BAYESIAN COMMON SPATIAL PATTERNS WITH PITMAN-YOR PROCESS PRIORS

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, Korea
{paanguin, seungjin}@postech.ac.kr

ABSTRACT

Common spatial patterns (CSP) and probabilistic CSP (PCSP) are popular methods for extracting discriminative features from electroencephalography (EEG), but they are trained on a subject-by-subject basis so that inter-subject information is neglected. When only a few training samples are available for each subject, the performance is degraded. In this paper we present a method for Bayesian CSP with Pitman-Yor process (PYP) priors, in which spatial patterns (corresponding to basis vectors) are simultaneously learned and clustered across subjects using variational inference, allowing for a flexible mixture model where the number of components are also learned. Spatial patterns in the same cluster share the hyperparameters of their prior distributions, so that the information transfer is encouraged between subjects involving similar spatial patterns. Numerical experiments on BCI competition IV 2a dataset demonstrate the high performance of our method, compared to existing PCSP and Bayesian CSP with a single prior distribution.

Index Terms— Common spatial patterns, EEG classification, nonparametric Bayesian methods

1. INTRODUCTION

Multi-subject EEG classification considers EEG from multiple subjects, each of whom undergoes the same mental task, so that such brain waves reflect task-specific and subject-specific characteristics, as well as inter-subject variations. Common spatial patterns (CSP) is a popular discriminative EEG feature extraction method, which is useful for learning a subject-specific spatial filter [1]. Learning common spatial patterns was also cast into a probabilistic framework leading to probabilistic CSP (PCSP) [2], i n which linear Gaussian generative models of two classes with a shared basis matrix are jointly learned to infer *spatial pattern vectors* corresponding to column vectors of the shared basis matrix. However, CSP and PCSP are subject-specific methods, so other subjects' information involving the same task as the subject of interest is not considered. In the case of a subject with much fewer training samples, the performance of CSP and PCSP are deteriorated.

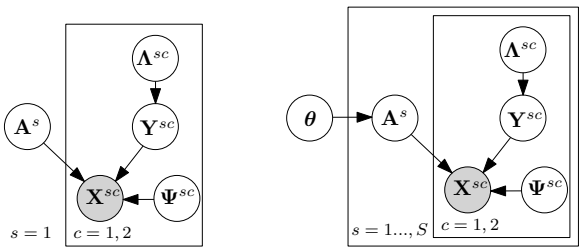
In this paper we propose a Bayesian CSP model where we exploit multi-subject EEG data to learn spatial patterns for a target subject, encouraging information transfer between subjects involving similar spatial patterns. To this end, we present a Bayesian CSP with Pitman-Yor process (PYP) priors, referred to as BCSP-PYP, in which we develop a variational inference algorithm to learn as well as to group spatial pattern vectors, so that spatial pattern vectors in the same group share the hyperparameters of their prior distributions. Coupling similar spatial patterns in the same cluster by sharing the hyperparameters encourages information transfer between subjects involving similar spatial patterns, while information transfer is discouraged between dissimilar subjects. BCSP-PYP is an extension of our recent Bayesian CSP (BCSP) [3] where we assign a single prior distribution to all spatial pattern vectors, regardless of subjects. Our method is motivated by task-clustering methods in the multi-task learning framework [4, 5], where similar tasks are identified and information is transferred between tasks in the same group.

2. RELATED WORK

In this section we briefly review two probabilistic models for CSP, i.e., probabilistic CSP (PCSP) [2] and Bayesian CSP (BCSP) [3]. We denote by $\boldsymbol{X}^{sc} = [\boldsymbol{x}_1^{sc}, \ldots, \boldsymbol{x}_{T_{sc}}^{sc}] \in \mathbb{R}^{D \times T_{sc}}$ a collection of EEG signals measured at D electrodes over trials (T_{sc} is the number of samples recorded for a pre-defined number of trials) for subject $s \in \{1, \ldots, S\}$ who undergoes the mental task involving class $c \in \{1, 2\}$. PCSP or BCSP assumes that $\boldsymbol{X}^{sc}$ is generated by

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

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 pattern vectors* shared across classes, $\boldsymbol{Y}^{sc} = [\boldsymbol{y}_1^{sc}, \ldots, \boldsymbol{y}_{T_{sc}}^{sc}] \in \mathbb{R}^{M \times T_{sc}}$ the *coefficient matrix (latent variables)*, $\boldsymbol{E}^{sc} = [\boldsymbol{\epsilon}_1^{sc}, \ldots, \boldsymbol{\epsilon}_{T_{sc}}^{sc}] \in \mathbb{R}^{D \times T_{sc}}$ is the *noise matrix*. It is assumed that each row of



(a) Probabilistic CSP (b) Bayesian CSP

Fig. 1. Graphical representations of PCSP model [2] and BCSP model [3].

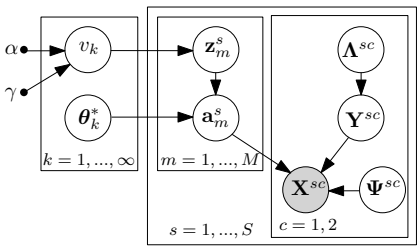


Fig. 2. Graphical representation of Bayesian CSP with PYP priors.

$\boldsymbol{X}^{sc}$ is already centered (zero mean). Coefficients and noise are assumed to be drawn from zero-mean Gaussian distributions:

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

where $\mathbf{\Lambda}^{sc} \in \mathbb{R}^{M \times M}$ and $\mathbf{\Psi}^{sc} \in \mathbb{R}^{D \times D}$ are diagonal precision matrices for $s = 1, \ldots, S$ and $c = 1, 2$, whose diagonal entries are given as $\{\lambda_1^{sc}, \ldots, \lambda_M^{sc}\}$ and $\{\psi_1^{sc}, \ldots, \psi_D^{sc}\}$, respectively. In the case of $S = 1$, the model (1) is equivalent to the PCSP model, as shown in Fig. 1(a), where maximum likelihood estimates of spatial pattern vectors $\boldsymbol{A}^s$ are learned by the expectation maximization [2].

In the case where a sufficient number of training samples is not available for some subjects, the performance of PCSP is degraded. Bayesian multi-task learning enforces spatial patten vectors across subjects to share hyperparameters of their prior distributions, allowing for learning from each other subjects. In BCSP (see Fig. 1(b)) [3], Gaussian prior was placed on the basis matrix $\boldsymbol{A}^s$, sharing the hyperparameters (mean vector and precision matrix) across subjects:

$$p(\boldsymbol{A}^s) = \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{a}_m^s | \boldsymbol{\mu}, \mathbf{\Omega}^{-1}),$$

for $s = 1, \ldots, S$ and the mean vector and the precision matrix are assumed to follow Gaussian-Wishart distribution

$$p(\boldsymbol{\mu}, \mathbf{\Omega}) = \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{m}_0, (\beta_0 \mathbf{\Omega})^{-1}) \mathcal{W}(\mathbf{\Omega} | \boldsymbol{W}_0, \nu_0),$$

where $\mathcal{W}(\mathbf{\Omega} | \boldsymbol{W}_0, \nu_0)$ denotes Wishart distribution parameterized by $\boldsymbol{W}_0$ and ν_0. Gamma distributions are assumed for precision parameters $\mathbf{\Lambda}^{sc}$ and $\mathbf{\Psi}^{sc}$,

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

Posterior distributions over $\boldsymbol{A}^s$ and $\boldsymbol{Y}^{sc}$ are approximately computed by Bayesian variation inference method to calculate CSP features [3].

3. BAYESIAN CSP WITH PYP PRIORS

In this section we present the main contribution of this paper, which is Bayesian CSP with PYP priors, referred to as BCSP-PYP. BCSP [3], as shown in Fig. 1(b), assumes that all spatial pattern vectors $\{\boldsymbol{a}_m^s\}$ share the hyperparameters, without proximity between spatial patterns. This restriction might bring negative effect such that information transfer is enforced even between subjects whose spatial patterns are very different. Motivated by the idea of task clustering in the multi-task learning framework [4, 5], we incorporate a infinite mixture model with PYP priors [6] into Bayesian CSP, as shown in Fig. 2, so that grouping spatial pattern vectors $\boldsymbol{a}_m^s$ and learning the model (1) are performed simultaneously. In this way, only spatial pattern vectors in the same cluster share the hyperparameters.

The Pitman-Yor process (PYP) [7, 8] is a two-parameter generalization of the Dirichlet process [9], which relaxes the "rich-get-richer" property of the Dirichlet process by setting an additional parameter. Invoking the linear model (1), BCSP-PYP assumes that spatial pattern vectors $\{\boldsymbol{a}_m^s\}$ are drawn from the distributions $p(\boldsymbol{a}_m^s | \theta_m^s)$ parameterized by $\{\theta_m^s\}$ that are independently drawn from a random measure G from a PYP with a scaling parameter α and a discount parameter γ and a base distribution G_0:

$$G \sim \text{PYP}(\alpha, \gamma, G_0), \quad \theta_m^s \sim G, \quad \boldsymbol{a}_m^s \sim p(\boldsymbol{a}_m^s | \theta_m^s), \qquad (2)$$

for $m = 1, \ldots, M$ and $s = 1, \ldots, S$. The process is defined with $\alpha > -\gamma$ and $0 \leq \gamma < 1$. Spatial pattern vectors $\{\boldsymbol{a}_m^s\}$ generated by this model are partitioned according to the distinct values of the parameters $\{\theta_m^s\}$. Parameter θ_m^s takes one of distinct values in $\{\theta_k^*\}$ ($k = 1, \ldots, MS$).

The stick-breaking representation [10] [7] of the random measure G is given by

$$\pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, \qquad (3)$$

where v_k and θ_k^* are independent random variables drawn from Beta distribution and the base measure G_0, respectively

$$v_k \sim \text{Beta}(v_k|1-\gamma, \alpha+k\gamma), \quad \theta_k^* \sim G_0.$$

The stick-breaking representation (3) makes it clear that G is an atomic random measure (with probability one), in which mixing proportions $\{\pi_k\}$ are given by successively breaking a unit-length stick into an infinite number of pieces. An independent draw v_k from a Beta$(1-\gamma, \alpha+k\gamma)$ distribution is re-scaled, proportional to the rest of stick, leading to the size of the broken piece π_k corresponding to the mixing proportion.

We introduce cluster indicator vectors $\boldsymbol{z}_m^s \in \mathbb{R}^{MS}$, the k-th entry of which, denoted by $z_m^s(k)$, equals 1 if $\theta_m^s = \theta_k^*$, otherwise zero. In the BCSP-PYP model, $\theta_k^* = (\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*)$, and spatial pattern vectors are assumed to be drawn from Gaussian distribution parameterized by the mean vector $\boldsymbol{\mu}_k^*$ and the precision matrix $\boldsymbol{\Omega}_k^*$. The base measure G_0 is chosen as Gaussian-Wishart distribution that is the conjugate prior for Gaussian likelihood $\mathcal{N}(\boldsymbol{a}_m^s|\boldsymbol{\mu}_k^*, (\boldsymbol{\Omega}_k^*)^{-1})$. BCSP-PYP also considers the same generative model (1) with the following parameterization:

$$\begin{aligned}
v_k &\sim \text{Beta}(v_k|1-\gamma, \alpha+k\gamma),\\
\theta_k^* = (\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*) &\sim \mathcal{N}(\boldsymbol{\mu}_k^*|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Omega}_k^*)^{-1})\mathcal{W}(\boldsymbol{\Omega}_k^*|\boldsymbol{W}_0, \nu_0),\\
p(z_m^s(k)=1) &= v_k \prod_{j=1}^{k-1}(1-v_j),\\
\boldsymbol{a}_m^s|(z_m^s(k)=1) &\sim \mathcal{N}(\boldsymbol{a}_m^s|\boldsymbol{\mu}_k^*, (\boldsymbol{\Omega}_k^*)^{-1}),\\
\boldsymbol{y}_t^{sc} &\sim \mathcal{N}(\boldsymbol{y}_t^{sc}|\boldsymbol{0}, (\boldsymbol{\Lambda}^{sc})^{-1}),\\
\boldsymbol{x}_t^{sc} &\sim \mathcal{N}(\boldsymbol{x}_t^{sc}|\boldsymbol{A}^s\boldsymbol{y}_t^{sc}, (\boldsymbol{\Psi}^{sc})^{-1}),\\
\lambda_m^{sc} &\sim \text{Gamma}(\lambda_m^{sc}|a_0^\lambda, b_0^\lambda),\\
\psi_d^{sc} &\sim \text{Gamma}(\psi_d^{sc}|a_0^\psi, b_0^\psi).
\end{aligned}$$

We employ the variational inference method [11] to approximately compute the posterior distributions over spatial pattern vectors as well as latent variables. As in variational inference for DP mixture models [11], we also consider the truncated stick breaking representation with the truncation level K. The variational inference considers a lower-bound on the marginal log-likelihood

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

where the Jensen's inequality was used and $\mathcal{F}(q)$ denotes the *variational lower-bound* to be maximized. $\boldsymbol{\Theta}$ denotes the set of variables to be inferred, where the variational distribution $q(\boldsymbol{\Theta})$ is factorized as

$$\begin{aligned}
q(\boldsymbol{\Theta}) &= q(\{\boldsymbol{A}^s\})\, q(\{\boldsymbol{z}_m^s\})\, q(\{\boldsymbol{Y}^{sc}\})\, q(\{v_k\})\\
&\quad q(\{\boldsymbol{\Lambda}^{sc}\})\, q(\{\boldsymbol{\Psi}^{sc}\})\, q(\{(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*)\}).
\end{aligned}$$

Optimal variational posterior distributions are computed by alternatively maximizing the variational lower-bound $\mathcal{F}(q)$, which is summarized in Table 1. The hyperparameters, $\{\alpha, \gamma, \beta_0, \nu_0, \boldsymbol{W}_0, \boldsymbol{m}_0, a_0^\psi, b_0^\psi, a_0^\lambda, b_0^\lambda\}$, are also estimated by maximizing the variational lower-bound $\mathcal{F}(q)$, which is summarized in Table 2.

Given a test data $\boldsymbol{X}^s \in \mathbb{R}^{D\times T}$, we compute the CSP feature vector $\boldsymbol{f} \in \mathbb{R}^{2n}$ as follows. We first compute the posterior mean matrices $\left\{\overline{\boldsymbol{Y}}^{sc}\right\}$ for $c=1,2$,

$$\overline{\boldsymbol{Y}}^{sc} = \boldsymbol{\Sigma}^{sc}\left\langle \boldsymbol{A}^{s\top}\right\rangle \langle \boldsymbol{\Psi}^{sc}\rangle\, \boldsymbol{X}^s,$$

which corresponds to η_t^{sc} in Table 1 and $\langle\cdot\rangle$ denotes the statistical expectation. Considering the class conditional probability as $p(\boldsymbol{X}^s \in c) = T_{sc}/(T_{s1}+T_{s2})$ for $c=1,2$, we compute $\overline{\boldsymbol{Y}}^s = \sum_{c=1}^2 p(\boldsymbol{X}^s \in c)\cdot \overline{\boldsymbol{Y}}^{sc}$. Treating columns in $\overline{\boldsymbol{Y}}^s$ as projected variables in CSP, we compute a M-dimensional vector $\widehat{\boldsymbol{f}}^s \in \mathbb{R}^M$, the m-th entry of which is calculated as

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

where $\boldsymbol{1}_T \in \mathbb{R}^T$ is the vector of all ones. We select $2n$ entries from $\left\{\widehat{f}^s(m)\right\}$ for m associated with top n and bottom n expected precision ratio $\left\{\langle\lambda_m^{s1}\rangle/\langle\lambda_m^{s2}\rangle\right\}$, to construct the CSP feature vector $\boldsymbol{f}^s \in \mathbb{R}^{2n}$.

4. NUMERICAL EXPERIMENTS

We compared the classification performances of the PCSP, BCSP and BCSP-PYP 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 of left/right hand movement to consider binary classification problem. Each imagery movement consists of 144 trials. Every trial was divided into $T=500$ times points, which corresponds to the time interval from 3.5s to 5.5s after the visual cue (250 Hz). The data was recorded with 22 electrodes ($D=22$). Every trial was bandpass-filtered to emphasize important frequency bands for the motor imagery task.

For all models, basis matrices are set to square matrix ($M=D$) and feature vectors $\boldsymbol{f}^s \in \mathbb{R}^{2n}$ (with $n=$ chosen) are constructed by PCSP, BCSP and BCSP-PYP. Linear discriminant analysis (LDA) is applied to transform these feature vectors down to scalar values which are fed into a minimum distance classifier. The classification accuracy was obtained by the ratio of the number of correctly classified test trials compared to the total number of test trials. We selected half of the trials in each subject as the test trials, and randomly selected some of the remaining trials as the training trials. At each experiment, a subject s in the dataset is chosen as the

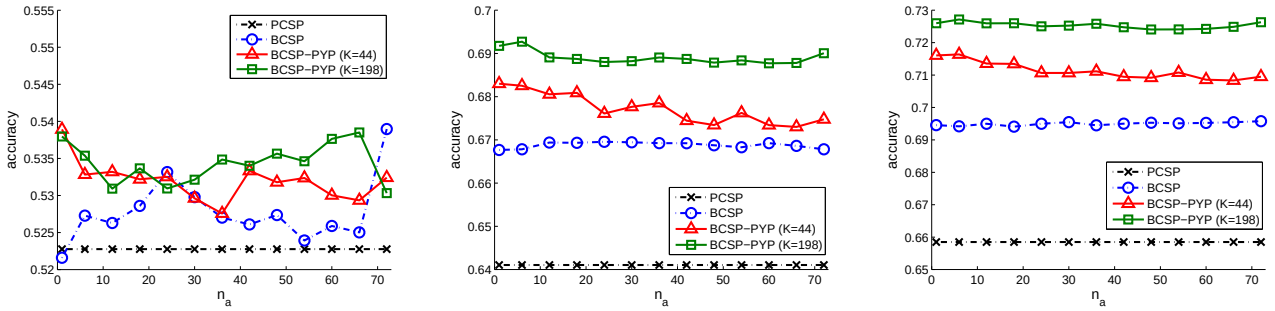
[1] http://www.bbci.de/competition/iv/

Table 1. Variational posteriors and corresponding updating equations in BCSP-PYP are summarized. Denote by $\langle\cdot\rangle$ the statistical expectation with respect to corresponding variational posterior distributions. The (i,j)-element of a matrix is denoted by $[\cdot]_{i,j}$, and $[\cdot]_{i,:}$ represents the i-th row of a matrix. The trace operator is denoted by $\mathrm{tr}(\cdot)$, and $\mathrm{diag}(\boldsymbol{x})$ represents the diagonal matrix whose diagonal entries are given by the vector $\boldsymbol{x}$. $\mathrm{Multinomial}(\boldsymbol{x}|\boldsymbol{p})$ represents the multinomial distribution such that $p(x_k=1)=p_k$.

Variational posterior distributions	Updating equations for variational parameters
$q(\boldsymbol{A}^s)=\prod_{d=1}^{D}\mathcal{N}\left([\boldsymbol{A}^s]_{d,:}\middle\vert\bar{\boldsymbol{\nu}}_d^s,\boldsymbol{\Phi}_d^s\right)$	$(\boldsymbol{\Phi}_d^s)^{-1}=\sum_{k=1}^{K}\langle[\boldsymbol{\Omega}_k^*]_{d,d}\rangle\,\mathrm{diag}\left(\langle\bar{\boldsymbol{z}}_{:}^{sk}\rangle\right)+\sum_{c=1}^{2}\langle\psi_d^{sc}\rangle\left\langle\boldsymbol{Y}^{sc}\boldsymbol{Y}^{sc\top}\right\rangle,$ $\bar{\boldsymbol{\nu}}_d^s=\left\{\sum_{c=1}^{2}\langle\psi_d^{sc}\rangle\left[\boldsymbol{X}^{sc}\right]_{d,:}\left\langle\boldsymbol{Y}^{sc\top}\right\rangle+\sum_{k=1}^{K}\left(\langle[\boldsymbol{\Omega}_k^*]_{d,:}\boldsymbol{\mu}_k\rangle\left\langle\bar{\boldsymbol{z}}_{:}^{sk}\right\rangle\right.\right.$ $\left.\left.-\,\mathrm{diag}\left(\langle\bar{\boldsymbol{z}}_{:}^{sk}\rangle\right)\sum_{j\neq d}\langle[\boldsymbol{\Omega}_k^*]_{d,j}\rangle\langle[\boldsymbol{A}^s]_{j,:}\rangle^{\top}\right)\right\}\boldsymbol{\Phi}_d^s,$ $\bar{\boldsymbol{z}}_{:}^{sk}=[z_1^s(k)\ \ldots\ z_M^s(k)]^{\top}.$
$q(\boldsymbol{z}_m^s)=\mathrm{Multinomial}(\boldsymbol{z}_m^s\vert\boldsymbol{r}_m^s)$	$\boldsymbol{r}_m^s(k)\propto\exp\left\{\frac{1}{2}\langle\log\vert\boldsymbol{\Omega}_k^*\vert\rangle-\frac{1}{2}\sum_{d=1}^{D}\langle[\boldsymbol{\Omega}_k^*]_{d,d}\rangle\left\langle([\boldsymbol{A}^s]_{d,m})^2\right\rangle\right.$ $-\frac{1}{2}\sum_{i\neq j}\langle[\boldsymbol{A}^s]_{i,m}\rangle\langle[\boldsymbol{\Omega}_k^*]_{i,j}\rangle\langle[\boldsymbol{A}^s]_{j,m}\rangle+\left\langle\boldsymbol{a}_m^{s\top}\right\rangle\langle\boldsymbol{\Omega}_k^*\boldsymbol{\mu}_k^*\rangle$ $\left.-\frac{1}{2}\left\langle\boldsymbol{\mu}_k^{\top}\boldsymbol{\Omega}_k^*\boldsymbol{\mu}_k^*\right\rangle+\langle\log v_k\rangle+\sum_{j=1}^{k-1}\langle\log(1-v_j)\rangle\right\}.$
$q(\boldsymbol{Y}^{sc})=\prod_{t=1}^{T_{sc}}\mathcal{N}\left(\boldsymbol{y}_t^{sc}\middle\vert\eta_t^{sc},\boldsymbol{\Sigma}^{sc}\right)$	$(\boldsymbol{\Sigma}^{sc})^{-1}=\langle\boldsymbol{\Lambda}^{sc}\rangle+\sum_{d=1}^{D}\langle\psi_d^{sc}\rangle\left\langle[\boldsymbol{A}^s]_{d,:}^{\top}[\boldsymbol{A}^s]_{d,:}\right\rangle,$ $\eta_t^{sc}=\boldsymbol{\Sigma}^{sc}\left\langle\boldsymbol{A}^{s\top}\right\rangle\langle\boldsymbol{\Psi}^{sc}\rangle\,\boldsymbol{x}_t^{sc}.$
$q(v_k)=\mathrm{Beta}(v_k\vert a_k^v,b_k^v)$	$a_k^v=1-\gamma+\langle L_k\rangle,\quad b_k^v=\alpha+k\gamma+\sum_{s=1}^{S}\sum_{m=1}^{M}\sum_{j=k+1}^{K}\langle z_m^s(j)\rangle.$
$q\left(\boldsymbol{\Lambda}^{sc}\right)=\prod_{m=1}^{M}\mathrm{Gamma}(\lambda_m^{sc}\vert a_m^{\lambda sc},b_m^{\lambda sc})$	$a_m^{\lambda sc}=a_0^{\lambda}+\frac{T_{sc}}{2},\quad b_m^{\lambda sc}=b_0^{\lambda}+\frac{1}{2}\left[\left\langle\boldsymbol{Y}^{sc}\boldsymbol{Y}^{sc\top}\right\rangle\right]_{m,m}.$
$q(\boldsymbol{\Psi}^{sc})=\prod_{d=1}^{D}\mathrm{Gamma}(\psi_d^{sc}\vert a_d^{\psi sc},b_d^{\psi sc})$	$a_d^{\psi sc}=a_0^{\psi}+\frac{T_{sc}}{2},$ $b_d^{\psi sc}=b_0^{\psi}+\frac{1}{2}\left[\boldsymbol{X}^{sc}\boldsymbol{X}^{sc\top}-2\boldsymbol{X}^{sc}\langle\boldsymbol{Y}^{sc\top}\rangle\langle\boldsymbol{A}^{s\top}\rangle+\left\langle\boldsymbol{A}^s\boldsymbol{Y}^{sc}\boldsymbol{Y}^{sc\top}\boldsymbol{A}^{s\top}\right\rangle\right]_{d,d}$
$q(\boldsymbol{\mu}_k^*,\boldsymbol{\Omega}_k^*)=\mathcal{N}\left(\boldsymbol{\mu}_k^*\middle\vert\boldsymbol{m}_k,(\beta_k\boldsymbol{\Omega}_k^*)^{-1}\right)$ $\cdot\,\mathcal{W}\left(\boldsymbol{\Omega}_k^*\middle\vert\boldsymbol{W}_k,\nu_k\right)$	$\beta_k=\beta_0+\langle L_k\rangle,$ $\nu_k=\nu_0+\langle L_k\rangle,$ $\boldsymbol{m}_k=\frac{1}{\beta_k}(\beta_0\boldsymbol{m}_0+\langle L_k\rangle\widehat{\boldsymbol{a}}_k),$ $(\boldsymbol{W}_k)^{-1}=(\boldsymbol{W}_0)^{-1}+\langle L_k\rangle\widehat{\boldsymbol{Y}}_k+\frac{\beta_0\langle L_k\rangle}{\beta_k}(\boldsymbol{m}_0-\widehat{\boldsymbol{a}}_k)(\boldsymbol{m}_0-\widehat{\boldsymbol{a}}_k)^{\top},$ $\widehat{\boldsymbol{a}}_k=\frac{1}{\langle L_k\rangle}\sum_{s=1}^{S}\sum_{m=1}^{M}\langle z_m^s(k)\rangle\langle\boldsymbol{a}_m^s\rangle,$ $\widehat{\boldsymbol{Y}}_k=\frac{1}{\langle L_k\rangle}\sum_{s=1}^{S}\sum_{m=1}^{M}\langle z_m^s(k)\rangle\langle\boldsymbol{a}_m^s\boldsymbol{a}_m^{s\top}\rangle-\widehat{\boldsymbol{a}}_k\widehat{\boldsymbol{a}}_k^{\top}.$

Table 2. Updating equations for hyperparameters $\{\alpha,\gamma,\beta_0,\nu_0,\boldsymbol{W}_0,\boldsymbol{m}_0,a_0^{\psi},b_0^{\psi},a_0^{\lambda},b_0^{\lambda}\}$ are summarized: (a) stationary point equations for $\{\alpha,\gamma,a_0^{\psi},a_0^{\lambda},\nu_0\}$, which do not have closed-form solutions, are numerically solved to update corresponding hyperparameters; (b) updating equations for $\{\beta_0,\boldsymbol{W}_0,\boldsymbol{m}_0,b_0^{\psi},b_0^{\lambda}\}$.

(a)	$f(\alpha)=\sum_{k=1}^{K-1}\left\{\psi(\alpha+(k-1)\gamma)-\psi(\alpha+k\gamma)+\langle\log(1-v_k)\rangle\right\}=0,$ $f(\gamma)=\sum_{k=1}^{K-1}\left\{(k-1)\psi(\alpha+1+(k-1)\gamma)+\psi(1-\gamma)-k\psi(\alpha+k\gamma)\right\}=0,$ $f(a_0^{\psi})=\log(a_0^{\psi})-\psi(a_0^{\psi})+\frac{1}{2SD}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{d=1}^{D}\langle\log\psi_d^{sc}\rangle-\log\left(\frac{1}{2SD}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{d=1}^{D}\langle\psi_d^{sc}\rangle\right)=0,$ $f(a_0^{\lambda})=\log(a_0^{\lambda})-\psi(a_0^{\lambda})+\frac{1}{2SM}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{m=1}^{M}\langle\log\lambda_m^{sc}\rangle-\log\left(\frac{1}{2SD}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{m=1}^{M}\langle\lambda_m^{sc}\rangle\right)=0,$ $f(\nu_0)=D\log\nu_0-\sum_{i=1}^{D}\psi\left(\frac{\nu_0+1-i}{2}\right)-\log\left\vert\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\rangle\right\vert+\frac{1}{K}\sum_{k=1}^{K}\langle\log\vert\boldsymbol{\Omega}_k^*\vert\rangle+D(\log K-\log 2)=0.$
(b)	$\beta_0=\frac{KD}{\sum_{k=1}^{K}\langle(\boldsymbol{m}_0-\boldsymbol{\mu}_k^*)^{\top}\boldsymbol{\Omega}_K^*(\boldsymbol{m}_0-\boldsymbol{\mu}_k^*)\rangle},\quad\boldsymbol{m}_0=\left(\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\rangle\right)^{-1}\left(\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\boldsymbol{\mu}_k^*\rangle\right),$ $\boldsymbol{W}_0=\frac{1}{\nu_0 K}\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\rangle,\quad b_0^{\psi}=\frac{a_0^{\psi}\cdot 2SD}{\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{d=1}^{D}\langle\psi_d^{sc}\rangle},\quad b_0^{\lambda}=\frac{a_0^{\lambda}\cdot 2SM}{\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{m=1}^{M}\langle\lambda_m^{sc}\rangle},$



(a) $n_t = 1$ (b) $n_t = 24$ (c) $n_t = 60$

Fig. 3. 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 $n_t = 1, 24, 60$, where n_t denotes the number of training samples for target subject for each class.

target subject. We randomly selected n_t trials from each class of the target subject as the training trials ($T_{sc} = T \cdot n_t$). We randomly selected n_a trials from each class of the non-target subjects ($T_{ic} = T \cdot n_a, i \neq s$) as the additional training trials. The classification accuracy was evaluated using the test trials of the target subjects only. We repeated the experiments 10 times for $s = 1, ..., S$, and averaged the accuracies to represent the classification performance of models on the given (n_t, n_a) setting. Note that PCSP cannot exploit the additional training trials so that the classification performance of which does not vary by n_a. As shown in Fig. 3, the performance of BCSP-PYP is better than PCSP and BCSP, in terms of classification accuracy when features computed by these methods are used. These results demonstrate that our proposed method BCSP-PYP is quite effective in exploiting non-target subjects' data, as compared to BCSP.

5. CONCLUSIONS

We have presented a Bayesian CSP model with PYP priors to tackle multi-subject EEG classification, where the infinite mixture model partitions the spatial pattern vectors into several groups and at the same time spatial pattern vectors are learned in Bayesian framework. Spatial patter vectors in the same group are coupled through sharing the hyperparameters of their prior distributions, such that information transfer between subjects involving similar spatial patterns is encouraged while information transfer between dissimilar subjects is discouraged. Numerical experiments on BCI competition IV 2a dataset confirmed the useful behavior of our BCSP-PYP, compared to existing probabilistic models such as PCSP and BCSP.

Acknowledgments: This work was supported by National Research Foundation (NRF) of Korea (2011-0018283, 2011-0018284), MKE and NIPA "IT Consilience Creative Program" (C1515-1121-0003), and NRF World Class University Program (R31-10100).

6. REFERENCES

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

[2] 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.

[3] 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.

[4] B. Bakker and T. Heskes, "Task clustering and gating for Bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, pp. 83–99, 2003.

[5] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, "Multi-task learning for classification with Dirichlet process priors," *Journal of Machine Learning Research*, vol. 8, pp. 35–63, 2007.

[6] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *Proceedings of Coling/ACL*, 2006.

[7] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, vol. 25, no. 2, pp. 855–900, 1997.

[8] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

[9] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[10] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[11] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.