Bayesian Multi-Task Learning for Common Spatial Patterns

Hyohyeong Kang
Department of Computer Science
Pohang University of Science and Techonology
Pohang, Korea
Email: paanguin@postech.ac.kr

Seungjin Choi
Department of Computer Science
Division of IT Convergence Engineering
Pohang University of Science and Techonology
Pohang, Korea
Email: seungjin@postech.ac.kr

***Abstract*—Common spatial pattern (CSP) is a widely-used feature extraction method for electroencephalogram (EEG) classification and corresponding probabilistic models were recently developed, adopting a linear generative model for each class. These models are trained on a subject-by-subject basis so that inter-subject information is neglected. Moreover when only a few training samples are available for each subject, the performance is degraded. In this paper we employ Bayesian multi-task learning so that subject-to-subject information is transferred in learning the model for a subject of interest. We present two probabilistic models where precision parameters of multivariate or matrix-variate Gaussian prior for the dictionary are shared across subjects. Numerical experiments on the BCI competition IV 2a dataset confirm that our methods improve classification performance over the standard CSP (on a subject-by-subject basis), especially in the case of subjects with fewer number of training samples.**

***Keywords*-Bayesian multi-task learning; brain computer interface; common spatial patterns;**

I. Introduction

Electroencephalogram (EEG) is multivariate brain wave signals recorded at multiple electrodes placed on scalp, reflecting electrical potentials induced by subjects' brain activities. EEG classification allows computers to translate a subject's intention or mind into a control signal for a device, which is important for brain computer interface (BCI) [1].

Common spatial pattern (CSP) is a popular discriminative EEG feature extraction methods, which is proven to be a useful subject-specific spatial filter [4]. Recently a probabilistic model was proposed for CSP, where two linear Gaussian generative models with the dictionary shared are jointly learned to infer common spatial patterns [6]. CSP is a subejct-specific spatial filter, which does not consider other subjects' information involving the same task as the subject of interest. In the case of a subject with much fewer training samples, the performance of CSP is much deteriorated. Thus it is desirable to transfer useful information of subjects involving the same task to the subject of interest with fewer training samples, which is known as *subject-to-subject transfer* [3], [5]. In this paper we present two Bayesian multi-task extensions of probabilistic CSP, where precision parameters of multivariate or matrix-variate Gaussian prior for the dictionary are shared across subjects.

II. Methods

We denote by $\boldsymbol{X}^{sc} = [\boldsymbol{x}_1^{sc}, \ldots, \boldsymbol{x}_{N_{sc}}^{sc}] \in \mathbb{R}^{D \times N_{sc}}$ a collection of EEG signals measured from D electrodes over trials (N_{sc} is the number of samples) for subject $s \in \{1, \ldots, S\}$ who undergoes the mental task involving class $c \in \{1, 2\}$. The multi-task extension of the probabilistic CSP model assumes that $\boldsymbol{X}^{sc}$ is generated by

$$\boldsymbol{X}^{sc} = \boldsymbol{W}^{s} \boldsymbol{Z}^{sc} + \boldsymbol{E}^{sc}, \tag{1}$$

where $\boldsymbol{W}^{s} = [\boldsymbol{w}_1^{s}, \ldots, \boldsymbol{w}_m^{s}] \in \mathbb{R}^{D \times M}$ is the *dictionary* for subject 's', containing M basis vectors, which is shared across two classes, $\boldsymbol{Z}^{sc} = [\boldsymbol{z}_1^{sc}, \ldots, \boldsymbol{z}_{N_c}^{sc}] \in \mathbb{R}^{M \times N_c}$ the *coefficient matrix*, and $\boldsymbol{E}^{sc} = [\boldsymbol{\epsilon}_1^{sc}, \ldots, \boldsymbol{\epsilon}_{N_c}^{sc}] \in \mathbb{R}^{D \times N_c}$ is the noise matrix. Each row of $\boldsymbol{X}^{sc}$ is already centered (zero mean).

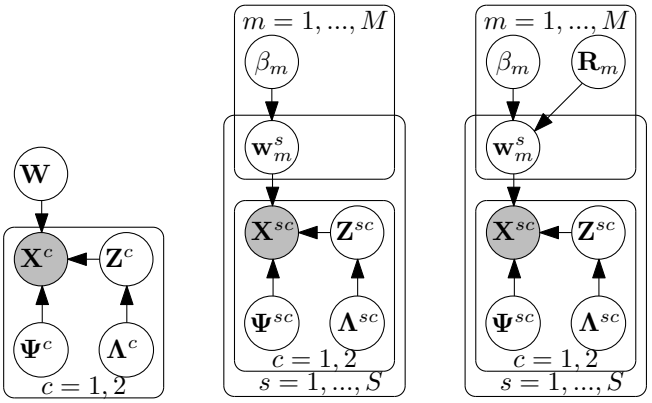
Coefficients and noise are assumed to be zero-mean Gaussians:

$$\begin{aligned} \boldsymbol{z}_t^{sc} &\sim \mathcal{N}(\boldsymbol{z}_t^{sc} \,|\, \boldsymbol{0}, (\boldsymbol{\Lambda}^{sc})^{-1}), \\ \boldsymbol{\epsilon}_t^{sc} &\sim \mathcal{N}(\boldsymbol{\epsilon}_t^{sc} \,|\, \boldsymbol{0}, (\boldsymbol{\Psi}^{sc})^{-1}), \end{aligned}$$

where $\boldsymbol{\Lambda}^{sc} = \mathrm{diag}(\lambda_1^{sc}, \ldots, \lambda_M^{sc}) \in \mathbb{R}^{M \times M}$ and $\boldsymbol{\Psi}^{sc} = \mathrm{diag}(\psi_1^{sc}, \ldots, \psi_D^{sc}) \in \mathbb{R}^{D \times D}$ are diagonal precision matrices for $s = 1, \ldots, S$ and $c = 1, 2$, which are assumed to follow Gamma distributions:

$$\begin{aligned} p(\boldsymbol{\Lambda}^{sc}) &= \prod_{m=1}^{M} \mathrm{Gamma}\,(\lambda_m^{sc} \,|\, a_0^{\lambda}, b_0^{\lambda}), \\ p(\boldsymbol{\Psi}^{sc}) &= \prod_{d=1}^{D} \mathrm{Gamma}\,(\psi_d^{sc} \,|\, a_0^{\psi}, b_0^{\psi}). \end{aligned}$$

In the case of $S = 1$ (subject-specific model), the model (1) reduces to the probabilistic CSP model in [6] (see Fig. 1(a)). We describe two multi-task learning methods for probabilistic CSP in the following two subsections, where multivariate Gaussian prior (in MTL-CSP1) or matrix-variate Gaussian prior (in MTL-CSP2) for $\{\boldsymbol{W}^{s}\}$ is defined, with some hyperparameters shared.



(a) CSP (b) MTL-CSP1 (c) MTL-CSP2

Figure 1. Graphical representations for three probabilistic CSP models: (a) probabilistic CSP on a subject-by-subject basis [6]; (b) multi-task learned CSP (proposed method 1); (c) multi-task learned CSP (proposed method 2).

A. MTL-CSP1: Sharing Multivariate Gaussian Prior

We consider the Gaussian prior distributions for the dictionaries $\{\boldsymbol{W}^s\}$ which are common across subjects:

$$\begin{aligned} p(\boldsymbol{W}^s \,|\, \boldsymbol{\beta}) &= \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{w}_m^s \,|\, \boldsymbol{0}, \beta_m^{-1} \boldsymbol{I}_D), \\ p(\beta_m) &= \text{Gamma}\,(\beta_m \,|\, a_0^\beta, b_0^\beta). \end{aligned}$$

Each basis vector $\boldsymbol{w}_m^s$ follows the isotropic Gaussian prior with the hyperparameter β_m that is shared across subjects, leading to an indirect sharing of information across subjects (see Fig. 1(b)). The indirect information transfer by sharing hyperparameters in Bayesian framework was motivated by the empirical Bayesian multi-task learning method [2]. Hyperparameters are estimated by the empirical variational Bayesian method, which is explained in Section II-C.

B. MTL-CSP2: Sharing Matrix-Variate Gaussian Prior

The MTL-CSP1 model does not consider correlations between subjects so that the negative information transfer (from subjects with negative correlations) might degrade the generalization performance. To overcome this negative effect, we present the MTL-CSP2 (see Fig. 1(c)) model which also takes inter-subject correlations into account. To this end, we consider the matrix-variate Gaussian prior distribution. The matrix-variate Gaussian distribution for a random matrix $\boldsymbol{Y} \in \mathbb{R}^{M \times N}$, denoted by $\mathcal{N}_{M,N}(\boldsymbol{Y} \,|\, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, is defined by the probability density function which takes the form:

$$\begin{aligned} &\mathcal{N}_{M,N}(\boldsymbol{Y} \,|\, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) = (2\pi)^{-\frac{MN}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} |\boldsymbol{\Omega}|^{-\frac{M}{2}} \\ &\exp\left\{ -\frac{1}{2} \text{tr}\Big(\boldsymbol{\Omega}^{-1} (\boldsymbol{Y} - \boldsymbol{M})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{M}) \Big) \right\}, \end{aligned}$$

where $|\cdot|$ denotes the determinant and $\text{tr}(\cdot)$ denotes the trace operator, $\boldsymbol{M} \in \mathbb{R}^{M \times N}$ is the mean matrix, $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$ and $\boldsymbol{\Omega} \in \mathbb{R}^{N \times N}$ are covariance matrices.

In MTL-CSP2 model, we consider the matrix-variate Gaussian distribution for $\overline{\boldsymbol{W}}_m = [\boldsymbol{w}_m^1, \ldots, \boldsymbol{w}_m^S] \in \mathbb{R}^{D \times S}$, in order to accomodate correlations between $\boldsymbol{w}_m^s$ and $\boldsymbol{w}_m^{s'}$ for $s, s' = 1, \ldots, S$:

$$p(\overline{\boldsymbol{W}}_m \,|\, \boldsymbol{R}_m, \beta_m) = \mathcal{N}_{D,S}(\overline{\boldsymbol{W}}_m | \boldsymbol{0}, \beta_m^{-1} \boldsymbol{I}_D, \boldsymbol{R}_m^{-1}),$$

where each $\boldsymbol{w}_m^s$ obeys the isotropic Gaussian as in the MTL-CSP1 model and correlations between $\boldsymbol{w}_m^s$ and $\boldsymbol{w}_m^{s'}$ are reflected in $\boldsymbol{R}_m \in \mathbb{R}^{S \times S}$ which is assumed to follow the Wishart distribution, $p(\boldsymbol{R}_m) = \mathcal{W}(\boldsymbol{R}_m | \boldsymbol{W}_0, \nu_0)$. Thus the prior distribution for $\{\boldsymbol{W}^s\}$ is given by

$$p(\{\boldsymbol{W}^s\} \,|\, \{\boldsymbol{R}_m\}, \{\beta_m\}) = \prod_{m=1}^{M} p(\overline{\boldsymbol{W}}_m | \boldsymbol{R}_m, \beta_m).$$

Note that when $\boldsymbol{R}_m = \boldsymbol{I}_S$, MTL-CSP2 reduces to MTL-CSP1. Learning $\boldsymbol{R}_m$ from training set enables us to capture positive or negative correlations between subjects so that the effect of negative transfer is reduced.

C. Variational Inference

We apply the variational inference method to compute the approximate posterior distribution over a set of variables, $\boldsymbol{\theta} = \big(\{\boldsymbol{W}^s\}, \{\boldsymbol{\Psi}^{sc}\}, \{\boldsymbol{Z}^{sc}\}, \{\boldsymbol{\Lambda}^{sc}\}, \{\boldsymbol{\beta}_m\}\big)$ for MTL-CSP1 or $\boldsymbol{\theta} = \big(\{\boldsymbol{W}^s\}, \{\boldsymbol{\Psi}^{sc}\}, \{\boldsymbol{Z}^{sc}\}, \{\boldsymbol{\Lambda}^{sc}\}, \{\boldsymbol{\beta}_m\}, \{\boldsymbol{R}_m\}\big)$ for MTL-CSP2. In MTL-CSP1, we assume that the variational distribution $q(\theta)$ factorizes over variables:

$$q(\boldsymbol{\theta}) = q(\{\boldsymbol{W}^s\}) q(\{\boldsymbol{\Psi}^{sc}\}) q(\{\boldsymbol{Z}^{sc}\}) q(\{\boldsymbol{\Lambda}^{sc}\}) q(\{\boldsymbol{\beta}_m\}),$$

where

$$\begin{aligned} q(\{\boldsymbol{W}^s\}) &= \prod_{s=1}^{S} \prod_{d=1}^{D} \mathcal{N}\left(\overline{\boldsymbol{w}}_d^{s\top} | \boldsymbol{\nu}_d^s, \boldsymbol{\Phi}_d^s\right), \\ q(\{\boldsymbol{\Psi}^{sc}\}) &= \prod_{s=1}^{S} \prod_{c=1}^{2} \prod_{d=1}^{D} \text{Gamma}(\psi_d^{sc} | a_d^{\psi sc}, b_d^{\psi sc}), \\ q(\{\boldsymbol{Z}^{sc}\}) &= \prod_{s=1}^{S} \prod_{c=1}^{2} \prod_{t=1}^{N_{sc}} \mathcal{N}(\boldsymbol{z}_t^{sc} | \boldsymbol{\mu}_t^{sc}, \boldsymbol{\Sigma}^{sc}), \\ q(\{\boldsymbol{\Lambda}^{sc}\}) &= \prod_{s=1}^{S} \prod_{c=1}^{2} \prod_{m=1}^{M} \text{Gamma}(\lambda_m^{sc} | a_m^{\lambda sc}, b_m^{\lambda sc}), \\ q(\{\beta_m\}) &= \prod_{m=1}^{M} \text{Gamma}(\beta_m | a_m^\beta, b_m^\beta), \end{aligned}$$

where $\overline{\boldsymbol{w}}_d^s$ represents the d-th row of $\boldsymbol{W}^s$. Variational parameters are estimated by maximizing the lower bound on

the marginal likelihood, $\mathcal{L}(q) = \mathbb{E}_q\left[\log\left(p(\boldsymbol{X}, \boldsymbol{\theta})/q(\boldsymbol{\theta})\right)\right]$,

$$
\begin{aligned}
(\boldsymbol{\Phi}_d^s)^{-1} &= \mathbb{E}[\boldsymbol{D}_\beta] + \sum_{c=1}^{2} \mathbb{E}[\psi_d^{sc}] \sum_{t=1}^{N_{sc}} \mathbb{E}\left[\boldsymbol{z}_t^{sc} \boldsymbol{z}_t^{sc\top}\right], \\
\boldsymbol{\nu}_d^s &= \boldsymbol{\Phi}_d^s \sum_{c=1}^{2} \mathbb{E}[\psi_d^{sc}] \sum_{t=1}^{N_{sc}} \boldsymbol{x}_t^{sc}(d) \mathbb{E}[\boldsymbol{z}_t^{sc}], \\
\boldsymbol{\Sigma}^{sc} &= \left(\mathbb{E}\left[\boldsymbol{\Lambda}^{sc}\right] + \mathbb{E}\left[\boldsymbol{W}^{s\top} \boldsymbol{\Psi}^{sc} \boldsymbol{W}^s\right]\right)^{-1}, \\
\boldsymbol{\mu}_t^{sc} &= \boldsymbol{\Sigma}^{sc} \mathbb{E}\left[\boldsymbol{W}^{s\top}\right] \mathbb{E}\left[\boldsymbol{\Psi}^{sc}\right] \boldsymbol{x}_t^{sc}, \\
a_d^{\psi sc} &= a_0^{\psi sc} + \frac{N_{sc}}{2}, \quad a_m^{\lambda sc} = a_0^{\lambda sc} + \frac{N_{sc}}{2}, \\
b_d^{\psi sc} &= b_0^{\psi sc} + \frac{1}{2} \sum_{t=1}^{N_{sc}} \mathbb{E}\left[(\boldsymbol{x}_t^{sc}(d) - \bar{\boldsymbol{w}}_d^s \boldsymbol{z}_t^{sc})^2\right], \\
b_m^{\lambda sc} &= b_0^{\lambda sc} + \frac{1}{2} \sum_{t=1}^{N_{sc}} \mathbb{E}\left[(\boldsymbol{z}_t^{sc}(m))^2\right], \\
a_m^{\beta} &= a_0^{\beta} + \frac{DS}{2}, \quad b_m^{\beta} = b_0^{\beta} + \frac{1}{2} \sum_{s=1}^{S} \mathbb{E}\left[\boldsymbol{w}_m^{s\top} \boldsymbol{w}_m^s\right],
\end{aligned}
$$

where $\boldsymbol{D}_\beta \in \mathbb{R}^{M \times M}$ is the diagonal matrix with diagonal entries β_m for $m = 1, ..., M$.

For MTL-CSP2, we also assume that the variational distribution $q(\theta)$ factorizes over variables:

$$q(\boldsymbol{\theta}) = q(\{\boldsymbol{W}^s\}) q(\{\boldsymbol{\Psi}^{sc}\}) q(\{\boldsymbol{Z}^{sc}\}) q(\{\boldsymbol{\Lambda}^{sc}\}) q(\{\boldsymbol{\beta}_m\}) q(\{\boldsymbol{R}_m\}),$$

where $q(\{\boldsymbol{R}_m\}) = \prod_{m=1}^{M} \mathcal{W}(\boldsymbol{R}_m | \boldsymbol{W}_m, \nu_m)$, and variational distributions over other variables are the same as MTL-CSP1. Variational parameters are updated:

$$
\begin{aligned}
(\boldsymbol{\Phi}_d^s)^{-1} &= \mathbb{E}[\boldsymbol{D}_\beta \boldsymbol{R}(s,s)] + \sum_{c=1}^{2} \sum_{t=1}^{N_{sc}} \mathbb{E}[\psi_d^{sc}] \mathbb{E}[\boldsymbol{z}_t^{sc} \boldsymbol{z}_t^{sc\top}], \\
\boldsymbol{\nu}_d^s &= \boldsymbol{\Phi}_d^s \Bigg(\sum_{c=1}^{2} \mathbb{E}[\psi_d^{sc}] \sum_{t=1}^{N_{sc}} \boldsymbol{x}_t^{sc}(d) \mathbb{E}[\boldsymbol{z}_t^{sc}] \\
&\qquad - \sum_{j \neq s} \mathbb{E}[\boldsymbol{D}_\beta \boldsymbol{R}(s,j)] \mathbb{E}[\overline{\boldsymbol{w}}_d^{j\top}] \Bigg), \\
\nu_m &= \nu_0 + D, \\
(\boldsymbol{W}_m)^{-1} &= \boldsymbol{W}_0^{-1} + \mathbb{E}[\beta_m] \mathbb{E}\left[\overline{\boldsymbol{W}}_m^{\top} \overline{\boldsymbol{W}}_m\right],
\end{aligned}
$$

where $\boldsymbol{R}(i,j)$ is the $M \times M$ diagonal matrix with diagonal entries $[\boldsymbol{R}_m]_{ij}, (m = 1, ..., M)$. The expectations $\mathbb{E}$ appearing in updating equations for variational parameters are taken with respect to the variational distribution $q(\cdot)$, the values of which are left out due to the space limit. Hyperparameters are also estimated by maximizing the lower-bound $\mathcal{L}(q)$.

III. Numerical Experiments

We tested our models on BCI competition IV [1] 2a dataset. The dataset contains EEG measurements of 9 subjects with 22 electrodes. Although there were 4 target classes in the data set, we only used the trials that correspond to the imagery left hand and right hand classes. For every trial, we picked the interval from 3.5 seconds to 5.5 seconds after each cue. Each subject performed 144 trials per class so that the dataset has 288 trials per subject.

We randomly selected only one trial from each class per subject as a training data. We extracted features and trained classifiers using the training data. The remaining trials are considered as the test data. We calculated log-variances projected by 6 CSP filters that corresponds to 3 largest and 3 smallest eigenvalues. Compared to original CSP, our MTL-CSP models can exploit every subject's data. When we extracting features for a subject s with MTL-CSP models, we randomly chose N trials of each class from the other subjects, and used them as an additional source of information. After the models were inferred, we calculated the log value of the variance of $\mathbb{E}[\boldsymbol{z}_t]$ within each trial to obtain features. By the using $\mathbb{E}[\lambda_m^{s1}]/\mathbb{E}[\lambda_m^{s2}]$, we selected 3 largest-valued and 3 smallest-valued features. In every case, we trained the logistic regression classifier, and predicted the labels of test data using these features. We repeated the experiment for each value of N for 10 times, and averaged the results. The classification accuracy was estimated by the ratio of the number of accurately classified test trials compared to the total number of test trials (Fig. 2, 3).

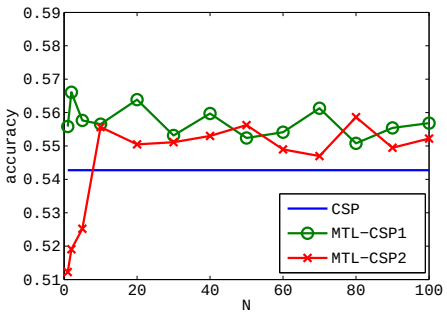


Figure 3. The classification performance averaged over every subject.

IV. Conclusions

We have presented Bayesian multi-task learning extensions of the probabilistic CSP method, allowing subject-to-subject transfer by sharing parameters of multivariate or matrix-variate Gaussian prior for the dictionary. We have presented two models, MTL-CSP1 and MTL-CSP2, developing variational inference algorithms to compute posterior distributions over variables. Numerical experiments on BCI competition IV 2a dataset demonstrated that our models improved EEG classification performance over the standard probabilistic CSP which is subject-specific. There were improvements for 5 subjects out of the total 9 subjects (Fig. 2). Although the performances for remaining subjects were

[1] http://www.bbci.de/competition/iv/index.html

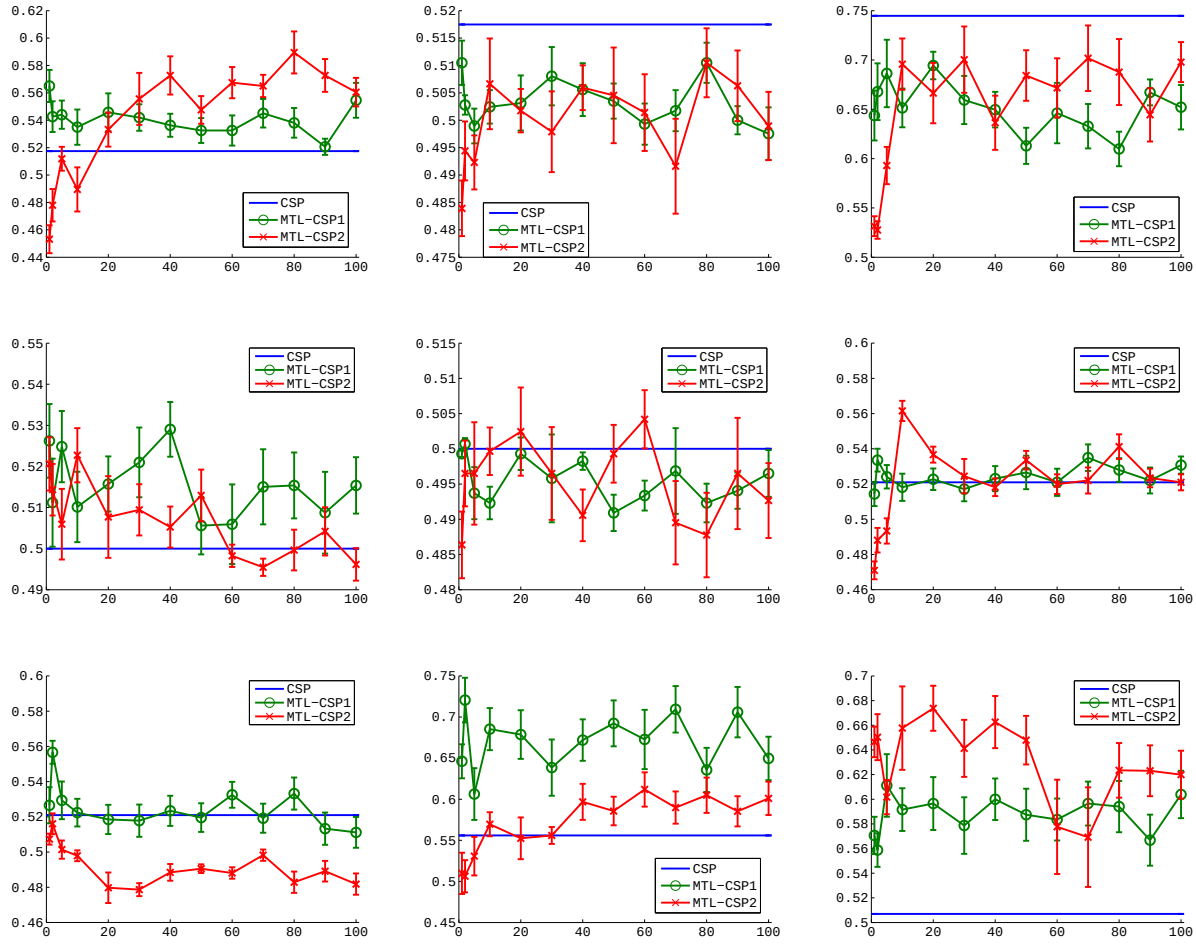


Figure 2. Classification accuracies of all subjects in the BCI competition IV 2a dataset (mean value $\pm$ 1 standard error). The first row shows the graphes for the subjects 1,2 and 3, starting from the left. The second row shows the graphes of the subject 4,5,6, and the third row shows that of the subject 7,8,9. For the subject 1, 4, 8, 6 and 9, there were some improvements by the information transfer, while other subjects suffer from negative side effects.

decreased, the sum of improvement were generally greater than that of the decrease. Thus the overall performance was increased, so that the experiment confirmed effectiveness of our methods (Fig. 3). Because collecting large number of training samples is often difficult in BCI applications, this result will be helpful for increasing recognition rate by better data utilization.

Acknowledgments

This work was supported by National Research Foundation (NRF) of Korea (No. 2010-0018828 and 2010-0018829) and WCU Program (R31-2010-000-10100-0).

References

[1] A. Cichocki, Y. Washizawa, T. Rutkowski, H. Bakardjian, A. H. Phan, S. Choi, H. Lee, Q. Zhao, L. Zhang, and Y. Li, "Noninvasive BCIs: Multiway signal-processing array decompositions," *IEEE Computer*, vol. 41, no. 10, pp. 34–42, 2008.

[2] T. Heskes, "Empirical Bayes for learning to learn," in *Proceedings of the International Conference on Machine Learning (ICML)*, San Francisco, CA, 2000.

[3] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, 2009.

[4] Z. J. Koles, "The quantitative extraction and topographic mapping of the abnormal components," *EEG and Clinical Neurophysilology*, vol. 79, pp. 440–447, 1991.

[5] F. Lotte and G. Cuntai, "Learning from other subjects helps reducing brain-computer interface calibration time," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, 2010.

[6] W. Wu, Z. Chen, S. Gao, and E. N. Brown, "A probabilistic framework for learning robust common spatial patterns," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, 2009.