Nonnegative Matrix Factorization for Motor Imagery EEG Classification

Hyekyoung Lee †, Andrzej Cichocki ‡, and Seungjin Choi †

† Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea
{leehk,seungjin}@postech.ac.kr

‡ Laboratory for Advanced Brain Signal Processing
Brain Science Institute, RIKEN
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan
cia@brain.riken.jp

Abstract. In this paper, we present a method of feature extraction for motor imagery single trial EEG classification, where we exploit nonnegative matrix factorization (NMF) to select discriminative features in the time-frequency representation of EEG. Experimental results with motor imagery EEG data in BCI competition 2003, show that the method indeed finds meaningful EEG features automatically, while some existing methods should undergo cross-validation to find them.

1 Introduction

Brain computer interface (BCI) is a system that is designed to translate a subject's intention or mind into a control signal for a device such as a computer, a wheelchair, or a neuroprosthesis [1]. BCI provides a new communication channel between human brain and computer and adds a new dimension to human computer interface (HCI). It was motivated by the hope of creating new communication channels for disabled persons, but recently draws attention in multimedia communication, too [2].

The most popular sensory signal used for BCI is electroencephalogram (EEG) which is the multivariate time series data where electrical potentials induced by brain activities are recorded in a scalp. Exemplary spectral characteristics of EEG involving motor, might be μ rhythm (8-12 Hz) and β rhythm (18-25 Hz) which decrease during movement or in preparation for movement (event-related desynchronization, ERD) and increase after movement and in relaxation (event-related synchronization, ERS) [1]. ERD and ERS could be used as relevant features for the task of motor imagery EEG classification. However those phenomena might happen in a different frequency band for some subjects, for instance, in 16-20 Hz, not in 8-12 Hz [3]. Moreover, it is not guaranteed that a subject always concentrates on imagination during experiments. Thus, it is

desirable to determine appropriate activated frequencies and associated features for each subject, during motor imagery experiments.

In this paper we present a method of discriminative feature extraction where we exploit the sparseness, L_1 norm, and nonnegative matrix factorization (NMF). Morlet wavelets are used to construct a nonnegative data matrix from the time-domain EEG data. We use the NMF with α-divergence that was recently proposed in [4–6]. The method is applied to the task of single-trial online classification of imaginary left and right hand movements using Data Set III of BCI competition 2003. As in [7], we use Gaussian probabilistic models for classification, where Gaussian class-conditional probabilities for a single point in time t are integrated temporally by taking the expectation of the class probabilities with respect to the discriminative power at each point in time. Numerical experiments show that our NMF-based method learns basis vectors indicating discriminative frequencies and determine useful features for the task of single-trial online classification of imaginary left and right hand movements.

2 Nonnegative Matrix Factorization

NMF is one of widely-used multivariate analysis methods for nonnegative data, which has many potential applications in pattern recognition and machine learning [8–10]. Suppose that N observed m-dimensional data points, $\{\boldsymbol{x}(t)\}$, $t = 1, \ldots, N$ are available. Denote the data matrix by $\boldsymbol{X} = [\boldsymbol{x}(1) \cdots \boldsymbol{x}(N)] = [X_{ij}] \in \mathbb{R}^{m \times N}$. NMF seeks a decomposition of the nonnegative data matrix $\boldsymbol{X}$ that is of the form:

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ contains basis vectors in its columns and $\boldsymbol{S} \in \mathbb{R}^{n \times N}$ is the associated encoding variable matrix. Both matrices $\boldsymbol{A}$ and $\boldsymbol{S}$ are restricted to have only nonnegative elements in the decomposition.

Various error measures for the factorization (1) with nonnegativity constraints, can be considered. Recently, Amari's α-divergence and its multiplicative algorithm were proposed in [5, **?**]. The α-divergence between $\boldsymbol{X}$ and $\boldsymbol{A}\boldsymbol{S}$ is given by

$$D_\alpha[\boldsymbol{X} \parallel \boldsymbol{A}\boldsymbol{S}] = \frac{1}{\alpha(1-\alpha)} \sum_{i,j} \left[\alpha X_{ij} + (1-\alpha)[\boldsymbol{A}\boldsymbol{S}]_{ij} - X_{ij}^{\alpha} [\boldsymbol{A}\boldsymbol{S}]_{ij}^{1-\alpha} \right]. \tag{2}$$

The α-divergence is a parametric family of divergence functional, including several well-known divergence measure: (1) KL divergence of $\boldsymbol{X}$ from $\boldsymbol{A}\boldsymbol{S}$ for $\alpha = 0$; (2) Hellinger divergence for $\alpha = 1/2$; (3) KL divergence of $\boldsymbol{A}\boldsymbol{S}$ from $\boldsymbol{X}$ for $\alpha = 1$; (4) χ^2-divergence for $\alpha = 2$. The parameter α is associated with the characteristics of a learning machine, in the sense that the model distribution is more inclusive (as α goes to ∞) more exclusive (as α approaches $-\infty$). The multiplicative algorithm regarding the minimization of the α-divergence of $\boldsymbol{A}\boldsymbol{S}$ from

$\boldsymbol{X}$ in (2), is given by

$$S_{ij} \leftarrow S_{ij} \left[\frac{\Sigma_k [A_{ki} (X_{kj} / [\boldsymbol{AS}]_{kl})^{\alpha}]}{\Sigma_l A_{li}} \right]^{\frac{1}{\alpha}}, \tag{3}$$

$$A_{ij} \leftarrow A_{ij} \left[\frac{\Sigma_k [S_{jk} (X_{ik} / [\boldsymbol{AS}]_{ik})^{\alpha}]}{\Sigma_l S_{jl}} \right]^{\frac{1}{\alpha}}. \tag{4}$$

More details on algorithms (3) and (4) can be found in [**?**].

3 Proposed Method

The overall structure of our proposed single trial EEG classification is illustrated in Fig. 1, where the method consists of three steps: (1) preprocessing involving wavelet transform; (2) NMF-based feature extraction; (3) probabilistic model-based classification. Each of these steps is described in detail.

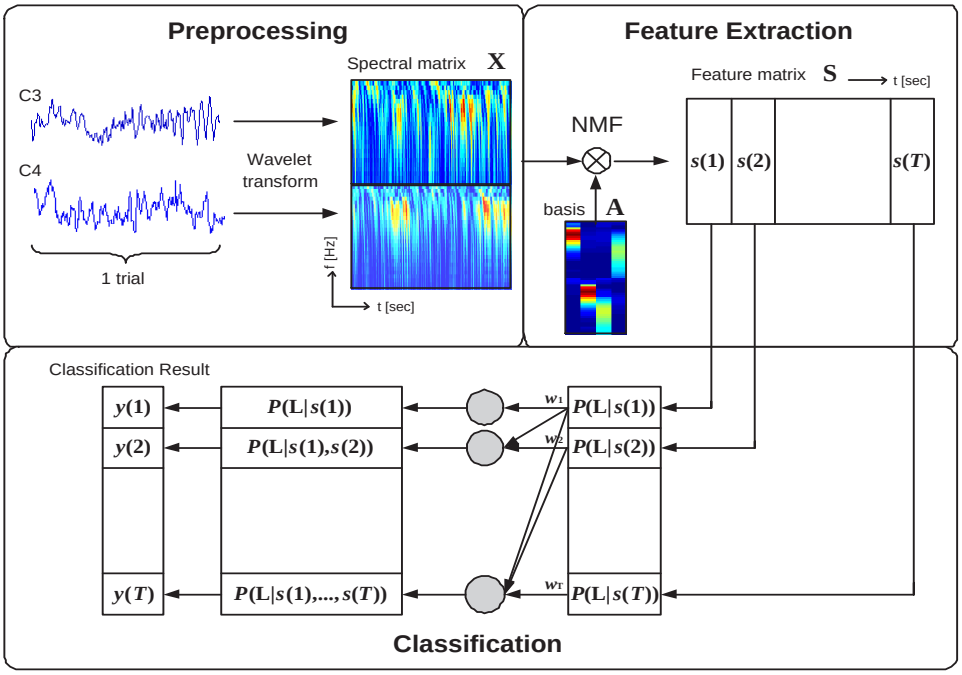


Fig. 1. The overall structure of the proposed EEG classification method is shown. In preprocessing, time-domain EEG waveforms are transformed into time-frequency representation by the Morlet wavelet transform. NMF is applied to determine representative basis vectors and associated with discriminant features. A probabilistic model-based classifier takes NMF-based features as inputs to make a decision.

3.1 Data Description

For our empirical study, we used one of BCI competition 2003 data sets, which was provided by the Department of Medical Informatics, Institute for Biomedical Engineering, Graz University of Technology, Austria [11]. The data set involves left/right imagery hand movements and consists of 140 labelled trials for training and 140 unlabelled trials for test. Each trial has a duration of 9 seconds, where a visual cue (arrow) is presented pointing to the left or the right after 3-second preparation period and imagination task is carried out for 6 seconds. It contains EEG acquired from three different channels (with sampling frequency 128 Hz) C_3, C_z and C_4. In our study we use only two channels, C_3 and C_4, because ERD has contralateral dominance and C_z channel contains little information for discriminant analysis.

3.2 Preprocessing

We obtain the time-frequency representation of the EEG data, by filtering it with complex Morlet wavelets, where the mother wavelet is given by

$$\Psi_0(\eta) = \pi^{-1/4} e^{iw_0\eta} e^{-\eta^2/2}, \tag{5}$$

where w_0 is the characteristic eigenfrequency (generally taken to be 6). Scaling and temporal shifting of the mother wavelet, leads to $\Psi_{\tau,d(f)}$ controlled by the factor $\eta = (t-\tau)/d(f)$ where

$$d(f) = \frac{w_0 + \sqrt{2+w_0^2}}{4\pi f}, \tag{6}$$

where f is the main receptive frequency.

We denote by $C_{3,k}(t)$ and $C_{4,k}(t)$ the EEG waveforms measured from C_3 and C_4 channels, in the kth trial. The wavelet transform of $C_{i,k}(t)$ $(i = 3, 4)$ at time τ and frequency f is their convolution with scaled and shifted wavelets. The amplitude of the wavelet transform, $x_{i,k}(f, \tau)$, is given by

$$x_{i,k}(f,\tau) = \| \, C_{i,k}(t) * \Psi_{\tau,d(f)}(t) \, \|, \tag{7}$$

for $i = 3, 4$ and $k = 1, \ldots, K$ where K is the number of trials. Concatenating those amplitudes for $i = 3, 4$ and $(f_1, \ldots, f_{27}) = [4, \ldots, 30]$ Hz, leads to the vector $\boldsymbol{x}_k(t) \in \mathbb{R}^{54}$ that is of the form

$$\boldsymbol{x}_k(t) = [x_{3,k}(f_1,t) \cdots x_{3,k}(f_{27},t) \quad x_{4,k}(f_1,t) \cdots x_{4,k}(f_{27},t)]^\top. \tag{8}$$

Incorporating with T data points in each trial, we construct

$$\boldsymbol{X}_k = [\boldsymbol{x}_k(1) \cdots \boldsymbol{x}_k(T)] \in \mathbb{R}^{54\times T}. \tag{9}$$

Collecting K trials leads to the data matrix

$$\boldsymbol{X} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_K] \in \mathbb{R}^{54\times KT}. \tag{10}$$

Labelled and unlabelled data are distinguished by $\boldsymbol{X}_{train}$ and $\boldsymbol{X}_{test}$, respectively.

3.3 Feature Extraction

We extract feature vectors by applying NMF to the data matrix $\boldsymbol{X}$ constructed from the wavelet transform of EEG over the frequency range $f \in [4, \ldots, 30]$ Hz. The data matrix $\boldsymbol{X} \in \mathbb{R}^{54 \times KT}$ contains a large number of data vectors reflecting K trials and T data points of EEG. Instead of using the whole data vectors, we first select candidate vectors which are expected to be more discriminative, then use only those candidate vectors as inputs to NMF, in order to determine the basis matrix $\boldsymbol{A}$. The power spectrum in the localized frequency range such as μ or β band of C_3 and C_4 channels, is activated during the imagination of movement. Thus, we investigate the power and sparseness of each data vector to select candidate vectors. We use the sparseness measure proposed by Hoyer [12], described by

$$\xi(\boldsymbol{x}) = \frac{\sqrt{m} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{m} - 1}, \tag{11}$$

where x_i is the ith element of the m-dimensional vector $\boldsymbol{x}$.

The candidate vector selection is performed in the following way. First, we compute the power of each column of $\boldsymbol{X}$, by summing its elements. For example, the power of $\boldsymbol{x}_i$, $\phi(\boldsymbol{x}_i)$, is the sum of all elements in $\boldsymbol{x}_i$, i.e., $\phi(\boldsymbol{x}_i) = \sum_{j=1}^{54} x_{ji}$ where x_{ji} is the jth element of the vector $\boldsymbol{x}_i$. The average power $\bar{\phi}$ is computed by

$$\bar{\phi} = \frac{1}{KT} \sum_{i=1}^{KT} \phi(\boldsymbol{x}_i). \tag{12}$$

The sparseness is computed for C_3 and C_4 channels, and each averaged sparseness is added, leading to the average sparseness. Data contributed by C_3 channel, corresponds to first 27 row vectors of $\boldsymbol{X}$ and the rest of row vectors are related to C_4 channels. For each column of $\boldsymbol{X}$, the sparseness is calculated for C_3 and C_4 channels, by considering the first 27 rows and the last 27 rows of $\boldsymbol{X}$, respectively. Averaged sparseness values for each channel are computed, then they are added, leading to the final average sparseness. We select candidate vectors from $\boldsymbol{X}$ if the data vector has the power greater than the average power and has the sparseness greater than 70% of the average sparseness.

We apply the NMF algorithm in (3) and (4), to the candidate data matrix $\widetilde{\boldsymbol{X}}$, leading to $\widetilde{\boldsymbol{X}} = \boldsymbol{A}\widetilde{\boldsymbol{S}}$. Then the basis matrix $\boldsymbol{A}$ is used to infer associated features $\boldsymbol{S}$, by applying the algorithm (3) to the original data matrix $\boldsymbol{X}$ with $\boldsymbol{A}$ fixed. In other words, the candidate matrix $\widetilde{\boldsymbol{X}}$ is used to determine the basis matrix $\boldsymbol{A}$ and the encoding variable matrix $\boldsymbol{S}$ (feature vectors) is inferred using the original data matrix $\boldsymbol{S}$. In our experiments, about 31% of data vectors were selected as candidate vectors. In our empirical study, basis vectors determined by the NMF of candidate vectors, showed better characteristics than those computed by the NMF of whole data vectors.

3.4 Classification

We denote by $y_k \in \{\mathrm{L}, \mathrm{R}\}$ the class label for the left or the right in the kth trial. Feature vectors $\boldsymbol{S} \in \mathbb{R}^{54 \times KT}$ consists of $\boldsymbol{s}_k(t)$ for $k = 1, \ldots, K$ and $t = 1, \ldots, T$.

For classification, we use the probabilistic model-based classifier proposed in [7], where Gaussian class-conditional densities for a single data point in time t are integrated temporally by taking the expectation of the class probabilities with respect to the discriminative power at each point in time. We assume feature vectors $\boldsymbol{s}(t)$ (the subscript k associated with trials, is left out if not necessary) follow Gaussian distribution at any time point $t \in [3, 9]$ sec, i.e.,

$$p(\boldsymbol{s}(t) \mid y) = \frac{1}{\mid 2\pi \boldsymbol{\Sigma}_{y,t} \mid^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{s}(t) - \boldsymbol{\mu}_{y,t}\right)^{\top} \boldsymbol{\Sigma}_{y,t}^{-1} \left(\boldsymbol{s}(t) - \boldsymbol{\mu}_{y,t}\right) \right\}, \quad (13)$$

where $\boldsymbol{\mu}_{y,t}$ and $\boldsymbol{\Sigma}_{y,t}$ are the mean vector and the covariance matrix for each class labelled by $y \in \{L, R\}$. These are estimated using features associated with labelled data, i.e.,

$$\boldsymbol{\mu}_{y,t} = E[\boldsymbol{s}_{y,k}(t)], \quad (14)$$

$$\boldsymbol{\Sigma}_{y,t} = E[(\boldsymbol{s}_{y,k}(t) - \boldsymbol{\mu}_{y,t})(\boldsymbol{s}_{y,k}(t) - \boldsymbol{\mu}_{y,t})^T]. \quad (15)$$

The prediction of the class label at time t, is performed using the posterior probability determined by Bayes rule:

$$p(y \mid \boldsymbol{s}(t)) = \frac{p(\boldsymbol{s}(t) \mid y)}{p(\boldsymbol{s}(t) \mid L) + p(\boldsymbol{s}(t) \mid R)}. \quad (16)$$

This posterior probability allows us to make a decision for the class label, at a single point in time. However, it is more desirable to take information across time into account. To this end, we consider

$$p(y \mid \boldsymbol{s}(1), \ldots, \boldsymbol{s}(t_0)) = \frac{\sum_{t \leq t_0} w_t p(y \mid \boldsymbol{s}(t))}{\sum_{t \leq t_0} w_t}, \quad (17)$$

where w_t are weights reflecting the discriminant power that is determined by minimizing Bayes misclassification error.

The Bayes error is defined by

$$p(\text{error}) = \int p(\text{error} \mid s(t)) p(s(t)) ds, \quad (18)$$

where

$$p(\text{error} \mid s(t)) = \min\left[p(L \mid s(t)), p(R \mid s(t))\right], \quad (19)$$

Following from the Chernoff bound

$$\min[a, b] \leq a^{\beta} b^{1-\beta}, \quad a, b \geq 0, \quad 0 \leq \beta \leq 1, \quad (20)$$

its upper-bound is given by

$$p(\text{error}) \leq \int \{p(L \mid s(t))p(s(t))\}^{\beta_t} \{p(R \mid s(t))p(s(t))\}^{1-\beta_t} ds$$
$$= p(L)^{\beta_t} p(R)^{1-\beta_t} \int p(s(t) \mid L)^{\beta_t} p(s(t) \mid R)^{1-\beta_t} ds. \quad (21)$$

The larger the discriminant power is , the smaller the Bayes error is. Thus, weights are determined by

$$2w_t = 1 - \min_{0 \leq \beta_t \leq 1} \int p(s(t) \mid L)^{\beta_t} p(s(t) \mid R)^{1-\beta_t} ds. \quad (22)$$

The class label y by combining the information through t_0 is determined by

$$y = \begin{cases} L & \text{if } p(L \mid \boldsymbol{s}(1), \ldots, \boldsymbol{s}(t_0)) > p(R \mid \boldsymbol{s}(1), \ldots, \boldsymbol{s}(t_0)), \\ R & \text{otherwise.} \end{cases} \quad (23)$$

4 Numerical Experiments

We apply the proposed method to the single-trial online classification of imaginary left and right hand movements in BCI competition 2003 (Data Set III). The time-domain EEG data is transformed into the time-frequency representation by complex Morlet wavelets with $w_0 = 6$, $f = [4, \ldots, 30]$ Hz, and $\tau = [3, \ldots, 9]$ sec using (7). We select candidate spectral vectors using the method described in Sec. 3.3. Then we apply the NMF algorithm in (4) and (3) with $\alpha = 0.5, 1, 2$ and $n = 2, 4, 5, 6$ (the number of basis vectors), in order to estimate basis vectors that are shown in Fig. 2. As the number of basis vector increases, the spectral components such as μ rhythm (8-12 Hz), β rhythm (18-22 Hz), and sensori-motor rhythm (12-16 Hz), appear in the order of their importance. All rhythms have the property of contralateral dominance, so they are present in basis vectors associated with C_3 or C_4 channel, separately.

In our empirical study, the best performance was achieved when $\alpha = 0.5 or 1$ and $n = 5$ (5 basis vectors). The single trial on-line classification result, is shown in Fig. 3, where the classification accuracy is shown in (a) and the mutual information between the true class label and the estimated class label is plotted in (b). The classification accuracy is suddenly raised from 3.43 sec. The maximal classification accuracy is 88.57 % at 6.05 sec, which is higher than the result without the data selection step in the training phase (86.43 % at 7.14 sec). The mutual information (MI) hits the maximum, 0.6549 bit, which occurs at 6.05 sec. The result is better than the one achieved by the BCI competition 2003 winner (0.61 bit). Table 1 show the maximum mutual information in the time courses per a trial varying the value of α and the number of basis. The smaller the value of α, the better the mutual information, however, α is not critical of determining the performance.

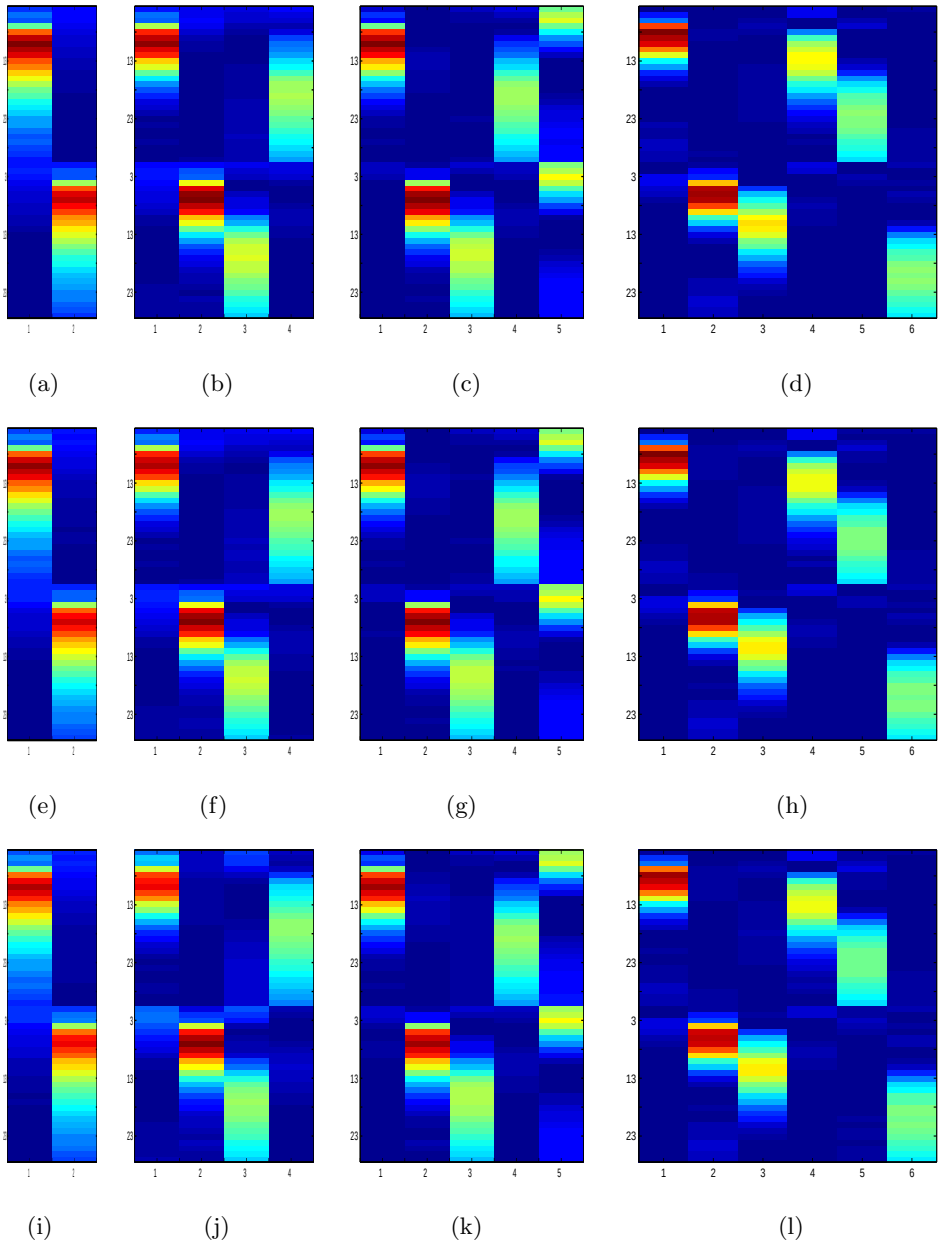
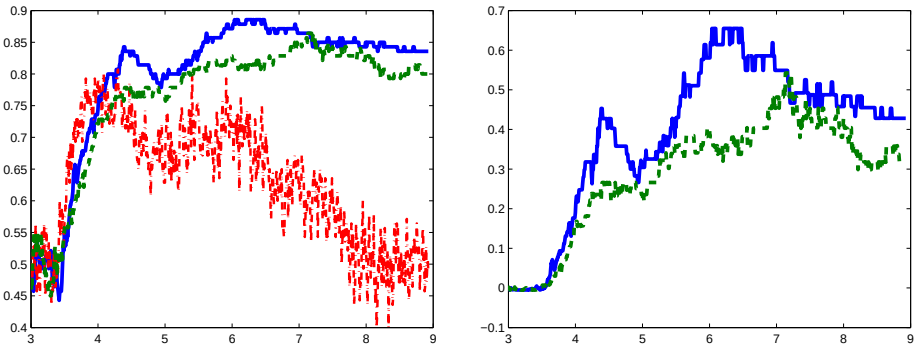


Fig. 2. Basis vectors determined by NMF are shown in the case of $\alpha = 0.5, 1, 2$ (from top to bottom) and $n = 2, 4, 5, 6$ (from left to right). In each plot, top $1/2$ is associated with C_3 and bottom $1/2$ is contributed by C_4. In each of those, the vertical axis represents frequencies between 4 and 30 Hz, the horizon axis is related to the number of basis vectors. Basis vectors reveals some useful characteristics: (1) μ rhythm (8-12 Hz); (2) β rhythm (18-22 Hz); (3) sensori-motor rhythm (12-16 Hz). ERD has the contralateral dominance, hence each rhythm occurs in each channel separately. Different values of α do not have much influence on basis vectors. However, it is observed that the lager the value of α is, the more smooth the distribution of basis vector is.



(a) (b)

Fig. 3. The on-line classification result is shown in terms of: (a) the classification accuracy; (b) the mutual information between the true class label and the estimated class label. In both plots, dotted lines (green color) are results without candidate data selection and solid lines (blue color) are results with the proposed data selection method. The data selection method improves the classification accuracy as well as the mutual information. The dot-dashed line (red color) in (a) is the result of the classifier based on the Gaussian probabilistic model taking a single time point into account. Combining the information across time, really improves the classification accuracy.

Table 1. Mutual information for different values of α and for different number of basis vectors.

α	number of basis				
	2	4	5	6	7
0.5	0.5545	0.5803	0.6549	0.6256	0.5875
1	0.5545	0.5803	0.6549	0.6256	0.5803
2	0.5408	0.5745	0.6404	0.6256	0.5803

5 Conclusion

We have presented an NMF-based method of feature extraction for on-line classification of motor imagery EEG data. We have also introduced a method of data selection where the power and the sparseness was exploited. Empirical results confirmed that the data selection scheme really improved the classification accuracy by 2.14 % and the mutual information by 0.1127 bit. Existing methods should undergo the cross-validation several times, in order to select discriminative frequency features. However, we have shown that our NMF-based method could find discriminative and representative basis vectors (which reflected appropriate spectral characteristics) without the cross-validation, which improved the on-line classification accuracy. Our method improved the mutual information achieved by BCI competition 2003 winner, by 0.0449 bit, where two frequencies (10 and 22 Hz) were selected using the leave-one-out cross validation. The value

of α in the NMF algorithm, was not critical in our empirical study. However, it was confirmed that the parameter α is associated with the characteristics of a learning machine, showing that distributions of basis vectors become more smooth, as α goes to ∞.

Acknowledgments: This work was supported by KOSEF International Cooperative Research Program and KOSEF Basic Research Program (grant R01-2006-000-11142-0).

References

1. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clinical Neurophysiology **113** (2002) 767–791
2. Ebrahimi, T., Vesin, J.F., Garcia, G.: Brain-computer interface in multimedia communication. IEEE Signal Processing Magazine **20** (2003) 14–24
3. Lal, T.N., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., Schölkopf, B.: Support vector channel selection in BCI. Technical Report 120, Max Planck Institute for Biological Cybernetics (2003)
4. Cichocki, A., Zdunek, R., Amari, S.: Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. In: Proc. Int'l Conf. Independent Component Analysis and Blind Signal Separation, Charleston, South Carolina (2006)
5. Cichocki, A., Zdunek, R., Amari, S.: New algorithms for non-negative matrix factorization in applications to blind source separation. In: Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, Toulouse, France (2006)
6. Cichocki, A., Choi, S.: Nonnegative matrix factorization with α-divergence. Pattern Recognition Letters (2006) submitted.
7. Lemm, S., Schäfer, C., Curio, G.: BCI competition 2003-data set III: Probabilistic modeling of sensorimotor μ rhythms for classification of imaginary hand movements. IEEE Trans. Biomedical Engineering **51** (2004)
8. Paatero, P., Tapper, U.: Least squares formulation of robust non-negative factor analysis. Chemometrics Intelligent Laboratory Systems **37** (1997) 23–35
9. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
10. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. Volume 13., MIT Press (2001)
11. Blankertz, B., Müller, K.R., Curio, G., Vaughan, T.M., Schalk, G., Wolpaw, J.R., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schroder, M., Birbaumer, N.: The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. IEEE Trans. Biomedical Engineering **51** (2004)
12. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research **5** (2004) 1457–1469