This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

Probabilistic Common Spatial Patterns for Multichannel EEG Analysis

Wei Wu, *Member, IEEE,* Zhe Chen, *Senior Member, IEEE,* Xiaorong Gao, *Member, IEEE,* Yuanqing Li, *Member, IEEE,* Emery N. Brown, *Fellow, IEEE,* and Shangkai Gao, *Fellow, IEEE*

Abstract—Common spatial patterns (CSP) is a well-known spatial filtering algorithm for multichannel electroencephalogram (EEG) analysis. In this paper, we cast the CSP algorithm in a probabilistic modeling setting. Specifically, *probabilistic CSP* (P-CSP) is proposed as a generic EEG spatio-temporal modeling framework that subsumes the CSP and regularized CSP algorithms. The proposed framework enables us to resolve the overfitting issue of CSP in a principled manner. We derive statistical inference algorithms that can alleviate the issue of local optima. In particular, an efficient algorithm based on eigendecomposition is developed for *maximum a posteriori* (MAP) estimation in the case of isotropic noise. For more general cases, a variational algorithm is developed for group-wise sparse Bayesian learning for the P-CSP model and for automatically determining the model size. The two proposed algorithms are validated on a simulated data set. Their practical efficacy is also demonstrated by successful applications to single-trial classifications of three motor imagery EEG data sets and by the spatio-temporal pattern analysis of one EEG data set recorded in a Stroop color naming task.

Index Terms—common spatial patterns, Fukunaga-Koontz transform, sparse Bayesian learning, variational Bayes, electroencephalogram, brain-computer interface.

1 Introduction

Electroencephalography (EEG) is a non-invasive imaging modality that is widely used to measure the electrical activities of the brain. Multichannel EEG simultaneously measures coordinated brain activities at multiple sites on the scalp at millisecond temporal resolution, which makes it valuable for cognitive and neural engineering studies and for clinical applications [1]. However, the analysis of EEG remains challenging because the volume-conducted EEG suffers from a low spatial resolution, such that the signal recorded at each individual channel is a mixture of attenuated activities from more than one brain region, and it frequently suffers interference from various (e.g., cardiac, muscular, and ocular) artifacts. To address these challenges, one must enhance the signal-to-noise ratio (SNR) and isolate the overlapping activities via *spatial filtering* — that is, linearly combining the EEG signals at multiple channels such that the sources of interest are enhanced and the unwanted sources are suppressed [2].

Among the various EEG spatial filtering methods, the *common spatial patterns* (CSP) algorithm [3] has attracted considerable attention as an effective method for the concurrent analysis of multichannel EEG signals recorded under two conditions. Under the name of *Fukunaga-Koontz transform*, CSP was first proposed as a supervised learning method that was an extension of *principal component analysis* (PCA) for feature extraction [4]. Since then, it has become popular in a diverse range of applications [5]–[7]. Notably, CSP has been successful in extracting sensorimotor rhythms for *brain-computer interfaces* (BCIs), as evidenced in international BCI competitions [8]–[10].

Consider two conditions of multichannel EEG signals $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times L\times 2}$, where $\mathbf{X}_{\cdot,\cdot,k} = [\mathbf{X}_{\cdot,1,k} \quad \cdots \quad \mathbf{X}_{\cdot,L,k}]$ is the data matrix for condition k, which consists of the vectors of N-channel EEG signals with L sample points [1]. For conciseness, we let $\mathbf{X}_k \triangleq \mathbf{X}_{\cdot,\cdot,k}$. CSP is aimed at finding a set of linear transforms (*spatial filters*) to maximize the ratio of the transformed data's variance between the two conditions. Mathematically, the spatial filters are the *stationary points* of the following optimization problem [3]:

$$\max_{\mathbf{w}} J(\mathbf{w}) \triangleq \frac{\mathbf{w}^\top \hat{\mathbf{R}}_1 \mathbf{w}}{\mathbf{w}^\top \hat{\mathbf{R}}_2 \mathbf{w}} \qquad s.t. \quad \|\mathbf{w}\|_2 = 1, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^N$ denotes a spatial filter, and $\hat{\mathbf{R}}_k \in \mathbb{R}^{N\times N} \triangleq \mathbf{X}_k\mathbf{X}_k^\top/L$ denotes the estimated spatial covariance matrix for condition k. Since $J(\mathbf{w})$ is a Rayleigh quotient, the stationary points can be obtained collectively in a closed form as the eigenvectors of a generalized eigendecomposition: $\hat{\mathbf{R}}_1\mathbf{w} = \lambda\hat{\mathbf{R}}_2\mathbf{w}$, where λ denotes the eigenvalue associated with $\mathbf{w}$. Since $\lambda = J(\mathbf{w})$, it is often presumed to be a good measure of the separability between the spatially filtered signals of two conditions.

Nonetheless, as a multivariate algorithm, CSP is known to suffer from *overfitting*, which may yield poor generalization performance [3], [11], [12]. We use the term "overfitting" when a statistical model or algorithm describes noise rather than the underlying data structure. In the literature, overfitting has mainly been tackled by *regularization*, i.e., by incorporating an additional penalty term in the cost function in

- *W. Wu and Y. Li are with School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: auweiwu@scut.edu.cn).*
- *Z. Chen is with Department of Psychiatry, Department of Neuroscience and Physiology, New York University School of Medicine, New York, NY 10016, USA.*
- *E. N. Brown is with Department of Brain and Cognitive Sciences and Division of Health Science and Technology, Massachusetts Institute of Technology-Harvard University, Cambridge, MA 02139, USA.*
- *X. Gao and S. Gao are with Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China.*

1. Without loss of generality, the EEG signal at each channel is assumed to have a zero mean hereafter.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

(1) to restrict the search space of the unknown spatial filters. More specifically, the *regularized CSP* [13], [14], motivated by Tikhonov regularization in the context of linear inverse problems, uses a weighted ℓ_2-norm penalty to enforce the smoothness of the entries in the (weighted) spatial filters (see also [15] for a comprehensive review of the group of regularized CSP algorithms):

$$\max_{\mathbf{w}} J_r(\mathbf{w}) \triangleq \frac{\mathbf{w}^\top \hat{\mathbf{R}}_1 \mathbf{w}}{\mathbf{w}^\top \hat{\mathbf{R}}_2 \mathbf{w} + \rho \mathbf{w}^\top \mathbf{H} \mathbf{w}} \qquad s.t. \quad \|\mathbf{w}\|_2 = 1, \quad (2)$$

where ρ is the regularization parameter, and $\mathbf{H}$ is a symmetric positive semi-definite matrix. By contrast, the *sparse CSP* [16]–[18] uses an ℓ_1-norm penalty to impose sparsity on the spatial filters:

$$\max_{\mathbf{w}} J_s(\mathbf{w}) \triangleq \frac{\mathbf{w}^\top \hat{\mathbf{R}}_1 \mathbf{w}}{\mathbf{w}^\top \hat{\mathbf{R}}_2 \mathbf{w} + \rho \|\mathbf{w}\|_1} \qquad s.t. \quad \|\mathbf{w}\|_2 = 1, \quad (3)$$

where the formulation in [17] is employed. Multiple filters are found sequentially via the deflation method [16].

Despite that the various regularization strategies may ameliorate CSP's overfitting, the algorithms were designed primarily for classifying instead of modeling the EEG data — in a way akin to classical modeling techniques such as factor analysis and independent component analysis (ICA) [19] — and therefore not specifically designed for exploring the underlying spatio-temporal dynamics. To the best of our knowledge, a principled modeling methodology to address the CSP overfitting issue remains missing to date.

1.1 Contributions

The contributions of this paper consist of both theoretical and algorithmic levels:

• We establish the *probabilistic CSP* (P-CSP) model as a general framework to characterize multichannel EEG under two experimental conditions (Section 3). Specifically, we show that CSP and regularized CSPs can be subsumed under the proposed framework in that they can be derived from the P-CSP model as a special case in the noiseless and square mixing scenario. Formulating an existing algorithm within a probabilistic framework is beneficial for both theoretical and practical reasons. From a statistical perspective, this approach allows us to examine when the algorithm will perform well or poorly. From a practical standpoint, associating a probability model with an algorithm allows us to assess the uncertainty of the data analysis results, and opens the possibility of improving the algorithmic performance by model refinement.

• We develop effective algorithms to address the overfitting issue of CSP (Section 4). Two inference algorithms: MAP-CSP and VB-CSP, are derived from the P-CSP framework to alleviate the local optima problem of conventional *maximum a posteriori* (MAP)-based iterative updating algorithms. Specifically, MAP-CSP assumes additive isotropic noise and is suited for real-time EEG classification due to its computational efficiency. VB-CSP performs approximate Bayesian inference for more general noise conditions. The algorithm is capable of automatically inferring the component number, and can be used for the exploratory analysis of EEG spatio-temporal patterns in neurophysiologically-driven studies when there is no obvious performance metric as in classification tasks. We also provide detailed analyses to examine their properties. Finally, we apply these algorithms to analyze the synthetic data and experimental EEG datasets (Section 5).

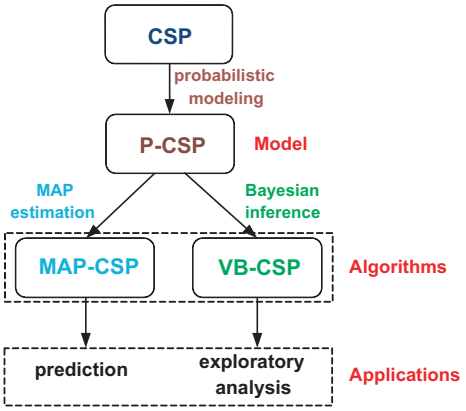

Fig. 1. P-CSP modeling framework. The models and algorithms proposed in this paper are based on a probabilistic modeling reformulation of the CSP algorithm.

As a roadmap of the paper, the P-CSP modeling framework is depicted in Fig. 1. For notations, the set of multi-condition multichannel EEG signals can be mathematically viewed as a three-way (channel×time×condition) tensor; we make this fact explicit and denote it by $\boldsymbol{\mathcal{X}}$. Similarly, the set of multi-condition component signals is denoted by a three-way (component×time×condition) tensor $\boldsymbol{\mathcal{Z}}$. Throughout the paper, scalars are denoted by italic normal letters, matrices and vectors are denoted by upright boldface letters, $\mathbf{I}$ denotes the identity matrix, and the superscript $^\top$ denotes the transpose operator.

2 Related Work

Recent years have witnessed a growing number of sophisticated CSP variants in the literature, particularly in the BCI field. We present a brief review below, in addition to the regularized CSP algorithms described in Section 1.

One important line of CSP-related algorithmic advancements concerns the automatic learning of the optimal temporal or spectral filters in conjunction with the spatial filters. Lemm et al. [20] and Dornhege et al. [12] exploited the idea of variance ratio maximization to optimize both the spatial and temporal filters Tomioka et al. [21] and Wu et al. [22] proposed iterative algorithms that alternate between CSP and other learning criteria (Fisher ratio maximization in the former and the maximum margin in the latter) for the simultaneous optimization of spatial and spectral filters. Zhao et al. [23] generalized CSP to high-dimensional spaces within a tensor analysis framework. Zhang et al. [24] considered a spatio-spectral filtering network, in which multiple CSPs were embedded within a filter bank, with each targeting a distinct frequency subband. More recently, Higashi et al. [25] proposed a discriminative algorithm to design spatio-temporal filters by optimizing a modified CSP cost function. Suk et al. [26] presented a particle-based Bayesian spatio-spectral filter optimization algorithm.

Along other lines, CSP has been extended from binary to multi-class case by several groups [27]–[29]. To handle the setting of a small labeled sample size, Li et al. [30] proposed an EM algorithm for joint extraction and classification of CSP features, where unknown labels of the data were

treated as latent variables. Wu et al. [31] presented a hierarchical Bayesian method to model the inter-trial variability of the EEG signals. Alternatively, the non-stationarity issue has also been addressed within a regularization framework in [32], [33], and via a cluster-based approach in [34]. Finally, several robust CSP algorithms have been developed to alleviate the sensitivity to noise and outliers [35]–[38].

3 Probabilistic Generative Model Formulation of the CSP Algorithm

The discriminative formulation of CSP in (1) is motivated by maximizing the separability between two conditions. In this section, we present a generative view of CSP, which casts the solution as a maximum likelihood (ML) estimate from a probability model of the multichannel EEG signals. The probability model consists of two *coupled* latent linear models, with each modeling EEG signals derived from one condition:

$$\mathbf{X}_k = \mathbf{A}\mathbf{Z}_k, \quad \mathbf{Z}_{\cdot,l,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_k), \tag{4}$$

where $\mathbf{Z}_k \triangleq \mathbf{Z}_{\cdot,\cdot,k} \in \mathbb{R}^{M\times L}$ consists of the vectors of component signals (latent variables) for condition k, and $\mathbf{A} \in \mathbb{R}^{N\times M}$ is the non-degenerate square mixing matrix that contains *spatial patterns* (i.e., scalp maps of the components) as columns. Three assumptions are made in the model (4): 1) $\mathbf{X}_k$ and $\mathbf{Z}_k$ are identically and independently distributed (IID) across time; 2) $M = N$; 3) The component signals are mutually uncorrelated, i.e., $\mathbf{\Lambda}_k = \mathrm{diag}(\lambda_{mk})$ is a diagonal matrix.

The connection between model (4) and CSP is revealed by the following theorem [3], [31], [39], [40]:

Theorem 1. *Let* $\mathbf{W} \triangleq [\mathbf{w}_1, \cdots, \mathbf{w}_N]^\top$, *where* $\mathbf{w}_{1:N}$ *are the stationary points for (1). Then* $\mathbf{W} = \hat{\mathbf{A}}^{-1}$, *where* $\hat{\mathbf{A}}$ *is the ML estimate of* $\mathbf{A}$ *in model (4).*

The proof of the theorem is detailed in [31]. Non-stationarity has been suggested as a generic criterion for blind source separation (BSS) in Pham's pioneering work [41]. Under the generative setting, CSP is derived as an algorithm for BSS by utilizing the non-stationarity of EEG data between conditions.

Remark 1. *In model (4), the mixing matrix* $\mathbf{A}$ *is identical for the two conditions, i.e., the spatial patterns are common to both conditions, hence justifying the name "CSP". It shall be stressed that while sharing common spatial patterns, the two conditions are differentiated by assessing the variance ratios of the associated time courses* $\mathbf{Z}_k$ *(see (1)) — the initial motivation for applying CSP to discriminative EEG analysis. Theorem 1 states that by fitting model (4) to the EEG signals derived from two conditions, the optimal spatial filters are the "dual" of* $\hat{\mathbf{A}}$ *in that they can be obtained by taking the inverse of the latter.*

In cognitive neuroscience, a standard practice for testing a hypothesis regarding a condition of interest is to contrast it with a control condition so that the confounding effects of extraneous variables can be eliminated [42]. The generative algorithmic formulation provides a theoretical ground for CSP as a spatio-temporal decomposition method (i.e., *modeling* the data as apposed to merely *classifying* the data).

3.1 Revisiting the Overfitting Issue

Probabilistic model (4) may shed light on the overfitting issue of CSP. Specifically, there are two situations in which CSP is prone to overfitting:
1) When the true number of underlying components is less than the number of the channels N, CSP necessarily produces spurious components due to the inflated component number assumed in its generative model [11]. The overfitted components often possess large variance ratios between conditions because they may fit the noise component in one condition that is only weakly present in the other condition.
2) Overfitting also occurs when N is large relative to the amount of data available [43]. Here the amount of data need not be taken literally as L; the effective amount of data is considerably smaller when high temporal correlations exist within the samples. In model (4), the number of free parameters is $N+N^2$, which may easily outnumber L, even when the channel number is moderate.

The overfitting issue of CSP stems from the square mixing and noiseless assumptions. The noiseless assumption implies that the EEG data are fully characterized by the estimated components and the mixing matrix. This assumption does not take into account of random factors, such as the amplifier noise. The square mixing assumption is closely linked to the noiseless assumption in that if we relax the square mixing assumption by using a smaller number of components, a model mismatch will automatically arise between the best linear fit and the EEG data.

3.2 Connections with Regularized CSPs

Learning CSP filters relies on estimating the spatial covariance matrix $\mathbf{R}_k$ for the EEG data of each condition. As such, regularized CSP (see (2)) has attempted to alleviate the overfitting issue by using more robust estimates of $\mathbf{R}_k$. In this section, we present a unified view for the group of regularized CSP algorithms based on the probabilistic model (4).

The following theorem asserts that by imposing conjugate priors on the spatial covariance matrices for the two conditions, various regularized CSP algorithms can be cast into the P-CSP framework as specific algorithms computing the MAP estimates of the model parameters with different priors.

Theorem 2. *Regularized CSPs yield the joint MAP estimate of* $\mathbf{A}$ *and* $\mathbf{\Lambda}_k$ *in model (4), with the following inverse-Wishart prior on* $\mathbf{R}_k$*:*

$$p(\mathbf{R}_k) \triangleq \frac{T_k |\mathbf{G}_k|^{(\nu_k - N - 1)/2}}{|\mathbf{R}_k|^{\nu_k/2}} \exp\big(-\mathrm{tr}\big[\mathbf{R}_k^{-1}\mathbf{G}_k\big]/2\big), \tag{5}$$

where $\mathbf{G}_k \in \mathbb{R}^{N\times N}$ *is a positive-definite scale matrix,* ν_k *is a degree-of-freedom parameter, and* T_k *is a normalization constant.* $\mathrm{tr}[\cdot]$ *denotes the trace operator.*

See Appendix B for the proof. Various regularized CSP algorithms can be differentiated by their specific choice of $\mathbf{G}_k$ in the inverse-Wishart prior. For instance, $\mathbf{G}_k$ can be proportional to the estimated covariance matrices from other subjects [14], [44], to the estimated covariance matrix of the noise source [13], or to the identity matrix [15]. Borrowing information from other subjects can potentially benefit subject-to-subject transfer, however, care must be taken to allow for the large between-subject variability in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

the EEG signals. Estimating the covariance matrix of the noise source requires additional EEG signals to be recorded beyond the experimental conditions.

The P-CSP modeling framework presented in the next section takes a different perspective by imposing joint priors on the underlying spatio-temporal patterns to obtain a parsimonious representation of the EEG signals. The derived algorithms regularize the common spatial patterns in a group-wise fashion, in which an "optimal" tradeoff between data fitting and regularization can be automatically learned from the data within the probabilistic framework (Section 4.2.2).

4 P-CSP Modeling of Multichannel EEG

4.1 Basic Model

The P-CSP model for multichannel EEG signals is a noise-corrupted *Bayesian* latent linear model:

$$\mathbf{X}_k = \mathbf{A}\mathbf{Z}_k + \mathbf{E}_k \tag{6}$$
$$\mathbf{A}_{n,\cdot} \sim \mathcal{N}(\mathbf{0}^\top, \mathbf{\Xi}), \mathbf{Z}_{\cdot,l,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_k), \mathbf{E}_{\cdot,l,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}_k).$$

Here, $\mathbf{X}_k$, $\mathbf{Z}_k$, and $\mathbf{A}$ are similarly defined as in model (4). $\mathbf{\Xi} \triangleq \mathrm{diag}[\boldsymbol{\xi}] \in \mathbb{R}^{M\times M}$. $\mathbf{E}_k \in \mathbb{R}^{N\times L}$ is the matrix of additive Gaussian noise for condition k, with the covariance matrix $\mathbf{\Psi}_k \triangleq \mathrm{diag}[\boldsymbol{\psi}_k] \in \mathbb{R}^{N\times N}$.

As opposed to the noise-free model (4) with $M = N$, model (6) assumes that $M \leq N$, i.e., there are no more component signals than EEG signals (*over-determined*). Intuitively, this fact means that the dynamics of the EEG signals from both conditions can be represented by a smaller number of independent component signals, with spatial patterns identical across conditions. In addition, $\mathbf{E}_k$ accounts for the mismatch between the component space and EEG space.

Remark 2. *Model (6) imposes priors on both the spatial and temporal patterns in the component space. With the row-wise IID Gaussian priors, $\mathbf{A}$ is placed on an equal probabilistic footing with $\mathbf{Z}_k$, which are endowed with column-wise IID Gaussian priors.*

4.2 Inference Algorithms

For model inference, one may compute the MAP estimates of $\{\mathbf{A}, \boldsymbol{\mathcal{Z}}\}$ in model (6) via alternate updates to increase the posterior. Nonetheless, there are two limitations that remain to be resolved. First, the MAP estimation via alternate updates is known to be susceptible to local optima, since it fails to account for uncertainties when making hard decisions throughout the update process [19]. Second, in practice $\mathbf{\Lambda}_k, \mathbf{\Psi}_k, \mathbf{\Xi}$ are unknown *a priori*, and the proper determination of these unknowns is crucial because they serve to control the model capacity to prevent overfitting.

To address the above limitations, we present two algorithms for model inference: MAP-CSP and VB-CSP. For a given M, MAP-CSP is able to compute the MAP estimates of $\{\mathbf{A}, \boldsymbol{\mathcal{Z}}\}$ in a closed form when $\mathbf{\Psi}_k$ are isotropic, thus mitigating the issue of local optima. Nonetheless, the model size must be specified in advance. VB-CSP is an approximate fully Bayesian inference algorithm that computes the variational distributions of $\{\mathbf{A}, \boldsymbol{\mathcal{Z}}\}$ by integrating over all of the other unknowns while simultaneously achieving automatic model selection. We also provide an analysis to show that VB-CSP can be understood as a sparse learning algorithm.

4.2.1 MAP-CSP: A Fast MAP Estimation Algorithm

MAP-CSP seeks the joint MAP estimates for $\{\mathbf{A}, \boldsymbol{\mathcal{Z}}\}$ in the following hierarchical Bayesian model with additive isotropic noise:

$$\mathbf{X}_k = \mathbf{A}\mathbf{Z}_k + \mathbf{E}_k \tag{7}$$
$$\mathbf{A}_{n,\cdot} \sim \mathcal{N}(\mathbf{0}^\top, \mathbf{\Xi}), \mathbf{Z}_{\cdot,l,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_k), \mathbf{E}_{\cdot,l,k} \sim \mathcal{N}(\mathbf{0}, \psi_k \mathbf{I})$$
$$\mathbf{\Lambda}_k \sim \prod_m \mathcal{G}a^{-1}(\alpha, \beta), \psi_k \sim \mathcal{G}a^{-1}(\alpha, \beta),$$

where $\mathcal{G}a^{-1}(x|\alpha,\beta) \triangleq \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1}\exp(-\beta/x)$ is the inverse-gamma distribution. We assume that $\mathbf{\Psi}_k$ are isotropic so that the number of unknown parameters in the noise covariances is reduced to two, thereby the local optima arising from estimating full $\mathbf{\Psi}_k$ are largely avoided.

To let the data speak for themselves, we further assume that $\alpha \to 0, \beta \to 0$ and $\boldsymbol{\xi} \to \infty$ to render the priors on $\mathbf{A}$ and $\mathbf{\Lambda}_k$ non-informative [45], yielding flat priors on $\mathbf{A}$ and $p(\lambda_{mk}) \propto 1/\lambda_{mk}$. In contrast to the flat prior on $\mathbf{A}$ (which does not lead to the orthogonality between the columns of $\mathbf{A}$ since the prior makes no contribution in the MAP estimation of $\{\mathbf{A}, \boldsymbol{\mathcal{Z}}\}$ regardless of the orthogonality between the columns of $\mathbf{A}$), the hierarchical prior on $\mathbf{Z}_k$ enforces the belief that the component signals are mutually uncorrelated (i.e., the rows of $\mathbf{Z}_k$ are orthogonal) as in CSP (see also Theorem 3 below along with its proof in Appendix C for how to achieve the orthogonality). More specifically, the hierarchical prior on $\mathbf{Z}_k$ is equivalent to the Student-t distributions on $\mathbf{Z}_{m,\cdot,k}$ ([46]; see also Appendix A for an integral representation of the Student-t distribution):

$$p(\mathbf{Z}_{m,\cdot,k}) = \lim_{\alpha,\beta\to 0} \frac{\Gamma(\alpha+\frac{L}{2})}{\Gamma(\alpha)[2\pi]^{\frac{L}{2}}} \beta^\alpha \left[\beta + \frac{\|\mathbf{Z}_{m,\cdot,k}\|_2^2}{2}\right]^{-(\alpha+\frac{L}{2})}$$
$$\propto 1/\|\mathbf{Z}_{m,\cdot,k}\|_2^L. \tag{8}$$

As $\alpha \to 0, \beta \to 0$, the resulting distributions are heavily tailed and sharply peaked at the origin, thereby favoring sparsity.

Remark 3. *$p(\lambda_{mk}) \propto 1/\lambda_{mk}$ is an improper probability distribution. It was argued in [47] that an improper distribution for the prior variance parameters does not yield a proper posterior distribution in several types of hierarchical models. To avoid the improperness issue, in the following we assume small nonzero values for α and β, e.g., $\alpha = \beta = 10^{-8}$. Moreover, the result of sensitivity analysis presented in Section 5.1.2 demonstrates that the VB solution is relatively insensitive to the choice of the values for α and β.*

In a nutshell, MAP-CSP is an iterative algorithm, with each iteration consisting of two phases. The first phase identifies a low-dimensional subspace common to the two conditions, and the second phase finds axes on which the data from the two conditions are jointly decorrelated. The algorithm is formalized below.

Theorem 3. *For given ψ_k, the joint MAP estimates of $\{\mathbf{A}, \boldsymbol{\mathcal{Z}}\}$ in (7) can be obtained by solving*

$$\min_{\mathbf{A},\mathbf{Z}_k} \sum_k \psi_k^{-1} \|\mathbf{X}_k - \mathbf{A}\mathbf{Z}_k\|_{\mathrm{F}}^2 \tag{9}$$
$$s.t.\ \mathbf{Z}_k\mathbf{Z}_k^\top \in \mathbb{D}^+,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the matrix Frobenius norm. $\mathbb{D}^+$ is the manifold of real diagonal matrices with nonnegative diagonal entries.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

Algorithm 1 The MAP-CSP Algorithm

Input: multichannel EEG data $\boldsymbol{\mathcal{X}}$ that are recorded from two experimental conditions
Output: MAP estimates $\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}, \hat{\psi}_k$
1: **Initialization:** set M; $\hat{\psi}_k = 1$
2: **repeat**
3: solve $\sum_k \hat{\psi}_k^{-1}\mathbf{X}_k\mathbf{X}_k^\top\mathbf{U} = \mathbf{U}\mathbf{D}$;
% perform eigendecomposition of $\sum_k \hat{\psi}_k^{-1}\mathbf{X}_k\mathbf{X}_k^\top$
% eigenvalues are sorted in descending order in
% the main diagonal of $\mathbf{D}$
4: $\mathbf{B} \leftarrow \mathbf{U}_{\cdot,1:M}$, $\mathbf{Y}_k \leftarrow (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{X}_k$
5: $\hat{\psi}_k \leftarrow (\|\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\|_{\mathrm{F}}^2 + 2\beta)/[N \cdot (L + 2\alpha)]$
6: **until** Convergence
7: solve $\mathbf{Y}_1\mathbf{Y}_1^\top\mathbf{V} = \mathbf{Y}_2\mathbf{Y}_2^\top\mathbf{V}\mathbf{D}$
% perform generalized eigendecomposition of $\mathbf{Y}_k\mathbf{Y}_k^\top$
8: $\hat{\mathbf{Z}}_k \leftarrow \mathbf{V}^\top\mathbf{Y}_k$, $\hat{\mathbf{A}} \leftarrow \mathbf{B}\mathbf{V}^{-\top}$

The following result is required to solve (9):

Theorem 4. *Problem (9) is equivalent to*

$$\min_{\mathbf{B},\mathbf{Y}_k} \sum_k \psi_k^{-1}\|\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\|_{\mathrm{F}}^2, \tag{10}$$

where $\mathbf{B} \in \mathbb{R}^{N\times M}$, $\mathbf{Y}_k \in \mathbb{R}^{M\times L}$.

Theorem 4 asserts that the orthogonality constraint on $\mathbf{Z}_k$ can be effectively removed when we solve (9), leading to an unconstrained optimization problem. Now let $\{\mathbf{B}^*, \mathbf{Y}_k^*\} \triangleq \arg\min_{\mathbf{B},\mathbf{Y}_k} \sum_k \psi_k^{-1}\|\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\|_{\mathrm{F}}^2$; then, we have

Theorem 5.

$$\begin{aligned}\mathbf{B}^* &= \mathbf{U}_{\cdot,1:M} \cdot \mathbf{G}\\ \mathbf{Y}_k^* &= (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{X}_k,\end{aligned} \tag{11}$$

where $\mathbf{U}_{\cdot,1:M} \in \mathbb{R}^{N\times M}$ *is the matrix with columns being the eigenvectors of* $\sum_k \psi_k^{-1}\mathbf{X}_k\mathbf{X}_k^\top$ *associated with the* M *largest eigenvalues, and* $\mathbf{G} \in \mathbb{R}^{M\times M}$ *is an arbitrary invertible matrix.*

Theorems 3–5 provide the theory to identify a low-dimensional subspace common to the two conditions. See Appendices C–E for the proofs. In light of the results presented above, the two phases per iteration of MAP-CSP are as follows:

1) Employ an iterative procedure to optimize $\{\mathbf{B}, \boldsymbol{\mathcal{Y}}\}$ and ψ_k in an alternate manner until convergence. More specifically, for given ψ_k, $\{\mathbf{B}, \boldsymbol{\mathcal{Y}}\}$ can be optimized according to Theorem 5, and ψ_k can in turn be updated as (see Appendix C)

$$\psi_k = (\|\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\|_{\mathrm{F}}^2 + 2\beta)/[N \cdot (L + 2\alpha)]. \tag{12}$$

2) Let the generalized eigendecomposition of $\mathbf{Y}_k\mathbf{Y}_k^\top$ be $\mathbf{Y}_1\mathbf{Y}_1^\top\mathbf{V} = \mathbf{Y}_1\mathbf{Y}_2^\top\mathbf{V}\mathbf{D}$. $\mathbf{Z}_k$ and $\mathbf{A}$ can then be estimated using

$$\mathbf{Z}_k = \mathbf{V}^\top\mathbf{Y}_k, \quad \mathbf{A} = \mathbf{B}\mathbf{V}^{-\top}. \tag{13}$$

The pseudocode of MAP-CSP is provided in Algorithm 1. We initialize the algorithm by setting M and ψ_k. In our implementation, $\psi_k = 1$. Each iteration of MAP-CSP involves an eigendecomposition and matrix inversion, requiring $\mathcal{O}(N^3 + M^3)$ flops. Moreover, the algorithm is guaranteed to converge to a stationary point typically within a few iterations, since the noise covariance matrices are parameterized by only two parameters. Hence, MAP-CSP can be implemented efficiently. The convergence can be checked by evaluating whether the relative change of the parameters between adjacent iterations is less than a predefined tolerance η.

Model Selection MAP-CSP assumes that the number of underlying components is known, which hardly holds in practice. Classical statistical model selection criteria, such as the Bayesian information criterion (BIC) [19], are not applicable to resolve this difficulty, since they require the number of parameters to be fixed, whereas in MAP-CSP, the dimensions of $\mathbf{Z}_k$ vary with the number of data points. Nonetheless, in EEG classification cross-validation can be used for model selection in a straightfoward manner based on predictive accuracy; the increased computational cost should not be a concern due to the fast running speed of MAP-CSP.

Alternatively, we can take a regularization approach by placing group-sparse priors on both $\mathbf{A}$ and $\mathbf{Z}_k$ (by contrast, placing the group-sparse prior on $\mathbf{Z}_k$ alone with no penalty on $\mathbf{A}$, as in MAP-CSP, cannot achieve model selection due to the scaling ambiguity between $\mathbf{A}$ and $\mathbf{Z}_k$ — the full model is always preferred, since the entries of $\mathbf{Z}_k$ can be made arbitrarily small by exchanging their amplitudes with those of $\mathbf{A}$). This approach has the advantage that the model order can be automatically determined with a proper inference procedure. We proceed to the detail in the next subsection.

4.2.2 VB-CSP: A Variational Bayesian Inference Algorithm

In more general cases where the additive noise is non-isotropic for each condition, MAP-CSP no longer preserves the optimality. In this section, we propose a fully Bayesian method for model inference. The developed algorithm is an approximate Bayesian inference algorithm that is capable of inferring the "optimal" model capacity.

In contrast to pursuing the mode of $p(\mathbf{A}, \boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}})$ as in MAP-CSP, Bayesian inference attempts to estimate the full posterior distribution $p(\mathbf{A}, \boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}})$ for the following model:

$$\begin{aligned}\mathbf{X}_k &= \mathbf{A}\mathbf{Z}_k + \mathbf{E}_k \qquad (14)\\ \mathbf{A}_{n,\cdot} \sim \mathcal{N}(\mathbf{0}^\top, \boldsymbol{\Xi}), \mathbf{Z}_{\cdot,l,k} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_k), \mathbf{E}_{\cdot,l,k} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_k)\\ \boldsymbol{\Xi} \sim \prod_m \mathcal{G}a^{-1}(\alpha,\beta), \boldsymbol{\Lambda}_k &\sim \prod_m \mathcal{G}a^{-1}(\alpha,\beta), \boldsymbol{\Psi}_k \sim \prod_n \mathcal{G}a^{-1}(\alpha,\beta).\end{aligned}$$

Analogous to $\mathbf{Z}_k$, with the inverse-gamma hyperprior on the covariance of $\mathbf{A}$, $\mathbf{A}$ is now endowed with a column-sparse prior (with α and β close to zeros):

$$\begin{aligned}p(\mathbf{A}_{\cdot,m}) &= \lim_{\alpha,\beta\to 0} \frac{\Gamma(\alpha + \frac{N}{2})}{\Gamma(\alpha)[2\pi]^{\frac{N}{2}}}\beta^\alpha\left[\beta + \frac{\|\mathbf{A}_{\cdot,m}\|_2^2}{2}\right]^{-(\alpha+\frac{N}{2})}\\ &\propto 1/\|\mathbf{A}_{\cdot,m}\|_2^N.\end{aligned} \tag{15}$$

This particular model specification is inspired by the influential idea of *automatic relevance determination* (ARD) [19], which has been widely used in the machine learning community. Intuitively, $\boldsymbol{\Xi}$ and $\boldsymbol{\Lambda}_k$ comprise hyperparameters that govern the amplitude of $\mathbf{A}_{\cdot,m}$ and $\mathbf{Z}_{m,\cdot,k}$, respectively; a component with small hyperparameters will be effectively zeroed out. Note that although the nuisance parameters $\boldsymbol{\Xi}$ and $\boldsymbol{\Lambda}_k$ are estimated by evidence maximization (also known as type-II ML) in ARD, we instead follow a fully Bayesian path by integrating out these parameters.

Remark 4. *Model (14) can be viewed as a Bayesian matrix co-factorization model [48] for* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *due to the symmetry*

between $\mathbf{A}$ *and* $\mathbf{Z}_k$.

Exact Bayesian inference is not viable for model (14) due to intractable integrations. The problem arises from the product coupling of $\mathbf{A}$ and $\mathbf{Z}_k$ in the likelihood, as well as the inconvenient form of the sparse priors. Instead, we devise the *variational Bayesian* CSP (VB-CSP) algorithm for approximate inference. The key ingredients of the algorithm are the two differing variational techniques that are employed to bound the marginal likelihood. The first technique seeks a "surrogate" probability to globally approximate the posterior probability. The second variational technique, which has gained popularity in recent years [49]–[51], is based on Fenchel's duality theorem. We use it for locally approximating the sparse priors. Below, we describe how these techniques are integrated in VB-CSP.

First, we seek a *variational distribution* $q^*(\mathbf{A}, \boldsymbol{\mathcal{Z}})$ in a structured probability space $\mathcal{Q}$ that finds the optimal approximation of the true posterior (in the Kullback-Leibler (KL) divergence sense) $p(\mathbf{A}, \boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}})$ [19]:

$$q^*(\mathbf{A}, \boldsymbol{\mathcal{Z}}) \triangleq \min_{q} \mathrm{KL}\big[q(\mathbf{A}, \boldsymbol{\mathcal{Z}}) \| p(\mathbf{A}, \boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}})\big]. \tag{16}$$

We make use of the mean-field approximation by assuming that the distributions in $\mathcal{Q}$ are factorable such that $\mathbf{A}$ and $\boldsymbol{\mathcal{Z}}$ are probabilistically decoupled: $q(\mathbf{A}, \boldsymbol{\mathcal{Z}}) = q(\mathbf{A})q(\boldsymbol{\mathcal{Z}})$. The marginal log-likelihood is given by

$$\begin{aligned}\log p(\boldsymbol{\mathcal{X}}) &= -\mathcal{F}(\boldsymbol{\mathcal{X}}, q(\mathbf{A}, \boldsymbol{\mathcal{Z}})) + \mathrm{KL}\big[q(\mathbf{A}, \boldsymbol{\mathcal{Z}}) \| p(\mathbf{A}, \boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}})\big] \\ &\geq -\mathcal{F}(\boldsymbol{\mathcal{X}}, q(\mathbf{A}, \boldsymbol{\mathcal{Z}})),\end{aligned} \tag{17}$$

with $\mathcal{F}$ being the *variational free energy*:

$$\mathcal{F}(\boldsymbol{\mathcal{X}}, q(\mathbf{A}, \boldsymbol{\mathcal{Z}})) \triangleq -\langle \log p(\boldsymbol{\mathcal{X}}, \mathbf{A}, \boldsymbol{\mathcal{Z}}) \rangle_q - \mathcal{H}[q(\mathbf{A}, \boldsymbol{\mathcal{Z}})], \tag{18}$$

where $\langle \cdot \rangle_q$ is the expectation with respect to the variational distribution and $\mathcal{H}[\cdot]$ is the differential entropy. As observed from (17), minimizing the KL divergence between the variational distribution and posterior distribution is equivalent to minimizing $\mathcal{F}$, which is an upper bound for the negative marginal log-likelihood.

Next, $\mathcal{F}$ can be further upper bounded by using a convex representation of the Student-t distribution (see Appendix A):

$$\mathcal{F} \leq \tilde{\mathcal{F}}, \tag{19}$$

where $\tilde{\mathcal{F}} \triangleq \min_{\boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k, \boldsymbol{\Xi}} (L/2+\alpha) \sum_k \log|\boldsymbol{\Psi}_k| + (L/2+\alpha) \sum_k \log|\boldsymbol{\Lambda}_k| + (N/2+\alpha)\log|\boldsymbol{\Xi}| + 1/2 \sum_k \Big\langle \mathrm{tr}\Big[\boldsymbol{\Psi}_k^{-1}\big[(\mathbf{X}_k - \mathbf{A}\mathbf{Z}_k)(\mathbf{X}_k - \mathbf{A}\mathbf{Z}_k)^\top + 2\beta\mathbf{I}\big]\Big]\Big\rangle_q + 1/2 \sum_k \Big\langle \mathrm{tr}\Big[\boldsymbol{\Lambda}_k^{-1}\big(\mathbf{Z}_k\mathbf{Z}_k^\top + 2\beta\mathbf{I}\big)\Big]\Big\rangle_q + 1/2 \Big\langle \mathrm{tr}\Big[\boldsymbol{\Xi}^{-1}\big(\mathbf{A}^\top\mathbf{A} + 2\beta\mathbf{I}\big)\Big]\Big\rangle_q + \Big\langle \log q(\mathbf{A})\Big\rangle_q + \sum_k \Big\langle \log q(\mathbf{Z}_k)\Big\rangle_q$.

VB-CSP is aimed at inferring $q(\mathbf{A}, \boldsymbol{\mathcal{Z}})$ by minimizing $\tilde{\mathcal{F}}$:

$$\min_{q(\mathbf{A}, \boldsymbol{\mathcal{Z}})} \tilde{\mathcal{F}}(\boldsymbol{\mathcal{X}}, q(\mathbf{A}, \boldsymbol{\mathcal{Z}})). \tag{20}$$

The problem can be tackled by alternately updating the variational distributions $q(\mathbf{A})$ and $q(\mathbf{Z}_k)$, and the variational parameters $\boldsymbol{\Lambda}_k, \boldsymbol{\Xi}, \boldsymbol{\Psi}_k$ via coordinate descent. Derivation of VB-CSP is provided in Appendix F. The pseudocode is provided in Algorithm 2, in which $\hat{\mathbf{A}}$ and $\hat{\mathbf{Z}}_k$ are the variational means of $\mathbf{A}$ and $\mathbf{Z}_k$, respectively. In our implementation, $\hat{\mathbf{A}}$, $\boldsymbol{\Xi}$, and $\boldsymbol{\Lambda}_k$ are initialized using the estimates from CSP. Moreover, $\boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} = \boldsymbol{\Psi}_k = 10^{-8}\mathbf{I}$. However, it is empirically observed that the algorithmic performance is only slightly affected by initialization (see sensitivity analysis in Section 5.1.2). The main cost of VB-CSP is the computation of matrix inversions at each iteration, which requires $\mathcal{O}(K \cdot L \cdot M^3 + N \cdot M^3)$ flops. Convergence is guaranteed to a stationary point and can be determined by checking whether the decrease of $\tilde{\mathcal{F}}$ between adjacent iterations is less than a pre-defined tolerance η.

Algorithm 2 The VB-CSP Algorithm

Input: multichannel EEG data $\boldsymbol{\mathcal{X}}$ that are recorded from two experimental conditions

Output: variational parameters $\{\boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k\}_{k=1,2}, \boldsymbol{\Xi}$; variational distributions $q^*(\mathbf{Z}_k), q^*(\mathbf{A})$

1: **Initialization**: $M = N$; set $\hat{\mathbf{A}}$ and $\boldsymbol{\Lambda}_k$ by calling CSP; $\boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} = \boldsymbol{\Psi}_k = 10^{-8}\mathbf{I}$, $\xi_m = \|\hat{\mathbf{A}}_{\cdot,m}\|_2^2$

2: **repeat**

3: $q(\mathbf{Z}_k) = \prod_l q(\mathbf{Z}_{\cdot,l,k}) \leftarrow \prod_l \mathcal{N}(\hat{\mathbf{Z}}_{\cdot,l,k}, \boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}})$ where $\hat{\mathbf{Z}}_{\cdot,l,k} \triangleq \boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}} \hat{\mathbf{A}}^\top \boldsymbol{\Psi}_k^{-1} \mathbf{X}_{\cdot,l,k}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}} \triangleq \big[\hat{\mathbf{A}}^\top \boldsymbol{\Psi}_k^{-1} \hat{\mathbf{A}} + \sum_n \psi_{nk}^{-1} \boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} + \boldsymbol{\Lambda}_k^{-1}\big]^{-1}$

4: $q(\mathbf{A}) = \prod_n q(\mathbf{A}_{n,\cdot}) \leftarrow \prod_n \mathcal{N}(\hat{\mathbf{A}}_{n,\cdot}, \boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}})$ where $\hat{\mathbf{A}}_{n,\cdot} \triangleq \sum_{n,k} \psi_{kn}^{-1} x_{nlk} \hat{\mathbf{Z}}_{\cdot,l,k}^\top \boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}}$ and $\boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} \triangleq \big[\boldsymbol{\Xi}^{-1} + \sum_{l,k} \psi_{nk}^{-1} \big(\boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}} + \hat{\mathbf{Z}}_{\cdot,l,k} \hat{\mathbf{Z}}_{\cdot,l,k}^\top\big)\big]^{-1}$

5: $\boldsymbol{\Lambda}_k \leftarrow \frac{1}{L+2\alpha} \sum_l \big(\mathrm{diag}\big[\boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}} + \hat{\mathbf{Z}}_{\cdot,l,k} \hat{\mathbf{Z}}_{\cdot,l,k}^\top\big] + 2\beta\mathbf{I}\big)$

6: $\boldsymbol{\Psi}_k \leftarrow \frac{1}{L+2\alpha} \sum_l \big(\mathrm{diag}\big[\mathbf{X}_{\cdot,l,k}\mathbf{X}_{\cdot,l,k}^\top - 2\mathbf{X}_{\cdot,l,k} \hat{\mathbf{Z}}_{\cdot,l,k}^\top \hat{\mathbf{A}}^\top + \big\langle \mathbf{A}[\boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}} + \hat{\mathbf{Z}}_{\cdot,l,k} \hat{\mathbf{Z}}_{\cdot,l,k}^\top]\mathbf{A}^\top \big\rangle_q\big] + 2\beta\mathbf{I}\big)$

7: $\boldsymbol{\Xi} \leftarrow \frac{1}{N+2\alpha} \sum_n \big(\mathrm{diag}\big[\boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} + \hat{\mathbf{A}}_{n,\cdot}^\top \hat{\mathbf{A}}_{n,\cdot}\big] + 2\beta\mathbf{I}\big)$

8: **until** Convergence

Analysis: VB-CSP as a Sparse Bayesian Learning Algorithm In order to gain deeper insight into VB-CSP, we provide analysis to show that the algorithm induces sparsity to the approximate Bayesian solution in a MAP-like manner. This portion of the study is partially inspired by the seminal work concerning sparse Bayesian learning [51], [52].

To simplify analysis, we assume that $\alpha = \beta = 0$, $\boldsymbol{\Psi}_k$ are known, and $\boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} \boldsymbol{\Sigma}_{\mathbf{Z}_{\cdot,l,k}} = \mathbf{0}$. Let $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\Lambda}_k, \boldsymbol{\Xi}\}$; then, we have

Theorem 6. *The VB inference problem (20) can be rephrased using the following MAP setup:*

$$\min_{\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}} \mathcal{L}(\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}) + \mathcal{R}_{\mathrm{VB}}(\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}), \tag{21}$$

with the loss term $\mathcal{L}$ *defined by* $\mathcal{L}(\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}) \triangleq \sum_k \mathrm{tr}\Big[\boldsymbol{\Psi}_k^{-1}\big(\mathbf{X}_k - \hat{\mathbf{A}}\hat{\mathbf{Z}}_k\big)\big(\mathbf{X}_k - \hat{\mathbf{A}}\hat{\mathbf{Z}}_k\big)^\top\Big]$, *and the regularization term* $\mathcal{R}_{\mathrm{VB}}$ *defined by* $\mathcal{R}_{\mathrm{VB}}(\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}) \triangleq \min_{\boldsymbol{\Theta}} \sum_k \mathrm{tr}\Big[\boldsymbol{\Lambda}_k^{-1} \hat{\mathbf{Z}}_k \hat{\mathbf{Z}}_k^\top\Big] + \mathrm{tr}\Big[\boldsymbol{\Xi}^{-1} \hat{\mathbf{A}}^\top \hat{\mathbf{A}}\Big] + \sum_n \log|\mathbf{U}_n| + L \sum_k \log|\mathbf{V}_k|$, *where* $\mathbf{U}_n \triangleq \sum_k \psi_{nk}^{-1} \boldsymbol{\Xi} \hat{\mathbf{Z}}_k \hat{\mathbf{Z}}_k^\top + \mathbf{I}$ *and* $\mathbf{V}_k \triangleq \boldsymbol{\Lambda}_k \hat{\mathbf{A}}^\top \boldsymbol{\Psi}_k^{-1} \hat{\mathbf{A}} + \mathbf{I}$.

Proof of Theorem 6 is provided in Appendix G. The $\mathcal{R}_{\mathrm{VB}}$ has the following desirable properties:

- Given $\hat{\boldsymbol{\mathcal{Z}}}$, $\mathcal{R}_{\mathrm{VB}}$ is a concave, non-decreasing function of $\Big[\big\|\hat{\mathbf{A}}_{\cdot,1}\big\|_2, \cdots, \big\|\hat{\mathbf{A}}_{\cdot,M}\big\|_2\Big]$, a hallmark of sparsity-inducing regularizer [53]. Likewise, given $\hat{\mathbf{A}}$, $\mathcal{R}_{\mathrm{VB}}$ is a concave, non-decreasing function of $\Big[\big\|\hat{\mathbf{Z}}_{1,\cdot,k}\big\|_2, \cdots, \big\|\hat{\mathbf{Z}}_{M,\cdot,k}\big\|_2\Big]$.
- For either $\hat{\mathbf{A}}$ or $\hat{\mathbf{Z}}_k$, $\mathcal{R}_{\mathrm{VB}}$ imposes stronger sparsity than the group norm $\|\hat{\mathbf{A}}\| \triangleq \sum_m \big\|\hat{\mathbf{A}}_{\cdot,m}\big\|_2$ or $\|\hat{\mathbf{Z}}_k\| \triangleq \sum_m \big\|\hat{\mathbf{Z}}_{m,\cdot,k}\big\|_2$, while producing far less local minima than the regularizers associated with (15) or (8). Therefore, VB

inference (21) is less susceptible to local minima than using a conventional numerical algorithm, e.g., coordinate descent, to seek the locally optimal MAP solution to model (14).

• The sparse priors associated with $\mathcal{R}_{\mathrm{VB}}$ for $\hat{\mathbf{A}}$ and $\hat{\mathbf{Z}}_k$ are mutually coupled. Such interdependency guarantees that the optima are not affected by the scaling indeterminacy between $\hat{\mathbf{A}}$ and $\hat{\mathbf{Z}}_k$ [51].

It is also suggested by (21) that $\boldsymbol{\Psi}_k$ play the role of regularization parameters that balance between $\mathcal{L}$ and $\mathcal{R}_{\mathrm{VB}}$. In VB-CSP, $\boldsymbol{\Psi}_k$ are optimized via a variational learning rule analogous to those for $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Xi}$.

In contrast to the sparse CSP algorithm (see (3)), which aims to sparsify each individual spatial filter by forcing many coefficients to zero, the sparsity in our VB-CSP algorithm is targeted at the group level in the component space, i.e., using as few components as possible to represent the multichannel EEG signal; the redundant components are automatically zeroed out within the Bayesian framework. Moreover, VB-CSP allows us to inspect spatio-temporal patterns in the component space, while it is generally unclear how to relate the spatial filters optimized by the sparse CSP to the components' spatial patterns (see [54] for a discussion of the difference between spatial filters and spatial patterns). Finally, it shall be cautioned that the estimate from VB inference may be biased, and its variance is often underestimated (due to the mean-field approximation) [55].

4.3 Spatial Filtering via MAP-CSP/VB-CSP

Similar to CSP, MAP-CSP and VB-CSP use label information to guide learning; thus, in classification tasks, they can be used to generate discriminative features as the inputs to the classifier by discarding inessential components, resulting in enhanced predictive accuracies.

For MAP-CSP, since $\boldsymbol{\Psi}_k$ are isotropic, the following linear transformation matrix $\mathbf{W}$ can be estimated from the training set of EEG data to map from the EEG space to the component space:

$$\mathbf{W} = \left[\hat{\mathbf{A}}^\top \hat{\mathbf{A}}\right]^{-1} \hat{\mathbf{A}}^\top. \tag{22}$$

Each row of $\mathbf{W}$ defines a spatial filter. Likewise, the following pair of linear transformation matrices can be estimated for VB-CSP (according to the third step in Algorithm 2):

$$\mathbf{W}_k = \left[\hat{\mathbf{A}}^\top \boldsymbol{\Psi}_k^{-1} \hat{\mathbf{A}} + \sum_n \psi_{nk}^{-1} \boldsymbol{\Sigma}_{\mathbf{A}_{n,\cdot}} + \boldsymbol{\Lambda}_k^{-1}\right]^{-1} \hat{\mathbf{A}}^\top \boldsymbol{\Psi}_k^{-1}. \tag{23}$$

The discriminative filters can be selected via similar measures as conventionally employed for selecting CSP filters, e.g., filters associated with large component variance ratios between conditions. One can also use other well-established feature selection criteria [19] to select discriminative filters.

Although both MAP-CSP and VB-CSP can be employed in single-trial EEG classification, MAP-CSP is more suited for real-time applications, such as BCI decoding, due to its low computational overhead (Table 1), whereas VB-CSP is more suited for off-line exploratory data analysis with the advantage of automatic model size determination.

5 Experiments

In this section, we test the performance of the proposed algorithms using both synthetic and experimental EEG data. In the synthetic experiment where the ground truth is known, we compare CSP, MAP-CSP, and VB-CSP in terms of the recovery accuracy of the spatio-temporal patterns via Monte Carlo simulations. We also assess VB-CSP's capability in model selection and sensitivity to hyperprior selection and algorithmic initialization. In the analysis of high-density EEG data, we demonstrate the effectiveness of MAP-CSP and VB-CSP in the single-trial classification of several motor imagery EEG data sets. In addition, we apply VB-CSP as an exploratory tool to analyze spatio-temporal EEG patterns in a Stroop task [56]. In all experiments, we set $\alpha = \beta = 10^{-8}$ unless otherwise specified, and the same tolerance $\eta = 10^{-8}$ is used for determining the convergence of MAP-CSP and VB-CSP.

5.1 Synthetic Experiment

5.1.1 Description

The experiment consists of 50 independent Monte Carlo runs. In each run, $N = 40$ channels of synthetic EEG signals $\mathbf{X}$ are randomly generated according to model (6) for two experimental conditions:

1) Two sets of $M = 6$ mutually uncorrelated component signals $\mathbf{Z}_k$ are generated, with each set corresponding to a single condition. Each component signal comprises L IID Gaussian samples. The variances of the 6 component signals summed over the two conditions are integers from 2 to 7.

2) A mixing matrix $\mathbf{A}$ of size 40×6 is randomly generated, with each entry having a standard Gaussian distribution.

3) For each condition, additive white Gaussian noise with non-isotropic covariance is simulated to produce different levels of SNR. The SNR is defined per condition and channel as the ratio of the variance of the noiseless EEG and the variance of the additive noise in the same channel. The experiment is run repeatedly under a variety of settings: $L \in \{500, 100\}$ and $\mathsf{SNR} \in \{10, 5, 0, -5\}$ dBs.

The noisy multichannel EEG signals are fed into CSP, MAP-CSP, and VB-CSP, from which we obtain their estimated spatio-temporal patterns $\{\hat{\mathbf{A}}, \hat{\boldsymbol{\mathcal{Z}}}\}$. The component number is assumed to be known for MAP-CSP.

First, we evaluate the ability of VB-CSP to uncover the number of underlying components (MAP-CSP is excluded for evaluation since the component number must be specified beforehand rather than being estimated). The procedure for determining the effective component number M_e is described as follows. For each component m, we divide $\hat{\mathbf{Z}}_{m,\cdot,k}$ by the scaling coefficient $s_m = \left[\left\|\hat{\mathbf{Z}}_{m,\cdot,1}\right\|_2^2 + \left\|\hat{\mathbf{Z}}_{m,\cdot,2}\right\|_2^2\right]^{1/2}$ such that their l_2-norms sum to one. We then multiply $\hat{\mathbf{A}}_{\cdot,m}$ with s_m such that $\hat{\mathbf{A}}_{\cdot,m} \cdot \hat{\mathbf{Z}}_{m,\cdot,k}$ remains unchanged. With the scaling applied on each component, M_e is set to represent the number of components with $\left\|\hat{\mathbf{A}}_{\cdot,m}\right\|_2$ larger than a given threshold τ (we use $\tau = 10^{-6}$).

Second, to assess the estimated model, the Amari index [57] is used to quantify the reconstruction fidelity of the components' spatio-temporal patterns:

$$D = (D_{\mathbf{A}} + D_{\mathbf{Z}_1} + D_{\mathbf{Z}_2})/3, \tag{24}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

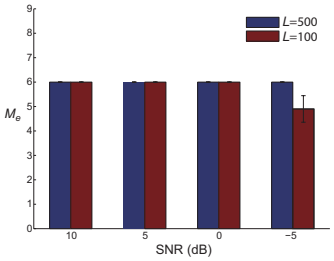

Fig. 2. Effective component number M_e determined by VB-CSP under different settings of L and SNR.

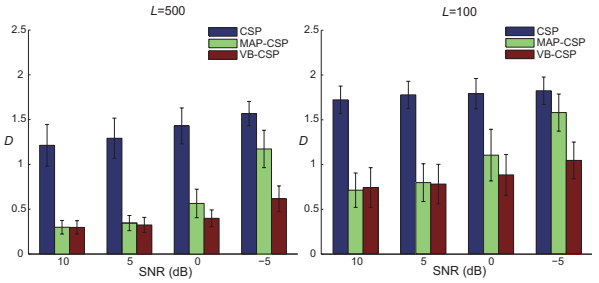

Fig. 3. The average Amari indices obtained under varying SNRs and sample sizes. left panel: $L = 500$; right panel: $L = 100$.

where

$$D_{\mathbf{A}} \triangleq \frac{1}{2M}\left[\sum_{i=1}^{M}\frac{\sum_{j=1}^{M}|b_{ij}|}{\max_j |b_{ij}|}+\sum_{j=1}^{M}\frac{\sum_{i=1}^{M}|b_{ij}|}{\max_i |b_{ij}|}-2M\right]$$

$$D_{\mathbf{Z}_k} \triangleq \frac{1}{2M}\left[\sum_{i=1}^{M}\frac{\sum_{j=1}^{M}|h_{ijk}|}{\max_j |h_{ijk}|}+\sum_{j=1}^{M}\frac{\sum_{i=1}^{M}|h_{ijk}|}{\max_i |h_{ijk}|}-2M\right]$$

$$b_{ij} \triangleq \left[(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\hat{\mathbf{A}}\right]_{ij}, \quad h_{ijk} \triangleq \left[(\mathbf{Z}_k^\top\mathbf{Z}_k)^{-1}\mathbf{Z}_k^\top\hat{\mathbf{Z}}_k\right]_{ij}.$$

As a prerequisite for computing the Amari index, the sizes of $\hat{\mathbf{A}}$ and $\hat{\mathbf{Z}}_k$ must be identical with those of $\mathbf{A}$ and $\mathbf{Z}_k$. To meet this requirement, for CSP and VB-CSP we select 6 components associated with the 6 largest $\left\|\hat{\mathbf{A}}_{\cdot,m}\right\|_2$ and discard the others. In the case of VB-CSP, in certain runs with $L = 100$ and $\mathsf{SNR} = -5$ dB the effective component number is less than 6, leaving the calculation of the Amari index problematic. Because the likelihood of such events is relatively low ($< 10\%$), we discard these runs when computing the statistics for simplicity.

5.1.2 Results

Figure 2 presents the results of component number estimation using VB-CSP. For $L = 500$ and $L = 100$, M_e is correctly identified to be 6 when $\mathsf{SNR} = 10, 5, 0, -5$ dBs and $\mathsf{SNR} = 10, 5, 0$ dBs, respectively. For $L = 100$ and $\mathsf{SNR} = -5$ dB, M_e deviates slightly from 6.

The Amari indices computed from CSP, MAP-CSP, and VB-CSP are displayed in Fig. 3. Two observations are in order. First, the benefit of using sparse learning is evident. As a general trend, MAP-CSP and VB-CSP substantially outperform CSP under all settings (three-way repeated-measure analysis of variance (ANOVA), with the algorithm, L, and SNR as the factors, indicates a significant main effect for the algorithm factor: $F(1, 49) = 1,793.172, P < 10^{-8}$ for VB-CSP vs. CSP; $F(1, 49) = 1,611.934, P < 10^{-8}$ for MAP-CSP vs. CSP). For each specific L and SNR, the decrease of the Amari indices for VB-CSP is greater than two-fold compared to those for CSP. Second, MAP-CSP is comparable to VB-CSP in performance, at relatively high SNRs (10 and 5 dBs), which is not unexpected because MAP-CSP finds the globally optimal MAP solution (up to the unknown and non-isotropic noise covariances). However, the performance gap between these two algorithms increases at low SNRs (0 and -5 dBs) since the increasing non-isotropic effect of the additive noise is not captured by MAP-CSP.

To provide an intuitive example, Fig. 4 presents the Hinton diagrams of $\mathbf{A}$ and $\hat{\mathbf{A}}$ from a simulated result with $L = 500$ and $\mathsf{SNR} = 0$ dB. In this specific run, we obtain $D = 1.3332, 0.5094$, and 0.3239 for CSP, MAP-CSP, and VB-CSP, respectively. Here, redundant spatio-temporal patterns

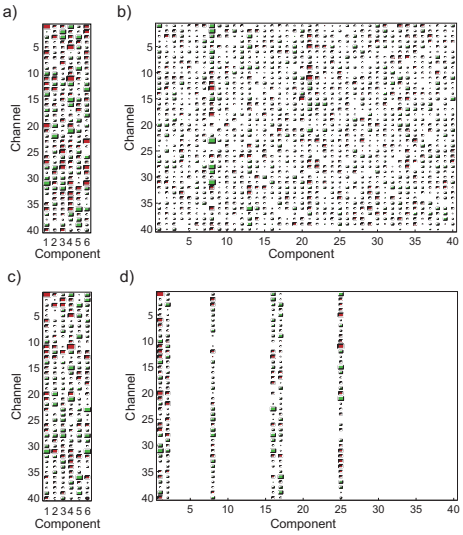

Fig. 4. Hinton diagrams of $\mathbf{A}$ and $\hat{\mathbf{A}}$ estimated by different algorithms for an exemplary run with $L = 500$ and $\mathsf{SNR} = 0$ dB (columns are in random order). The magnitude of each entry in the matrices is proportional to the square size (red: positive, green: negative). a) Non-square mixing matrix $\mathbf{A}$, b) Estimated square matrix $\hat{\mathbf{A}}$ from CSP ($D = 1.3332$), c) Estimated non-square matrix $\hat{\mathbf{A}}$ from MAP-CSP ($D = 0.5094$), d) Estimated sparse matrix $\hat{\mathbf{A}}$ from VB-CSP ($D = 0.3239$).

are shrunk to negligible values in VB-CSP. By contrast, it is hard to tell which columns are the redundant patterns in $\hat{\mathbf{A}}$ obtained from CSP, confirming that CSP is insufficient for component number determination.

Sensitivity analysis We conduct Monte Carlo simulations to assess the sensitivity of VB-CSP ($L = 500$ and $\mathsf{SNR} = 0$ dB). Specifically, sensitivity to hyperprior selection is studied by sampling α and β uniformly from $(0, 10^{-3})$, whereas sensitivity to algorithmic initialization is studied by sampling the elements of $\hat{\mathbf{A}}$ from IID standard Gaussian distributions, and the diagonal elements of $\boldsymbol{\Xi}$ uniformly from $(0, 1)$. VB-CSP inference is then performed over 100 repetitions for each Monte Carlo simulation, yielding $D = 0.2899 \pm 0.0091$ for hyperprior selection and $D = 0.3026 \pm 0.0192$ for initialization. These results demonstrate that the performance of VB-CSP is only slightly affected by hyperprior selection and initialization.

Computational speed Table 1 provides the runtime of the three algorithms in one representative Monte Carlo run for two setups: $L = 500$, $\mathsf{SNR} = 10$ dB and $L = 100$, $\mathsf{SNR} =$

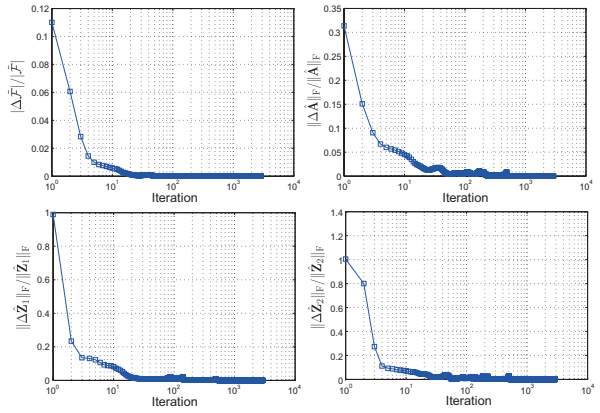

Fig. 5. Examples of convergence curves of $\tilde{\mathcal{F}}$, $\hat{\mathbf{A}}$, and $\hat{\mathbf{Z}}_k$ at $L = 100$.

10 dB. The algorithms are implemented in MATLAB©7.10 on a PC with 3.4 GHz Intel Core (TM) i7-3770 CPU and 8 GB RAM. The numbers of iterations needed to reach convergence for MAP-CSP and VB-CSP are shown in parentheses. In this example, MAP-CSP converges in far fewer iterations than VB-CSP. Note that in the course of VB-CSP, we monitor not only the convergence of $\tilde{\mathcal{F}}$, but also the convergence of the variational means of the model parameters. Results indicate that the variational means converge in a regular manner as $\tilde{\mathcal{F}}$ decreases (see Fig. 5 for the convergence curves at $L = 100$). In terms of the runtime, MAP-CSP is only 10 times slower than CSP but more than 10^3 times faster than VB-CSP, making it highly appealing for online applications, such as BCI decoding.

TABLE 1
Comparison of the runtime (second) from CSP, MAP-CSP and VB-CSP. The numbers in parentheses indicate the number of iterations to reach convergence.

L	CSP	MAP-CSP	VB-CSP
500	1.61×10^{-3}	2.17×10^{-2} (4)	1.20×10^{2} (3693)
100	1.40×10^{-3}	2.06×10^{-2} (5)	8.58×10 (2909)

5.2 Single-Trial Classification of Motor Imagery EEG Data

5.2.1 Description

A single-trial binary EEG classification experiment is conducted on three motor imagery (MI) EEG data sets. It has been well-documented in literature that imagined movements give rise to an attenuation of the sensorimotor rhythms in specific regions of the sensorimotor cortices, a phenomenon known as event-related desynchronization (ERD) [1] (e.g., imagined left or right hand movements generate ERD over hand regions in the contralateral motor cortices). The fact that ERD can be examined by evaluating the variance change of EEG spatial patterns across conditions provides good justifications to apply CSP and the related algorithms for MI EEG data analysis.

Among the three data sets, two were from BCI Competition III Data Set IIIa and Data Set IVa [2]. The third data set was collected in the Laboratory of Neural Engineering at Tsinghua University. Data set 1 consists of 60-channel EEG data from 3 subjects recorded for the left-hand, right-hand, foot, and tongue MI tasks (sampling rate: 250 Hz). There are 90, 60, and 60 trials per task for subjects $k3$, $l1$, and $k6$, respectively, with equal number of training and test trials. Data set 2 consists of 118-channel EEG data from 5 subjects recorded for the right-hand and right-foot MI tasks (sampling rate: 100 Hz). A total of 140 trials per task were collected for each subject, with varying percentages of training and test trials (168, 224, 84, 56, and 28 training trials for subject *aa*, *al*, *av*, *aw*, and *ay*, respectively). Data set 3 consists of 32-channel EEG data from 20 subjects for the left- and right-hand MI tasks (sampling rate: 256 Hz). For each subject, a total of 240 trials (120 per task) were split into equal number of training and test trials.

Since data set 1 contains the EEG data recorded under multi-class MI tasks, we construct smaller data sets for each possible pair of MIs for the purpose of binary classification, resulting in 6 data sets for left-hand vs. righthand, left-hand vs. foot, left-hand vs. tongue, right-hand vs. foot, right-hand vs. tongue, and foot vs. tongue MI tasks, respectively.

We compare the classification performance of CSP, CSP with Tikhonov regularization (TR-CSP, with $\mathbf{H} = \mathbf{I}$ in (2)) [15], MAP-CSP, and VB-CSP on a total of $18 + 5 + 20 = 43$ subsets of multichannel EEG signals as described above. TR-CSP is chosen from the existing regularized CSP algorithms as a benchmark due to its excellent classification performance, as demonstrated earlier [15]. Sparse CSP is not included for comparison for three reasons: 1) The major use of sparse CSP is channel selection, which is not the concern of this paper; 2) According to the results reported in [16], [17] on experimental EEG data, in general, sparse CSP yielded a degraded performance compared with CSP using a full set of channels; 3) The deflation procedure proposed in [16] for optimizing multiple spatial filters is not theoretically well-grounded since it does not preserve the positive semi-definiteness of the data covariance matrices when applied to a sparse spatial filter [58]. Addressing this issue is beyond the scope of this paper.

To avoid any potential bias towards any algorithm, we apply identical preprocessing settings to the data for channel selection (all EEG channels are used), bandpass filtering (8-30 Hz bandpass filtered using a 5th order Chebyshev Type-I filter. The frequency range is known to encompass the ERD effect [1]), and time windowing (0.5-3.5 sec rectangular window relative to the initiation of the MI tasks) before the use of each algorithm. To form the proper input to each algorithm, the training data for the k-th condition is concatenated across trials along the time axis to yield the data matrices $\mathbf{X}_k$ for each subset.

As suggested in [3], [15], the feature vector of each trial is formed as the log-variances of the estimated component signals obtained by the 6 spatial filters (see Section 4.3) associated with the 3 largest and 3 smallest variance ratios between the first and second experimental tasks. The 3 largest/smallest variance ratios correspond to large response in the first/second task. *Fisher linear discriminant analysis* (FLDA) is employed as a classifier due to its computational efficiency. The use of log-variance features helps FLDA to attain optimality due to the Gaussian-like distribution. The hyperparameters are determined using 10-fold CV on the training sets. For MAP-CSP, the component number M is sought within $\{10, 15, 20, \cdots, 60\}$ for data

2. Downloadable at http://www.bbci.de/competition/iii/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

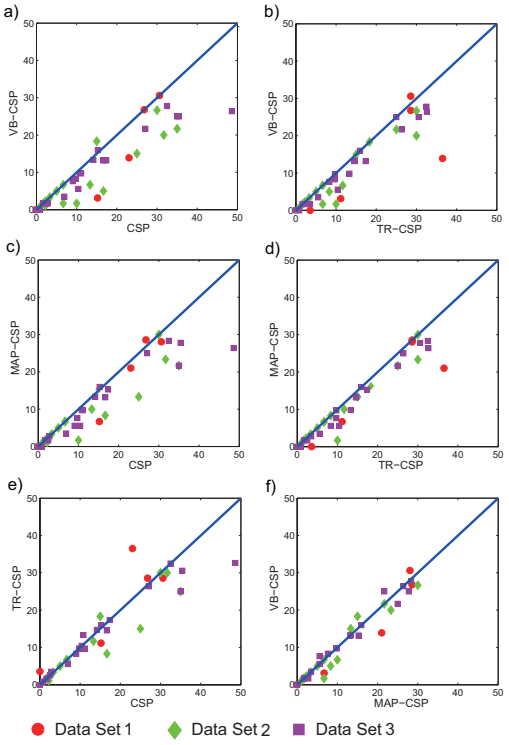

Fig. 6. Test errors (%) on three motor imagery BCI data sets (43 subsets) for four tested algorithms. Each point provides the result on one subset of EEG data. a) VB-CSP vs. CSP, b) VB-CSP vs. TR-CSP, c) MAP-CSP vs. CSP, d) MAP-CSP vs. TR-CSP, e) TR-CSP vs. CSP, f) VB-CSP vs. MAP-CSP

set 1 and 3, and $\{10, 20, \cdots, 110, 118\}$ for data set 2. For TR-CSP, the regularization parameter ρ is sought within $\{10^{-10}, 10^{-9}, \cdots, 10^{-1}\}$ as suggested in [15].

5.2.2 Results

Test errors for all 43 subsets of EEG data are summarized in Fig. 6. A point beneath the diagonal indicates the superiority of the algorithm on the y-axis over the one on the x-axis. The results indicate that VB-CSP and MAP-CSP have a superior or equal performance compared to CSP and TR-CSP for the majority of subsets. The exceptions are that CSP is slightly superior to VB-CSP and MAP-CSP on two and three subsets, respectively, and that TR-CSP is slightly superior to VB-CSP on one subset. The overall average test errors for VB-CSP, MAP-CSP, TR-CSP, and CSP are 10.36, 11.03, 13.07, and 14.22, respectively. Paired T-tests indicate that VB-CSP and MAP-CSP significantly outperform CSP and TR-CSP (VB-CSP vs. CSP: $P = 1.61 \times 10^{-5}$, VB-CSP vs. TRCSP: $P = 9.96 \times 10^{-5}$; MAP-CSP vs. CSP: $P = 7.54 \times 10^{-5}$, MAP-CSP vs. TR-CSP: $P = 8.73 \times 10^{-5}$), but TR-CSP does not significantly improve CSP ($P = 0.10$). There is no significant difference between the test errors from VB-CSP and MAP-CSP ($P = 0.0532$). Due to the markedly lower computational load of MAP-CSP compared with VB-CSP, the use of MAP-CSP in single-trial classification tasks is encouraged.

To facilitate reproducibility, the test errors for the publicly available data sets 1 and 2 are reported in Table 2, with the lowest test error highlighted in boldface for each subject. The values of the hyperparameters (ρ and M are determined by 10-fold CVs on the training sets; see Section 5.1.1 for calculating M_e) listed in the parentheses for TR-CSP, MAP-CSP, and VB-CSP. The results again indicate that VB-CSP and MAP-CSP have the overall best performance in that they yield the lowest test errors for all subjects. Furthermore, M and M_e are much lower than the channel number for a substantial portion of the subsets, suggesting that it is beneficial to use a sparse model to fit the higly redundant multichannel EEG data. M and M_e differ considerably for some subsets since they are estimated according to different model selection criteria. M is determined based on 10-fold CV classification errors, whereas M_e is the number of remaining components when a full Bayesian approach is applied to learning sparse models.

For subject $k3$, there is little room for improvement due to the low baseline test error as achieved by CSP. The improvement for subjects $k6$ (left-hand vs. right-hand, left-hand vs. foot, left-hand vs. tongue, right-hand vs. foot) and aw are the most prominent, with more than 8% decline in the test errors from VB-CSP and MAP-CSP compared with CSP. For subject aw the training set consists of only 56 trials, which CSP has a tendency to overfit. In contrast, VB-CSP and MAP-CSP alleviate the overfitting by using only 53 and 50 components, respectively, to characterize the 118-channel EEG data.

For data set 2, it is noteworthy that the difficulty level varies differently among subjects for classifying various combinations of MI tasks. For example, whereas it is easy to discriminate the foot from the tongue MIs for subjects $k3$ and $k6$, the classification performance deteriorates for subject $l1$. By contrast, the left- and right-hand MIs can be reliably discriminated for subjects $k3$ and $l1$ but not for subject $k6$. This result suggests that it is worthwhile to determine the optimal sets of MI tasks for subject-specific BCI systems.

The results of CSP on data set 2 differ substantially from the winning entries of the competition. Through personal communications with the winner, we are aware of several factors that may contribute to the superiority of the winning entries: 1) Intensive manual tuning was performed previously to obtain the optimal preprocessing settings. Relevant parameters included the EEG channels to be used, passband of the spectral filter, and time window, et al.; 2) Apart from ERD/ERS, two additional features, namely the autoregressive coefficients and temporal waves of the readiness potential, were employed previously for classification; 3) a semi-supervised tactic was applied to use part of the test data in the training stage. Our current paper focuses on the CSP algorithm, and it is beyond our scope to investigate the effect of the factors mentioned above. We stress that our comparison of the tested algorithms is fair, since except for spatial filtering, all of the other settings are identical for the tested algorithms.

5.3 Analysis of Stroop EEG Data

5.3.1 Description

Next, we consider a neurophysiologically-driven example and illustrate how VB-CSP can be used for exploratory EEG analysis in this context. The data set contains EEG recordings from two male subjects ($s1$ and $s2$) participating in a *Stroop color naming* task, in which they were instructed to name the colors of Chinese characters. There were two

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2
Comparison of the test errors (%) from four algorithms on two publicly available data sets from BCI Competition III.

Data Set	Subject	CSP	TR-CSP	MAP-CSP	VB-CSP
Data set IIIa left-hand vs. right-hand	$k3$	**3.33**	**3.33** ($\rho = 10^{-2}$)	**3.33** ($M = 30$)	**3.33** ($M_e = 34$)
	$k6$	31.67	30.00 ($\rho = 10^{-2}$)	23.33 ($M = 20$)	**20.00** ($M_e = 20$)
	$l1$	**6.67**	**6.67** ($\rho = 10^{-10}$)	**6.67** ($M = 60$)	**6.67** ($M_e = 35$)
Data set IIIa left-hand vs. foot	$k3$	**1.11**	**1.11** ($\rho = 10^{-10}$)	**1.11** ($M = 60$)	**1.11** ($M_e = 30$)
	$k6$	25.00	15.00 ($\rho = 10^{-2}$)	**13.33** ($M = 30$)	15.00 ($M_e = 15$)
	$l1$	16.67	8.33 ($\rho = 10^{-1}$)	8.33 ($M = 40$)	**5.00** ($M_e = 29$)
Data set IIIa left-hand vs. tongue	$k3$	**1.11**	**1.11** ($\rho = 10^{-2}$)	**1.11** ($M = 35$)	**1.11** ($M_e = 33$)
	$k6$	10.00	10.00 ($\rho = 10^{-4}$)	**1.67** ($M = 25$)	**1.67** ($M_e = 26$)
	$l1$	6.67	6.67 ($\rho = 10^{-10}$)	6.67 ($M = 55$)	**1.67** ($M_e = 30$)
Data set IIIa right-hand vs. foot	$k3$	**1.11**	**1.11** ($\rho = 10^{-10}$)	**1.11** ($M = 15$)	**1.11** ($M_e = 16$)
	$k6$	35.00	25.00 ($\rho = 10^{-1}$)	**21.67** ($M = 15$)	**21.67** ($M_e = 30$)
	$l1$	13.33	11.67 ($\rho = 10^{-4}$)	10.00 ($M = 25$)	**6.67** ($M_e = 26$)
Data set IIIa right-hand vs. tongue	$k3$	**1.11**	**1.11** ($\rho = 10^{-10}$)	**1.11** ($M = 55$)	**1.11** ($M_e = 30$)
	$k6$	**15.00**	18.33 ($\rho = 10^{-2}$)	**15.00** ($M = 45$)	18.33 ($M_e = 23$)
	$l1$	**5.00**	**5.00** ($\rho = 10^{-10}$)	**5.00** ($M = 60$)	**5.00** ($M_e = 26$)
Data set IIIa foot vs. tongue	$k3$	**2.22**	**2.22** ($\rho = 10^{-10}$)	**2.22** ($M = 60$)	**2.22** ($M_e = 33$)
	$k6$	**1.67**	**1.67** ($\rho = 10^{-2}$)	**1.67** ($M = 50$)	**1.67** ($M_e = 30$)
	$l1$	30.00	30.00 ($\rho = 1$)	30.00 ($M = 30$)	**26.67** ($M_e = 24$)
Data set IVa right-hand vs. right-foot	aa	**26.79**	28.57 ($\rho = 10^{-1}$)	28.57 ($M = 80$)	**26.79** ($M_e = 81$)
	al	**0**	3.57 ($\rho = 10^{-10}$)	**0** ($M = 30$)	**0** ($M_e = 32$)
	av	30.61	28.57 ($\rho = 10^{-10}$)	**28.06** ($M = 50$)	30.61 ($M_e = 50$)
	aw	15.18	11.16 ($\rho = 10^{-10}$)	6.70 ($M = 50$)	**3.12** ($M_e = 53$)
	ay	23.02	36.51 ($\rho = 10^{-10}$)	21.03 ($M = 20$)	**13.89** ($M_e = 19$)
Mean±SD		13.19 ± 11.85	12.47 ± 11.60	10.41 ± 10.18	$\mathbf{9.32 \pm 9.93}$

experimental conditions: *congruent* versus *incongruent*. In the congruent condition, the color and the meaning of the characters were consistent (e.g., the Chinese character for "red" in red), whereas the color and meaning differed in the incongruent condition. The experiment was comprised of 4 sessions. A total of 144 trials of 64-channel EEG data were collected per condition in each session. Each trial lasted for 1 sec. Signals were down-sampled to 200 Hz offline. For preprocessing, the EEG signals were band-pass filtered between 1 and 40 Hz. The filtered signals were then temporally concatenated across trials to feed into VB-CSP.

5.3.2 Results

The Hinton diagrams of $\hat{\mathbf{A}}$ obtained from VB-CSP are shown in the upper panel of Fig. 7. Among the 60 possible components, only 11 and 12 are retained by VB-CSP for the two subjects, respectively. By visually inspecting the spatio-temporal patterns of the retained components for each subject, we are able to identify two components that are neurophysiologically meaningful. The respective spatio-temporal patterns are shown in Fig. 8. For each subject, the left panel presents the *event-related potentials* (ERPs) of the two components. For each component, the ERP is calculated by averaging the time courses of the entire 72 trials. As indicated by the shaded bars, the incongruent condition elicits stronger negative potentials than the congruent condition within the time intervals 400-600 msec and 700-900 msec. The differences are significant as can be observed from the 95% *posterior credible intervals* [19] (light-colored curves) derived from the variational posterior distributions. The corresponding spatial patterns for 400-600 msec and 700-900 msec are concentrated on the fronto-central and fronto-polar scalp regions, respectively.

The results are consistent with the previous findings reported in [56]. Through the source localization of the ERP components in the brain, it was suggested that the enhanced negativity for the incongruent condition was likely to stem from activation of the anterior cingulate cortex (ACC), reflecting its role in the detection of interference between the character meaning and color (400-600 msec interval) and in the engagement of central executive processes (700-900 msec interval). It is important to notice that our current analysis is conducted on an individual subject basis, as opposed to previous instances of grand-averaging over multiple subjects [56]. For comparison, CSP is also applied to the same EEG data set. As expected, the estimated mixing matrices are not sparse (see the lower panel of Fig. 7), and no meaningful spatio-temporal patterns are observed.

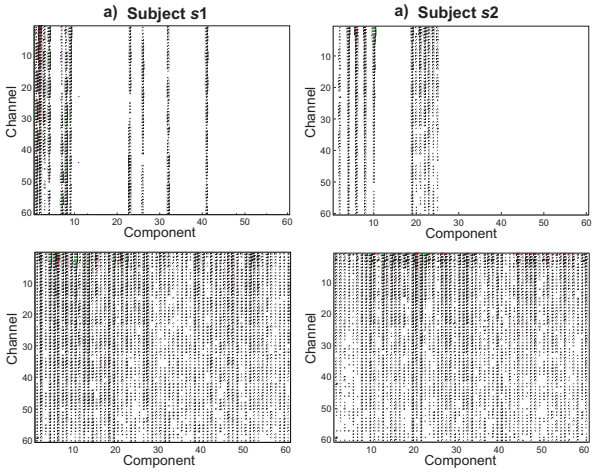

Fig. 7. Hinton diagrams of the estimated sparse matrix $\hat{\mathbf{A}}$ (columns are in random order). Upper panel: the variational mean estimates from VB-CSP; lower panel: the estimates from CSP. a) subject $s1$, b) subject $s2$

6 Conclusion

With the motivation of overcoming the CSP's overfitting problem, here we presented a Bayesian framework for modeling multichannel EEG signals from two experimental con-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2330598, IEEE Transactions on Pattern Analysis and Machine Intelligence

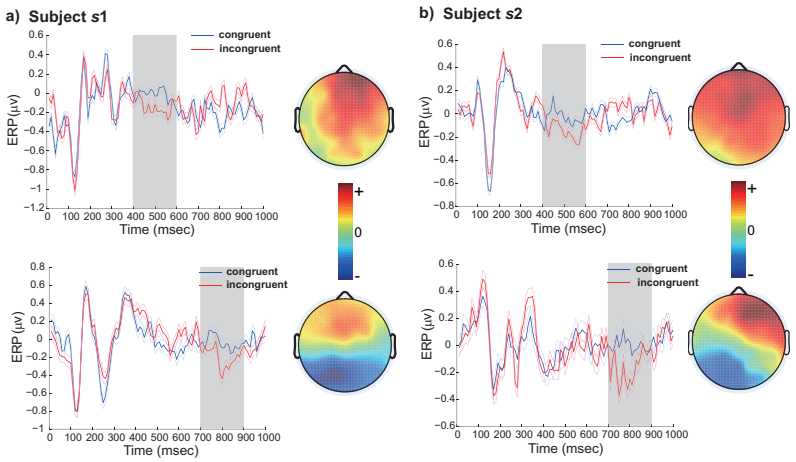

Fig. 8. The spatio-temporal patterns of selected two meaningful components. For each subject, the left panel presents the ERPs of the components (dark curves: variational means; light curves: 95% posterior credible intervals), and the right panel presents the spatial patterns of the components as scalp maps. a) subject $s1$, b) subject $s2$

ditions. The proposed framework encompasses the existing CSP and regularized CSP algorithms as special cases, which addresses overfitting in a principled manner by using the sparse Bayesian learning technique. Under this framework, we developed the MAP-CSP and VB-CSP algorithms for use in real-time single-trial EEG classification and exploratory EEG analysis, respectively. Their algorithmic efficacy and superiority were demonstrated by the successful analyses of synthetic and experimental EEG data sets.

Questions that remain to be addressed in the future include the following: 1) Although the models presented in this paper exploit the spatial structure of the multichannel EEG signals, temporal dynamics are not fully characterized by the IID assumption. More sophisticated time-series modeling techniques are required to account for the temporal dependency; 2) It may not always be viable to use a unimodal variational posterior as the proxy for a potentially multimodal posterior distribution. Thus it would be interesting to know the degree of multimodality of the true posterior distribution under model (14), and to make comparison with the unimodal VB solution. This can be empirically assessed using the Markov chain Monte Carlo (MCMC) strategy, which is capable of numerically representing the true posterior distribution, as opposed to the approximate nature of VB [59]; 3) Further effort is required to improve the slow convergence of VB-CSP.

ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities (2013ZM0076), Specialized Research Fund for the Doctoral Program of Higher Education of China (20130172120032), Guangdong Natural Science Foundation (S2013010013445), National High-tech R&D Program of China (863 Program) under grant 2012AA011601, the National Natural Science Foundation of China under grants 60825306 and 91120305, High Level Talent Project of Guangdong Province, and the US National Institutes of Health (NIH) under grants DP1-OD003646. The authors are grateful to Klaus-Robert Müller, Benjamin Blankertz, Gabriel Curio, Gert Pfurscheller, Alois Schlögl, and Wenjing Gao for providing the EEG data sets used in this paper. All correspondence should be directed to W. Wu.

REFERENCES

[1] E. Niedermeyer and F. L. D. Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, 5th ed. Lippincot Williams and Wilkins, 2004.

[2] S. Sanei and J. A. Chambers, *EEG Signal Processing*. Wiley-Interscience, 2007.

[3] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, pp. 41 – 56, 2008.

[4] F. Fukunaga and W. Koontz, "Applications of the Karhunen-Loève expansion to feature selection and ordering," *IEEE Trans. Comput.*, vol. 19, pp. 311 – 318, 1970.

[5] S. Zhang and T. Sim, "Discriminant subspace analysis: a Fukunaga-Koontz approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 1732 – 1745, 2007.

[6] X. Huo, "A statistical analysis of Fukunaga-Koontz transform," *IEEE Signal Process. Lett.*, vol. 11, pp. 123 – 126, 2004.

[7] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topo.*, vol. 2, pp. 275 – 284, 1990.

[8] B. Blankertz, K. R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, vol. 51, pp. 1044 – 1051, 2004.

[9] B. Blankertz, K. R. Müller, D. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, and N. Birbaumer, "The BCI competition III: validating alternative approaches to actual BCI problems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, pp. 153 – 159, 2006.

[10] M. Tangermann, K. R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI competition IV," *Front. Neurosci.*, vol. 6, 2012, doi: 10.3389/fnins.2012.00055.

[11] N. J. Hill, T. N. Lal, T. H. M. Schröder, G. Widman, G. E. Elger, B. Schölkopf, and N. Birbaumer, "Classifying event-related desynchronization in EEG, ECoG and MEG signals," in *Towards Brain-Computer Interfacing*, G. Dornhege, J. del R. Millán, T. Hinterbeger, D. McFarland, and K. R. Müller, Eds. MIT Press, 2007.

[12] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2274 – 2281, 2006.

[13] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K. R. Müller, "Invariant common spatial patterns: alleviating nonstationarity in brain-computer interfacing," in *Advances in Neural Information Processing Systems*, vol. 20, 2008, pp. 113 – 120.

[14] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Process. Lett.*, vol. 16, pp. 683 – 686, 2009.

[15] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, pp. 355 – 362, 2010.

[16] J. Farquhar, N. Hill, T. Lal, and B. Schölkopf, "Regularised CSP for sensor selection in BCI," in *3rd International BCI Workshop*, 2006.

[17] X. Yong, R. Ward, and G. Birch, "Sparse spatial filter optimization for EEG channel reduction in brain-computer interface," in *ICASSP*, 2008, pp. 417 – 420.

[18] I. Onaran, N. F. Ince, and A. E. Cetin, "Sparse spatial filter via a novel objective function minimization with smooth l_1 regularization," *Biomedical Signal Processing and Control*, vol. 8, pp. 282 – 288, 2013.

[19] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. The MIT Press, 2012.

[20] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improved classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541 – 1548, 2005.

[21] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, and K.-R. Müller, "Spectrally weighted common spatial pattern algorithm for single trial EEG classification," Dept. of Mathematical Engineering, The University of Tokyo, Technical Report 40, 2006.

[22] W. Wu, X. Gao, B. Hong, and S. Gao, "Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL)," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 6, pp. 1733 – 1743, 2008.

[23] Q. Zhao, L. Zhang, and A. Cichocki, "Multilinear generalization of common spatial patterns," in *ICASSP*, 2009, pp. 525 – 528.

[24] H. Zhang, C. Z. Yang, K. K. Ang, C. Guan, and C. Wang, "Optimum spatio-spectral filtering network for brain-computer interface," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 52 – 63, 2011.

[25] H. Higashi and T. Tanaka, "Simultaneous design of FIR filter banks and spatial patterns for EEG signal classification," *IEEE Trans. Biomed. Eng.*, vol. 60, pp. 1100 – 1110, 2013.

[26] H. Suk and S. Lee, "A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 286 – 299, 2013.

[27] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993 – 1002, 2004.

[28] W. Wu, X. Gao, and S. Gao, "One-versus-the-rest (OVR) algorithm: an extension of common spatial patterns (CSP) algorithm to multi-class case," in *Proc. 27th Int. Conf. IEEE-EMBS*, 2005, pp. 2387 – 2390.

[29] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991 – 2000, 2008.

[30] Y. Li and C. Guan, "An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces," *Neural Computation*, vol. 18, no. 11, pp. 2730 – 2761, 2006.

[31] W. Wu, Z. Chen, S. Gao, and E. N. Brown, "A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG," *NeuroImage*, vol. 56, pp. 1929 – 1945, 2011.

[32] W. Samek, C. Vidaurre, K. R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *Journal of Neural Engineering*, vol. 9, p. 026013, 2012.

[33] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface," *IEEE Trans. Neural Networks and Learning Systems*, vol. 24, pp. 610 – 619, 2013.

[34] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multi-subject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, vol. 2011, p. 217987, 2011.

[35] W. Wu, Z. Chen, S. Gao, and E. N. Brown, "A probabilistic framework for learning robust common spatial patterns," in *Proc. 31st Int. Conf. IEEE-EMBS*, 2009, pp. 4658 – 4661.

[36] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653 – 662, 2012.

[37] W. Samek, D. Blythe, K. R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems*, vol. 26, 2013, pp. 1007 – 1015.

[38] M. Kawanabe, W. Samek, K. R. Müller, and C. Vidaurre, "Robust common spatial filters with a maxmin approach," *Neural Computation*, vol. 26, no. 2, pp. 1 – 28, 2014.

[39] C. Gouy-Pailler, M. Congedo, C. Brunner, C. Jutten, and G. Pfurtscheller, "Nonstationary brain source separation for multiclass motor imagery," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 2, pp. 469 – 478, 2010.

[40] L. C. Parra and P. Sajda, "Blind source separation via generalized eigenvalue decomposition," *Journal of Machine Learning Research*, vol. 4, pp. 1261 – 1269, 2003.

[41] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of non-stationary sources," *IEEE Trans. Signal Process.*, vol. 49, pp. 1837 – 1848, 2001.

[42] B. J. Baars and N. M. Gage, *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*, 2nd ed. Academic Press, 2010.

[43] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[44] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial patterns with aggregation for EEG classification in small-sample setting," *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 2936 – 2946, 2010.

[45] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. Springer, 2007.

[46] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society*, vol. 36, pp. 99 – 102, 1974.

[47] A. Gelman, "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, vol. 1, pp. 515 – 533, 2006.

[48] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 880 – 887.

[49] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao, "Variational EM algorithms for non-Gaussian latent variable models," in *Advances in Neural Information Processing Systems*, vol. 18, 2006, pp. 1059 – 1066.

[50] M. W. Seeger and D. P. Wipf, "Variational bayesian inference techniques," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 81 – 91, 2010.

[51] D. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, pp. 6236 – 6255, 2011.

[52] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 317 – 329, 2010.

[53] K. Kreutz-Delgado and B. D. Rao, "A general approach to sparse basis selection: majorization, concavity, and affine scaling," Center for Information Engineering, UCSD, Technical Report UCSD-CIE-97-7-1, 1997.

[54] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K. R. Müller, "Single-trial analysis and classification of ERP components — a tutorial," *NeuroImage*, vol. 56, pp. 814 – 825, 2011.

[55] R. Turner and M. Sahani, "Two problems with variational expectation maximisation for time series models," in *Bayesian Time Series Models*. Cambridge Univ. Press, 2011.

[56] S. Hanslmayr, B. Pastötter, K. H. Bäuml, S. Gruber, M. Wimber, and W. Klimesch, "The electrophysiological dynamics of interference during the Stroop task," *Journal of Cognitive Neurosciences*, vol. 20, pp. 215 – 225, 2008.

[57] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*, vol. 8, 1996, pp. 757 – 763.

[58] L. Mackey, "Deflation methods for sparse PCA," in *Advances in Neural Information Processing Systems*, vol. 21, 2009, pp. 1017 – 1024.

[59] A. Nummenmaa, T. Auranen, M. S. Hämäläinen, I. P. Jääskeläinen, J. Lampinen, M. Sams, and A. Vehtari, "Hierarchical Bayesian estimates of distributed MEG sources: theoretical aspects and comparison of variational and MCMC methods," *NeuroImage*, vol. 35, pp. 669 – 685, 2007.

Wei Wu (S'05, M'12) received the Ph.D. degree in biomedical engineering in 2012 from Tsinghua University, China. From 2008 to 2010, he was a visiting student at the Neuroscience Statistics Laboratory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology. Since 2012, he has been with the School of Automation Science and Engineering, South China University of Technology, as an Associate Professor. He works in the field of neural signal processing and neural engineering, specializing in particular on developing statistical models and algorithms for the analysis and decoding of brain signals. He is an associate editor of Neurocomputing, and a member of IEEE Biomedical Signal Processing Technical Committee.

Zhe Chen (S'99, M'09, SM'10) received the Ph.D. degree in Electrical and Computer Engineering in 2005 from McMaster University, Canada. He joined RIKEN Brain Science Institute in June 2005 as a research scientist. From March 2007 he worked in Harvard Medical School and Massachusetts Institute of Technology, subsequently as a Harvard Research Fellow, Instructor, and Senior Research Scientist. Since April 2014, he became an Assistant Professor at the New York University School of Medicine, with joint appointment at the Department of Psychiatry and Department of Neuroscience and Physiology. His research interests include computational neuroscience, neural engineering, neural signal processing, machine learning, and Bayesian modeling. He is the lead author of the book "Correlative Learning" (Wiley, 2007). He is an Early Career Award recipient from the Mathematical Biosciences Institute. He is also a principal investigator for an NSF-CRCNS (Collaborative Research in Computational Neuroscience) grant.

Xiaorong Gao (M'04) received the B.S. degree in biomedical engineering from Zhejiang University, China, in 1986, the M.S. degree in biomedical engineering from Peking Union Medical College, China, in 1989, and the Ph.D. degree in biomedical engineering from Tsinghua University, China, in 1992. He is currently a Professor in the Department of Biomedical Engineering, Tsinghua University. His main research interest is biomedical signal processing.

Yuanqing Li received the B.S. degree in applied mathematics from Wuhan University, China, in 1988, the M.S. degree in applied mathematics from South China Normal University, China, in 1994, and the Ph.D. degree in control theory and applications from South China University of Technology, China, in 1997. Since 1997, he has been with South China University of Technology, where he became a Full Professor in 2004. From 2002 to 2004, he was a Researcher at the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan. From 2004 to 2008, he was a Research Scientist at the Laboratory for Neural Signal Processing, Institute for Inforcomm Research, Singapore. He is the author of more than 70 scientific papers in journals and conference proceedings. His research interests include blind signal processing, sparse representation, machine learning, BCI, and brain data analysis.

Emery N. Brown (M'01, SM'06, F'08) received the B.A. degree from Harvard College, the M.D. degree from Harvard Medical School, and the A.M. and Ph.D. degrees in statistics from Harvard University. He is the Edward Hood Taplin Professor of Medical Engineering in the Institute for Medical Engineering and Science and a professor of computational neuroscience in the Department of Brain and Cognitive Sciences at Massachusetts Institute of Technology; the Warren M. Zapol Professor of Anaesthesia at Harvard Medical School; and an Anesthesiologist in the Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital. His research interests include signal processing algorithms for the study of neural systems, functional neuroimaging, and electrophysiological approaches to study general anesthesia. Dr. Brown is a member of the Association of University of Anesthesiologists, a Fellow of the American Statistical Association, a Fellow of the American Association for the Advancement of Science, a member of the Institute of Medicine, and a Fellow of the American Academy of Arts and Sciences. He was a recipient of a 2007 NIH Directors Pioneer Award, the 2011 recipient of the National Institute of Statistical Sciences Sacks Award for Outstanding Cross-Disciplinary Research and a 2012 NIH Directors Transformative Research Award. Dr. Brown is a member of the NIH BRAIN Initiative Working Group.

Shangkai Gao (SM'94, F'07) graduated in 1970 from the Department of Electrical Engineering in Tsinghua University, China, where she also received the M.E. degree of biomedical engineering in 1982. She is now a professor at the Department of Biomedical Engineering in Tsinghua University. Her research interests include neural engineering and medical imaging, especially the study of brain-computer interface. She is also a fellow of American Institute for Medical and Biological Engineering. She is now the Editorial Board Member of IEEE Transactions on Biomedical Engineering, Journal of Neural Engineering and Physiological Measurement, as well as the senior editor of IEEE Transactions on Neural Systems and Rehabilitation Engineering.