LETTER Communicated by Stefan Haufe

Discriminative Learning of Propagation and Spatial Pattern for Motor Imagery EEG Analysis

Xinyang Li
a0068297@nus.edu.sg
NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore 119613

Haihong Zhang
hhzhang@i2r.a-star.edu.sg
Cuntai Guan
ctguan@i2r.a-star.edu.sg
*Institute for Infocomm Research, A*STAR, Singapore 138632*

Sim Heng Ong
eleongsh@nus.edu.sg
Department of Electrical and Computer Engineering and Department of Bioengineering, National University of Singapore 119613

Kai Keng Ang
kkang@i2r.a-star.edu.sg
Yaozhang Pan
yzpan@i2r.a-star.edu.sg
*Institute for Infocomm Research, A*STAR, Singapore 138632*

Effective learning and recovery of relevant source brain activity patterns is a major challenge to brain-computer interface using scalp EEG. Various spatial filtering solutions have been developed. Most current methods estimate an instantaneous demixing with the assumption of uncorrelatedness of the source signals. However, recent evidence in neuroscience suggests that multiple brain regions cooperate, especially during motor imagery, a major modality of brain activity for brain-computer interface. In this sense, methods that assume uncorrelatedness of the sources become inaccurate. Therefore, we are promoting a new methodology that considers both volume conduction effect and signal propagation between multiple brain regions. Specifically, we propose a novel discriminative algorithm for joint learning of propagation and spatial pattern with an iterative optimization solution. To validate the new methodology, we conduct experiments involving 16 healthy subjects and perform numerical analysis of the proposed algorithm for EEG classification in motor imagery brain-computer interface. Results from extensive analysis

Neural Computation **25, 2709–2733** (2013)
doi:10.1162/NECO_a_00500

validate the effectiveness of the new methodology with high statistical significance.

1 Introduction

Scalp EEG signals are stochastic, nonlinear, and nonstationary (Guler, Kiymik, Akin, & Alkan, 2001) and have relatively low spatial resolution. Therefore, it has been a considerable challenge to compute discriminative and robust features for detecting the brain activity of interest, especially in single-trial brain-computer interface (BCI) studies(Li & Guan, 2006; Llera, Gomez, & Kappen, 2012). In this letter, we consider BCI using motor imagery, although the general methodology can be applied to other brain signals. Motor imagery is a dynamic brain state that can induce the same motor representation internally as motor execution (Jeannerod, 1995). In particular, distinctive brain signals of event-related desynchronization (ERD) and event-related synchronization (ERS) are detectable from EEG during motor imagery (Stavrinou, Moraru, Cimponeriu, Stefania, & Bezerianos, 2007; Pfurtscheller, Brunner, Schlogl, & da Silva, 2006). Therefore, motor imagery becomes an important modality in developing BCI systems (Lo et al., 2010; Ang, Chin, Zhang, & Guan, 2008; Vidaurre, Sannelli, Muller, & Blankertz, 2011).

To improve the signal-to-noise ratio, spatial filtering has been widely used to counter volume conduction effects (Blankertz, Tomika, Lemm, Kawanabe, & Muller, 2008). In motor imagery, EEG classification, probably the most recognized technique, is a common spatial pattern (CSP) (Ramoser, Muller-Gerking, & Pfurtscheller, 2000). In CSP, the desired spatial filters are designed to extract prominent ERD/ERS by maximizing the variance of the projected signal under one condition while minimizing it under the other (Koles, 1991; Gerkinga, Pfurtscheller, & Flyvbjergc, 1999). Various methods have been proposed to improve the performance of CSP by addressing the problem of selecting proper time segments or frequency bands of EEG. In Lemm, Blankertz, Curio, and Muller (2005), common spatiospectral pattern (CSSP) optimizes a simple filter by adding a one-time-delayed sample to have more channels. In Dornhege et al. (2006), common sparse spectral spatial pattern (CSSSP) extends CSSP by adding the optimization of a complete global spatial-temporal filter into CSP. In Ang, Chin, Wang, Guan, and Zhang (2012), Ang, Chin, Zhang, and Guan (2012), and Thomas, Guan, Lau, Vinod, and Ang (2009), EEG signals are decomposed into several frequency bands, CSP is applied to different bands independently, and feature fusion or classifier fusion is introduced to produce final classification results. These methods either implicitly or explicitly assume that raw scalp EEG waveforms are generated by uncorrelated source signals, and subsequently, they may not account for more complicated brain signal dynamics such as causal propagation between different brain regions.

Recently brain activities during motor imagery other than ERD/ERS have been observed in multifunctional areas using functional magnetic resonance imaging (fMRI) or EEG (Formaggio, Storti, Cerini, Fiaschi, & Manganotti, 2010; Chen, Yang, Liao, Gong, & Shen, 2009). In particular, the analysis of neural connectivity is gaining more attention in neuroscience because it describes the general functioning of the brain and communication among its different regions (Astolfi et al., 2006; Ewald, Marzetti, Zappasodi, Meinecke, & Nolte, 2012). For example, causal connectivity is found in motor-related core regions such as the primary motor cortex (M1) and supplementary motor area (SMA) during motor imagery (Chen et al., 2009). The causal flow or time-lagged correlation is beyond volume conduction and is caused by possible neuronal propagation (Gomez-Herrero, Atienza, Egiazarian, & Cantero, 2008). To investigate such propagation effects, directed transfer function (DTF) has been used to evaluate causal flow between any given pair of channels in a multichannel EEG in frequency domain, which was introduced in Baccala and Sameshima (2001), Kaminski and Blinowska (1991), and Kaminski, Ding, Truccolo, and Bressle (2001). This estimation of DTF is based on a multivariate autoregressive model (MVAR), and, more importantly it has been applied to EEG data of voluntary finger movement and motor imagery for event-related causal flow investigation (Ginter, Blinowska, Kaminski, & Durka, 2001; Schlogl & Supp, 2006). Kus, Kaminski, and Blinowska (2004) found that there is a rapid increase of information outflow from electrodes Fc3 and C3 caused by ERS, and propagation of β-synchronization from Fc3 and Fc1 to C3, C1, Cz, Cp3 and Cp1 exists, which gives evidence of communication among sensorimotor areas. However, looking at only the time profiles of ERD/ERS, it is difficult to determine the primary source of activity; hence, existing instantaneous demixing models are not capable of modeling signal propagation among underlying ERD/ERS sources.

In the presence of neuronal propagation and causal relationship during motor imagery, conventional spatial filter design methodology is not sufficient to capture the underlying brain activities (Dyrholm, Makeig, & Hansen, 2007; Bahramisharif, van Gerven, Schoffelen, Ghahramani, & Heskes, 2012). We would like to note that although some of the connectivity measurements mentioned above have been explored already (Wei, Wang, Gao, & Gao, 2007; Gysels & Celka, 2007), only scalp connectivity and intrachannel synchronization measurements are directly used as features, whereas volume conduction effects are not rigorously addressed. One consequence would be that bandpower variations are misinterpreted as changes in connectivity (Grosse-Wentrup, 2009).

Therefore, rather than ignoring the connectivity or propagation between sources in spatial filter design or using scalp connectivity directly as features, we would like to promote a computational model that can more accurately describe the underlying processes by considering both neuronal propagation and volume conduction effects.

In this work, we devise a novel discriminative learning model for motor imagery EEG based on a multivariate convolutive process with an analysis of the spurious effects in classifying ERD/ERS based on an instant linear mixture model. The effectiveness of introducing a time-lagged demixing matrix to produce time-decorrelated data is analyzed theoretically from the perspective of background noise elimination. Furthermore, the demixing matrices accounting for propagation and volume conduction are estimated jointly and iteratively in the proposed unified model. From the experimental study, we evaluate the efficiency of the new methodology in terms of classification accuracy in the two-class motor imagery EEG classification problem. We also analyze the effectiveness of the proposed method for background noise elimination using the Küllback-Leibler divergence measure.

This letter is organized as follows. In section 2, we discuss limitations of conventional spatial filter design and the necessity of considering the causal propagation. Then we give the details of the proposed discriminative learning of propagation and spatial pattern. In section 3, the validity of the proposed method is verified by experimental studies on two-class motor imagery classification. Our concluding remarks are in section 4.

2 Discriminative Learning of Propagation and Spatial Pattern

2.1 Data Model and Problem Formulation. Let $X(t)$ be the timeseries of a multichannel EEG signal, with each component in $X(t)$ representing a particular EEG channel measured at time t. Considering the complex temporal dynamics, especially the latent causal relations in $X(t)$, we describe the observed data $X(t)$ as an m-dimension linear convolutive mixture process of order l (Dyrholm et al., 2007; Mørup, Madsen, & Hansen, 2009),

$$X(t) = \sum_{\tau=1}^{l} \Phi(\tau) S(t-\tau), \tag{2.1}$$

where $S(t)$ is the source signal of interest, $\Phi(\tau)$ is the projection matrix of the order τ, and l is the maximum time-lagged order. When $l = 0$, the observed data $X(t)$ is an instant mixing process. For simplicity of description, the additive EEG noise can be described by an component in $S(t)$. Conventionally, it is assumed in motor imagery EEG classification that $X(t)$ is an instant linear mixture of source signals. This leads to an instant demixing solution to the estimation of $S(t)$,

$$\hat{S}(t) = WX(t), \tag{2.2}$$

where W is the projection or demixing matrix containing m rows, and each row of W is effectively a spatial filter $\mathbf{w}$.

Interestingly, we note that the estimate $\hat{S}(t)$ given by equation 2.2 is also a mixture of the time-lagged components,

$$\hat{S}(t) = \sum_{\tau} \Phi_{\mathbf{w}}(\tau) S(t-\tau), \tag{2.3}$$

where $\Phi_{\mathbf{w}}(\tau) = W\Phi(\tau)$ is a mixing matrix.

A perfect solution would be that $\Phi_{\mathbf{w}}(\tau)$ takes an identity matrix form for $\tau = 0$ and a zero matrix form for any $\tau \neq 0$. This is generally impossible except in the exceptional case that $\Phi(\tau) = 0$ for $\tau \neq 0$, or, in other words, when the convolutive mixture model in equation 2.1 reduces to an instant mixing model.

Remark 1. In discriminative analysis, the spatial filter W is designed to extract the most discriminative signal $\hat{S}(t)$. However, due to the time-lagged relationships, discriminative signals are still mixed with nondiscriminative ones in $\hat{S}(t)$. Therefore, it is necessary to take the causal flow into consideration, together with spatial filter design in a unified model, to have a better estimation of $S(t)$, which is the motivation of this letter.

Solving the reconstruction problem of $S(t)$ from equation 2.1 may lead to a solution in the form of an infinite impulse response (IIR) filter. As we will elaborate shortly and also for practical use, we simplify the problem into a finite impulse response (FIR) filter given by

$$S(t) = W(X(t) - \sum_{\tau=1}^{p} A(\tau) X(t-\tau)), \tag{2.4}$$

where $A(\tau)$ is the demixing matrix of the order τ that accounts for the time-lagged propagation effect.

Remark 2. The manipulation of simplifying the IIR form into the FIR form is for the convenience of practical implementation. Practically, this mixing effect can be accounted for by a finite number of orders, while the rest can be ignored. Although not rigorously proven, the feasibility of this simplification in the discriminative problem will be discussed and validated by the experimental results in section 3.

For the convenience of presentation and analysis, we divide the reconstruction problem of $S(t)$ into two parts. First, we define

$$\widetilde{X}(t) = X(t) - \sum_{\tau=1}^{p} A(\tau) X(t-\tau), \tag{2.5}$$

where $\widetilde{X}(t)$ is the signal processed by a finite multivariate FIR filter of order p. We refer to it as the time-decorrelated data in the following discussion. The source signal can be recovered from the time-decorrelated data $\widetilde{X}(t)$ by

$$S(t) = W\widetilde{X}(t). \tag{2.6}$$

It is interesting that reconstructing $S(t)$ based on equations 2.6 and 2.5 resembles the classical causal connectivity estimation based on MVAR analysis (Dyrholm et al., 2007; Gomez-Herrero et al., 2008; Haufe, Tomioka, Nolte, Muller, & Kawanabe, 2010), where the process $S(t)$ is usually defined as a temporally and spatially uncorrelated time sequence. Different from the connectivity identification, the objective in this letter lies in discriminative learning. Therefore, rather than modeling the signals, the demixing matrix $A(\tau)$ is used to construct the ERD/ERS sources from the measurements. Moreover, $S(t)$ does correspond not to the innovation process but to the ERD/ERS sources, which we explain in detail in the appendix. The objective in estimating $A(\tau)$ is the variance difference between two classes but not the independence of the source, so that the discriminative power of $S(t)$ is maintained. Based on the convolutive model, possible propagation effects can be addressed in the discriminative model. Details of joint estimation of $A(\tau)$ and W in equations 2.6 and 2.5 for the objective of classification are introduced in the following section.

2.2 Joint Estimation of Propagation and Spatial Pattern. We introduce the principle of CSP in the design of joint estimation of propagation and spatial pattern. As CSP can be viewed as a spatial transformation, the principle lies in maximizing the power of the transformed signal for one class while minimizing it for the other. The normalized sample covariance matrix R_i of trial i is obtained as

$$R_i = \frac{X_i X_i^T}{\mathrm{tr}(X_i X_i^T)}, \tag{2.7}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix. In this letter, we consider only the binary classification problem, and the two classes are indexed by $c = \{0, 1\}$. Let $\mathcal{Q}_c$ denote the set of trials that belong to class c such that $\mathcal{Q}_0 \bigcap \mathcal{Q}_1 = \varnothing$. The average covariance matrix for each class is then calculated as

$$R^{(c)} = \frac{1}{|\mathcal{Q}_c|} \sum_{i \in \mathcal{Q}_c} R_i, \tag{2.8}$$

where $|\mathcal{Q}_c|$ denotes the total number of samples belonging to set $\mathcal{Q}_c$. Suppose the signal power is to be maximized for class 0; the objective function in CSP is given by

$$\max_{\mathbf{w}} \mathbf{w} R^{(0)} \mathbf{w}^T \quad \text{s.t.} \quad \mathbf{w}(R^{(0)} + R^{(1)})\mathbf{w}^T = 1. \tag{2.9}$$

Note that the dependence of EEG signals (in equation 2.8 and onwards) on time is implied unless otherwise stated. The idea of discriminating the EEG signals of two different motor imagery classes in terms of power (the variance of the projected signal) in equation 2.9 is directly related to the nature of ERD/ERS. Therefore, we deal with the estimation of $S(t)$ in the proposed model by adopting variance differentiation as the objective. To embed the estimation of $A(\tau)$ in equation 2.4 into the objective function, equation 2.9, we rewrite equation 2.5 to make the relationship between raw EEG data X and the time-decorrelated data $\widetilde{X}$ more compact by defining

$$\hat{A}(\tau) = \begin{cases} I, & \tau = 0 \\ -A(\tau), & \tau > 0 \end{cases}, \tag{2.10}$$

which we refer to as the time-lagged demixing matrix for the simplicity. Therefore, $\widetilde{X}(t)$ in equation 2.5 becomes

$$\widetilde{X}(t) = \sum_{\tau=0}^{p} \hat{A}(\tau) X(t-\tau). \tag{2.11}$$

Similarly, the covariance matrix of $\widetilde{X}(t)$ is

$$\widetilde{R}_i = \frac{\widetilde{X}_i \widetilde{X}_i^T}{\mathrm{tr}(\widetilde{X}_i \widetilde{X}_i^T)}, \tag{2.12}$$

and the average covariance based on $\widetilde{X}(t)$ for each class is

$$\widetilde{R}^{(c)} = \frac{1}{|\mathcal{Q}_c|} \sum_{i \in \mathcal{Q}_c} \widetilde{R}_i. \tag{2.13}$$

Replacing $R^{(c)}$ in equation 2.9 with $\widetilde{R}^{(c)}$ and considering equations 2.11 and 2.12, the optimization problem becomes

$$\max_{\mathbf{w}, \hat{A}(\tau)} \mathbf{w} \left(\sum_{\tau_1=0}^{p} \sum_{\tau_2=0}^{p} \hat{A}(\tau_1) R^{(0)}(\tau_\Delta) \hat{A}(\tau_2) \right) \mathbf{w}^T, \quad \text{s.t.}$$

$$\mathbf{w} \left(\sum_{\tau_1=0}^{p} \sum_{\tau_2=0}^{p} \hat{A}(\tau_1) (R^{(0)}(\tau_\Delta) + R^{(1)}(\tau_\Delta)) \hat{A}(\tau_2) \right) \mathbf{w}^T = 1, \tag{2.14}$$

where $R^{(c)}(\tau_\Delta) = \frac{1}{|\mathcal{Q}_c|} \sum_{i \in \mathcal{Q}_c} X_i(t-\tau_1)(X_i(t-\tau_2))^T$. In this way, the estimation of model 2.4 is achieved by solving the optimization problem in equation 2.14. Moreover, as shown in equation 2.14, only one $\hat{A}(\tau)$, as a part

of the feature extraction model, is obtained on the completion of the optimization since the calculation is conducted with the averaged covariance matrix $R^{(c)}(\tau_\Delta)$ over all the trials. This is very different from the regression model in connectivity analysis, in which the estimated models are different for different trials.

Because the above objective function can be highly nonlinear, we adopt an iteration procedure to estimate $\mathbf{w}$ and $\hat{A}(\tau)$. Since both of the estimations of the spatial filter $\mathbf{w}$ and the time-lagged demixing matrix $\hat{A}(\tau)$ depend on each other, the iterative method alternatively updates one while fixing the other. To be specific, the spatial filter $\mathbf{w}$ can be obtained based on a fixed $\hat{A}(\tau)$ by solving equation 2.9. For $\hat{A}(\tau)$, we calculate the jth column of $\hat{A}(\tau)$, $[\hat{a}_{1j}, \hat{a}_{2j}, \ldots, \hat{a}_{Cj}]^T$, separately based on the fixed spatial filter and $[\hat{a}_{1k}, \hat{a}_{2k}, \ldots, \hat{a}_{Ck}]^T$ $(k = 1, \ldots, C \text{ and } k \neq j)$ from the last iteration. In this way, the information flow from different channels is optimized individually, and the update of $\hat{A}(\tau)$ finishes on the completion of estimating $[\hat{a}_{1j}, \hat{a}_{2j}, \ldots, \hat{a}_{Cj}]^T$ for $j = 1, \ldots, C$. The implementation of the proposed discriminative learning algorithm of propagation and spatial patterns is summarized in algorithm 1. The loop will not stop until the convergence criteria are met. Note that during the optimization, only one spatial filter $\mathbf{w}$ is used. On completion of the optimization, $\widetilde{X}$ can be obtained from equation 2.11, and subsequently $\widetilde{R}^{(c)}$ can be obtained based on equation 2.12. With R_c substituted with $\widetilde{R}^{(c)}$, the optimization problem in equation 2.9 is equivalent to solving the eigenvalue decomposition problem,

$$W\widetilde{R}^{(0)} = \Lambda W \widetilde{R}^{(1)}, \tag{2.15}$$

where Λ is the diagonal matrix containing the eigenvalues of $(\widetilde{R}^{(1)})^{-1}\widetilde{R}^{(0)}$. With the projection matrix W, we select r pairs of spatial filters corresponding to the r largest or smallest components in Λ as in the usual CSP procedure. And the feature $\widetilde{F}_i$ for trial i is obtained from $\widetilde{X}_i$ as

$$\widetilde{F}_i = \log \frac{\mathbf{w}_j \widetilde{X}_i \widetilde{X}_i^T \mathbf{w}_j^T}{\sum_j \mathbf{w}_j \widetilde{X}_i \widetilde{X}_i^T \mathbf{w}_j^T}, \quad j = 1, \ldots, r, N - r + 1, \ldots, N. \tag{2.16}$$

2.3 Background Noise Separation. In this section, we investigate the effectiveness of introducing the time-lagged demixing matrix $\hat{A}(\tau)$ into the estimation of the ERD/ERS source, combined with spatial filter design. To further analyze and evaluate the proposed model, the difference between the time-decorrelated EEG signal $\widetilde{X}(t)$ (see equation 2.5) and original EEG data $X(t)$ is investigated. Suppose $X(t)$ is described by the following MVAR model,

$$X(t) = \sum_{\tau=1}^{q} B(\tau) X(t - \tau) + N(t), \tag{2.17}$$

Algorithm 1: Discriminative Learning of Propagation and Spatial Pattern

Input:

Training EEG data that comprises N sample blocks of X, with each block having a specific class label;

Output:

Spatial filter **w** and time-lagged correlation estimates $\hat{A}(\tau)$.

begin

Set the initial parameters of the spatiotemporal filters $\hat{A}(\tau)$ as zero matrices;

for $k = 1 : n_k$ **do**

Compute $\widetilde{X}$ based on $\hat{A}(\tau)$ using equation 2.11;

Compute **w** by solving the optimization problem in equation 2.9;

% *Update the spatial filter* ***w***

for $j = 1 : C$ **do**

Compute $[\hat{a}_{1j}, \hat{a}_{2j}, \ldots, \hat{a}_{mj}]^T$ based on the updated spatial filter **w** by solving the optimization problem in equation 2.14;

% *Update* $\hat{A}(\tau)$.

Compute the change in the norm $\hat{A}(\tau)$ by $\delta = \|\hat{A}(\tau)^k\| - \|\hat{A}(\tau)^{k-1}\|$;

if $\delta < \zeta$ *(ζ is a small preset constant)* **then**

Stop.

where $N(t)$ is the prediction error. It is also regarded as the innovation process because it is spontaneous and cannot be totally predicted by past observations (Gomez-Herrero et al., 2008). Note that $B(\tau)$ is the mixing matrix based on the regression model, which is different from $A(\tau)$ estimated in the proposed model for discriminative purposes and q is the order of the

MVAR model. Similarly, equation 2.17 is rearranged in the following form to make the input-output relationship more compact,

$$N(t) = \sum_{\tau=0}^{q} \hat{B}(\tau) X(t-\tau), \tag{2.18}$$

where

$$\hat{B}(\tau) = \begin{cases} -I, & \tau = 0; \\ B(\tau), & \tau > 0. \end{cases} \tag{2.19}$$

Transforming equation 2.18 into the frequency domain yields

$$N(f) = B(f)X(f), \tag{2.20}$$

$$B(f) = \sum_{\tau=0}^{q} \hat{B}(\tau) e^{-i2\pi f\tau}, \tag{2.21}$$

where f is the frequency. Therefore, the transfer function of the system $H(f)$ can be described by

$$H(f) = B^{-1}(f), \tag{2.22}$$

such that $X(f) = H(f)N(f)$.

By substituting equation 2.17 into 2.5 and following the steps from equation 2.20 to 2.22, we obtain

$$\widetilde{X}(f) = (I - A(f))X(f) \tag{2.23}$$

$$= \left(H(f) - \frac{A(f)}{B(f)}\right) N(f), \tag{2.24}$$

where

$$A(f) = \sum_{\tau=0}^{p} \hat{A}(\tau) e^{-i2\pi f\tau}. \tag{2.25}$$

Let $\widetilde{H}(f) = H(f) - \frac{A(f)}{B(f)}$, which is the transfer function from $N(f)$ to $\widetilde{X}$. Since the causal flow measurement DTF is defined based on the transfer function (Kaminski et al., 2001), we see that the proposed method changes the information flow by changing the transfer function from $H(f)$ to $\widetilde{H}(f)$. Moreover, comparison of the transfer functions of $\widetilde{X}$ and X in equation 2.23 shows its similarity to the classical signal-plus-noise (SPN) model. In

particular, in Xu et al. (2009), the observed EEG data containing ERP $X_E(f)$ is usually formulated as

$$X_E(f) = \Phi S_E(f) + Z(f) \tag{2.26}$$

where $S_E(f)$ is the ERP of interest and $Z(f)$ is the background noise or the ongoing activity.

Remark 3. As Xu et al. (2009) discussed, the background noise is not noise despite its noise-like appearance but represents ongoing brain activity rich in oscillatory content. In the light of the above discussion, we can interpret equation 2.23 from a similar perspective. As indicated in equation 2.23, the frequency component removed from X is an oscillatory signal with a transfer function $\frac{A(f)}{B(f)}$, and it can be regarded as an estimate of ongoing activity. In other words, this ongoing activity constitutes part of the MVAR process of X with the portion as $\frac{A(f)}{B(f)}$. In this way, the ERD/ERS components are enhanced in the proposed model with the oscillatory background noise attenuated.

The Küllback-Leibler (KL) divergence is a measure of probability divergence given two probability distributions, and it has been used to evaluate nonstationarity in motor imagery EEG classification problem (Arvaneh, Guan, Ang, & Quek, 2013a, 2013b; Bamdadian, Guan, Ang, & Xu, 2012). Therefore, to verify that the component removed from X is the background noise, we adopt the KL divergence as the criterion.

As the gaussian model is usually used to model EEG data, we consider the KL divergence between two gaussian distributions. In particular, the KL divergence between two gaussian distributions with the means and nonsingular covariance matrices corresponding to distribution $\mathcal{N}_0/\mathcal{N}_1$ as μ_0/μ_1 and Σ_0/Σ_1 is

$$D_{\mathrm{KL}}(\mathcal{N}_0||\mathcal{N}_1) = \frac{1}{2}\Bigg(\mathrm{tr}\big(\Sigma_1^{-1}\Sigma_0\big) - (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - \ln\left(\frac{\det\Sigma_0}{\det\Sigma_1} - k\right)\Bigg). \tag{2.27}$$

It is reasonable to assume that the improved separation of background noise will result in more stationary data with fewer within-class dissimilarities. We therefore adopt KL divergence to measure such within-class dissimilarities. The smaller the KL divergences within trials from the same class, the less the variation of the data, which generally relates to better classification results. Since EEG data are usually processed to be centered and the dimension k of the distribution is the number of channel m, for every trial i in class c, we use $D_{\mathrm{KL}}(\mathcal{N}(0,R_i)||\mathcal{N}(0,R^{(c)}))$ to measure the dissimilarity of

the distribution of this trial from the mean distribution of the class c as

$$D_{\mathrm{KL}}(\mathcal{N}(0,R_i)||\mathcal{N}(0,R^{(c)})) = \frac{1}{2}\left(\mathrm{tr}(R_i^{-1}R^{(c)}) - \ln\left(\frac{\det R_i}{\det R^{(c)}}\right) - m\right), \tag{2.28}$$

and subsequently we obtain an average probability divergence D for EEG data X as

$$D = \sum_{c=0,1}\frac{1}{|\mathcal{Q}_c|}\sum_{i\in\mathcal{Q}_c} D_{\mathrm{KL}}(\mathcal{N}(0,R_i)||\mathcal{N}(0,R^{(c)})). \tag{2.29}$$

Similarly, we obtain $\widetilde{D}$ based on $\widetilde{X}$ as

$$\widetilde{D} = \sum_{c=0,1}\frac{1}{|\mathcal{Q}_c|}\sum_{i\in\mathcal{Q}_c} D_{\mathrm{KL}}(\mathcal{N}(0,\widetilde{R}_i)||\mathcal{N}(0,\widetilde{R}^{(c)})). \tag{2.30}$$

In this way, by comparing D and $\widetilde{D}$, we can evaluate the quality of X and $\widetilde{X}$ in terms of within-class dissimilarities.

Remark 4. It is worth noting that the proposed method addresses a more complicated dynamics of motor imagery EEG but does not depend on the very critical explanation of the generation of ERD/ERS. On the one hand, it is possible that propagation effects that contribute to the generation of ERD/ERS exist. On the other hand, discriminative sources could correlate with noise in a convolutive way. Blind source separation or connectivity estimation methodology, as discussed before, may not be effective for classification problems because it is difficult to differentiate between two kinds of propagation effects. The proposed model, which is formulated in a phenomenological form, equation 2.23, takes both cases into consideration.

3 Experimental Results and Discussion

3.1 Data Description and Processing. Sixteen subjects participated in the study with informed consent. Ethics approval was obtained beforehand from the Institutional Review Board of the National University of Singapore. EEGs from the full 27 channels were obtained using Nuamps EEG acquisition hardware with unipolar Ag/AgCl electrodes channels. The sampling rate was 250 Hz with a resolution of 22 bits for the voltage range of $\pm$130 mV. A bandpass filter of 0.05 to 40 Hz was set in the acquisition hardware.

In the experiment, the training and test sessions were recorded on different days with the subjects performing motor imagery. During the EEG recording process, the subjects were asked to avoid physical movement and eye blinking. In addition, they were instructed to perform kinesthetic motor imagery of the chosen hand in two runs. During the rest state, they did mental counting to make the resting EEG signal more consistent. Each run lasted approximately 16 minutes and consisted of 40 trials of motor imagery and 40 trials of rest state. Each training session consisted of two runs, while the test session consisted of two or three runs.

We select the time segments from 0.5 s to 2.5 s after the cue (Arvaneh, Guan, Ang, & Quek, 2011). The raw data are prefiltered by an 8 Hz to 35 Hz bandpass filter that covers rhythms related to motor imagery. The filtered training data are used to train the feature extraction model based on the proposed method as described in section 2.2. The numbers of spatial filters in W are chosen as 2 and 3 ($r = 2, 3$ in equation 2.16). Finally, the extracted training features are used to train a support vector machine (SVM) classifier.

3.2 Investigation on the Order of the Time-Lagged Demixing Matrix. To determine the order p of $\hat{A}(\tau)$ in equation 2.11, we fit the MVAR model to EEG data as in equation 2.17. Although the orders p and q have different meanings, the analysis of the order q of the mixing matrix $B(\tau)$ in equation 2.14 provides the information at which time-lagged level the propagation effects are stronger. Based on equation 2.20 and the analysis given in section 2.3, as $\hat{A}(\tau)$ corresponds to certain components of $B(\tau)$ in frequency domain, it is reasonable to choose the order p of $\hat{A}(\tau)$ in accordance with q, the order of $B(\tau)$. Therefore, the analysis of the mixing matrix $B(\tau)$ can be used to initialize the order p of $\hat{A}(\tau)$ in the proposed model. The Swartz Bayesian criterion is used to automatically select the model order that best matches the data (Schneider & Neumaier, 2001). We found that for every subject, the order 5 for q is selected for most of the trials and the order 4 or 6 is selected for the remaining of the trials. Therefore, we restrict the investigation on the order 4, 5, or 6.

Figure 1 illustrates the result of one subject in the data set introduced in section 3.1. The y-axis indicates the value of the norm of mixing matrix $B(\tau)$ in equation 2.17 of different orders, and the x-axis indicates the order τ. The coefficient matrices are obtained under MVAR models with q equal to 4, 5, or 6 and averaged over the training set and test set, respectively, resulting in the six lines in Figure 1. We see that in all six cases, the norms of the coefficient matrices of orders 2 and 3 are the highest, which means that the data at time t are most influenced by the data at time $t - 2$ and time $t - 3$. Therefore, the order p of $\hat{A}(\tau)$ should include these two time lags, and subsequently the proposed discriminative learning model addresses the most influential propagation effects. Furthermore, we focus on investigating the feasibility of the proposed model with orders 4 and below.

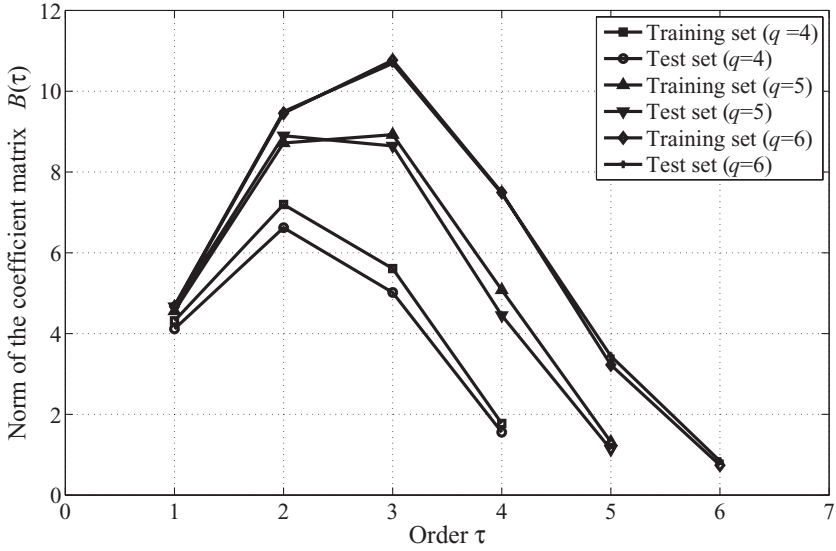


Figure 1: Norms of coefficient matrices under the MVAR model. The x-axis represents the order τ, and y-axis represents the norm of $B(\tau)$. Three MVAR models with order q from 4 to 6 are used to fit EEG data of training and test sets separately, yielding six lines. And the peak points of the six lines correspond to either $\tau = 2$ or $\tau = 3$.

3.3 Classification Results. Tables 1 and 2 summarize the performance of the proposed feature extraction method, compared with CSP as the baseline. In these two tables, we refer to the proposed method as discriminative propagation and spatial pattern analysis (DPSP). Tables 1 and 2 correspond to $r = 2$ and $r = 3$, respectively, and in both tables, results of DPSP with $p = 1, 2, \ldots, 4$ are included.

According to the results, the proposed feature extraction method improves the performance of the classifier, and the improvements are significant when the order of $\hat{A}(\tau)$ in DPSP is 2 or 3 regardless of the value of r, which is in agreement with the previous analysis based on the MAVR model. Specifically, the average classification accuracy for order 2 is 68.30%, and the accuracy for order 3 is 67.91% when $r = 2$, both of which are higher than that of CSP (65.56%). The paired t-test confirms the significance of the improvement at a 5% level with p-values equaling 0.008 and 0.040, corresponding to the cases of $p = 2$ and $p = 3$, respectively. Similar to the results based on two pairs of spatial filters, the average classification accuracy is 68.98% for $p = 2$ and 68.75% for $p = 3$ of DPSP when $r = 3$, higher than that of CSP (66.48%). Also, the significance of the improvement is confirmed by t-test with p-values of 0.027 and 0.022, corresponding to the cases of $p = 2$

Table 1: Session-to-Session Transfer Test Results for $r = 2$ (%).

		DPSP			
Subject	CSP	$p = 1$	$p = 2$	$p = 3$	$p = 4$
1	65.00	65.41	62.91	66.66	67.08
2	51.25	51.25	54.17	52.08	52.08
3	55.00	55.00	57.50	55.83	55.00
4	66.67	66.67	70.41	71.25	77.08
5	54.58	54.16	67.08	70.41	58.33
6	67.08	67.50	72.50	69.16	69.58
7	77.08	77.08	77.92	76.66	72.5
8	94.16	94.16	92.50	96.25	95.41
9	74.58	75.00	75.83	75.83	74.58
10	61.66	61.25	60.41	60.83	60.00
11	46.25	46.67	49.16	53.33	47.08
12	77.00	77.08	81.25	79.58	73.33
13	51.25	51.25	54.58	51.25	50.00
14	72.08	72.08	79.16	73.75	74.58
15	65.83	65.58	67.50	64.16	64.58
16	69.58	69.60	70.00	68.75	65.00
Mean	65.56	65.59	68.30	67.91	66.01
SD	12.26	12.28	11.57	11.79	12.35
p-value	–	0.64	0.008*	0.040*	0.63

*$p \leq 0.05$.

and $p = 3$, respectively. The accuracy for order 4 is 66.01% when $r = 2$ and 66.41% when $r = 3$, which are not significant. Interestingly, the accuracy for order 1 is almost the same as that of CSP in both tables, which also confirms our previous analysis: it is necessary and sufficient for $\hat{A}(\tau)$ to cover the major components of $\hat{B}(\tau)$. The propagation effect is strongest at orders 2 and 3, and the optimization based on $\hat{A}(\tau)$ for order 1 has very limited effect and results in almost the same result. The optimization based on $\hat{A}(\tau)$ of order 4 accounts for most of the propagation effect, but more parameters pose a risk of overfitting. In other words, ideally the higher the order of $\hat{A}(\tau)$, the better the results should be, since more propagation effects are taken into consideration. However, for a higher order, the increased number of parameters would cause overfitting, which makes the classification results deteriorate. To keep a balance between accounting for the propagation effects and overfitting, it is effective to cover as few major components of propagation as possible, which come from orders 2 and 3 in this experiment.

Figure 2 is used to show the comparison result in a more intuitive way. Each plot in Figure 2 shows the test accuracy under DPSP with order p against that under CSP. The x-axis represents the accuracy results under

Table 2: Session-to-Session Transfer Test Results for $r = 3$ (%).

		DPSP			
Subject	CSP	$p = 1$	$p = 2$	$p = 3$	$p = 4$
1	70.41	70.41	71.66	73.33	73.33
2	54.58	54.58	57.08	60.83	54.16
3	56.66	56.66	57.50	55.83	55.00
4	75.41	76.66	76.66	74.16	75.41
5	53.33	53.33	67.08	66.67	54.16
6	68.33	68.33	71.66	71.66	70.83
7	72.50	72.50	75.00	72.92	71.66
8	94.58	94.58	91.66	94.58	95.00
9	76.25	76.58	77.91	76.25	72.50
10	57.50	60.83	60.41	61.67	60.00
11	47.50	47.50	50.41	47.92	47.08
12	75.83	75.41	80.83	81.25	72.05
13	49.58	49.58	51.25	50.00	49.58
14	74.16	74.16	80.41	74.58	75.41
15	64.16	64.16	64.58	65.00	72.08
16	72.91	72.91	68.75	72.08	68.75
Mean	66.48	66.52	68.98	68.74	66.14
SD	12.51	12.04	11.51	11.70	12.34
p-value	–	0.53	0.027*	0.022*	0.55

*$p \leq 0.05$.

CSP, and the y-axis represents that under DPSP. In each plot, a circle above the diagonal line marks a subject for which DPSP outperforms CSP.

Figure 3 shows $A(\tau)$ for two subjects. For a better comparison of differences between the proposed method and the MVAR model, mixing matrices $B(\tau)$ based on the MVAR model of the two subjects are also provided. As shown in Figure 3, the diagonal elements of $B(\tau)$ are much higher than the off-diagonal elements, because the self-spectrum of the signal is usually stronger than the cross-spectrum between the EEG signals. However, there are no large differences between diagonal and off-diagonal elements of $A(\tau)$, and diagonal elements are not significantly higher, which means the self-spectrum of the signal is not modulated radically by $A(\tau)$. Moreover, since elements of higher values concentrate in certain columns, higher weights are given to tune propagation from certain channels.

3.4 Analysis of Background Noise Separation. To further verify the validity of DPSP, we have evaluated the classwise KL divergence (see section 2.3). Results averaged among all subjects are shown in Table 3 and Figure 4. Note that for the computation of D_{KL} of both the training set and the test set, the average covariance matrix $R/\widetilde{R}$ is the mean of the training set since under the single-trial analysis setting, we cannot obtain the mean

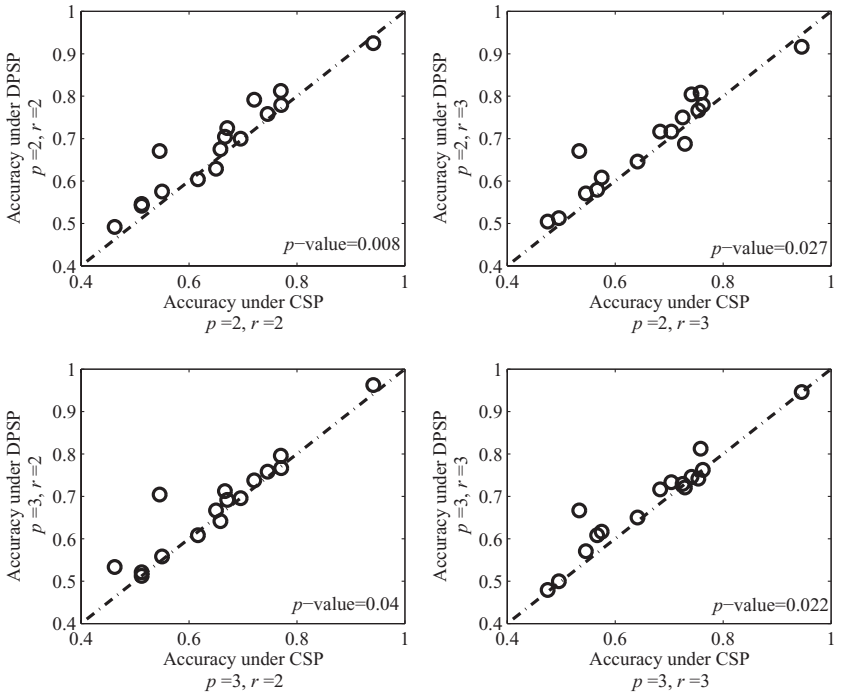
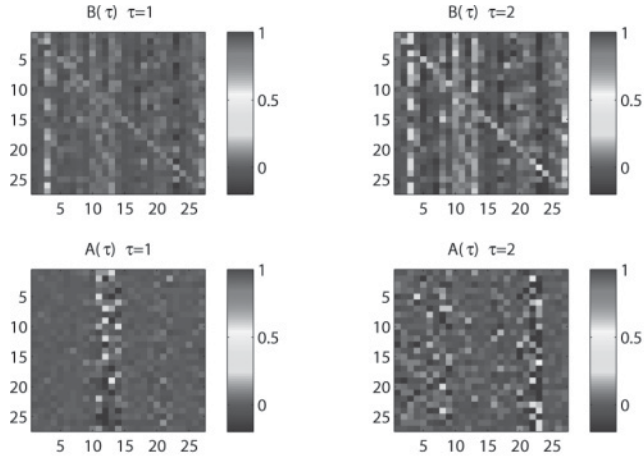


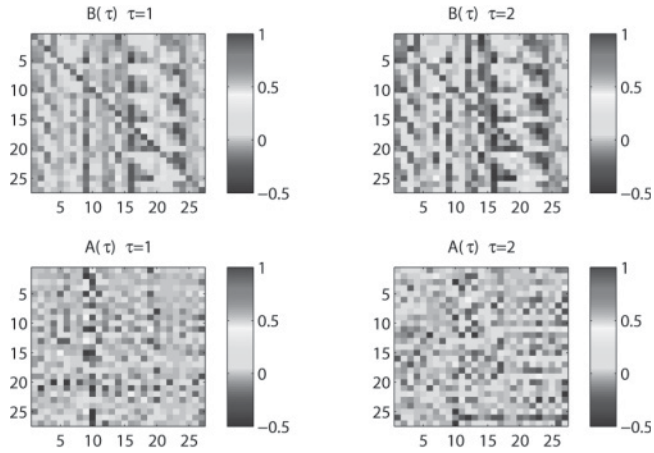
Figure 2: Session-to-session transfer test accuracy. The x-axis represents the accuracy results under CSP, and the y-axis represents that under DPSP with different orders p and numbers of spatial filters r. The $y = x$ line is denoted by a dotted-dashed line. In each plot, a circle above the $y = x$ line marks a subject for which DPSP outperforms CSP. It can be seen from the plots that improvements of DPSP for order 2 and 3 are significant.

of the test set. Therefore, the fact that the average divergence D of the test set is larger than that of the training set in all cases reflects the differences between the test set and the training set, as indicated by Table 3. This is mainly caused by the session-to-session transfer effects. According to the results, the proposed DPSP algorithm decreases the KL divergence within the same class for both the training set and the test set, which means that compared to the raw EEG data X, data processed by DPSP $\widetilde{X}$ are more stationary. A more significant decrease is achieved for the test set, which means that the proposed method is more robust to the session-to-session transfer effects. Moreover, the comparison between different orders indicates that better performance is achieved with the order 2, which is in accord with the accuracy results.

Figure 5 illustrates the correlation between the decrease of KL divergence and the increase of the classification accuracy at the subject level. The linear



(a) Comparison between $A(\tau)$ and $B(\tau)$ for subject 7



(b) Comparison between $A(\tau)$ and $B(\tau)$ for subject 14

Figure 3: Comparison of coefficient matrices obtained by the proposed method, $A(\tau)$, and the mixing matrices in MVAR, $B(\tau)$. For both subjects, the diagonal elements of $B(\tau)$ are much higher than the off-diagonal elements. For $A(\tau)$, elements of higher values are found in certain columns.

correlation coefficient r_c equals 0.30 and 0.31, corresponding to $p = 2$ and $p = 3$, respectively. Due to the large variety across subjects, their KL divergence may lie in different feature spaces. The decrease of KL divergence and the increase of classification performance may not correlate linearly. As illustrated in Figure 5, almost all the points lie in the first quadrant, indicating

Table 3: Decrease of KL Divergence (%).

	$p=2$		$p=3$			$p=4$	
	D	$\widetilde{D}$	$1-\frac{\widetilde{D}}{D}$	$\widetilde{D}$	$1-\frac{\widetilde{D}}{D}$	$\widetilde{D}$	$1-\frac{\widetilde{D}}{D}$
Training set	4.96	4.09	17.68%	4.25	14.39%	4.84	2.55%
Test set	64.3	25.2	60.84%	36.68	42.98%	57.09	11.24%

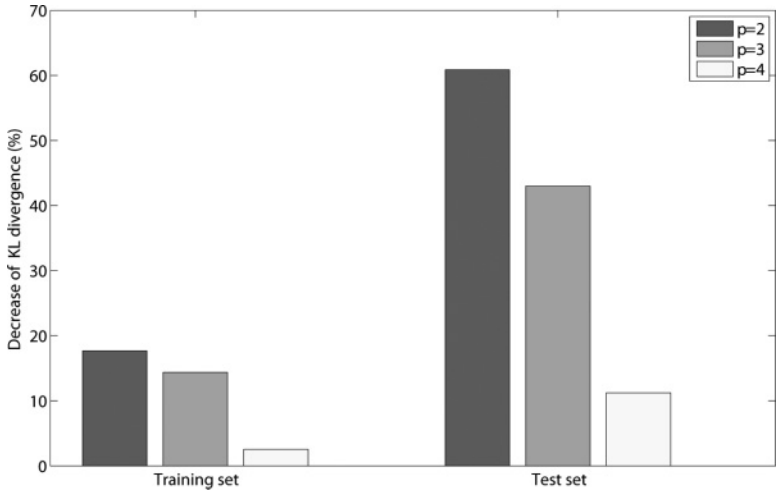
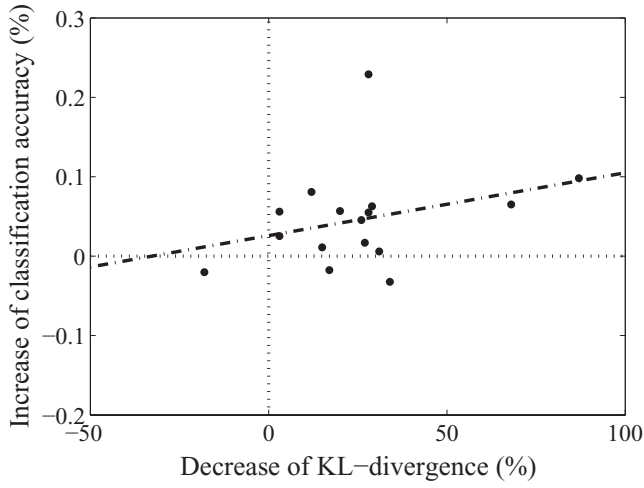


Figure 4: Decrease of the KL divergence. The decreases of the KL divergence in $\widetilde{X}$ of different orders compared to X are shown as percentages. A great decrease in the KL divergence indicates that $\widetilde{X}$ is more stationary than X. Therefore, the proposed DPSP algorithm is more robust toward varying background noise and session-to-session transfer effects.

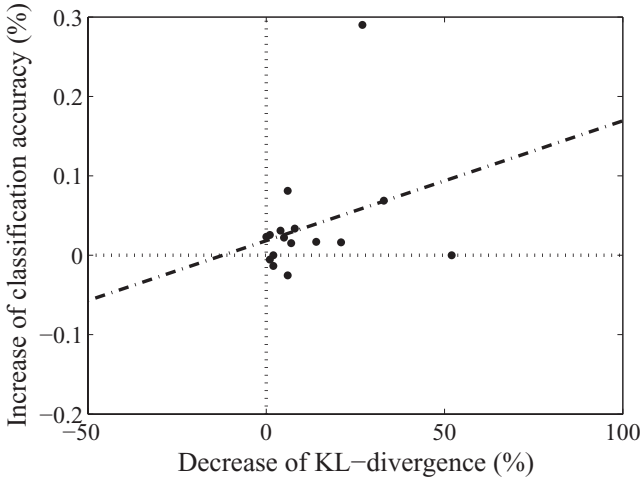
that the decreased KL divergence is positively correlated with the increased classification accuracy. Therefore, the decrease of the KL divergence contributes to the increase of the classification accuracy to a certain extent. Nevertheless, the reason for the increase of the classification could be more complicated so that KL divergence cannot completely represent it. We will investigate this issue in the future work.

4 Conclusion

The coexistence of brain connectivity and volume conduction may have complicated effects in EEG measurements and poses technical challenge to detecting specific brain activities of interest. Conventional linear spatial filters design methods with the assumption of unconnectedness of sources



(a) $r_c = 0.30$ ($p = 2$)



(b) $r_c = 0.31$ ($p = 3$)

Figure 5: Correlation between the decrease of the KL divergence and the increase of the classification accuracy. The *x*-axis represents the decrease of the KL divergence, and the *y*-axis represents the increase of the classification accuracy. Panels a and b correspond to $p = 2$ and $p = 3$ respectively.

are not sufficient in addressing such complicated dynamics. Due to the causal relationship, reconstructed ERD/ERS signals based on instantaneous demixing may not be the optimized results in terms of discrimination.

Moreover, the propagation effects are closely related to the background noise and nonstationarity of EEG. It is possible that an electrode that contains no discriminative information could be given a high weight due to information flow from signals containing ERD/ERS, and such dependence could be very unstable compared with original ERD/ERS source. This analysis is the motivation to propose the unified model for discriminative learning of propagation and spatial patterns.

Therefore, we have reported in this letter a novel computational model that accounts for both time-lagged correlations between signals and the volume conduction effect. Different from the sparsely connected sources analysis (SCSA) model in Haufe et al. (2010) and MVAR-ICA model in Gomez-Herrero et al. (2008), the proposed computational model is designed from discriminative analysis but also takes propagation into account. Besides, an iteration procedure–based algorithm is implemented for the estimation of the proposed discriminative model. Experiment results have shown statistically significant improvement in classification accuracy under the proposed learning method. Moreover, the effectiveness of the background noise attenuation is also confirmed with a significant decrease of KL divergence of EEG data of the same class, especially for test data. This indicates that the proposed method is more robust than conventional methods against the session-to-session nonstationarity in EEG.

Appendix: Relations Between the Convolutive Model and the Instantaneous Model with Connected Sources

Based on the model in Haufe et al. (2010) and Gomez-Herrero et al. (2008), $X(t)$ can be assumed to be generated as a linear instantaneous mixture of source signal $S(t)$, which follows an multivariant autoregression (MVAR) model,

$$X(t) = MS(t), \tag{A.1}$$

$$S(t) = \sum_{\tau} B_s(\tau) S(t-\tau) + \epsilon(t), \tag{A.2}$$

where $B_s(\tau)$ is the coefficient matrix of the MVAR model and it represents the connectivity between sources (Ginter et al., 2001; Schlogl & Supp, 2006). From equation A.1, the innovation process $\epsilon(t)$ can be written as

$$\begin{aligned}\epsilon(t) &= M^{-1}X(t) - \sum_{\tau} B_s(\tau) M^{-1} X(t-\tau) \\ &= \sum_{\tau} \hat{B}_s(\tau) X(t-\tau),\end{aligned} \tag{A.3}$$

where

$$\hat{B}_s(\tau) = \begin{cases} M^{-1}, & \tau = 0 \\ -B_s(\tau)M^{-1}, & \tau > 0 \end{cases}. \tag{A.4}$$

Equation A.3 shows the equivalence between this model and the convolutive model in Dyrholm et al. (2007) and Mørup et al. (2009) and the proposed approach, with the underlying convolutive sources replaced by innovations. Because the objective in Haufe et al. (2010) and Gomez-Herrero et al. (2008) is connectivity analysis, the estimation of $B_s(\tau)$ and M is based on the nongaussianity assumption of $\epsilon(t)$. In the proposed model, $S(t)$ represents the discriminative sources related to ERD/ERS, and thus the estimation of the FIR matrix $\hat{A}(\tau)$ in equation 2.11 and spatial filter $\mathbf{w}$ is based on maximizing the variance difference between the two classes. With the discriminative objective, it is preferable to apply the convolutive model to impose the variance difference as the prior information of the source. Moreover, since the two models are equivalent, it is also possible to build a discriminative model based on the instantaneous mixing model with connected sources in equations A.1 and A.2. In future work, we would like to explore possible discriminative learning approach to study the connectivity that contains class information.

References

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., & Zhang, H. (2012). Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, *6*(39).

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *Proceedings of the IEEE International Joint Conference on Neural Networks and Computational Intelligence* (pp. 2390–2397). Piscataway, NJ: IEEE.

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2012). Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs. *Pattern Recognition*, *45*(6), 2137–2144.

Arvaneh, M., Guan, C., Ang, K. K., & Quek, C. (2011). Optimizing the channel selection and classification accuracy in EEG-based BCI. *IEEE Transactions on Biomedical Engineering*, *58*(6), 1865–1873.

Arvaneh, M., Guan, C., Ang, K. K., & Quek, C. (2013a). EEG data space adaptation to reduce inter-session non-stationarity in brain-computer interface. *Neural Computation*, *25*, 2146–2171.

Arvaneh, M., Guan, C., Ang, K. K., & Quek, C. (2013b). Optimizing spatial filters by minimizing within-class dissimilarities in EEG-based BCI. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(4), 610–619.

Astolfi, L., Cincotti, F., Mattia, D., de Vico Fallani, F., Salinari, S., Ursino, M., & Babiloni, F. (2006). Estimation of the cortical connectivity patterns during the intention of limb movements. *IEEE Engineering in Medicine and Biology Magazine*, *25*(4), 32–38.

Baccala, L. A., & Sameshima, K. (2001). Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics*, *84*, 463–474.

Bahramisharif, A., van Gerven, M. A. J., Schoffelen, J. M., Ghahramani, Z., & Heskes, T. (2012). The dynamic beamformer. Machine learning and interpretation in neuroimaging. In G. Langs, I. Rish, M. Grosse-Wentrup, & B. Murphy (Eds.), *Lecture notes in computer science* (vol. 7263, pp. 148–155). Berlin: Springer.

Bamdadian, A., Guan, C., Ang, K. K., & Xu, J. (2012). Online semi-supervised learning with kl distance weighting for motor imagery-based BCI. In *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 2732–2735). Piscataway, NJ: IEEE.

Blankertz, B., Tomika, R., Lemm, S., Kawanabe, M., & Muller, K. R. (2008). Optimizing spatial filters for robust EEG single trial-trial analysis. *IEEE Signal Processing Magazine*, *25*(1), 41–56.

Chen, H., Yang, Q., Liao, W., Gong, Q., & Shen, S. (2009). Evaluation of the effective connectivity of supplementary motor areas during motor imagery using Granger causality mapping. *NeuroImage*, *47*(4), 1844–1853.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., & Muller, K.-R. (2006). Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Transactions on Biomedical Engineering*, *53*(11), 2274–2281.

Dyrholm, M., Makeig, S., & Hansen, L. K. (2007). Convolutive ICA for spatiotemporal analysis of EEG. *Neural Computation*, *19*, 934–955.

Ewald, A., Marzetti, L., Zappasodi, F., Meinecke, F. C., & Nolte, G. (2012). Estimating true brain connectivity from EEG/MEG data invariant to linear and static transformations in sensor space. *NeuroImage*, *60*(1), 476–488.

Formaggio, E., Storti, S. F., Cerini, R., Fiaschi, A., & Manganotti, P. (2010). Brain oscillatory activity during motor imagery in EEG-fMRI coregistration. *Magnetic Resonance Imaging*, *28*(10), 1403–1412.

Gerkinga, J. M., Pfurtscheller, G., & Flyvbjergc, H. (1999). Designing optimal spatial filters for single-trial EEG classiffication in a movement task. *Clinical Neurophysiology*, *110*, 787–798.

Ginter, J. Jr., Blinowska, K. J., Kaminski, M., & Durka, P. J. (2001). Phase and amplitude analysis in time-requency space-pplication to voluntary finger movement. *Journal of Neuroscience Methods*, *110*(1–2), 113–124.

Gomez-Herrero, G., Atienza, M., Egiazarian, K., & Cantero, J. L. (2008). Measuring directional coupling between EEG sources. *NeuroImage*, *43*(3), 497–508.

Grosse-Wentrup, M. (2009). Understanding brain connectivity patterns during motor imagery for brain-computer interfacing. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*, *21* (pp. 561–568). Cambridge, MA: MIT Press.

Guler, I., Kiymik, M. K., Akin, M., & Alkan, A. (2001). AR spectral analysis of EEG signals by using maximum likelihood estimation. *Computers in Biology and Medicine*, *31*(6), 441–450.

Gysels, E., & Celka, P. (2007). Phase synchronization for the recognition of mental tasks in a braincomputer interface. *IEEE Transactions on Rehablitation Engineering*, *12*(4), 406–415.

Haufe, S., Tomioka, R., Nolte, G., Muller, K.-R., & Kawanabe, M. (2010). Modeling sparse connectivity between underlying brain sources for EEG/MEG. *IEEE Transactions on Biomedical Engineering*, *57*(8), 1954–1963.

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia*, *33*, 1419–1432.

Kaminski, M., & Blinowska, K. (1991). A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, *65*, 203–210.

Kaminski, M., Ding, M., Truccolo, W. A., & Bressle, S. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, *85*, 145–157.

Koles, Z. J. (1991). The quantitative extraction and toporraphic mapping of the abnormal components in the clinical EEG. *Electroencephalography and Clinical Neurophysiology*, *79*, 440–447.

Kus, R., Kaminski, M., & Blinowska, K. J. (2004). Determination of EEG activity propagation: Pair-wise versus multichannel estimate. *IEEE Transactions on Biomedical Engineering*, *51*(9), 1501–1510.

Lemm, S., Blankertz, B., Curio, G., & Muller, K.-R. (2005). Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Transactions on Biomedical Engineering*, *52*(9), 1541–1548.

Li, Y., & Guan, C. (2006). An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces. *Neural Computation*, *18*, 2730–2761.

Llera, A., Gomez, V., & Kappen, H. J. (2012). Adaptive classification on brain-computer interfaces using reinforcement signals. *Neural Computation*, *24*, 2900–2923.

Lo, A. C., Guarino, P. D., Richards, L. G., Haselkorn, J. K., Wittenberg, G. F., Federman, D. G., & Peduzzi, P. (2010). Robot-assisted therapy for long-term upper-limb impairment after stroke. *New England Journal of Medicine*, *362*, 1772–1783.

Mørup, M., Madsen, K. H., & Hansen, L. K. (2009). Latent causal modelling of neuroimaging data. In *NIPS Workshop on Connectivity Inference in Neuroimaging*.

Pfurtscheller, G., Brunner, C., Schlogl, A., & da Silva, F. H. L. (2006). Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*, *31*(1), 153–159.

Ramoser, H., Muller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, *8*(4), 441–446.

Schlogl, A., & Supp, G. (2006). Analyzing event-related EEG data with multivariate autoregressive parameters. *Progress in Brain Research*, *159*, 135–147.

Schneider, T., & Neumaier, A. (2001). Algorithm 808: Arfit—a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, *6*, 58–65.

Stavrinou, M., Moraru, L., Cimponeriu, L., Stefania, P. D., & Bezerianos, A. (2007). Evaluation of cortical connectivity during real and imagined rhythmic finger tapping. *Brain Topography*, *19*(3), 137–145.

Thomas, K. P., Guan, C., Lau, C. T., Vinod, A. P., & Ang, K. K. (2009). A new discriminative common spatial pattern method for motor imagery brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, *56*(11), 2730–2733.

Vidaurre, C., Sannelli, C., Muller, K.-R., & Blankertz, B. (2011). Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural Computation*, *23*, 791–816.

Wei, Q., Wang, Y., Gao, X., & Gao, S. (2007). Amplitude and phase coupling measures for feature extraction in an EEG-based brain¨ccomputer interface. *Journal of Neural Engineering*, *4*, 120–129.

Xu, L., Stoica, P., Li, J., Bressler, S. L., Shao, X., & Ding, M. (2009). Aseo: A method for the simultaneous estimation of single-trial event-related potentials and ongoing brain activities. *IEEE Transactions on Biomedical Engineering*, *56*(1), 111–121.

Received January 28, 2013; accepted May 5, 2013.