WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia
IJCNN

Bayesian Common Spatial Patterns with Dirichlet Process Priors for Multi-Subject EEG Classification

Hyohyeong Kang
Department of Computer Science,
Pohang University of Science and Technology,
Pohang 790-784, Korea
Email: paanguin@postech.ac.kr

Seungjin Choi
Department of Computer Science and
Division of IT Convergence Engineering,
Pohang University of Science and Technology,
Pohang 790-784, Korea
Email: seungjin@postech.ac.kr

***Abstract*—Multi-subject electroencephalography (EEG) classification involves the categorization of brain waves measured from multiple subjects, each of whom undergoes the same mental task. Common spatial patterns (CSP) or probabilistic CSP (PCSP) are widely used for extracting discriminative features from EEG, although they are trained on a subject-by-subject basis and inter-subject information is neglected. Moreover, the performance is degraded when only a few training samples are available for each subject. In this paper, we present a method for *Bayesian CSP* with *Dirichlet process (DP) priors,* where spatial patterns (corresponding to basis vectors) are simultaneously learned and clustered across subjects using variational Bayesian inference, which facilitates 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, which means information transfer is facilitated among subjects with similar spatial patterns. Numerical experiments using the BCI competition IV 2a dataset demonstrated the high performance of our method, compared with existing PCSP and Bayesian CSP methods with a single prior distribution.**

I. Introduction

Electroencephalography (EEG) is the recording of electrical potentials using multiple sensors placed on the scalp, to collect multivariate time series data that reflects 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) [4], [7], [8]. Multi-subject EEG classification considers brain signals from multiple subjects, each of whom undergoes the same mental task, where the 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 has proved to be a useful subject-specific spatial filter [7]. The learning of common spatial patterns was placed into a probabilistic framework, leading to probabilistic CSP (PCSP) [11], where two linear Gaussian generative models with a shared basis matrix are jointly learned to infer *spatial pattern vectors* that correspond to the column vectors of the shared basis matrix. CSP is a subject-specific spatial filter, which does not consider information from other subjects involved in the same task as the subject of interest.

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. Bayesian CSP [6] works better than PCSP (on subject-by-subject basis), although similarities among spatial patterns are neglected because all spatial patterns are forced to share the same hyperparameters.

In this paper, we present a more flexible Bayesian model where we exploit multi-subject EEG data to learn spatial patterns for a target subject, which facilitates information transfer among subjects with similar spatial patterns. To this end, we present a Bayesian CSP with Dirichlet process (DP) priors, referred to as BCSP-DP, in which we develop a variational inference algorithm to learn and group spatial pattern vectors. This means 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 hyperparameters facilitates information transfer among subjects with similar spatial patterns, whereas information transfer is prevented among dissimilar subjects. Our method is motivated by task-clustering methods in a multi-task learning framework [2], [12], where similar tasks are identified and information is transferred between tasks in the same group. Numerical experiments using the BCI competition IV 2a dataset demonstrated the high performance of BCSP-DP compared with PCSP and BCSP.

II. Related Work

In this section, we briefly review two probabilistic models for CSP, i.e., probabilistic CSP (PCSP) [11] and Bayesian CSP (BCSP) [6]. 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 using D electrodes over trials (T_{sc} is the number of samples recorded for a pre-defined number of trials) for a subject $s \in \{1, \ldots, S\}$ who undergoes a mental task involving class $c \in \{1, 2\}$. PCSP or BCSP assume that $\boldsymbol{X}^{sc}$ is generated by

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

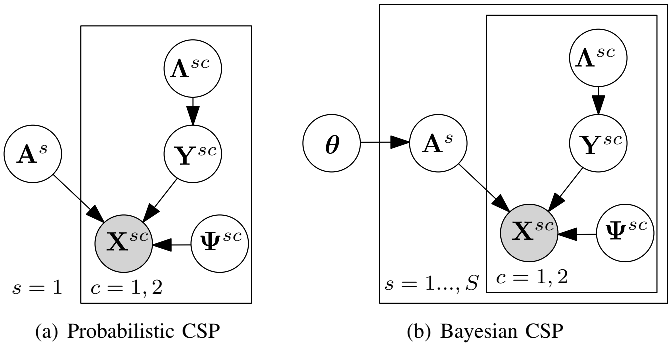
where $\boldsymbol{A}^s = [\boldsymbol{a}_1^s, \ldots, \boldsymbol{a}_M^s] \in \mathbb{R}^{D\times M}$ is the *basis matrix* for subject 's', containing M *spatial pattern vectors* shared among classes, $\boldsymbol{Y}^{sc} = [\boldsymbol{y}_1^{sc}, \ldots, \boldsymbol{y}_{T_{sc}}^{sc}] \in \mathbb{R}^{M\times T_{sc}}$ is the *coefficient matrix (latent variables)*, and $\boldsymbol{E}^{sc} = [\boldsymbol{\epsilon}_1^{sc}, \ldots, \boldsymbol{\epsilon}_{T_{sc}}^{sc}] \in \mathbb{R}^{D\times T_{sc}}$ is the *noise matrix*. It is assumed that each row of $\boldsymbol{X}^{sc}$ is already centered (zero mean). Coefficients and noise are assumed to be drawn from zero-mean Gaussian distributions:

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

where $\boldsymbol{\Lambda}^{sc}$ and $\boldsymbol{\Psi}^{sc}$ are diagonal precision matrices for $s = 1, \ldots, S$ and $c = 1, 2$,

$$\begin{aligned} \boldsymbol{\Lambda}^{sc} &= \mathrm{diag}(\lambda_1^{sc}, \ldots, \lambda_M^{sc}) \in \mathbb{R}^{M\times M}, \\ \boldsymbol{\Psi}^{sc} &= \mathrm{diag}(\psi_1^{sc}, \ldots, \psi_D^{sc}) \in \mathbb{R}^{D\times D}. \end{aligned}$$

In the case of $S = 1$ (subject-specific model), the model (1) reduces to the probabilistic CSP model, as shown in Fig. 1(a), where maximum likelihood estimates of spatial pattern vectors $\boldsymbol{A}^s$ are learned by expectation maximization [11].



(a) Probabilistic CSP (b) Bayesian CSP

Fig. 1. Graphical representations of the PCSP [11] and BCSP models [6].

The performance of PCSP is degraded if a sufficient number of training samples is not available for some subjects. Bayesian multi-task learning enforces spatial patten vectors across subjects to share hyperparameters of their prior distributions, facilitating learning from other subjects. In BCSP (see Fig. 1(b)) [6], a Gaussian prior is placed in the basis matrix $\boldsymbol{A}^s$, thereby sharing the hyperparameters (mean vector and precision matrix) among subjects:

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

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

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

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

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

Posterior distributions over $\boldsymbol{A}^s$ and $\boldsymbol{Y}^{sc}$ are approximately computed using a Bayesian variation inference method to calculate CSP features [6].

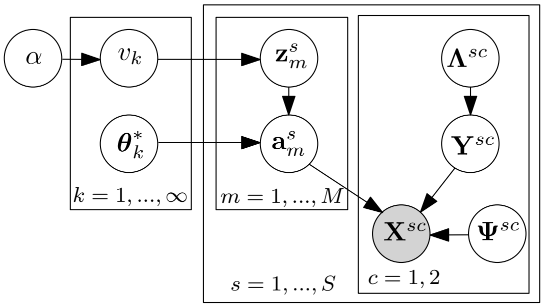


Fig. 2. Graphical representation of Bayesian CSP with DP priors.

III. Bayesian CSP with DP Priors

In this section, we present the main contribution of this paper, i.e., Bayesian CSP with DP priors, which is referred to as BCSP-DP. As shown in Fig. 1(b), BCSP [6] assumes that all spatial pattern vectors $\boldsymbol{a}_m^s$ for $m = 1, \ldots, M$ and $s = 1, \ldots, S$ share the hyperparameters, without proximity between spatial patterns. This restriction might have a negative effect because information transfer would also be facilitated among subjects whose spatial patterns are very different. It is desirable to facilitate (or prevent) information transfer among similar (or dissimilar) patterns. Motivated by the concept of task clustering in a multi-task learning framework [2], [12], we incorporate a DP mixture model [1], [9] into Bayesian CSP, as shown in Fig. 2, such that the grouping of spatial pattern vectors $\boldsymbol{a}_m^s$ and model learning (1) are performed simultaneously. Thus, only spatial pattern vectors in the same cluster share the hyperparameters.

A. BCSP-DP Model

We present a detailed description of our BCSP-DP model shown in Fig. 2. Invoking a linear model (1), BCSP-DP assumes that spatial pattern vectors $\{\boldsymbol{a}_m^s\}$ are drawn from distributions $p(\boldsymbol{a}_m^s | \theta_m^s)$ parameterized by $\{\theta_m^s\}$, which are drawn independently from a random measure G in a DP with a positive scaling parameter α and a base distribution G_0:

$$G \sim \mathrm{DP}(\alpha, G_0), \quad \theta_m^s \sim G, \quad \boldsymbol{a}_m^s \sim p(\boldsymbol{a}_m^s | \theta_m^s), \tag{2}$$

for $m = 1, \ldots, M$ and $s = 1, \ldots, S$. 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 a distinct value in $\{\theta_k^*\}$ $(k = 1, \ldots, MS)$.

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

$$\begin{aligned} \pi_k &= v_k \prod_{j=1}^{k-1}(1 - v_j), && (3) \\ G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, && (4) \end{aligned}$$

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

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

The stick-breaking representation (4) makes it clear that G is an atomic random measure (with probability one), where 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 $\text{Beta}(1, \alpha)$ distribution is rescaled in proportion to the remaining stick, so the size of the broken piece $\pi_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$ corresponds to the mixing proportion.

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

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

where $\boldsymbol{\Lambda}^{sc}$ and $\boldsymbol{\Psi}^{sc}$ are diagonal precision matrices and each diagonal entry is assumed to be drawn from Gamma distributions

$$\begin{aligned} \lambda_m^{sc} &\sim \text{Gamma}(\lambda_m^{sc}|a_0^{\lambda}, b_0^{\lambda}), \\ \psi_d^{sc} &\sim \text{Gamma}(\psi_d^{sc}|a_0^{\psi}, b_0^{\psi}). \end{aligned}$$

B. Variational Inference

We employ the variational inference method [3] to approximately compute the posterior distributions for spatial pattern vectors and latent variables. As in variational inference for DP mixture models [3], we also consider a truncated stick-breaking representation with a truncation level K.

We define a set of variables to be inferred as

$$\boldsymbol{\Theta} = \Big\{\{\boldsymbol{A}^s\}, \{\boldsymbol{z}_m^s\}, \{\boldsymbol{Y}^{sc}\}, \{v_k\}, \alpha, \{\boldsymbol{\Lambda}^{sc}\}, \{\boldsymbol{\Psi}^{sc}\}, \{(\boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*)\}\Big\}.$$

The variational inference method 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 Jensen's inequality was used and $\mathcal{F}(q)$ denotes the *variational lower-bound* to be maximized. We assume that the variational distribution $q(\boldsymbol{\Theta})$ is factorized as

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

The optimal variational posterior distributions are computed by alternatively maximizing the variational lower-bound $\mathcal{F}(q)$, which is summarized in Table I.

The hyperparameters, $\{\beta_0, \nu_0, \boldsymbol{W}_0, \boldsymbol{m}_0, a_0^{\psi}, b_0^{\psi}, a_0^{\lambda}, b_0^{\lambda}\}$, are also estimated by maximizing the variational lower-bound $\mathcal{F}(q)$. The stationary point equations for β_0 and $\boldsymbol{m}_0$ have closed-form solutions and the updating equations are given by

$$\begin{aligned} \beta_0 &= \frac{KD}{\sum_{k=1}^{K} \langle(\boldsymbol{m}_0 - \boldsymbol{\mu}_k^*)^\top \boldsymbol{\Omega}_K^*(\boldsymbol{m}_0 - \boldsymbol{\mu}_k^*)\rangle}, \\ \boldsymbol{m}_0 &= \left(\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\rangle\right)^{-1} \left(\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\boldsymbol{\mu}_k^*\rangle\right). \end{aligned}$$

However, the stationary point equations for ν_0, a_0^{ψ} and a_0^{λ} do not have closed-form solutions, so that the equations are solved numerically. The solution to $f(x) = 0$, x^*, is found by searching for a point where the sign of $f(x)$ is changed within a given interval $[x_a, x_b]$. The stationary point equations for ν_0, a_0^{ψ} and a_0^{λ} are given by

$$D \log \nu_0 - \sum_{i=1}^{D} \psi\left(\frac{\nu_0 + 1 - i}{2}\right) - \log\left|\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\rangle\right|$$

$$+\frac{1}{K}\sum_{k=1}^{K}\langle\log|\boldsymbol{\Omega}_k^*|\rangle + D(\log K - \log 2) = 0,$$

$$\log(a_0^{\psi}) - \psi(a_0^{\psi}) + \frac{1}{2SD}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{d=1}^{D}\langle\log\psi_d^{sc}\rangle$$

$$-\log\left(\frac{1}{2SD}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{d=1}^{D}\langle\psi_d^{sc}\rangle\right) = 0,$$

$$\log(a_0^{\lambda}) - \psi(a_0^{\lambda}) + \frac{1}{2SM}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{m=1}^{M}\langle\log\lambda_m^{sc}\rangle$$

$$-\log\left(\frac{1}{2SD}\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{m=1}^{M}\langle\lambda_m^{sc}\rangle\right) = 0,$$

TABLE I
VARIATIONAL POSTERIORS AND CORRESPONDING UPDATING EQUATIONS IN BCSP-DP 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 $\boldsymbol{x}$. $\mathrm{Multinomial}(\boldsymbol{x}|p)$ REPRESENTS THE MULTINOMIAL DISTRIBUTION SUCH THAT $p(x_k=1)=p_k$.

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

which are numerically solved for ν_0, a_0^ψ and a_0^λ, respectively. $\psi(\cdot)$ denotes a digamma function such that $\psi(x)=\frac{\partial}{\partial x}\log\Gamma(x)$. The updating equations for $\boldsymbol{W}_0$, b_0^ψ, and b_0^λ are given by the following closed-form solutions:

$$
\begin{aligned}
\boldsymbol{W}_0 &= \frac{1}{\nu_0 K}\sum_{k=1}^{K}\langle\boldsymbol{\Omega}_k^*\rangle,\\
b_0^\psi &= \frac{a_0^\psi\cdot 2SD}{\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{d=1}^{D}\langle\psi_d^{sc}\rangle},\\
b_0^\lambda &= \frac{a_0^\lambda\cdot 2SM}{\sum_{s=1}^{S}\sum_{c=1}^{2}\sum_{m=1}^{M}\langle\lambda_m^{sc}\rangle},
\end{aligned}
$$

which are dependent on ν_0, a_0^ψ, and a_0^λ, respectively.

Given 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 I. Given the class conditional probability $p(\boldsymbol{X}^s\in c)=\frac{T_{sc}}{T_{s1}+T_{s2}}$ for $c=1,2$, we compute

$$
\overline{\boldsymbol{Y}}^{s}=\sum_{c=1}^{2}\frac{T_{sc}}{T_{s1}+T_{s2}}\overline{\boldsymbol{Y}}^{sc}.
$$

Treating columns in $\overline{\boldsymbol{Y}}^{s}$ as projected variables in CSP, we compute an M-dimensional vector $\widehat{\boldsymbol{f}}^{s}\in\mathbb{R}^M$, where the m-th entry is calculated as

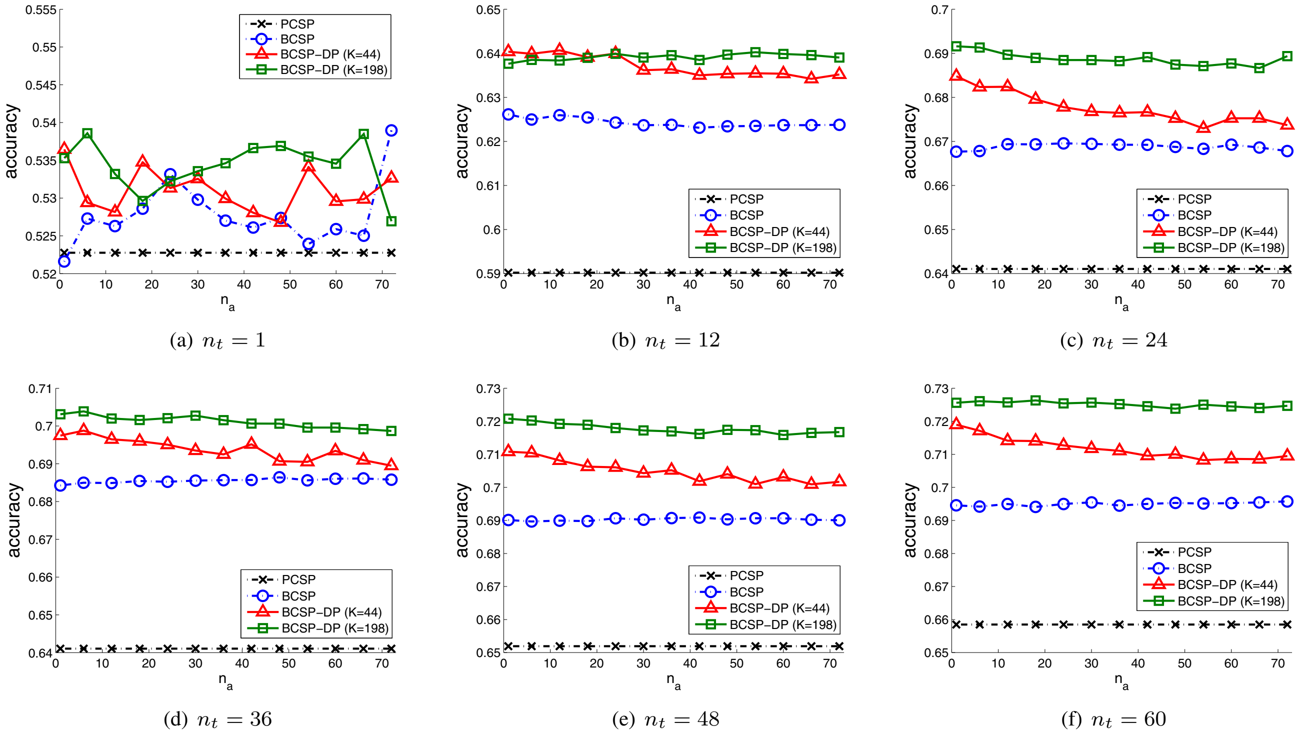
$$
\hat{f}^s(m)=\log\left(\frac{1}{T}\left[\overline{\boldsymbol{Y}}^{s}\overline{\boldsymbol{Y}}^{s\top}\right]_{m,m}-\left(\frac{1}{T}\left[\overline{\boldsymbol{Y}}^{s}\boldsymbol{1}_T\right]_m\right)^2\right),
$$

where $\boldsymbol{1}_T\in\mathbb{R}^T$ is a vector containing all ones. We select $2n$ entries from $\left\{\widehat{f}^s(m)\right\}$ for m associated with the 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 $\boldsymbol{f}^s\in\mathbb{R}^{2n}$.

IV. Numerical Experiments

We compared the classification performance of the PCSP, BCSP, and BCSP-DP methods using the BCI Competition IV[1]-2a data set. This data set contains nine subjects with four

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



(a) $n_t = 1$

(b) $n_t = 12$

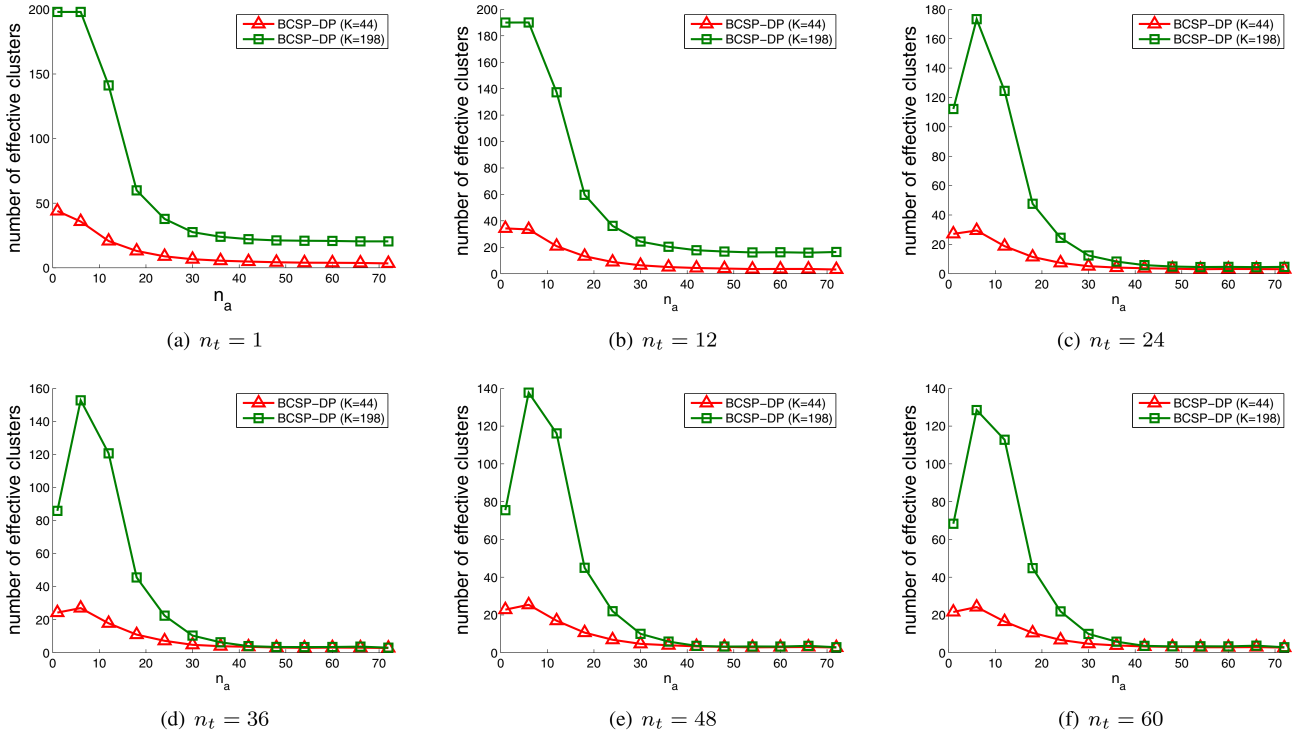
(c) $n_t = 24$

(d) $n_t = 36$

(e) $n_t = 48$

(f) $n_t = 60$

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. Six different plots are shown for $n_t = 1, 12, 24, 36, 48, 60$, where n_t denotes the number of training samples for target subject for each class.



(a) $n_t = 1$

(b) $n_t = 12$

(c) $n_t = 24$

(d) $n_t = 36$

(e) $n_t = 48$

(f) $n_t = 60$

Fig. 4. The number of effective clusters for target subjects is shown when the number of training samples for the non-target subjects, denoted by n_a, varies. Six different plots are shown for $n_t = 1, 12, 24, 36, 48, 60$, where n_t denotes the number of training samples for target subject for each class.

imaginary movements, i.e., left/right hand, right foot, tongue. We used the trials for left/right hand movements to consider the binary classification problem. Each imaginary movement consisted of 144 trials. Every trial was divided into $T = 500$ time points (250 Hz), which corresponded to a time interval 3.5 s to 5.5 s after the visual cue. The data was recorded using 22 electrodes ($D = 22$). Each trial was bandpass-filtered to emphasize important frequency bands in the motor imaginary task.

The BCSP-DP model was evaluated for two different truncation levels such that $K = 44$ and $K = 198$. Note that the maximum number of clusters was $M \cdot S = 198$, which occurred when every spatial pattern was assigned to its own cluster. We assumed that the basis matrices were square in all models ($M = D$) and we constructed feature vectors $\boldsymbol{f}^s \in \mathbb{R}^{2n}$ using PCSP, BCSP and BCSP-DP, where $n = 3$. We trained the Linear Discriminant Analysis using the feature vectors. The accuracy was determined as the ratio of the number of correctly classified test trials compared with the total number of test trials.

We selected half of the trials for each subject as the test trials and we randomly selected some of the remaining trials as training trials. During each run of the experiment, we selected a subject s from the S subjects in the dataset as the 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 additional training trials. The classification accuracy was evaluated using the test trials for the target subjects only. We repeated the experiments 10 times for $s = 1, ..., S$, and averaged the accuracies to represent the classification performance of the models in the given (n_t, n_a) setting. Note that PCSP cannot exploit the additional training trials so that its classification performance does not vary with n_a.

The classification performance of BCSP and BCSP-DP were higher than PCSP (Fig. 3). Moreover, the proposed BCSP-DP models had higher classification performance than BCSP, which shows that our proposed models were more effective at exploiting the non-target subject data compared with BCSP. The number of effective clusters, which was computed by counting the clusters with $\langle L_k \rangle > 0.5$, varied with the truncation level K (Fig. 4). The number of effective clusters of BCSP-DP with $K = 198$ was much higher than that for BCSP-DP with $K = 44$. A higher truncation level allowed more clusters which improved the classification performance of the BCSP-DP model, while the computational cost was also increased by updating more cluster parameters, because the cost was proportional to K.

V. Conclusions

We developed a Bayesian CSP model that uses DP priors to tackle multi-subject EEG classification, where the DP mixture model partitions the spatial pattern vectors into several groups, while the spatial pattern vectors are learned by the Bayesian framework at the same time. Spatial patter vectors in the same group are coupled by sharing the hyperparameters of their prior distributions, which facilitates information transfer among subjects with similar spatial patterns whereas information transfer among dissimilar subjects is prevented. Numerical experiments using the BCI competition IV 2a dataset confirmed the superior performance of our BCSP-DP method when compared with the existing probabilistic models, PCSP and BCSP.

Acknowledgments

This work was supported by National Research Foundation (NRF) of Korea (2011-0018283, 2011-0018284), the Converging Research Center Program funded by the Ministry of Education, Science, and Technology (2011K000673), and NRF World Class University Program (R31-10100).

References

[1] C. E. Antoniak, “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

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

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

[4] 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.

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

[8] 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.

[9] R. M. Neal, “Markov chain sampling methods for Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

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

[11] 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.

[12] 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.