Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Probabilistic Models for Common Spatial Patterns: Parameter-Expanded EM and Variational Bayes

Hyohyeong Kang[1] and **Seungjin Choi**[1,2,3]
[1] Department of Computer Science and Engineering,
[2] Division of IT Convergence Engineering,
[3] Department of Creative IT Excellence Engineering,
Pohang University of Science and Technology,
San 31, Hyoja-dong, Nam-gu, Pohang 790-784, Korea
{paanguin,seungjin}@postech.ac.kr

Abstract

Common spatial patterns (CSP) is a popular feature extraction method for discriminating between positive and negative classes in electroencephalography (EEG) data. Two probabilistic models for CSP were recently developed: probabilistic CSP (PCSP), which is trained by expectation maximization (EM), and variational Bayesian CSP (VBCSP) which is learned by variational approximation. Parameter expansion methods use auxiliary parameters to speed up the convergence of EM or the deterministic approximation of the target distribution in variational inference. In this paper, we describe the development of parameter-expanded algorithms for PCSP and VBCSP, leading to PCSP-PX and VBCSP-PX, whose convergence speed-up and high performance are emphasized. The convergence speed-up in PCSP-PX and VBCSP-PX is a direct consequence of parameter expansion methods. The contribution of this study is the performance improvement in the case of CSP, which is a novel development. Numerical experiments on the BCI competition datasets, III IV a and IV 2a demonstrate the high performance and fast convergence of PCSP-PX and VBCSP-PX, as compared to PCSP and VBCSP.

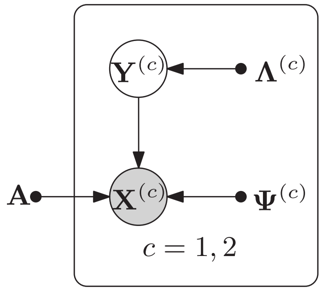
Introduction

Electroencephalography (EEG) is the recording of electrical potentials at multiple sensors placed on the scalp, leading to multivariate time series data reflecting brain activities. EEG classification is a crucial part of non-invasive brain computer interface (BCI) systems, enabling computers to translate a subject's intention or mind into control signals for a device such as a computer, wheelchair, or neuroprosthesis (Wolpaw et al. 2002; Ebrahimi, Vesin, and Garcia 2003; Cichocki et al. 2008).

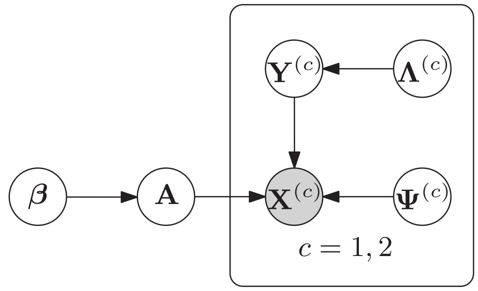
Common spatial patterns (CSP) is a widely-used discriminative EEG feature extraction method (Blankertz et al. 2008; Koles 1991; Müller-Gerking, Pfurtscheller, and Flyvbjerg 1999; Kang, Nam, and Choi 2009), also known as the Fukunaga-Koontz transform (Fukunaga and Koontz 1970), where we seek a discriminative subspace such that the variance for one class is maximized while the variance for the other class is minimized. CSP was recently cast into a probabilistic framework (Wu et al. 2009), where a linear Gaussian model for each of the positive/negative classes was considered and the maximum likelihood estimate of the basis matrix shared across two models (positive and negative class models) was shown to yield the same solution as CSP. Bayesian models were also proposed for CSP (Wu et al. 2010; Kang and Choi 2011), where posterior distributions over variables of interest are estimated by variational approximation.

We revisit two probabilistic models for CSP. One is probabilistic CSP (PCSP) (Wu et al. 2009) where the maximum likelihood estimate is determined by the expectation maximization (EM) optimization and the other is variational Bayesian CSP (VBCSP) (Kang and Choi 2011) where the posterior distributions over variables in the model are computed by variational inference in the framework of Bayesian multi-task learning (Heskes 2000). EM and variational inference, while successful, often suffer from slow convergence to the solution. Parameter eXpanded-EM (PX-EM) (Liu, Rubin, and Wu 1998) is a method for accelerating EM, using the over-parameterization of the model. The underlying idea in PX-EM is to use a covariance adjustment to correct the analysis of the M step, thereby exploiting extra information captured in the imputed complete data. Similarly, Parameter eXpanded-VB (PX-VB) (Qi and Jaakkola 2007) expands a model with auxiliary parameters to reduce the coupling between variables in the original model, so that it accelerates the deterministic approximation of the target distribution in variational Bayesian inference.

In this study, we employ the parameter-expansion methods of (Liu, Rubin, and Wu 1998; Qi and Jaakkola 2007; Luttinen and Ilin 2010) in order to develop parameter-expanded algorithms for PCSP and VBCSP, leading to PCSP-PX and VBCSP-PX. By capitalizing on the convergence speed-up by parameter-expansion methods, we show that the expanded models, PCSP-PX and VBCSP-PX, con-



(a) PCSP



(b) VBCSP

Figure 1: Graphical representation of PCSP and VBCSP models.

verge to solutions faster than PCSP and VBCSP. In addition, we show that the generalization performance of PCSP-PX and VBCSP-PX is better than that of PCSP and VBCSP. In PCSP and VBCSP, feature vectors are constructed using only variances of the expected latent variables so that the information on covariances is neglected. In contrast, the auxiliary parameters in PCSP-PX and VBCSP-PX reduce such information loss by simultaneous diagonalization of the second-order moments of the latent variables. This is found to improve the generalization performance in the classification. Numerical experiments on the BCI competition datasets, III IVa and IV 2a, confirmed that PCSP-PX and VBCSP-PX not only speed-up the computations but also improve the classification performances of the feature vectors, as compared to PCSP and VBCSP.

Probabilistic Models for CSP

We briefly review two probabilistic models, PCSP (Wu et al. 2009) and VBCSP (Kang and Choi 2011). The graphical representations of these models are shown in Fig. 1(a) and 1(b), respectively.

Suppose that EEG signals involving two different mental tasks ($c \in \{1, 2\}$) recorded at D electrodes over multiple trials constitute a D-dimensional vector $\boldsymbol{x}_t^{(c)}$ for $t = 1, \ldots, T_c$, where T_c represents the number of samples obtained over multiple trials. We denote the EEG data matrix by $\boldsymbol{X}^{(c)} = \left[\boldsymbol{x}_1^{(c)}, ..., \boldsymbol{x}_{T_c}^{(c)}\right] \in \mathbb{R}^{D \times T_c}$. The probabilistic model for CSP assumes that data matrices $\boldsymbol{X}^{(c)}$ are generated by

$$\boldsymbol{X}^{(c)} = \boldsymbol{A}\boldsymbol{Y}^{(c)} + \boldsymbol{E}^{(c)}, \tag{1}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M] \in \mathbb{R}^{D \times M}$ is the *basis matrix* shared across two classes ($c = 1, 2$),

$$\boldsymbol{Y}^{(c)} = \left[\boldsymbol{y}_1^{(c)}, \ldots, \boldsymbol{y}_{T_c}^{(c)}\right] \in \mathbb{R}^{M \times T_c},$$

is the *encoding matrix* (corresponding to latent variables), and $\boldsymbol{E}^{(c)} = \left[\boldsymbol{\epsilon}_1^{(c)}, \ldots, \boldsymbol{\epsilon}_{T_c}^{(c)}\right] \in \mathbb{R}^{D \times T_c}$ is the *noise matrix*. Latent variables and noise are assumed to follow zero-mean Gaussian distributions,

$$\begin{aligned} \boldsymbol{y}_t^{(c)} &\sim \mathcal{N}\left(\boldsymbol{y}_t^{(c)} \,\middle|\, 0, \left[\boldsymbol{\Lambda}^{(c)}\right]^{-1}\right), \\ \boldsymbol{\epsilon}_t^{(c)} &\sim \mathcal{N}\left(\boldsymbol{\epsilon}_t^{(c)} \,\middle|\, 0, \left[\boldsymbol{\Psi}^{(c)}\right]^{-1}\right), \end{aligned}$$

where $\boldsymbol{\Lambda}^{(c)} = \mathrm{diag}\left(\lambda_1^{(c)}, \ldots, \lambda_M^{(c)}\right) \in \mathbb{R}^{M \times M}$ and $\boldsymbol{\Psi}^{(c)} = \mathrm{diag}\left(\psi_1^{(c)}, \ldots, \psi_D^{(c)}\right) \in \mathbb{R}^{D \times D}$ are precision matrices for $c = 1, 2$.

PCSP

In PCSP, the basis matrix $\boldsymbol{A}$ is treated as a matrix of parameters, and their maximum likelihood estimate $\widehat{\boldsymbol{A}}_{ML}$ is determined by EM, where the E-step involves computing the expectation of the complete-data log-likelihood $\log p\left(\{\boldsymbol{X}^{(c)}\}, \{\boldsymbol{Y}^{(c)}\} \,\middle|\, \boldsymbol{A}, \{\boldsymbol{\Lambda}^{(c)}\}, \{\boldsymbol{\Psi}^{(c)}\}\right)$ with respect to the posterior distribution $p\left(\{\boldsymbol{Y}^{(c)}\} \,\middle|\, \{\boldsymbol{X}^{(c)}\}\right)$and the M-step re-estimates $\boldsymbol{A}$ as well as other model parameters $\left\{\boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}\right\}$. It was shown in (Wu et al. 2009) that $\widehat{\boldsymbol{A}}_{ML}^{-\top}$ is equal to the linear transformation matrix computed in CSP, in the case of zero noise limit and when $\boldsymbol{A}$ is a square matrix. CSP feature vectors are constructed by taking logarithms of top-n variances of the projected variables for each class within each trial. In the case of PCSP, CSP features are computed using logarithms of top-n variances for each class of posterior means over latent variables in each trial.

VBCSP

In VBCSP, the basis matrix $\boldsymbol{A}$ is treated as a matrix of random variables, and the automatic relevance determination (ARD) prior is applied to it, i.e.,

$$p\left(\boldsymbol{A}|\boldsymbol{D}_\beta\right) = \prod_{m=1}^{M} \mathcal{N}\left([\boldsymbol{A}]_{:,m}|0, \beta_m^{-1}\boldsymbol{I}_D\right), \tag{2}$$

where $\boldsymbol{I}_D \in \mathbb{R}^{D \times D}$ is the identity matrix, $\boldsymbol{D}_\beta \in \mathbb{R}^{M \times M} = \mathrm{diag}(\beta_1, \ldots, \beta_M)$, and the precision hyperparameters β_m are assumed to follow Gamma distribution:

$$\beta_m \sim \mathrm{Gam}\left(a_0^\beta, b_0^\beta\right). \tag{3}$$

Inferring posterior distributions over β_m leads us to predict an appropriate number of columns in $\boldsymbol{A}$. ARD priors are also applied to $\{\lambda_m^{(c)}\}$ and $\{\psi_d^{(c)}\}$ (which are diagonal entries of precision matrices $\boldsymbol{\Lambda}^{(c)}$ and $\boldsymbol{\Psi}^{(c)}$):

$$\lambda_m^{(c)} \sim \text{Gam}\left(a_0^{\lambda}, b_0^{\lambda}\right), \quad \psi_d^{(c)} \sim \text{Gam}\left(a_0^{\psi}, b_0^{\psi}\right). \tag{4}$$

Variational posterior distributions are determined by variational Bayesian inference (Kang and Choi 2011). Again, logarithms of top-n variances for each class of the variational posterior means of latent variables within each trial are used as the CSP features.

Parameter-Expanded Algorithms

In this section, we present the main contribution of this paper, parameter-expanded algorithms for both PCSP and VBCSP. First, we introduce the parameter-expanded models (PCSP-PX and VBCSP-PX), as shown in Fig. 2(a) and 2(b), inspired by (Luttinen and Ilin 2010). Then we develop a parameter-expanded EM algorithm for PCSP-PX and parameter-expanded variational Bayesian inference for VBCSP-PX.

We introduce an invertible matrix $\boldsymbol{R} \in \mathbb{R}^{M\times M}$. We define $\boldsymbol{A}_* = \boldsymbol{A}\boldsymbol{R}$ and $\boldsymbol{Y}_*^{(c)} = \boldsymbol{R}^{-1}\boldsymbol{Y}^{(c)}$. Then, we can write the model (1) as

$$\boldsymbol{X}^{(c)} = \boldsymbol{A}_*\boldsymbol{Y}_*^{(c)} + \boldsymbol{E}^{(c)}, \tag{5}$$

for $c = 1, 2$, since $\boldsymbol{A}_*\boldsymbol{Y}_*^{(c)} = \boldsymbol{A}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{Y}^{(c)}$. The invertible matrix $\boldsymbol{R}$ introduces a transformation of the encoding matrix $\left\{\boldsymbol{Y}^{(c)}\right\}$ while preserving the conditional distribution $p\left(\boldsymbol{X}^{(c)} \mid \boldsymbol{A}, \boldsymbol{Y}^{(c)}\right)$. We optimize the auxiliary parameters $\boldsymbol{R}$ such that the expected complete-data log-likelihood (for PCSP-PX) or the variational lower-bound (for VBCSP-PX) is maximized.

PCSP-PX

We present a parameter-expanded EM algorithm to estimate the model parameters $\left\{\boldsymbol{A}_*, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}\right\}$ as well as auxiliary parameters $\boldsymbol{R}$. The basis matrix $\boldsymbol{A}$ in the original model (1) is recovered by $\boldsymbol{A}_*\boldsymbol{R}^{-1}$. To this end, we consider the complete-data likelihood in the expanded model (5) (shown in Fig. 2(a)):

$$p\left(\left\{\boldsymbol{X}^{(c)}\right\}, \left\{\boldsymbol{Y}_*^{(c)}\right\} \mid \boldsymbol{A}_*, \left\{\boldsymbol{\Lambda}^{(c)}\right\}, \left\{\boldsymbol{\Psi}^{(c)}\right\}, \boldsymbol{R}\right)$$
$$= \prod_{c=1}^2 p\left(\boldsymbol{X}^{(c)} \mid \boldsymbol{Y}_*^{(c)}, \boldsymbol{A}_*, \boldsymbol{\Psi}^{(c)}\right) p\left(\boldsymbol{Y}_*^{(c)} \mid \boldsymbol{\Lambda}^{(c)}, \boldsymbol{R}\right),$$

where

$$p\left(\boldsymbol{X}^{(c)} \,\middle|\, \boldsymbol{Y}_*^{(c)}, \boldsymbol{A}_*, \boldsymbol{\Psi}^{(c)}\right) = \prod_{t=1}^{T_c} \mathcal{N}\left(\boldsymbol{x}_t^{(c)} \,\middle|\, \boldsymbol{A}_*\boldsymbol{y}_{*t}^{(c)}, \left[\boldsymbol{\Psi}^{(c)}\right]^{-1}\right),$$
$$p\left(\boldsymbol{Y}_*^{(c)} \,\middle|\, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{R}\right) = \prod_{t=1}^{T_c} \mathcal{N}\left(\boldsymbol{y}_{*t}^{(c)} \,\middle|\, 0, \left[\boldsymbol{R}^\top\boldsymbol{\Lambda}^{(c)}\boldsymbol{R}\right]^{-1}\right).$$

In the E-step, we compute the expected complete-data log-likelihood $\langle\mathcal{L}_c\rangle$

$$\sum_{c=1}^{2}\left\langle \log p\left(\boldsymbol{X}^{(c)}, \boldsymbol{Y}_*^{(c)} \mid \boldsymbol{\Theta}\right)\right\rangle,$$

where the expectation $\langle\cdot\rangle$ is taken with respect to the posterior distribution over latent variables $p\left(\boldsymbol{Y}_*^{(c)} \mid \boldsymbol{X}^{(c)}\right)$ given the current estimate of parameters $\boldsymbol{\Theta} = \left\{\boldsymbol{A}_*, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}, \boldsymbol{R}\right\}$. In the M-step, we re-estimate parameters $\boldsymbol{\Theta}$ that maximize $\langle\mathcal{L}_c\rangle$ computed in the E-step. The EM iteration for PCSP-PX, which is summarized in Algorithm 1, alternates between the E-step and M-step until convergence.

In contrast to PCSP, the auxiliary parameters $\boldsymbol{R}$ should also be optimized. The stationary point equation for $\boldsymbol{R}$ is given by

$$\begin{aligned}\frac{\partial\langle\mathcal{L}_c\rangle}{\partial\boldsymbol{R}} &= \left(\sum_{c=1}^2 T_c\right)\boldsymbol{R}^{-\top} - \sum_{c=1}^2 \boldsymbol{\Lambda}^{(c)}\boldsymbol{R}\left\langle\boldsymbol{Y}_*^{(c)}\boldsymbol{Y}_*^{(c)\top}\right\rangle \\ &= 0,\end{aligned} \tag{6}$$

leading to

$$\sum_{c=1}^2 \boldsymbol{\Lambda}^{(c)}\boldsymbol{R}\left\langle\boldsymbol{Y}_*^{(c)}\boldsymbol{Y}_*^{(c)\top}\right\rangle\boldsymbol{R}^\top = \left(\sum_{c=1}^2 T_c\right)\boldsymbol{I}_M, \tag{7}$$

which is solved for $\boldsymbol{R}$ by simultaneous diagonalization of the second-order moments of encodings $\left\langle\boldsymbol{Y}_*^{(1)}\boldsymbol{Y}_*^{(1)\top}\right\rangle$ and $\left\langle\boldsymbol{Y}_*^{(2)}\boldsymbol{Y}_*^{(2)\top}\right\rangle$, followed by re-scaling.

The posterior distribution over latent variables in the original model is easily computed by

$$p\left(\boldsymbol{y}_t^{(c)} \mid \boldsymbol{x}_t^{(c)}\right) = \mathcal{N}\left(\boldsymbol{y}_t^{(c)} \mid \boldsymbol{R}\boldsymbol{\mu}_t^{(c)}, \boldsymbol{R}\boldsymbol{\Sigma}^{(c)}\boldsymbol{R}^\top\right),$$

where $\boldsymbol{\mu}_t^{(c)}$ and $\boldsymbol{\Sigma}^{(c)}$ are calculated in the E-step in Algorithm 1. We compute CSP features using posterior means $\boldsymbol{R}\boldsymbol{\mu}_t^{(c)} = \boldsymbol{R}\boldsymbol{\Sigma}^{(c)}\boldsymbol{R}^\top\boldsymbol{A}^\top\boldsymbol{\Psi}^{(c)}\boldsymbol{x}_t^{(c)}$ that correspond to projected variables in CSP. Given test trial data $\boldsymbol{X} \in \mathbb{R}^{D\times T}$, we first compute T M-dimensional posterior mean vectors over latent variables, constructing posterior mean matrices $\overline{\boldsymbol{Y}}^{(c)} \in \mathbb{R}^{M\times T}$ for $c = 1, 2$,

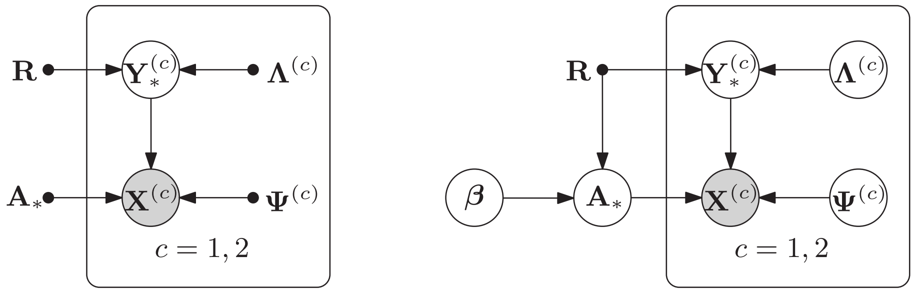
$$\overline{\boldsymbol{Y}}^{(c)} = \boldsymbol{R}\boldsymbol{\Sigma}^{(c)}\boldsymbol{R}^\top\boldsymbol{A}^\top\boldsymbol{\Psi}^{(c)}\boldsymbol{X}.$$

To model the project variables in CSP, we average $\overline{\boldsymbol{Y}}^{(c)}$ of the two classes considering the class prior probability as $p(\boldsymbol{X} \in (c)) = \frac{T_c}{T_1+T_2}$,

$$\overline{\boldsymbol{Y}} = \sum_{c=1}^2 \frac{T_c}{T_1+T_2}\overline{\boldsymbol{Y}}^{(c)}. \tag{8}$$

Then, we build a vector $\boldsymbol{z} \in \mathbb{R}^M$, the m-th entry of which is computed by

$$[\boldsymbol{z}]_m = \log\left(\frac{1}{T}\left[\overline{\boldsymbol{Y}}\overline{\boldsymbol{Y}}^\top\right]_{m,m} - \left(\frac{1}{T}\left[\overline{\boldsymbol{Y}}\boldsymbol{1}_T\right]_m\right)^2\right), \tag{9}$$



(a) PCSP-PX (b) VBCSP-PX

Figure 2: Graphical representation of PCSP-PX and VBCSP-PX models.

Algorithm 1 EM for PCSP-PX

Input: EEG data $\{\boldsymbol{X}^{(c)}\}$.

Output: estimate of parameters $\boldsymbol{\Theta} = \{\boldsymbol{A}_*, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}, \boldsymbol{R}\}$

initialize $\boldsymbol{\Theta} = \{\boldsymbol{A}_*, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}, \boldsymbol{R}\}$.

repeat

E-step Calculate the posterior distribution over latent variables $p\left(\boldsymbol{y}_{*t}^{(c)} \,\middle|\, \boldsymbol{x}_t^{(c)}\right)$:

$$
\begin{aligned}
p\left(\boldsymbol{y}_{*t}^{(c)} \,\middle|\, \boldsymbol{x}_t^{(c)}\right) &= \mathcal{N}\left(\boldsymbol{y}_{*t}^{(c)} \,\middle|\, \boldsymbol{\mu}_t^{(c)}, \boldsymbol{\Sigma}^{(c)}\right), \\
\boldsymbol{\mu}_t^{(c)} &= \boldsymbol{\Sigma}^{(c)} \boldsymbol{A}_*^\top \boldsymbol{\Psi}^{(c)} \boldsymbol{x}_t^{(c)}, \\
\left[\boldsymbol{\Sigma}^{(c)}\right]^{-1} &= \boldsymbol{R}^\top \boldsymbol{\Lambda}^{(c)} \boldsymbol{R} + \boldsymbol{A}_*^\top \boldsymbol{\Psi}^{(c)} \boldsymbol{A}_*.
\end{aligned}
$$

M-step Re-estimate $\boldsymbol{\Theta}$:

- Update $\{\boldsymbol{A}_*, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}\}$:

$$
\begin{aligned}
[\boldsymbol{A}_*]_{d,:} &= \left(\sum_{c=1}^{2} \psi_d^{(c)} [\boldsymbol{X}^{(c)}]_{d,:} \left\langle \boldsymbol{Y}_*^{(c)\top} \right\rangle\right) \\
&\quad \left(\sum_{c=1}^{2} \psi_d^{(c)} \left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle\right)^{-1}, \\
\left[\psi_d^{(c)}\right]^{-1} &= \frac{1}{T_c} \Big[\boldsymbol{X}^c \boldsymbol{X}^{(c)\top} - 2\boldsymbol{X}^{(c)} \left\langle \boldsymbol{Y}_*^{(c)\top} \right\rangle \boldsymbol{A}_*^\top \\
&\quad + \boldsymbol{A}_* \left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle \boldsymbol{A}_*^\top \Big]_{d,d}, \\
\left[\lambda_m^{(c)}\right]^{-1} &= \frac{1}{T_c} \left[\boldsymbol{R} \left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle \boldsymbol{R}^\top\right]_{m,m}.
\end{aligned}
$$

- Solve (7) for $\boldsymbol{R}$ by simultaneous diagonalization of $\left\langle \boldsymbol{Y}_*^{(1)} \boldsymbol{Y}_*^{(1)\top} \right\rangle$ and $\left\langle \boldsymbol{Y}_*^{(2)} \boldsymbol{Y}_*^{(2)\top} \right\rangle$ to update $\boldsymbol{R}$.

until convergence.

---

where $\mathbf{1}_T \in \mathbb{R}^T$ is the vector of all ones. We choose $2n$ $[\boldsymbol{z}]_m$'s with m corresponding to n largest and n smallest values of the ratio $\lambda_m^{(1)}/\lambda_m^{(2)}$ to construct the CSP feature vector $\boldsymbol{f} \in \mathbb{R}^{2n}$.

VBCSP-PX

We present a parameter-expanded variational Bayesian inference to speed up the deterministic approximation of the posterior distributions over variables $\mathcal{Z} = \left\{\boldsymbol{Y}_*^{(c)}, \boldsymbol{A}_*, \beta_m, \boldsymbol{\Psi}^{(c)}, \boldsymbol{\Lambda}^{(c)}\right\}$ with auxiliary parameters $\boldsymbol{R}$. Variational posterior distributions over $\boldsymbol{A}$ and $\boldsymbol{Y}^{(c)}$ in the original model are easily recovered by variable transformation: $\boldsymbol{A} = \boldsymbol{A}_* \boldsymbol{R}^{-1}$ and $\boldsymbol{Y}^{(c)} = \boldsymbol{R}\boldsymbol{Y}_*^{(c)}$. We write the joint distribution over $\left\{\boldsymbol{X}^{(c)}\right\}$ and $\mathcal{Z}$ in the expanded model (shown in Fig. 2(b)) with prior distributions defined in (2), (3), and (4) as

$$
\begin{aligned}
& p\left(\left\{\boldsymbol{X}^{(c)}\right\}, \left\{\boldsymbol{Y}_*^{(c)}\right\}, \boldsymbol{A}_*, \{\beta_m\}, \left\{\boldsymbol{\Lambda}^{(c)}\right\}, \left\{\boldsymbol{\Psi}^{(c)}\right\} \,\middle|\, \boldsymbol{R}\right) \\
&= \prod_{c=1}^{2} p\left(\boldsymbol{X}^{(c)} \,\middle|\, \boldsymbol{Y}_*^{(c)}, \boldsymbol{A}_*, \boldsymbol{\Psi}^{(c)}\right) p\left(\boldsymbol{Y}_*^{(c)} \,\middle|\, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{R}\right) \\
&\quad p\left(\boldsymbol{\Lambda}^{(c)}\right) p\left(\boldsymbol{\Psi}^{(c)}\right) p(\boldsymbol{A}_* | \boldsymbol{D}_\beta, \boldsymbol{R}) \prod_{m=1}^{M} p(\beta_m),
\end{aligned}
$$

where

$$
p\left(\boldsymbol{Y}_*^{(c)} \,\middle|\, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{R}\right) = \prod_{t=1}^{T_c} \mathcal{N}\left(\boldsymbol{y}_{*t}^{(c)} \,\middle|\, 0, \left[\boldsymbol{R}^\top \boldsymbol{\Lambda}^{(c)} \boldsymbol{R}\right]^{-1}\right).
$$

The prior distribution over $\boldsymbol{A}_* \in \mathbb{R}^{D \times M}$ is assumed to be *matrix-variate Gaussian* since it is not column-wise independent in contrast to (2). Thus we assume

$$
p\left(\boldsymbol{A}_* | \boldsymbol{D}_\beta, \boldsymbol{R}\right) = \mathcal{N}_{D \times M}\left(\boldsymbol{A}_* \,\middle|\, 0, \boldsymbol{I}_D \otimes \boldsymbol{R}^\top \boldsymbol{D}_\beta^{-1} \boldsymbol{R}\right),
$$

where matrix-variate Gaussian distribution for a random matrix $\boldsymbol{B} \in \mathbb{R}^{D \times M}$ with mean matrix $\boldsymbol{M} \in \mathbb{R}^{D \times M}$ and covariance matrix $\boldsymbol{\Omega}_D \otimes \boldsymbol{\Omega}_M$ ($\boldsymbol{\Omega}_D \in \mathbb{R}^{D \times D}$ and $\boldsymbol{\Omega}_M \in \mathbb{R}^{M \times M}$) takes the form

$$
\begin{aligned}
& \mathcal{N}_{D \times M}\left(\boldsymbol{B} \,|\, \boldsymbol{M}, \boldsymbol{\Omega}_D \otimes \boldsymbol{\Omega}_M\right) \\
&= (2\pi)^{-\frac{DM}{2}} |\boldsymbol{\Omega}_D|^{-\frac{M}{2}} |\boldsymbol{\Omega}_M|^{-\frac{D}{2}} \\
&\quad \exp\left[-\frac{1}{2} \mathrm{tr}\Big(\boldsymbol{\Omega}_D^{-1} (\boldsymbol{B} - \boldsymbol{M}) \boldsymbol{\Omega}_M^{-1} (\boldsymbol{B} - \boldsymbol{M})^\top\Big)\right].
\end{aligned}
$$

Note that the prior distribution over $\boldsymbol{A}$ in the original model, given in (2), can also be written as

$$
p\left(\boldsymbol{A} | \boldsymbol{D}_\beta\right) = \mathcal{N}_{D \times M}\left(\boldsymbol{A} \,\middle|\, 0, \boldsymbol{I}_D \otimes \boldsymbol{D}_\beta^{-1}\right).
$$

The variational inference involves the maximization of a lower-bound on the *marginal log-likelihood* given by

$$\begin{aligned}
& \log p\left(\left\{\boldsymbol{X}^{(c)}\right\} \mid \boldsymbol{R}\right) \\
= \;& \log \int p\left(\left\{\boldsymbol{X}^{(c)}\right\}, \mathcal{Z} \mid \boldsymbol{R}\right) d\mathcal{Z} \\
\geq \;& \int q(\mathcal{Z}) \log \frac{p\left(\left\{\boldsymbol{X}^{(c)}\right\}, \mathcal{Z} \mid \boldsymbol{R}\right)}{q(\mathcal{Z})} d\mathcal{Z} \\
\equiv \;& \mathcal{F}(q|\boldsymbol{R}),
\end{aligned} \tag{10}$$

where variational posterior distribution $q(\mathcal{Z})$ is assumed to factorize as

$$q(\mathcal{Z}) = q\left(\boldsymbol{A}_*\right) q\left(\left\{\boldsymbol{y}_{*t}^{(c)}\right\}\right) q\left(\left\{\beta_m\right\}\right) q\left(\left\{\psi_d^{(c)}\right\}\right) q\left(\left\{\lambda_m^{(c)}\right\}\right).$$

Variational posterior distributions over each variable are alternatively updated such that the lower-bound $\mathcal{F}(q|\boldsymbol{R})$ is maximized. Updating equations are summarized in Algorithm 2. In addition, auxiliary parameters $\boldsymbol{R}$ are optimized by the maximization of $\mathcal{F}(q|\boldsymbol{R})$, given the variational posterior distribution $q(\mathcal{Z})$. We consider the terms involving $\boldsymbol{R}$ in the lower-bound $\mathcal{F}(q|\boldsymbol{R})$, given by

$$\begin{aligned}
& \frac{\sum_{c=1}^{2} T_c - D}{2} \log\left(|\boldsymbol{R}|^2\right) \\
& -\frac{1}{2} \sum_{c=1}^{2} \operatorname{tr}\left(\left\langle \boldsymbol{\Lambda}^{(c)} \right\rangle \boldsymbol{R} \left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle \boldsymbol{R}^\top\right) \\
& -\frac{1}{2} \operatorname{tr}\left(\left\langle \boldsymbol{D}_\beta \right\rangle \boldsymbol{R}^{-\top} \left\langle \boldsymbol{A}_*^\top \boldsymbol{A}_* \right\rangle \boldsymbol{R}^{-1}\right).
\end{aligned} \tag{11}$$

Suppose a_0^β and b_0^β are set to small values so that $\langle \beta_m \rangle$ can be approximated as

$$\begin{aligned}
\langle \beta_m \rangle &= \frac{a_m^\beta}{b_m^\beta} = \frac{a_0^\beta + D/2}{b_0^\beta + \frac{1}{2}\left[\boldsymbol{R}^{-\top}\left\langle \boldsymbol{A}_*^\top \boldsymbol{A}_* \right\rangle \boldsymbol{R}^{-1}\right]_{m,m}} \\
&\simeq \frac{D}{\left[\boldsymbol{R}^{-\top}\left\langle \boldsymbol{A}_*^\top \boldsymbol{A}_* \right\rangle \boldsymbol{R}^{-1}\right]_{m,m}},
\end{aligned}$$

implying that

$$\begin{aligned}
& \operatorname{tr}\left(\left\langle \boldsymbol{D}_\beta \right\rangle \boldsymbol{R}^{-\top} \left\langle \boldsymbol{A}_*^\top \boldsymbol{A}_* \right\rangle \boldsymbol{R}^{-1}\right) \\
& = \sum_{m=1}^{M} \langle \beta_m \rangle \left[\boldsymbol{R}^{-\top}\left\langle \boldsymbol{A}_*^\top \boldsymbol{A}_* \right\rangle \boldsymbol{R}^{-1}\right]_{m,m} \\
& \simeq MD.
\end{aligned} \tag{12}$$

Applying the approximation (12) to (11), as in (Luttinen and Ilin 2010), we have the stationary point equation for $\boldsymbol{R}$ given by

$$\begin{aligned}
& \frac{\partial \mathcal{F}(q|\boldsymbol{R})}{\partial \boldsymbol{R}} \\
= \;& \left(\sum_{c=1}^{2} T_c - D\right) \boldsymbol{R}^{-\top} - \sum_{c=1}^{2} \left\langle \boldsymbol{\Lambda}^{(c)} \right\rangle \boldsymbol{R} \left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle \\
= \;& 0.
\end{aligned} \tag{13}$$

Again, we can solve the equation for $\boldsymbol{R}$ by simultaneous diagonalization of the second-order moments of encodings $\left\langle \boldsymbol{Y}_*^{(1)} \boldsymbol{Y}_*^{(1)\top} \right\rangle$ and $\left\langle \boldsymbol{Y}_*^{(2)} \boldsymbol{Y}_*^{(2)\top} \right\rangle$, followed by re-scaling, as in PCSP-PX.

Algorithm 2 Variational Bayesian Inference for VBCSP-PX

Input: EEG data $\{\boldsymbol{X}^{(c)}\}$

Output: approximated posterior $q(\mathcal{Z})$ for variables $\mathcal{Z} = \{\boldsymbol{A}_*, \boldsymbol{Y}_*^{(c)}, \boldsymbol{\Lambda}^{(c)}, \boldsymbol{\Psi}^{(c)}\}$ and the auxiliary parameter $\boldsymbol{R}$

initialize $q(\mathcal{Z})$.

repeat

- Update $q(\boldsymbol{A}_*) = \prod_{d=1}^{D} \mathcal{N}\left([\boldsymbol{A}_*]_{d,:} | \bar{\boldsymbol{\nu}}_d, \boldsymbol{\Omega}_d\right)$ by

$$\begin{aligned}
[\boldsymbol{\Omega}_d]^{-1} &= \boldsymbol{R}^{-1} \left\langle \boldsymbol{D}_\beta \right\rangle \boldsymbol{R}^{-\top} \\
&\quad + \sum_{c=1}^{2} \left\langle \psi_d^{(c)} \right\rangle \left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle, \\
\bar{\boldsymbol{\nu}}_d &= \sum_{c=1}^{2} \left\langle \psi_d^{(c)} \right\rangle [\boldsymbol{X}^{(c)}]_{d,:} \left\langle \boldsymbol{Y}_*^{(c)\top} \right\rangle \boldsymbol{\Omega}_d.
\end{aligned}$$

- Update $q\left(\boldsymbol{y}_{*t}^{(c)}\right) = \mathcal{N}\left(\boldsymbol{\mu}_t^{(c)}, \boldsymbol{\Sigma}^{(c)}\right)$ by

$$\begin{aligned}
\left[\boldsymbol{\Sigma}^{(c)}\right]^{-1} &= \boldsymbol{R}^\top \boldsymbol{\Lambda}^{(c)} \boldsymbol{R} + \left\langle \boldsymbol{A}_*^\top \boldsymbol{\Psi}^{(c)} \boldsymbol{A}_* \right\rangle, \\
\boldsymbol{\mu}_t^{(c)} &= \boldsymbol{\Sigma}^{(c)} \left\langle \boldsymbol{A}_*^\top \right\rangle \left\langle \boldsymbol{\Psi}^{(c)} \right\rangle \boldsymbol{x}_t^{(c)}.
\end{aligned}$$

- Update $q(\beta_m) = \text{Gam}\left(a_m^\beta, b_m^\beta\right)$ by

$$\begin{aligned}
a_m^\beta &= a_0^\beta + D/2, \\
b_m^\beta &= b_0^\beta + \frac{1}{2}\left[\boldsymbol{R}^{-\top}\left\langle \boldsymbol{A}_*^\top \boldsymbol{A}_* \right\rangle \boldsymbol{R}^{-1}\right]_{m,m}.
\end{aligned}$$

- Update $q\left(\psi_d^{(c)}\right) = \text{Gam}\left(a_d^{\psi(c)}, b_d^{\psi(c)}\right)$ by

$$\begin{aligned}
a_d^{\psi(c)} &= a_0^\psi + T_c/2, \\
b_d^{\psi(c)} &= b_0^\psi + \frac{1}{2}\Big[\boldsymbol{X}^{(c)} \boldsymbol{X}^{(c)\top} \\
&\quad -2 \langle \boldsymbol{A}_* \rangle \left\langle \boldsymbol{Y}_*^{(c)} \right\rangle \boldsymbol{X}^{(c)\top} \\
&\quad + \left\langle \boldsymbol{A}_* \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \boldsymbol{A}_*^\top \right\rangle\Big]_{d,d}.
\end{aligned}$$

- Update $q\left(\lambda_m^{(c)}\right) = \text{Gam}\left(a_m^{\lambda(c)}, b_m^{\lambda(c)}\right)$ by

$$\begin{aligned}
a_m^{\lambda(c)} &= a_0^\lambda + T_c/2, \\
b_m^{\lambda(c)} &= b_0^\lambda + \frac{1}{2}\left[\boldsymbol{R}\left\langle \boldsymbol{Y}_*^{(c)} \boldsymbol{Y}_*^{(c)\top} \right\rangle \boldsymbol{R}^\top\right]_{m,m}.
\end{aligned}$$

- Solve (13) for $\boldsymbol{R}$ by simultaneous diagonalization of $\left\langle \boldsymbol{Y}_*^{(1)} \boldsymbol{Y}_*^{(1)\top} \right\rangle$ and $\left\langle \boldsymbol{Y}_*^{(2)} \boldsymbol{Y}_*^{(2)\top} \right\rangle$ to update $\boldsymbol{R}$.

until convergence.

The approximated posterior over latent variables in the original model is computed by

$$q\left(\boldsymbol{y}_t^{(c)}\right) = \mathcal{N}\left(\boldsymbol{y}_t^{(c)} \,\middle|\, \boldsymbol{R}\boldsymbol{\mu}_t, \boldsymbol{R}\boldsymbol{\Sigma}^{(c)}\boldsymbol{R}^\top\right),$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}^{(c)}$ are calculated as in Algorithm 2. We compute a CSP feature vector $\boldsymbol{f}$ for a test trial $\boldsymbol{X} \in \mathbb{R}^{D\times T}$ using the posterior means $\boldsymbol{R}\boldsymbol{\mu}_t^{(c)} = \boldsymbol{R}\boldsymbol{\Sigma}^{(c)}\left\langle \boldsymbol{A}_*^\top \right\rangle \left\langle \boldsymbol{\Psi}^{(c)} \right\rangle \boldsymbol{x}_t^{(c)}$,

$$\overline{\boldsymbol{Y}}^{(c)} = \boldsymbol{R}\boldsymbol{\Sigma}^{(c)}\boldsymbol{R}^\top \left\langle \boldsymbol{A}^\top \right\rangle \left\langle \boldsymbol{\Psi}^{(c)} \right\rangle \boldsymbol{X}.$$

We compute $\overline{\boldsymbol{Y}}$ and $[\boldsymbol{z}]_m$ as in (8) and (9). Then we select $[\mathbf{z}]_m$'s with m corresponding to n-largest and n-smallest values of the ratio of $\left\langle \lambda_m^{(1)} \right\rangle / \left\langle \lambda_m^{(2)} \right\rangle$ to construct the CSP feature vector $\boldsymbol{f} \in \mathbb{R}^{2n}$.

Numerical Experiments

We compared the performances of PCSP, VBCSP, PCSP-PX, and VBCSP-PX on the BCI competition datasets, III IVa (Blankertz et al. 2006)[1] and IV 2a [2]. Both datasets consist of the EEG measurements of several subjects during motor imagery tasks. The EEG data was pre-processed by band-pass filtering, to emphasize important frequency bands for recognizing the motor imagery tasks. Every trial was divided into the same number of time intervals after each visual cue, which contains EEG variation caused by the imagination of the subject.

We extracted feature vectors $\boldsymbol{f} \in \mathbb{R}^{2n}$ using PCSP, VBCSP, PCSP-PX, and VBCSP-PX, and we applied the linear discriminant analysis (LDA) to transform these feature vectors down to scalar values which are fed into a minimum distance classifier. We set $D = M$ and $n = 3$ for every model. The classification performance of each model is represented by the prediction accuracy of the LDA classifier on the test trials. The accuracy was calculated as the ratio of the number of correctly classified trials to the total number of test trials. We repeated the experiments 10 times, varying the number of training trials, while the number of test trials was fixed. We selected half of the trials in each data as the test trials, and randomly selected some of the remaining trials as the training trials. The classes were strictly balanced by selecting the same number of trials from each class.

BCI competition III IVa dataset was collected from five subjects using 118 electrodes ($D = 118$) during the imagery movements of the right hand and right foot. The trials were separated by up to 3.5s after each cue, and we used the down-sampled version (100 Hz) of the data. 140 trials were conducted for each subject and each class. BCI competition IV 2a dataset contains 4 motor imagery tasks of 9 subjects, recorded using 22 electrodes ($D = 22$). We considered only the binary classification problem so that we selected the imagery left/right hand movement classes. The trials were separated by from 3.5s to 5.5s after each cue, and the sampling rate was 250 Hz. 144 trials were conducted for each subject and each class.

[1]http://www.bbci.de/competition/iii/

[2]http://www.bbci.de/competition/iv/

Compared to PCSP and VBCSP, parameter-expanded algorithms PCSP-PX and VBCSP-PX perform additional computation to optimize $\boldsymbol{R}$ at every iteration. However, PCSP-PX and VBCSP-PX converge in a smaller number of iterations; hence, they are faster than PCSP and VBCSP, respectively (Table 1). In general, the classification performance of PCSP-PX and VBCSP-PX was also higher than that of PCSP and VBCSP (Fig. 3).

Conclusions

We have presented two new parameter-expanded algorithms for PCSP and VBCSP, leading to PCSP-PX and VBCSP-PX, where we expanded the models using auxiliary parameters $\boldsymbol{R}$ to speed up the convergence as well as to improve the performance. The auxiliary parameters $\boldsymbol{R}$ were estimated by simultaneous diagonalization of $\left\langle \boldsymbol{Y}_*^{(1)}\boldsymbol{Y}_*^{(1)\top} \right\rangle$ and $\left\langle \boldsymbol{Y}_*^{(2)}\boldsymbol{Y}_*^{(2)\top} \right\rangle$, reducing the coupling so that the convergence was accelerated and the performance was improved, while CSP features determined by PCSP or VBCSP ignored off-diagonal entries of the empirical second-order moment matrix of posterior mean vectors. Numerical experiments on the BCI competition datasets, III IVa and IV 2a, demonstrated the high performance of PCSP-PX and VBCSP-PX, as compared to their counterparts PCSP and VBCSP.

Acknowledgments: This work was supported by National Research Foundation (NRF) of Korea (2011-0018283, 2011-0018284), MEST Converging Research Center Program (2011K000673), NIPA Program of Software Engineering Technologies Development and Experts Education, MKE and NIPA "IT Consilience Creative Program" (C1515-1121-0003), and NRF World Class University Program (R31-10100).

References

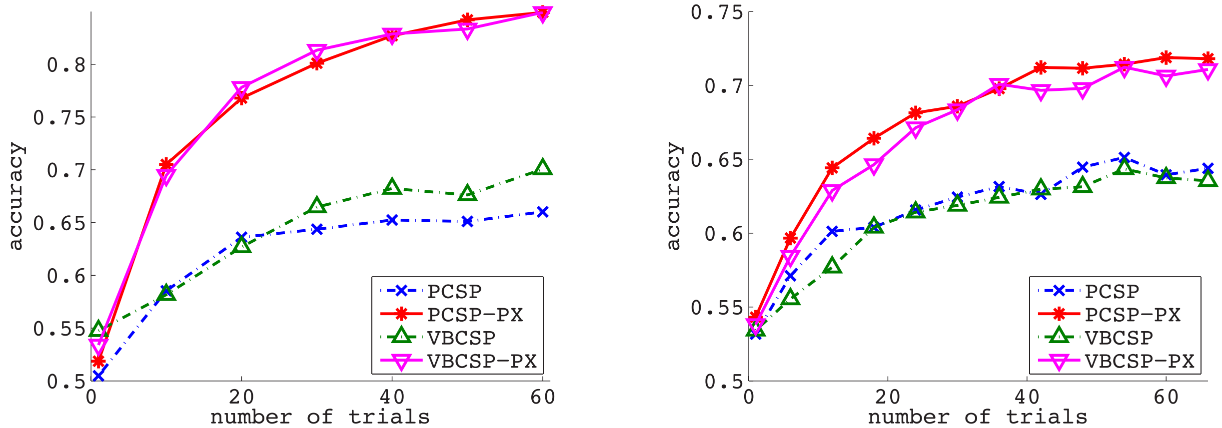
Blankertz, B.; Müller, K. R.; Krusierski, D. J.; Schalk, G.; Wolpaw, J. R.; Schlögl, A.; Pfurtscheller, G.; and Birbaumer, N. 2006. The BCI competition III: Validating alternative approaches to actual BCI problems. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14:153–159.

Blankertz, B.; Tomioka, R.; Lemm, S.; Kawanabe, M.; and Müller, K. R. 2008. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine* 41–56.

Cichocki, A.; Washizawa, Y.; Rutkowski, T.; Bakardjian, H.; Phan, A. H.; Choi, S.; Lee, H.; Zhao, Q.; Zhang, L.; and Li, Y. 2008. Noninvasive BCIs: Multiway signal-processing array decompositions. *IEEE Computer* 41(10):34–42.

Ebrahimi, T.; Vesin, J. F.; and Garcia, G. 2003. Brain-computer interface in multimedia communication. *IEEE Signal Processing Magazine* 20(1):14–24.

Fukunaga, K., and Koontz, W. L. G. 1970. Application of the Karhunen-Loève expansion to feature selection and ordering. *IEEE Transactions on Computers* 19(4):311–318.



(a) III IVa data

(b) IV 2a data

Figure 3: Classification performances of existing methods (PCSP and VBCSP) as well as our proposed methods (PCSP-PX and VBCSP-PX).

Table 1: Performance comparison in terms of number of iterations and run time. For each algorithm, the iterations stop when the variant of the expected complete-data log-likelihood (for PCSP and PCSP-PX) or the variational lower-bound (for VBCSP and VBCSP-PX) falls within a pre-defined value. The maximum number of iterations was set as 100 for each algorithm. The 'number of iterations', 'run time' and 'run time per iteration' were averaged over multiple runs with 10 different training sets.

data	measure	PCSP	PCSP-PX	VBCSP	VBCSP-PX
III IVa	number of iterations	86.6400 ± 24.0742	$\mathbf{15.7057 \pm 7.0065}$	54.9114 ± 26.1892	$\mathbf{13.8800 \pm 3.4429}$
	run time (sec)	14.3137 ± 4.5311	$\mathbf{3.0026 \pm 1.4251}$	37.1477 ± 16.7673	$\mathbf{12.2370 \pm 3.2140}$
	run time per iteration	$\mathbf{0.1631 \pm 0.0166}$	0.1901 ± 0.0246	$\mathbf{0.6917 \pm 0.1013}$	0.8804 ± 0.0638
IV 2a	number of iterations	58.0417 ± 27.4807	$\mathbf{34.2093 \pm 16.1989}$	18.5046 ± 7.9813	$\mathbf{13.5111 \pm 3.8341}$
	run time (sec)	0.1702 ± 0.0841	$\mathbf{0.1338 \pm 0.0668}$	0.1579 ± 0.0740	$\mathbf{0.1337 \pm 0.0427}$
	run time per iteration	$\mathbf{0.0029 \pm 0.0003}$	0.0039 ± 0.0004	$\mathbf{0.0084 \pm 0.0009}$	0.0098 ± 0.0011

Heskes, T. 2000. Empirical Bayes for learning to learn. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kang, H., and Choi, S. 2011. Bayesian multi-task learning for common spatial patterns. In *Proceedings of the IEEE International Workshop on Pattern Recognition in NeuroImaging (PRNI)*.

Kang, H.; Nam, Y.; and Choi, S. 2009. Composite common spatial pattern for subject-to-subject transfer. *IEEE Signal Processing Letters* 16(8):683–686.

Koles, Z. J. 1991. The quantitative extraction and topographic mapping of the abnormal components. *EEG and Clinical Neurophysilology* 79:440–447.

Liu, C.; Rubin, D. B.; and Wu, Y. N. 1998. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* 85(4):755–770.

Luttinen, J., and Ilin, A. 2010. Transformations in variational Bayesian factor analysis to speed up learning. *Neurocomputing* 73:1093–1102.

Müller-Gerking, J.; Pfurtscheller, G.; and Flyvbjerg, H. 1999. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology* 110:787–798.

Qi, Y., and Jaakkola, T. S. 2007. Parameter expanded variational Bayeisian methods. In *Advances in Neural Information Processing Systems (NIPS)*, volume 19. MIT Press.

Wolpaw, J. R.; Birbaumer, N.; McFarland, D. J.; Pfurtscheller, G.; and Vaughan, T. M. 2002. Brain-computer interfaces for communication and control. *Clinical Neurophysiology* 113:767–791.

Wu, W.; Chen, Z.; Gao, S.; and Brown, E. N. 2009. 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*.

Wu, W.; Chen, Z.; Gao, S.; and Brown, E. N. 2010. Hierarchical Bayesian modeling of inter-trial variability and variational Bayesian learning of common spatial patterns from multichannel EEG. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.