BAYESIAN MULTI-SUBJECT COMMON SPATIAL PATTERNS WITH INDIAN BUFFET PROCESS PRIORS

Hyohyeong Kang[1] *and Seungjin Choi*[1,2]

[1] Department of Computer Science and Engineering, POSTECH, Korea
[2] Division of IT Convergence Engineering, POSTECH, Korea
{paanguin, seungjin}@postech.ac.kr

ABSTRACT

Common spatial patterns (CSP) or its probabilistic counterpart, PCSP, is a popular discriminative feature extraction method for electroencephalography (EEG) classification. Models in CSP or PCSP are trained on a subject-by-subject basis so that inter-subject information is not used. In the case of *multi-subject EEG classification* where brain waves recorded from multiple subjects who undergo the same mental task are available, it is desirable to capture inter-subject relatedness in learning a model. In this paper we present a nonparametric Bayesian model for a multi-subject extension of CSP where subject relatedness is captured by assuming that spatial patterns across subjects share a latent subspace. Spatial patterns and the shared latent subspace are jointly learned by variational inference. We use an infinite latent feature model to automatically infer the dimension of the shared latent subspace, placing Indian Buffet process (IBP) priors on our model. Numerical experiments on BCI competition IV 2a dataset demonstrate the high performance of our method, compared to PCSP and existing Bayesian multi-task CSP models.

***Index Terms*—** Brain computer interface, common spatial patterns, EEG classification, Indian Buffet processes, nonparametric Bayesian methods

1. INTRODUCTION

Electroencephalography (EEG) is the recording of electrical potentials using multiple sensors placed on a scalp, to collect multivariate time series data involving brain activities. EEG classification allows computers to translate a subject's intention or mental status into a control signal for a device, which is important for brain-computer interfaces (BCI) [1, 9, 12]. Multi-subject EEG classification involves the categorization of brain waves measured from multiple subjects, each of whom undergoes the same mental task, so that task-specific and subject-specific characteristics as well as inter-subject variations need to be considered.

Common spatial patterns (CSP) seeks a subject-specific spatial filter to extract discriminative features from EEG [9]. In its probabilistic counterpart, PCSP [15], two linear Gaussian generative models with a shared basis matrix are jointly learned to infer *spatial patterns* corresponding to column vectors of the shared basis matrix. Models in CSP or PCSP do not consider inter-subject relatedness, so learning spatial patterns are performed on a subject-by-subject basis. In the case of a subject with much fewer training samples, the performance of PCSP is deteriorated.

Multi-subject extension of CSP can be found in [2, 8, 10, 11]. In regularized CSP methods, class-conditional covariance matrices for a subject of interest are regularized by a linear combination of other subjects' covariance matrices, in order to incorporate inter-subject relatedness [8, 10, 11]. Learning CSP is re-formulated as a risk minimization problem, where regularization is added to constrain spatial filters to become similar across subjects [2]. Those methods directly extend CSP to consider multiple subjects, whereas this work extends PCSP by assuming multi-subject priors.

Bayesian multi-task learning [5] deals with several related tasks at the same time, with the intention that the tasks will *learn from each other* by sharing hyperparameters (parameters of prior distributions). A Bayesian multi-task extension of CSP (BCSP) was recently developed in [6], where subject-to-subject information was transferred during the learning of model for a subject of interest by sharing hyperparameters across subjects, while treating subjects as tasks. BCSP [6] works better than PCSP, although similarities among spatial patterns are neglected because all spatial patterns are forced to share the same hyperparameters. Bayesian CSP with Dirichlet process (DP) priors (BCSP-DP) [7] jointly learns and groups spatial patterns, so that spatial patterns in the same group, determined by the DP mixture model, share the hyperparameters of their prior distributions. Coupling similar spatial patterns in the same cluster by sharing hyperparameters facilitates information transfer among subjects with similar spatial patterns, whereas information transfer is prevented among dissimilar subjects. However, information transfer across clusters is not possible, so that the common characteristics across all the subjects are not captured. We propose this work to alleviate those limitations in two previous Bayesian CSP models.

In this paper we present a nonparametric Bayesian model for a multi-subject extension of CSP where given multi-subject EEG data, allowing inter-subject relatedness to be captured by assuming that spatial patterns across subjects share a latent subspace. To this end, we develop Bayesian CSP with Indian Buffet process (IBP) priors [4], referred to as BCSP-IBP, where spatial patterns and the shared latent subspace are jointly learned by variational inference. Our method, BCSP-IBP, is motivated by multi-task learning methods, based on infinite latent feature models [13, 16]. Numerical experiments on BCI competition IV 2a dataset demonstrate the high performance of our method, compared to PCSP [15] and existing Bayesian multi-task CSP models [6, 7].

2. RELATED WORK

We briefly review PCSP [15] and our earlier work on Bayesian CSP (BCSP) [6]. We are given the data matrix $\boldsymbol{X}^{s,c} = [\boldsymbol{x}_1^{s,c}, \ldots, \boldsymbol{x}_{N_{sc}}^{s,c}] \in \mathbb{R}^{D \times N_{sc}}$ which is a collection of EEG signals measured from D electrodes over trials (N_{sc} is the number of samples recorded for a pre-defined number of trials) for subject $s \in \{1, \ldots, S\}$ who under-

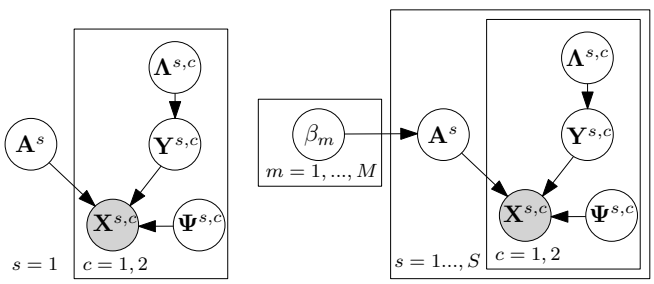
goes the mental task involving class $c \in \{1, 2\}$. The probabilistic model in PCSP or BCSP assumes that $\boldsymbol{X}^{s,c}$ is generated by

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

where $\boldsymbol{A}^s = [\boldsymbol{a}_1^s, \ldots, \boldsymbol{a}_M^s] \in \mathbb{R}^{D \times M}$ is the *basis matrix* for subject 's', containing M *spatial patterns* shared across classes, $\boldsymbol{Y}^{s,c} = [\boldsymbol{y}_1^{s,c}, \ldots, \boldsymbol{y}_{N_{sc}}^{s,c}] \in \mathbb{R}^{M \times N_{sc}}$ the *coefficient matrix*, and $\boldsymbol{E}^{s,c} = [\boldsymbol{\epsilon}_1^{s,c}, \ldots, \boldsymbol{\epsilon}_{N_{sc}}^{s,c}] \in \mathbb{R}^{D \times N_{sc}}$ is the noise matrix. Each row of $\boldsymbol{X}^{s,c}$ is assumed to be already centered (zero mean). Coefficients and noise are assumed to be zero-mean Gaussians:

$$\boldsymbol{y}_t^{s,c} \sim \mathcal{N}(\boldsymbol{y}_t^{s,c} \,|\, \boldsymbol{0}, (\boldsymbol{\Lambda}^{s,c})^{-1}), \quad \boldsymbol{\epsilon}_t^{s,c} \sim \mathcal{N}(\boldsymbol{\epsilon}_t^{s,c} \,|\, \boldsymbol{0}, (\boldsymbol{\Psi}^{s,c})^{-1}), \tag{2}$$

where $\boldsymbol{\Lambda}^{s,c}$ and $\boldsymbol{\Psi}^{s,c}$ are diagonal precision matrices for $s = 1, \ldots, S$ and $c = 1, 2$, each diagonal entry of which is assumed to follow Gamma distribution. When $S = 1$ (subject-specific model), the model (1) reduces to PCSP, where spatial patterns are learned by EM optimization [15] (see Fig. 1(a)).



(a) Probabilistic CSP (b) Bayesian CSP

Fig. 1. Graphical representations of PCSP and BCSP models.

Bayesian CSP [6], shown in Fig. 1(b), employed a Bayesian multi-task learning method [5], enforcing spatial patterns (isotropic Gaussian prior placed on) across subjects to share the hyperparameters (precision of Gaussian) of their prior distributions which are allowed for learning from each other subjects:

$$p(\boldsymbol{A}^s) \quad = \quad \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{a}_m^s | \boldsymbol{0}, \beta_m^{-1} \boldsymbol{I}_D),$$

for $s = 1, \ldots, S$ and the precision variables are assumed to follow gamma prior distribution, $p(\beta_m) = \text{Gam}(\beta_m | a_0^\beta, b_0^\beta)$. Variational inference was used to determine the variational posterior distribution over $\boldsymbol{A}^s$ [6].

3. BAYESIAN CSP WITH IBP PRIORS

In this section we present the main contribution of this paper. Bayesian CSP [6], shown in Fig. 1(b), enforces hyperparameters $\{\beta_m\}$ of the prior distributions over spatial patterns $\{\boldsymbol{a}_m^s\}$ to be shared across subjects, so that all spatial patterns are coupled through the shared hyperparameters. This often facilitates information transfer between subjects whose spatial patterns are much different, leading to the degradation of performance. In order to overcome this limitation, Bayesian CSP with Dirichet process (DP) priors [7] was developed, where DP mixture model was incorporated such that spatial patterns are simultaneously learned and clustered

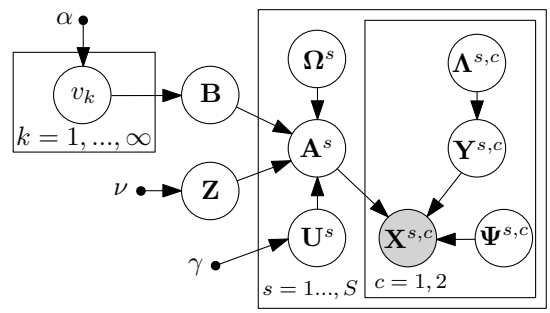


Fig. 3. Graphical representation for Bayesian CSP with IBP priors.

across subjects. Spatial patterns in the same cluster share the hyperparameters of their prior distributions, facilitating information transfer between subjects with similar spatial patterns. However, information is not transferred across clusters in this model.

In order to alleviate limitations existing in our earlier Bayesian CSP models [6, 7], we base our nonparametric Bayesian model on infinite latent feature models [13, 16], assuming that spatial patterns across subjects share a latent subspace to capture subject relatedness. In our model, spatial patterns and the shared latent subspace are jointly learned and the dimension of the latent subspace is automatically inferred. We assume that the subject-specific basis matrix $\boldsymbol{A}^s$ for each subject is generated by

$$\boldsymbol{A}^s \quad = \quad (\boldsymbol{B} \odot \boldsymbol{Z}) \boldsymbol{U}^s + \boldsymbol{\Xi}^s, \tag{3}$$

where the common basis matrix is modeled as $(\boldsymbol{B} \odot \boldsymbol{Z})$ of a binary matrix $\boldsymbol{B} \in \mathbb{R}^{D \times K}$ and a real-valued matrix $\boldsymbol{Z} \in \mathbb{R}^{D \times K}$. $\odot$ represents the Hadamard product (element-wise product), i.e., $[\boldsymbol{B} \odot \boldsymbol{Z}]_{d,k} = B_{d,k} Z_{d,k}$. One important issue in the model (3) is to determine the intrinsic dimension K of the shared latent subspace. Nonparametric Bayesian methods provide a flexible way to infer the dimension K, allowing it to be infinite. The Indian Buffet Process (IBP) [4] is a stochastic process which is served as a nonparametric Bayesian prior over infinite binary matrices. We place an IBP prior on the binary matrix $\boldsymbol{B}$, so that the dimensionality K of the shared latent subspace is automatically inferred. $\boldsymbol{U}^s \in \mathbb{R}^{K \times M}$ is the subject-specific coefficient matrix, and $\boldsymbol{\Xi}^s \in \mathbb{R}^{D \times M}$ represents the additive noise matrix. Gaussian priors are placed on $\boldsymbol{Z}$ and $\boldsymbol{U}^s$, and the noise is also assumed to be Gaussian, thus BCSP-IBP is formulated by

$$\begin{aligned}
\boldsymbol{B} &\sim \text{IBP}(\alpha), \\
Z_{d,k} &\sim \mathcal{N}(0, \nu^{-1}), \\
U_{k,m}^s &\sim \mathcal{N}(0, \gamma^{-1}), \\
\boldsymbol{a}_m^s &\sim \mathcal{N}([\boldsymbol{A}^s]_{:,m} | (\boldsymbol{B} \odot \boldsymbol{Z})[\boldsymbol{U}^s]_{:,m}, (\boldsymbol{\Omega}^s)^{-1}),
\end{aligned}$$

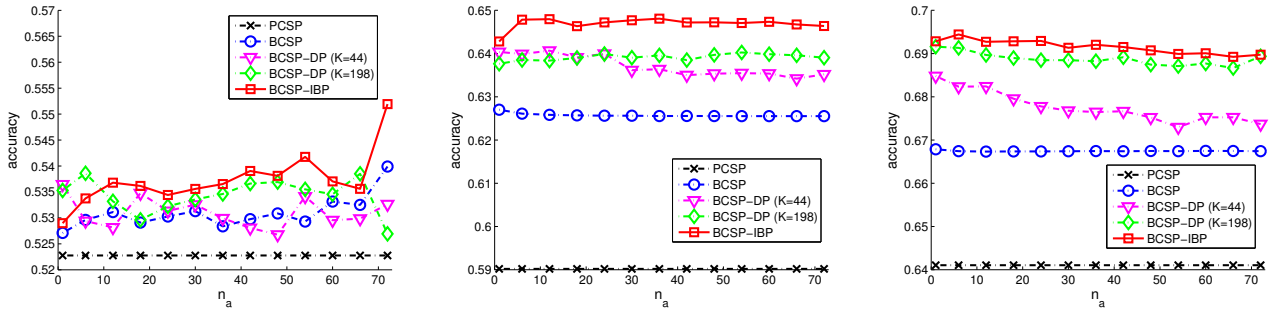
where $\boldsymbol{\Omega}^s \in \mathbb{R}^{D \times D}$ is a diagonal matrix. Diagonal entries of $\boldsymbol{\Lambda}^{s,c}$, $\boldsymbol{\Psi}^{s,c}$ and $\boldsymbol{\Omega}^s$ are assumed to be drawn from Gamma distributions:

$$\begin{aligned}
\lambda_m^{s,c} &\sim \text{Gam}(\lambda_m^{s,c} | a_0^\lambda, b_0^\lambda), \\
\psi_d^{s,c} &\sim \text{Gam}(\psi_d^{s,c} | a_0^\psi, b_0^\psi), \\
\omega_d^s &\sim \text{Gam}(\omega_d^s | a_0^\omega, b_0^\omega).
\end{aligned}$$

BCSP-IBP also uses the parameterizations in (1) and (2) (Fig. 3). The stick-breaking construction of the IBP prior [14] represents the probability of selecting the k-th entry in d-th row of $\boldsymbol{B}$ as breaking a unit-length stick into an infinite number of pieces successively,

Table 1. Updating equations for variational parameters in BCSP-IBP.

Variational posterior distributions	Updating equations for variational parameters
$q(\boldsymbol{A}^s) = \prod_{d=1}^{D} \mathcal{N}([\boldsymbol{A}^s]_{d,:}^\top \vert \boldsymbol{\nu}_d^s, \boldsymbol{\Phi}_d^s)$	$(\boldsymbol{\Phi}_d^s)^{-1} = \langle \omega_d^s \rangle \boldsymbol{I}_M + \sum_{c=1}^{2} \langle \boldsymbol{\psi}_d^{s,c} \rangle \left\langle \boldsymbol{Y}^{s,c} \boldsymbol{Y}^{s,c\top} \right\rangle,$ $\boldsymbol{\nu}_d^s = \boldsymbol{\Phi}_d^s \left\{ \langle \omega_d^s \rangle \left\langle \boldsymbol{U}^{s\top} \right\rangle \left\langle [\boldsymbol{B} \odot \boldsymbol{Z}]_{d,:}^\top \right\rangle + \sum_{c=1}^{2} \langle \psi_d^{s,c} \rangle \left[\langle \boldsymbol{Y}^{s,c} \rangle \boldsymbol{X}^{s,c\top} \right]_{:,d} \right\}$
$q(\boldsymbol{B}) = \prod_{D=1}^{D} \prod_{k=1}^{K} \text{Bern}(B_{d,k} \vert r_{d,k})$ $r_{d,k} = \frac{1}{1+\exp(-\bar{r}_{d,k})}$	$\bar{r}_{d,k} = -\frac{1}{2} \sum_{s=1}^{S} \langle \omega_d^s \rangle \left\{ \left\langle Z_{d,k}^2 \right\rangle \left\langle [\boldsymbol{U}^s \boldsymbol{U}^{s\top}]_{kk} \right\rangle \right.$ $+2 \langle Z_{d,k} \rangle \langle [\boldsymbol{U}^s]_{k,:} \rangle \left(\sum_{l \neq k} \left\langle [\boldsymbol{U}^s]_{l,:}^\top \right\rangle \langle B_{d,l} \rangle \langle Z_{d,l} \rangle - \left\langle [\boldsymbol{A}^s]_{d,:}^\top \right\rangle \right) \Big\}$ $+ \sum_{j=1}^{k} \left(\langle \log v_j \rangle - \pi_{kj} \langle \log(1-v_j) \rangle - \pi_{kj} \sum_{l=1}^{j-1} \langle \log v_l \rangle + \pi_{kj} \log \pi_{kj} \right)$
$q(\boldsymbol{Z}) = \prod_{D=1}^{D} \mathcal{N}([\boldsymbol{Z}]_{d,:}^\top \vert \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$	$(\boldsymbol{\Sigma}_d)^{-1} = \nu \boldsymbol{I}_K + \sum_{s=1}^{S} \langle \omega_d^s \rangle \left\langle \boldsymbol{U}^s \boldsymbol{U}^{s\top} \right\rangle \odot \left\langle [\boldsymbol{B}]_{d,:}^\top [\boldsymbol{B}]_{d,:} \right\rangle,$ $\boldsymbol{\mu}_d = \boldsymbol{\Sigma}_d \left\{ \sum_{s=1}^{S} \left(\langle \omega_d^s \rangle \langle \boldsymbol{U}^s \rangle \left\langle [\boldsymbol{A}^s]_{d,:}^\top \right\rangle \right) \odot \left\langle [\boldsymbol{B}]_{d,:}^\top \right\rangle \right\}$
$q(\boldsymbol{U}^s) = \prod_{m=1}^{M} \mathcal{N}([\boldsymbol{U}^s]_{:,m} \vert \boldsymbol{\tau}_m^s, \boldsymbol{\Delta}_m^s)$	$(\boldsymbol{\Delta}_m^s)^{-1} = \gamma \boldsymbol{I}_K + \left\langle (\boldsymbol{B} \odot \boldsymbol{Z})^\top \boldsymbol{\Omega}^s (\boldsymbol{B} \odot \boldsymbol{Z}) \right\rangle,$ $\boldsymbol{\tau}_m^s = \boldsymbol{\Delta}_m^s (\langle \boldsymbol{B} \rangle \odot \langle \boldsymbol{Z} \rangle)^\top \langle \boldsymbol{\Omega}^s \rangle \langle [\boldsymbol{A}^s]_{:,m} \rangle$
$q(\boldsymbol{v}) = \prod_{k=1}^{K} \text{Beta}(v_k \vert a_k^v, b_k^v)$	$a_k^v = \alpha + \sum_{j=K}^{K} D_j + \sum_{j=k+1}^{K} \sum_{l=k+1}^{j} (D - D_j) \pi_{j,l}, \quad D_k = \sum_{d=1}^{D} \langle B_{d,k} \rangle,$ $b_k^v = 1 + \sum_{j=k}^{K} (D - D_j \pi_{j,k}), \quad \pi_{k,l} \propto \exp\left(\langle \log(1-v_l) \rangle + \sum_{j=1}^{l-1} \langle \log v_j \rangle \right)$
$q(\boldsymbol{Y}^{s,c}) = \prod_{t=1}^{N_{sc}} \mathcal{N}(\boldsymbol{y}_t^{s,c} \vert \eta_t^{s,c}, \boldsymbol{\Sigma}^{s,c})$	$(\boldsymbol{\Sigma}^{s,c})^{-1} = \langle \boldsymbol{\Lambda}^{s,c} \rangle + \langle \boldsymbol{A}^s \boldsymbol{\Psi}^{s,c} \boldsymbol{A}^s \rangle, \quad \eta_t^{s,c} = \boldsymbol{\Sigma}^{s,c} \left\langle \boldsymbol{A}^{s\top} \right\rangle \langle \boldsymbol{\Psi}^{s,c} \rangle \boldsymbol{x}_t^{s,c}$
$q(\boldsymbol{\Lambda}^{s,c}) = \prod_{m=1}^{M} \text{Gam}(\lambda_m^{s,c} \vert a_m^{\lambda s,c}, b_m^{\lambda s,c})$	$a_m^{\lambda s,c} = a_0^\lambda + \frac{N_{sc}}{2}, \quad b_m^{\lambda s,c} = b_0^\lambda + \frac{1}{2} \left\langle [\boldsymbol{Y}^{s,c} \boldsymbol{Y}^{s,c\top}]_{m,m} \right\rangle$
$q(\boldsymbol{\Psi}^{s,c}) = \prod_{d=1}^{D} \text{Gam}(\psi_d^{s,c} \vert a_d^{\psi s,c}, b_d^{\psi s,c})$	$a_d^{\psi s,c} = a_0^\psi + \frac{N_{sc}}{2},$ $b_d^{\psi s,c} = b_0^\psi + \frac{1}{2} \Big[\boldsymbol{X}^{s,c} \boldsymbol{X}^{s,c\top} - \boldsymbol{X}^{s,c} \left\langle \boldsymbol{Y}^{s,c\top} \right\rangle \left\langle \boldsymbol{A}^{s\top} \right\rangle - \langle \boldsymbol{A}^s \rangle \langle \boldsymbol{Y}^{s,c} \rangle \boldsymbol{X}^{s,c}$ $+ \left\langle \boldsymbol{A}^s \boldsymbol{Y}^{s,c} \boldsymbol{Y}^{s,c\top} \boldsymbol{A}^{s\top} \right\rangle \Big]_{d,d}$
$q(\boldsymbol{\Omega}^s) = \prod_{d=1}^{D} \text{Gam}(\omega_d^s \vert a_d^{\omega s}, b_d^{\omega s})$	$a_d^{\omega s} = a_0^\omega + \frac{M}{2},$ $b_d^{\omega s} = b_0^\omega + \frac{1}{2} \Big[\left\langle \boldsymbol{A}^s \boldsymbol{A}^{s\top} \right\rangle - \langle \boldsymbol{A}^s \rangle \left\langle \boldsymbol{U}^{s\top} \right\rangle (\langle \boldsymbol{B} \rangle \odot \langle \boldsymbol{Z} \rangle)^\top - (\langle \boldsymbol{B} \rangle \odot \langle \boldsymbol{Z} \rangle) \langle \boldsymbol{U}^s \rangle \left\langle \boldsymbol{A}^{s\top} \right\rangle$ $+ \left\langle (\boldsymbol{B} \odot \boldsymbol{Z}) \boldsymbol{U}^s \boldsymbol{U}^{s\top} (\boldsymbol{B} \odot \boldsymbol{Z})^\top \right\rangle \Big]_{d,d}$



(a) $n_t = 1$ (b) $n_t = 12$ (c) $N_t = 24$

Fig. 2. Averaged classification accuracy for target subjects is shown when the number of training samples for non-target subjects, denoted by n_a, varies. Three different plots are shown for $n_t = 1, 12, 24$, where n_t denotes the number of training samples for target subject for each class.

such that $p(B_{d,k} = 1 | \{v_j\}) = \prod_{j=1}^{k} v_j$, where v_j are independent random variables drawn from Beta distribution $\text{Beta}(v_k | \alpha, 1)$. An independent draw v_k is re-scaled, proportional to the length of previous broken piece, $\prod_{j=1}^{k-1} v_j$.

BCSP-IBP provides a flexible model, compared to two previous Bayesian CSP methods, in the sense that the shared latent subspace in BCSP-IBP allows the relatedness across every subjects to be captured, while the subject-specific characteristics is reflected by subject-specific coefficients. BCSP-IBP allows the information sharing between all the spatial patterns from every subjects, whereas the information transfer between subjects in different clusters was prohibited in Bayesian CSP with DP priors [7]. Compared to BCSP [6], BCSP-IBP allows for subject variations encoded by subject-specific coefficients and noises, to alleviate negative effects caused by enforcing the prior distributions of all the spatial patterns to share the

common hyperparameters in BCSP.

We employ the variational inference method [3] to approximately compute the posterior distributions over spatial patterns, where the number of columns of $\boldsymbol{B}$ is limited by a truncation parameter K. We define a set of variables to be inferred as

$$\Theta = \Big\{\{\boldsymbol{A}^s\}, \boldsymbol{B}, \boldsymbol{Z}, \{\boldsymbol{U}^s\}, \{\boldsymbol{\Omega}^s\}, \{\boldsymbol{Y}^{s,c}\}, \boldsymbol{v}, \{\boldsymbol{\Lambda}^{s,c}\}, \{\boldsymbol{\Psi}^{s,c}\}\Big\}.$$

The variational inference considers a lower-bound on the marginal log-likelihood

$$\begin{aligned}\log p(\{\boldsymbol{X}^{s,c}\}) &= \log \int p(\{\boldsymbol{X}^{s,c}\}, \Theta) d\Theta \\ &\geq \int q(\Theta) \log \frac{p(\{\boldsymbol{X}^{s,c}\}, \Theta)}{q(\Theta)} d\Theta \equiv \mathcal{F}(q),\end{aligned}$$

where the Jensen's inequality was used and $\mathcal{F}(q)$ denotes the *variational lower-bound* to be maximized. We assumes that the variational distribution $q(\Theta)$ is factorized:

$$\begin{aligned}q(\Theta) &= q(\{\boldsymbol{A}^s\})\, q(\boldsymbol{B})\, q(\boldsymbol{Z})\, q(\{\boldsymbol{U}^s\})\, q(\{\boldsymbol{\Omega}^{s,c}\}) \\ &\quad q(\{\boldsymbol{Y}^{s,c}\})\, q(\boldsymbol{v})\, q(\{\boldsymbol{\Lambda}^{s,c}\})\, q(\{\boldsymbol{\Psi}^{s,c}\}).\end{aligned}$$

Most of the expectations in $\mathcal{F}(q)$ are easily computed, but the key difficulty lies in computing the expectations $\langle \log p(\boldsymbol{B}|\{v_k\})\rangle$, where $\langle\cdot\rangle$ denotes the statistical expectation with respect to the variational distribution $q(\cdot)$. $\langle \log p(\boldsymbol{B}|\{v_k\})\rangle$ contains the expectation $\left\langle \log\left(1 - \prod_{j=1}^k v_j\right)\right\rangle$, which cannot be computed analytically. We apply the local variational approach that induces a tractable lower-bound on $\mathcal{F}(q)$, following the technique similar to the one used in [3], with additional parameters $\{\pi_{k,l}\}$ such that

$$\begin{aligned}\left\langle \log\left(1 - \prod_{j=1}^k v_j\right)\right\rangle &= \left\langle \log\left(\sum_{l=1}^k \pi_{k,l} \frac{(1-v_l)\prod_{j=1}^{l-1} v_j}{\pi_{k,l}}\right)\right\rangle \\ &\geq \sum_{l=1}^k \pi_{k,l} \left\langle \log\left(\frac{(1-v_l)\prod_{j=1}^{l-1} v_j}{\pi_{k,l}}\right)\right\rangle \quad (4)\end{aligned}$$

where $\pi_{k,l} > 0$ and $\sum_{l=1}^k \pi_{k,l} = 1$ for $k = 1, ..., K$ and $l = 1, ..., k$. The tractable lower-bound on $\mathcal{F}(q)$, $\widetilde{\mathcal{F}}(q|\{\pi_{k,l}\})$, is given by replacing the expectation in $\mathcal{F}(q)$ with the right-side of (4). The learning algorithm iteratively optimizes q given $\{\pi_{k,l}\}$, and $\{\pi_{k,l}\}$ given q to maximize $\widetilde{\mathcal{F}}(q|\{\pi_{k,l}\})$.

Variational posterior distributions, $q(\cdot)$, are determined by maximizing the approximated variational lower-bound $\widetilde{\mathcal{F}}(q|\{\pi_{k,l}\})$, which is summarized in Table 1, with detailed derivations left out due to the space limitation. The hyperparameters of the priors $\{\alpha, \nu, \gamma, a_0^\lambda, b_0^\lambda, a_0^\psi, b_0^\psi, a_0^\omega, b_0^\omega\}$ were also updated to maximize $\widetilde{\mathcal{F}}(q|\{\pi_{k,l}\})$.

To compute feature vectors from test trials $\boldsymbol{X}^s \in \mathbb{R}^{D\times T}$, we computed the expected latent signals $\overline{\boldsymbol{Y}}^s$ using $\eta_t^{s,c}$ in Table 1

$$\overline{\boldsymbol{Y}}^s = \sum_{c\in\{1,2\}} \frac{N_{sc}}{N_s} \boldsymbol{\Sigma}^{s,c} \langle \boldsymbol{A}^{s\top}\rangle \langle \boldsymbol{\Psi}^{s,c}\rangle \boldsymbol{X}^s,$$

where $N_s = \sum_{c\in\{1,2\}} N_{sc}$. Note that the class prior probabilities $p(\boldsymbol{X} \in c)$ are assumed as N_{sc}/N_s. Then we took the log of the variance of each dimension of $\overline{\boldsymbol{Y}}^s$ such that

$$\mathbf{f}^*(m) = \log\left(\frac{1}{T}\left[\overline{\boldsymbol{Y}}^s \overline{\boldsymbol{Y}}^{s\top}\right]_{m,m} - \left(\frac{1}{T}\left[\overline{\boldsymbol{Y}}^s \mathbf{1}_T\right]_{m,m}\right)^2\right),$$

where $\mathbf{1}_T \in \mathbb{R}^T$ is the vector of all ones. We selected top-n and bottom-n dimensions of $\mathbf{f}^*$, according to the expected precision ratio between the classes $\langle\lambda_m^{s1}\rangle/\langle\lambda_m^{s2}\rangle$. We applied the Linear Discriminant Analysis to transform these feature vectors down to scalar values which are fed into a minimum distance classifier. The accuracy was obtained by the ratio of the number of correctly classified test trials compared to the total number of test trials.

4. NUMERICAL EXPERIMENTS

We compare the performance of our proposed model BCSP-IPB to existing models such as PCSP, BCSP [6], and BCSP-DP [7], on the BCI Competition IV[1]-2a data set. The data set contains 9 subjects with 4 imagery movements such that left/right hand, right foot, tongue, and we took trials for left/right hand movements to consider binary classification problem. Each imagery movement consists of 144 trials. Each trial was cut from 3.5s to 5.5s after the cue, and consists of $T = 500$ times points. The data was recorded with 22 electrodes so that $D = 22$. Every trials were bandpass-filtered from 8 Hz to 30 Hz before further processing. The basis matrices $\boldsymbol{A}^s$ were assumed to be square ($M = D$), and the dimensionality of the feature vectors were set to six ($n = 3$) in every models.

At each run of the experiments, we selected one of the subjects as a target. For the target subject, we randomly selected n_t labeled trials from each class as the training data ($N_{sc} = T \cdot n_t$). We also randomly selected n_a labeled trials from each class of the remaining subjects ($N_{sc} = T \cdot n_a$). We evaluated the classification accuracies of the trained models over test trials from the target subject only. The number of test trials was 72 for each class. We repeated each run 10 times and averaged the results from each setting of (n_t, n_a). Fig. 2 shows that the accuracy averaged by target subjects has been improved by using Bayesian CSPs. The results for BCSP-DP is presented for different values of truncation level 'K', which limits the maximum number of clusters. The maximum possible value for K in BCSP-DP was $M \cdot S = 198$. Compared to other Bayesian CSP models, the proposed model higher improvement as n_t increases.

5. CONCLUSIONS

We have presented a Bayesian CSP model with IBP priors for multi-subject EEG classification, where spatial patterns across subjects are assumed to share a latent subspace to capture subject relatedness. Spatial patterns are coupled through sharing common basis vectors but subject-specific characteristics is reflected by coefficients. Our nonparametric Bayesian model is more flexible compared to previous Bayesian CSP models, in the sense that the the dimension of the shared latent subspace is automatically inferred and information transfer is allowed between subjects, depending on their relatedness, without enforcing the hyperparameters shared across subjects or allowing information transfer inside clusters only. Numerical experiments on BCI competition IV 2a dataset confirmed the useful behavior of BCSP-IBP, compared to existing PCSP and other Bayesian Multi-task CSP models.

Acknowledgments: This work was supported by National Research Foundation (NRF) of Korea (2012-0005785 and 2012-0005786), POSTECH Rising Star Program, and NRF World Class University Program (R31-10100).

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

6. 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] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multisubject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, vol. 9, 2011.

[3] F. Doshi-Velez, K. T. Miller, J. V. Gael, and Y. W. Teh, "Variational inference for the Indian buffet process," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida, USA, 2009.

[4] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 18. MIT Press, 2006.

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

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

[7] ——, "Bayesian common spatial patterns with Dirichlet process priors for multi-subject EEG classification," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Brisbane, Australia, 2012.

[8] 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.

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

[10] F. Lotte and C. Guan, "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, Texas, USA, 2010.

[11] ——, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

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

[13] P. Rai and H. Daumé III, "Infinite predictor subspace models for multitask learning," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy, 2010.

[14] Y. W. Teh, D. Görür, and Z. Ghahramani, "Stick-breaking construction for the Indian buffet process," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Juan, Puerto Rico, 2007.

[15] 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.

[16] J. Zhang, Z. Ghahramani, and Y. Yang, "Flexible latent variable models for multi-task learning," *Machine Learning*, vol. 73, no. 3, 2008.