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 neuroprosthetic (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-
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 $x_t^{(c)}$ 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 $X^{(c)} = [x_1^{(c)}, \ldots, x_T^{(c)}] \in \mathbb{R}^{D \times T_c}$. The probabilistic model for CSP assumes that data matrices $X^{(c)}$ are generated by

$$X^{(c)} = A Y^{(c)} + E^{(c)},$$

where $A = [a_1, \ldots, a_M] \in \mathbb{R}^{D \times M}$ is the basis matrix shared across two classes ($c = 1, 2$), $Y^{(c)} = [y_1^{(c)}, \ldots, y_{T_c}^{(c)}] \in \mathbb{R}^{M \times T_c}$ is the encoding matrix (corresponding to latent variables), and $E^{(c)} = [\epsilon_1^{(c)}, \ldots, \epsilon_{T_c}^{(c)}] \in \mathbb{R}^{D \times T_c}$ is the noise matrix.

Latent variables and noise are assumed to follow zero-mean Gaussian distributions,

$$y_t^{(c)} \sim \mathcal{N} \left( y_t^{(c)} \big| 0, \left( \Lambda^{(c)} \right)^{-1} \right),$$

$$\epsilon_t^{(c)} \sim \mathcal{N} \left( \epsilon_t^{(c)} \big| 0, \left( \Psi^{(c)} \right)^{-1} \right),$$

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

**PCSP**

In PCSP, the basis matrix $A$ is treated as a matrix of parameters, and their maximum likelihood estimate $A_{ML}$ is determined by EM, where the E-step involves computing the expectation of the complete-data log-likelihood

$$p \left( \{X^{(c)}\}, \{Y^{(c)}\} \big| A, \{\Lambda^{(c)}\}, \{\Psi^{(c)}\} \right)$$

and the M-step re-estimates $A$ as well as other model parameters $\{\Lambda^{(c)}, \Psi^{(c)}\}$. It was shown in (Wu et al. 2009) that $A_{ML}$ is equal to the linear transformation matrix computed in CSP, in the case of zero noise limit and when $A$ 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 $A$ is treated as a matrix of random variables, and the automatic relevance determination (ARD) prior is applied to it, i.e.,

$$p (A | D_\beta) = \prod_{m=1}^M \mathcal{N} \left( [A]_{:,m} \big| 0, \beta_m^{-1} I_D \right),$$

where $I_D \in \mathbb{R}^{D \times D}$ is the identity matrix, $D_\beta \in \mathbb{R}^{M \times M} = \text{diag}(\beta_1, \ldots, \beta_M)$, and the precision hyperparameters $\beta_m$ are assumed to follow Gamma distribution:

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

Figure 1: Graphical representation of PCSP and VBCSP models.
Inferring posterior distributions over $\beta_m$ leads us to predict an appropriate number of columns in $A$. ARD priors are also applied to $\{\lambda_{m}^{(c)}\}$ and $\{\psi_{d}^{(c)}\}$ (which are diagonal entries of precision matrices $\Lambda^{(c)}$ and $\Psi^{(c)}$):

$$
\lambda_{m}^{(c)} \sim \text{Gam} \left( a_0, b_0 \right), \quad \psi_{d}^{(c)} \sim \text{Gam} \left( a_0^d, b_0^d \right).
$$

(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 $R \in \mathbb{R}^{M \times M}$. We define $A_{x} = A R$ and $Y_{x}^{(c)} = R^{-1} Y^{(c)}$. Then, we can write the model (1) as

$$
X^{(c)} = A_{x} Y_{x}^{(c)} + E^{(c)},
$$

(5)

for $c = 1, 2$, since $A_{x} Y_{x}^{(c)} = A R R^{-1} Y^{(c)}$. The invertible matrix $R$ introduces a transformation of the encoding matrix $\{Y^{(c)}\}$ while preserving the conditional distribution $p \left( X^{(c)} \mid A_{x}, Y^{(c)} \right)$. We optimize the auxiliary parameters $R$ 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 $\{A_{x}, \Lambda^{(c)}, \Psi^{(c)}\}$ as well as auxiliary parameters $R$. The basis matrix $A$ in the original model (1) is recovered by $A_{x}, R^{-1}$. To this end, we consider the complete-data likelihood in the expanded model (5) (shown in Fig. 2(a)):

$$
p \left( \{X^{(c)}\}, \{Y^{(c)}\}, A_{x}, \{\Lambda^{(c)}\}, \{\Psi^{(c)}\}, R \right) = \prod_{c=1}^{2} p \left( X^{(c)} \mid Y_{x}^{(c)}, A_{x}, \Psi^{(c)} \right) p \left( Y_{x}^{(c)} \mid \Lambda^{(c)}, R \right),
$$

where

$$
p \left( X^{(c)} \mid Y_{x}^{(c)}, A_{x}, \Psi^{(c)} \right) = \prod_{t=1}^{T_{c}} N \left( x_{t}^{(c)} \mid A_{x} y_{s_{t}}^{(c)}, \left[ \Psi^{(c)} \right]^{-1} \right),
$$

$$
p \left( Y_{x}^{(c)} \mid \Lambda^{(c)}, R \right) = \prod_{t=1}^{T_{c}} N \left( y_{s_{t}}^{(c)} \mid 0, \left[ R^{T} \Lambda^{(c)} R \right]^{-1} \right).
$$

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

$$
\sum_{c=1}^{2} \langle \log p \left( X^{(c)}, Y_{x}^{(c)} \mid \Theta \right) \rangle,
$$

where the expectation $\langle \cdot \rangle$ is taken with respect to the posterior distribution over latent variables $p \left( Y_{x}^{(c)} \mid X^{(c)} \right)$ given the current estimate of parameters $\Theta = \{A_{x}, \Lambda^{(c)}, \Psi^{(c)}, R\}$. In the M-step, we re-estimate parameters $\Theta$ 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 $R$ should also be optimized. The stationary point equation for $R$ is given by

$$
\frac{\partial \langle \mathcal{L}_{c} \rangle}{\partial R} = \left( \sum_{c=1}^{2} T_{c} \right) R^{-T} - 2 \sum_{c=1}^{2} \Lambda^{(c)} R \left( Y_{x}^{(c)} Y_{x}^{(c)T} \right) = 0,
$$

(6)

leading to

$$
\sum_{c=1}^{2} \Lambda^{(c)} R \left( Y_{x}^{(c)} Y_{x}^{(c)T} \right) R^{-T} = \left( \sum_{c=1}^{2} T_{c} \right) I_{M},
$$

(7)

which is solved for $R$ by simultaneous diagonalization of the second-order moments of encodings $\left\{Y_{x}^{(1)} Y_{x}^{(1)T}\right\}$ and $\left\{Y_{x}^{(2)} Y_{x}^{(2)T}\right\}$, followed by re-scaling.

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

$$
p \left( y_{t}^{(c)} \mid x_{t}^{(c)} \right) = N \left( y_{t}^{(c)} \mid R_{\mu_{t}^{(c)}}, R_{\Sigma^{(c)} R^{T}} \right),
$$

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

$$
Y^{(c)} = R \Sigma^{(c)} R^{T} A_{x}^{T} \Psi^{(c)} X.
$$

To model the project variables in CSP, we average $Y^{(c)}$ of the two classes considering the class prior probability as $p(X \in (c)) = \frac{T_{1}}{T_{1} + T_{2}}$,

$$
Y = \sum_{c=1}^{2} \frac{T_{c}}{T_{1} + T_{2}} Y^{(c)}.
$$

(8)

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

$$
[z]_{m} = \log \left( \frac{1}{T} \left[ Y Y^{T} \right]_{m,m} - \left( \frac{1}{T} \left[ Y 1_{T} \right]_{m} \right)^{2} \right),
$$

(9)
Algorithm 1 EM for PCSP-PX

Input: EEG data \( \{X^{(c)}\} \).
Output: estimate of parameters \( \Theta = \{A_s, \Lambda^{(c)}, \Psi^{(c)}, R\} \).

initialize \( \Theta = \{A_s, \Lambda^{(c)}, \Psi^{(c)}, R\} \).

repeat

E-step Calculate the posterior distribution over latent variables \( p \left( y_{st}^{(c)} \mid x_t^{(c)} \right) \):

\[
p \left( y_{st}^{(c)} \mid x_t^{(c)} \right) = \mathcal{N} \left( y_{st}^{(c)} \mid \mu_t^{(c)}, \Sigma_t^{(c)} \right),
\]

\[
\mu_t^{(c)} = \Sigma_t^{(c)} A_s^T \psi_t^{(c)},
\]

\[
\Sigma_t^{(c)}^{-1} = R^T \Lambda^{(c)} R + A_s^T \Psi^{(c)} A_s.
\]

M-step Re-estimate \( \Theta \):

- Update \( \{A_s, \Lambda^{(c)}, \Psi^{(c)}\} \):

\[
[A_s]_{d,:} = \left( \sum_{c=1}^{2} \psi_d^{(c)} X^{(c)} d,: \right) \left( \sum_{c=1}^{2} y_d^{(c)} Y^{(c)}^T \right)^{-1},
\]

\[
\left[ \psi_d^{(c)} \right]^{-1} = \frac{1}{T_c} \left[ X^{(c)} Y^{(c)}^T - 2 X^{(c)} Y^{(c)}^T A_s^T + A_s Y^{(c)} Y^{(c)}^T A_s^T \right]_{d,d},
\]

\[
\left[ \Lambda^{(c)} \right]^{-1} = \frac{1}{T_c} \left[ R^T Y^{(c)} Y^{(c)}^T R \right]_{m,m}.
\]

- Solve (7) for \( R \) by simultaneous diagonalization of \( Y^{(1)} Y^{(1)^T} \) and \( Y^{(2)} Y^{(2)^T} \) to update \( R \).

until convergence.

VBCSP-PX

We present a parameter-expanded variational Bayesian inference to speed up the deterministic approximation of the posterior distributions over variables \( Z = \{Y^{(c)}_s, A_s, \beta_m, \Psi^{(c)}, \Lambda^{(c)}\} \) with auxiliary parameters \( R \).

Variational posterior distributions over \( A \) and \( Y^{(c)} \) in the original model are easily recovered by variable transformation: \( A = A_s R^{-1} \) and \( Y^{(c)} = R Y^{(c)} \).

We write the joint distribution over \( \{X^{(c)}\} \) and \( Z \) in the expanded model (shown in Fig. 2(b)) with prior distributions defined in (2), (3), and (4) as

\[
p \left( \left\{ X^{(c)} \right\}, \left\{ Y^{(c)}_s \right\}, A_s, \beta_m, \left\{ \Lambda^{(c)} \right\}, \left\{ \Psi^{(c)} \right\} \mid R \right) = \prod_{c=1}^{2} p \left( x^{(c)} \mid y^{(c)}_s, A_s, \beta_m \right) p \left( y^{(c)}_s \mid \Lambda^{(c)}, R \right) p \left( \Lambda^{(c)} \right) p \left( \Psi^{(c)} \right) p(A_s \mid D_{\beta}, R) \prod_{m=1}^{M} p(\beta_m),
\]

where

\[
p \left( y^{(c)}_s \mid \Lambda^{(c)}, R \right) = \prod_{t=1}^{T_c} \mathcal{N} \left( y^{(c)}_{st} \mid 0, \left[ R^T \Lambda^{(c)} R \right]^{-1} \right).
\]

The prior distribution over \( A_s \) is assumed to be matrix-variate Gaussian since it is not column-wise independent in contrast to (2). Thus we assume

\[
p \left( A_s \mid D_{\beta}, R \right) = \mathcal{N}_{D \times M} \left( A_s \mid 0, I_D \otimes R^T D_{\beta}^{-1} R \right),
\]

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

\[
\mathcal{N}_{D \times M} \left( B \mid M, \Omega_D \otimes \Omega_M \right) = (2\pi)^{-\frac{D M}{2}} |\Omega_D|^{-\frac{D}{2}} |\Omega_M|^{-\frac{M}{2}} \exp \left[ -\frac{1}{2} \operatorname{tr} \left( \Omega_D^{-1} (B - M) \Omega_M^{-1} (B - M)^T \right) \right].
\]

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

\[
p \left( A \mid D_{\beta} \right) = \mathcal{N}_{D \times M} \left( A \mid 0, I_D \otimes D_{\beta}^{-1} \right).
\]
The variational inference involves the maximization of a lower-bound on the marginal log-likelihood given by

$$\log p \left( \{X^{(c)}\} \mid R \right) = \log \int p \left( \{X^{(c)}\}, Z \mid R \right) dZ \geq \int q(Z) \log p \left( \{X^{(c)}\}, Z \mid R \right) q(Z) dZ = \mathcal{F}(q|R),$$

where variational posterior distribution \( q(Z) \) is assumed to factorize as

$$q(Z) = q(A) q(y^{(c)}_t) q(\{\beta_m\}) q(\{\psi_d^{(c)}\}) q(\{\lambda_m^{(c)}\}).$$

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

$$\sum_{c=1}^{2T_e - D} \frac{1}{2} \sum_{c=1}^{2} \text{tr} \left( \Lambda^{(c)} R \left( Y_s^{(c)} Y_s^{(c)\top} \right) R^\top \right)$$

$$= \frac{1}{2} \text{tr} \left( \left( D_\beta \right) R^\top \left( A_s^\top A_s \right) R^{-1} \right).$$

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

$$\langle \beta_m \rangle = \frac{a_0^\beta}{b_0^\beta} = \frac{a_0^\beta + D/2}{b_0^\beta + \frac{1}{2} \left[ R^{-\top} \left( A_s^\top A_s \right) R^{-1} \right]_{m,m}}.$$

implying that

$$\text{tr} \left( \left( D_\beta \right) R^\top \left( A_s^\top A_s \right) R^{-1} \right) = \sum_{c=1}^{M} \langle \beta_m \rangle \left[ R^{-\top} \left( A_s^\top A_s \right) R^{-1} \right]_{m,m} \approx MD.$$

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

$$\frac{\partial \mathcal{F}(q|R)}{\partial R} = \left( \sum_{c=1}^{2} T_e - D \right) R^\top - \sum_{c=1}^{2} \langle \Lambda^{(c)} \rangle R \left( Y_s^{(c)} Y_s^{(c)\top} \right) = 0.$$

Again, we can solve the equation for \( R \) by simultaneous diagonalization of the second-order moments of encodings \( \{Y_s^{(1)} Y_s^{(1)\top}\} \) and \( \{Y_s^{(2)} Y_s^{(2)\top}\} \), followed by re-scaling, as in PCSP-PX.

**Algorithm 2 Variational Bayesian Inference for VBCSP-PX**

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

**Output:** approximated posterior \( q(Z) \) for variables \( Z = \{A_s, Y_s^{(c)}, \Lambda^{(c)}, \psi^{(c)}_d\} \) and the auxiliary parameter \( R \)

**Initialize** \( q(Z) \).

**Repeat**

- Update \( q(A_s) = \prod_{d=1}^{D} N([A_d], \mu, \Omega_d) \) by

$$\Omega_d^{-1} = R^{-1} \left( D_\beta \right) R^\top + \frac{2}{\mu} \left( \psi_d^{(c)} \right) \left( Y_s^{(c)} Y_s^{(c)\top} \right),$$

$$\mu_d = \frac{1}{\mu} \left( \psi_d^{(c)} \right) \left( Y_s^{(c)} \right)^\top \Omega_d,$$

- Update \( q(y^{(c)}_t) = N(\mu^{(c)}, \Sigma^{(c)}) \) by

$$\Sigma^{(c)}^{-1} = R^\top \Lambda^{(c)} R + \left( A_s^\top \psi^{(c)} \right) \left( A_s^\top \psi^{(c)} \right)^\top,$$

$$\mu^{(c)} = \Sigma^{(c)} \left( A_s^\top \psi^{(c)} \right) \left( A_s^\top \psi^{(c)} \right)^\top \Lambda^{(c)},$$

- Update \( q(\beta_m) = \text{Gam}(a^\beta_m, b^\beta_m) \) by

$$a^\beta_m = a_0^\beta + D/2,$$

$$b^\beta_m = b_0^\beta + \frac{1}{2} \left[ \left( R^\top \left( A_s^\top A_s \right) R^{-1} \right)_{m,m} \right].$$

- Update \( q(\psi_d^{(c)}) = \text{Gam}(a^\psi_d^{(c)}, b^\psi_d^{(c)}) \) by

$$a^\psi_d^{(c)} = a_0^\psi + T_e/2,$$

$$b^\psi_d^{(c)} = b_0^\psi + \frac{1}{2} \left[ X^{(c)} X^{(c)\top} \right]_{d,d} \left( A_s^{(c)} \right) \left( Y_s^{(c)} \right)^\top + \left( A_s^{(c)} Y_s^{(c)} \right)^\top \left( A_s^{(c)} Y_s^{(c)} \right)^\top R^{-1} \left( A_s^{(c)} Y_s^{(c)} \right)^\top \Lambda^{(c)}.$$

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

$$a_m^{\lambda^{(c)}} = a_0^{\lambda} + T_e/2,$$

$$b_m^{\lambda^{(c)}} = b_0^{\lambda} + \frac{1}{2} \left[ R \left( Y_s^{(1)} Y_s^{(1)\top} \right) \left( Y_s^{(2)} Y_s^{(2)\top} \right) \right]_{m,m}.$$

- Solve (13) for \( R \) by simultaneous diagonalization of \( \left\{ Y_s^{(1)} Y_s^{(1)\top} \right\} \) and \( \left\{ Y_s^{(2)} Y_s^{(2)\top} \right\} \) to update \( R \).

**Until** convergence.

The approximated posterior over latent variables in the
We compute $R \in \Sigma$ usual cue, which contains EEG variation caused by the imag- 
band-pass filtering, to emphasize important frequency bands 
motor imagery tasks. The EEG data was pre-
processed by IVa (Blankertz et al. 2006) 
We compared the performances of PCSP, VBCSP, PCSP-
f feature vector 
values of the ratio of $\mu$ 
original model is computed by 

$$
q \left( y_t^{(c)} \right) = N \left( y_t^{(c)}, R\mu_t, R\Sigma^{(c)}R^\top \right),
$$

where $\mu_t$ and $\Sigma^{(c)}$ are calculated as in Algorithm 2. We compute a CSP feature vector $f$ for a test trial 
$X \in \mathbb{R}^{D \times T}$ using the posterior means $R\mu_t^{(c)} = R\Sigma^{(c)} \left( A^*_c \right) \left( \tilde{\psi}(c) \right) x_t^{(c)}$, 

$$
\tilde{Y}^{(c)} = R\Sigma^{(c)}R^\top \left( A^*_c \right) \left( \tilde{\psi}(c) \right) X.
$$

We compute $\tilde{Y}$ and $|z|_m$ as in (8) and (9). Then we se-
select $|z|_m$’s with $m$ corresponding to $n$-largest and $n$-smallest 
values of the ratio of $\langle \lambda_m^{(1)} \rangle / \langle \lambda_m^{(2)} \rangle$ to construct the CSP 
feature vector $f \in \mathbb{R}^{2n}$.

**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) and IV 2a. 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 di-
vided into the same number of time intervals after each vi-
sual cue, which contains EEG variation caused by the imag-
ination of the subject.

We extracted feature vectors $f \in \mathbb{R}^{2n}$ using PCSP, VBCSP, PCSP-PX, and VBCSP-PX, and we applied the lin-
ard 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 rep-
resented 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 tri-
als 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 im-
egagy 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 im-
egagy left/right hand movement classes. The trials were sepa-
rated 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.

Compared to PCSP and VBCSP, parameter-expanded algorithms PCSP-PX and VBCSP-PX perform additional 
computation to optimize $R$ at every iteration. However, 
PCSP-PX and VBCSP-PX converge in a smaller number of iterations; hence, they are faster than PCSP and VBCSP, re-
spectively (Table 1). In general, the classification perfor-
mance 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 parame-
ters $R$ to speed up the convergence as well as to improve 
the performance. The auxiliary parameters $R$ were esti-
imated by simultaneous diagonalization of $\langle Y^{(1)}_* \rangle^\top$ and $\langle Y^{(2)}_* \rangle^\top$, reducing the coupling so that the conver-
gence was accelerated and the performance was improved, 
while CSP features determined by PCSP or VBCSP ignored 
of-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 Na-
tional Research Foundation (NRF) of Korea (2011-0018283, 
2011-0018284), MEXT Converging Research Center Pro-
gram (2011K000673), NIPA Program of Software Engi-
neering 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 Bir-
bauer, N. 2006. The BCI competition III: Validating al-
terative approaches to actual BCI problems. *IEEE Trans-
actions 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 Mag-
azine* 41–56.

Cichocki, A.; Washizawa, Y.; Ruttkowski, T.; Bakardjian, 
H.; Phan, A. H.; Choi, S.; Lee, H.; Zhao, Q.; Zhang, L.; and 
Li, Y. 2008. Noninvasive BCIs: Multiway signal-processing 

Ebrahimi, T.; Vesin, J. F.; and Garcia, G. 2003. Brain-

Fukunaga, K., and Koontz, W. L. G. 1970. Application of the Karhunen-Loève expansion to feature selection and 

1 http://www.bbci.de/competition/i/iiii/
2 http://www.bbci.de/competition/iv/
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.

<table>
<thead>
<tr>
<th></th>
<th>measure</th>
<th>PCSP</th>
<th>PCSP-PX</th>
<th>VBCSP</th>
<th>VBCSP-PX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>number of iterations</td>
<td></td>
<td></td>
<td>number of iterations</td>
</tr>
<tr>
<td>III IVa</td>
<td></td>
<td>86.6400 ± 24.0742</td>
<td>15.7057 ± 7.0065</td>
<td>54.9114 ± 26.1892</td>
<td>15.7057 ± 7.0065</td>
</tr>
<tr>
<td></td>
<td>run time per iteration</td>
<td>0.1631 ± 0.0166</td>
<td>0.1901 ± 0.0246</td>
<td>0.6917 ± 0.1013</td>
<td>0.1901 ± 0.0246</td>
</tr>
<tr>
<td></td>
<td>run time (sec)</td>
<td>0.1702 ± 0.0841</td>
<td>0.1338 ± 0.0668</td>
<td>0.1579 ± 0.0740</td>
<td>0.1338 ± 0.0668</td>
</tr>
<tr>
<td></td>
<td>run time per iteration</td>
<td>0.0029 ± 0.0003</td>
<td>0.0039 ± 0.0004</td>
<td>0.0084 ± 0.0009</td>
<td>0.0039 ± 0.0004</td>
</tr>
</tbody>
</table>

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


