis, G. Liaros, S. Nikolopoulos, I. Kompatsiaris are with the Information Technologies Institute, Centre for Research and Technology Hellas, CERTH-ITI, 6th km Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece. {viknmu, amaronidis, geoliaros, nikolopo,ikom}@iti.gr

A SSVEP is the brain response, evoked in occipital and occipital - parietal areas of the brain, by a visual stimulus flashing at a fixed frequency [6]. 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 the frequency characteristics of SSVEPs, Power Spectrum Density Analysis (PSDA)-based methods, such as Fast Fourier Transform (FFT), are widely used for frequency detection. Also, Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDA) are some of the classification schemes employed to detect SSVEPs. A comparison between different classification schemes is presented in [7]. Furthermore, time domain approaches based on Canonical Correlation Analysis (CCA) are proposed in the literature for detecting the SSVEP response. For a review of these methods the interested reader could refer to [8].

Usually to analyze SSVEP related time series, a feature extraction step takes place that extracts a set of frequency characteristics which are subsequently fed into the classification system. The above procedure is applied repetitively for each subject (subject dependent approach). The subject dependent approach designs a specific classifier for each subject. This approach has two drawbacks with respect to BCI applications. First, from a model perspective, the subject-to-subject variability is not taken into account in the decision process, and second, from an application point of view, it is not possible to construct a general purpose system (subject independent) that would only need minimal (or even zero) calibration time. From CCA - based methods only the classical CCA approach [9] has the capability to facilitate a subject-independent classification scheme. All other CCA - based approaches that use templates [8] are restricted to subject dependent analysis.

In our study, we adopt the Leave-One-Subject-Out Cross Validation approach (LOSO-CV) to train the proposed model. This approach generates subject - independent classifiers since the system is trained by using a known group of subjects while the testing is performed into an unknown subject. To solve the multiclass classification problem, linear regression models are used, under the Sparse Bayesian Learning (SBL) framework, in conjunction with k-NN classifier. This approach is advantageous since the bayesian framework help us to avoid overfitting of the model, to incorporate prior knowledge into it, and, finally, to avoid the

need for a cross validation procedure in order to determine the optimal model parameter(s), such as in the case of determining the cost C of SVM classifier.

In the following sections, at first, we describe the proposed classification scheme and how the proposed approach performs subject independent analysis by exploiting the structure of a kernel matrix. Then, experimental results are provided by using a publicly available SSVEP dataset. Also, a comparison with well known approaches, such as the SVM classifier and the CCA approach, is provided. Finally, we provide the conclusions of our work and future directions.

II. Methodology

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \Re^D$ be a set of EEG samples (feature vectors), where each sample is the concatenation of P temporal points from M channels, i.e. $D = M \times P$ and N is the number of training samples. 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 with each regression model trying to learn 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 necessary 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 samples (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 (or regression coefficients).

Instead of working on the original feature space described from 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 $\mathbf{y}_k = \sum_{n=1}^{N} w'_{kn} k(\mathbf{x}_k, \mathbf{x}_n) + \mathbf{e}_k = \mathbf{X}'\mathbf{w}'_k + \mathbf{e}_k$ where the matrix $\mathbf{X}'$ is an $N \times N$ 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 model parameters. 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 $O(N^3)$ [10]. 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 kernel feature space.

In the case where the features are coming from S distinct sources (i.e. different subjects), the similarity matrix $\mathbf{X}$ has the following form:

$$\mathbf{X}' = [\mathbf{X}'_1 \mathbf{X}'_2 \cdots \mathbf{X}'_S] \tag{2}$$

where $\mathbf{X}'_s \in \Re^{N \times N_s}, s = 1, \cdots, S$ contains the similarities of s-th source and N_s is the number of samples belonging to s-th sources ($\sum_{s=1}^{S} N_s = N$). By exploiting this particular structure of similarity matrix, and taking into account the above representation of multiclass SSVEP problem (see Eq. 1), we obtain the following K linear models

$$\begin{aligned} \mathbf{y}_k &= \sum_{s=1}^{S} \mathbf{X}'_{(s)} \mathbf{w}_k^{(s)} + \mathbf{e}_k, k = 1, \cdots, K, s = 1, \cdots, S \\ &= \sum_{m=1}^{M} \left(\mathbf{X}'_{(s)} \mathbf{w}_k^{(s)} + \mathbf{e}_k^{(s)} \right), \begin{matrix} k = 1, \cdots, K, \\ s = 1, \cdots, S \end{matrix} \end{aligned} \tag{3}$$

Furthermore, in the above equation, we have decomposed the noise component into S components. By assuming that observations $\mathbf{y}_k$ can be decomposed into S components, $\mathbf{y}_k^{(s)} = \frac{1}{S}\mathbf{y}_k, s = 1, \cdots, S$, and in conjunction with Eq. 3, we obtain the following $S \times K$ linear models

$$\mathbf{y}_k^{(s)} = \mathbf{X}'_{(s)} \mathbf{w}_k^{(s)} + \mathbf{e}_k^{(s)}, k = 1, \cdots, K, s = 1, \cdots, S \tag{4}$$

Our goal is to estimate the regression parameters $\mathbf{w}^{(s)}$ given the observations $\mathbf{y}_k^{(s)}$ and the similarity matrices $\mathbf{X}_{(s)}$. In the next subsection, we describe the estimation procedure. Furthermore, since the linear models are independent to each other, due to the assumptions that we have made, we will omit the subscript notations.

It is worth to note that the proposed decomposition of $\mathbf{X}'$ can not be applied into the original feature domain, but only into the kernel domain, since we exploit the structure of similarity matrix $\mathbf{X}'$ in order to perform the classification. This structure doesn't exist into the original feature domain but it is a by-product of kernel transformation. Also, the matrix $\mathbf{X}'_s$ contains the similarities of source (i.e. subject) s with respect to all other sources (i.e. subjects). Due to this property the proposed algorithm adopts a generalization ability among all subjects (see Algorithm 1 below).

A. Sparse Bayesian Learning

Our goal is to infer/learn the model parameters $\mathbf{w}$ and use them to make predictions about the class labels. 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 [11]. In our approach, we follow the VB framework since it provides us the ability to use prior (and hyperprior) distributions over all model parameters.

A useful choice for the prior distribution is the sparse prior [12], where 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 $\mathbf{w}$:

$$p(\mathbf{w}|\mathbf{a}) = \prod_{i=1}^{N_s} N(0, a_i^{-1}), \tag{5}$$

where N_s is the length of the vector $\mathbf{w}$. 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: $p(\mathbf{a}) = \prod_{i=1}^{N_s} Gamma(a_i; b_a, c_a)$. Furthermore, 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 is the scale and the shape of the Gamma distribution, respectively. So, the overall prior over model parameters $\{\mathbf{w}, \mathbf{a}, \beta\}$ is: $p(\mathbf{w}, \mathbf{a}, \beta) = p(\mathbf{w}|\mathbf{a}) \prod_{i=1}^{N_s} p(a_i) p(\beta)$. The likelihood of the data is given by:

$$p(\mathbf{y}|\mathbf{w}, \beta) = \frac{\beta^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \cdot \exp\left\{ - \frac{\beta}{2} (\mathbf{y} - \mathbf{X}'\mathbf{w})^T (\mathbf{y} - \mathbf{X}'\mathbf{w}) \right\} \tag{6}$$

To apply the VB methodology [11] 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: $q(\mathbf{w}, \mathbf{a}, \beta) = q(\mathbf{w}|\mathbf{a}) \prod_{i=1}^{N_s} q(a_i) q(\beta)$.

Applying the VB methodology, and taking into account the above factorization, the following posteriors are obtained: $q(\mathbf{w}) = N(\hat{\mathbf{w}}, \mathbf{C_w})$, $q(\beta) = Gamma(\beta; b', c')$, $q(\mathbf{a}) = \prod_{i=1}^{N_s} Gamma(a_i; b_{a_i}^{'}, c_{a_i}^{'})$, where

$$\mathbf{C_w} = (\hat{\beta} \mathbf{X'}^T \mathbf{X'} + \hat{\mathbf{A}})^{-1}, \tag{7}$$

$$\hat{\mathbf{w}} = (\hat{\beta} \mathbf{X'}^T \mathbf{X'} + \hat{\mathbf{A}})^{-1} \hat{\beta} \mathbf{X'}^T \mathbf{y}, \tag{8}$$

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

$$c_{a_i}^{'} = \frac{1}{2} + c_a, \tag{10}$$

$$\hat{a}_i = b_{a_i}^{'} c_{a_i}^{'}, \tag{11}$$

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

$$c_\beta^{'} = \frac{N}{2} + c, \tag{13}$$

$$\hat{\beta} = b_\beta^{'} c_\beta^{'}, \tag{14}$$

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

B. Overall Classification procedure

Given an unknown similarity 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 [13]. 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 $p(y|\mathbf{x}) = \mathcal{N}(\mathbf{x}^T \hat{\mathbf{w}}, \hat{\beta} + \mathbf{x}^T C_\mathbf{w} \mathbf{x})$ [13]. In our analysis we use the predictive mean $\mathbf{x}^T \hat{\mathbf{w}}$ as a new feature. More specifically, when a new unknown 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. The overall procedure is described in Algorithm 1.

Data: EEG data
Result: Classification output
Perform the kernel trick to obtain the similarities, i.e matrix $\mathbf{X}'$;
for *each subject* s **do**
  extract matrix $\mathbf{X}'_s$;
  for *each label* k **do**
    extract label vector $\mathbf{y}_k^{(s)}$;
    learn parameters $\mathbf{w}_k^{(s)}$ by using Eqs. (7) - (14);
  end
end
construct the augmented vector of parameters $\mathbf{w}_k = [\mathbf{w}_k^{(1)}, \mathbf{w}_k^{(2)}, \cdots, \mathbf{w}_k^{(S)}]$;
for each unknown similarity vector $\mathbf{x}$ construct the new feature vector $\mathbf{z} = [\mathbf{x}^T \mathbf{w}_1, \mathbf{x}^T \mathbf{w}_2, \cdots, \mathbf{x}^T \mathbf{w}_K]$;
Use k-NN classifier to classify vector $\mathbf{z}$.

Algorithm 1: Overall Classification Procedure

III. Results

In order to validate the performance of the proposed classification algorithm for SSVEP classification, we use the EEG dataset described in [8]. In this dataset, 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. All EEG data were down-sampled to 256Hz and then were band-pass filtered from 6Hz to 80Hz. Further information about the preprocessing steps on this dataset can be found in [8].

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. The goal of a SSVEP pattern recognition algorithm is to take as input the matrix $\mathcal{X}$ and assign it into one of K(=12) classes, where each class corresponds to a stimulation frequency $f_k, k = 1, \cdots, K$. In our study, we concatenate the matrix $\mathcal{X}$ into one large feature vector. Also, to train the classifiers, the Leave-One-Subject-Out (LOSO) cross - validation approach is adopted. This training procedure provides us with the ability to construct subject independent systems, since the classifiers are trained by using a group of subjects, and then, tested in subjects not belonging to the training group.

We compared the proposed classification scheme (name it from now on SBL-SIC) with the well - known SVM classifier and the classical CCA approach. The performance of all approaches has been compared in terms of classification accuracy and Information Transfer Rate (ITR) [8]. The ITR is a measure that takes into account the classification accuracy

of the algorithm as well as the data length (time) to achieve the corresponding accuracy. The above measures have been computed by using variable time lengths from the beginning of stimulus and various channels configurations. For the SVM classifier we adopt the one-vs-all scheme, a linear kernel, as the proposed method, and we set the parameter cost C equal to 1. All the experiments have been performed using the eeg-processing-toolbox [14].

In our study, we performed three series of experiments by using different configurations of EEG recordings. At the first experiment all channels (8) of the dataset are used. At the second experiment, we used 3 channels, the channel Oz and two other channels, which are based close to O1 and O2. In this experiment, the used channels are the classical channels of 10-20 international EEG system. Finally, at the third experiment, we used 2 channels (O1 and O2) where we have excluded the Oz from the previous 3 channels. This configuration corresponds to a device with small number of channels such as the EMOTIV EPOC device [15]. The obtained results are reported in Fig. 1 for the aforementioned approaches and channel configurations. For the 8-channel configuration the CCA approach provides us with the best accuracy at time length equal to 4secs, while the SVM approach present the better ITR at 1sec. However, in the other two configurations, 3-channel and 2-channel, the SBL-SIC method outperforms considerably the SVM and the CCA with respect to both measures. Furthermore, the difference between the SBL-SIC and the other two approaches is more obvious in the 2-channel configuration.

IV. Conclusions

In this work, a new classification scheme is proposed for subject independent classification. The proposed method has been compared with the SVM classifier and the CCA approach. To train the proposed method, the LOSO CV training procedure has been adopted. The presented results have shown that the proposed classification procedure is more suitable to analyze SSVEP data when the number of channels is small, a useful characteristic in cases where we use limited number of channels, for example when the EMOTIV EPOC device [15] is used. In the future, we intent to incorporate other types of sparse priors into the proposed model and ways on reducing the calibration time (i.e. less trials) of the proposed procedure. Also, various kernels (for the kernel-based approaches) should be investigated with respect to the properties of a SSVEP EEG signal (i.e. multichannel recordings with a dominant frequency related to the frequency of the visual stimulus).

V. Acknowledgements

This work is part of project MAMEM that has received funding from the European Unions 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. 145152, 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. 183186, 2006.

[4] F. Lotte, L. Bougrain, and M. Clerc, "Electroencephalography (eeg)-based brain-computer interfaces," *Wiley Encyclopedia of Electrical and Electronics Engineering*, p. 44, 2015.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for ssvep-based bcis," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 2610–2614, Dec 2006.

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

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

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

[13] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Stanford University, Stanford, California, USA, June 30 - July 3, 2000*, pp. 46–53, 2000.

[14] 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.

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

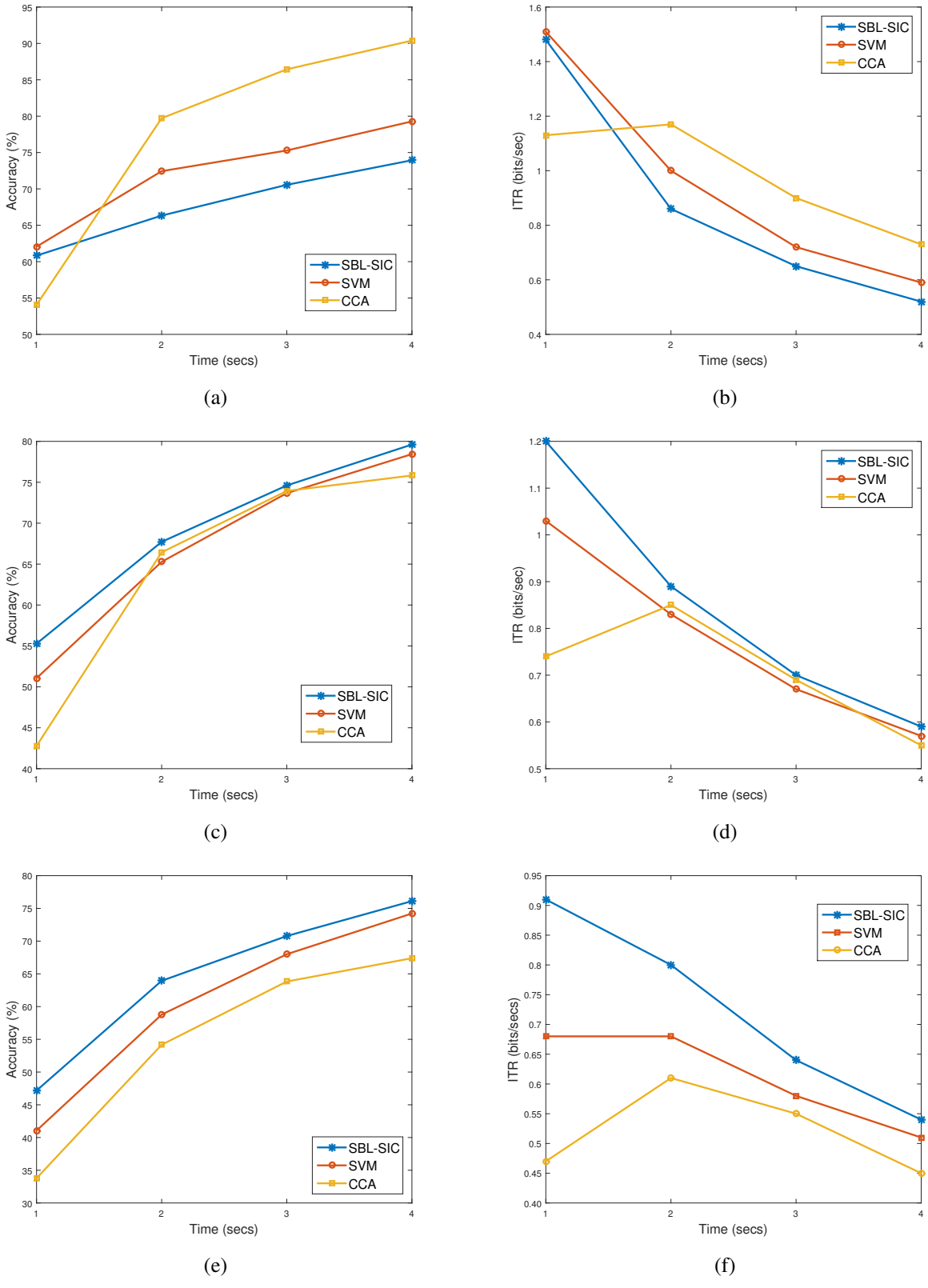

Fig. 1. Mean Accuracy and ITR for 3 methods using: 8 channels, 3 channels and 2 channels. The accuracy is shown in subfigures (a),(c) and (e) for 8-channels, 3-channels and 2-channels, respectively. The ITR is shown in subfigures (b),(d) and (f) for 8-channels, 3-channels and 2-channels,respectively.