Bayesian Learning for Spatial Filtering in an EEG-Based Brain–Computer Interface

Haihong Zhang, *Member, IEEE*, Huijuan Yang, *Member, IEEE*, and Cuntai Guan, *Senior Member, IEEE*

***Abstract*— Spatial filtering for EEG feature extraction and classification is an important tool in brain–computer interface. However, there is generally no established theory that links spatial filtering directly to Bayes classification error. To address this issue, this paper proposes and studies a Bayesian analysis theory for spatial filtering in relation to Bayes error. Following the maximum entropy principle, we introduce a gamma probability model for describing single-trial EEG power features. We then formulate and analyze the theoretical relationship between Bayes classification error and the so-called Rayleigh quotient, which is a function of spatial filters and basically measures the ratio in power features between two classes. This paper also reports our extensive study that examines the theory and its use in classification, using three publicly available EEG data sets and state-of-the-art spatial filtering techniques and various classifiers. Specifically, we validate the positive relationship between Bayes error and Rayleigh quotient in real EEG power features. Finally, we demonstrate that the Bayes error can be practically reduced by applying a new spatial filter with lower Rayleigh quotient.**

***Index Terms*— Bayes error, brain–computer interface, Rayleigh quotient, spatial filtering.**

I. Introduction

SINGLE-TRIAL electroencephalogram (EEG) feature extraction and classification [1] play a vital role in the emerging technology of brain–computer interface (BCI) [2]–[4]. BCI enables a user to interact with computers or machines just by means of brain activities. Thus, it is promising for restoration of lost communication and control functions in severely disabled people [5], [6], as well as in other potential applications [7], [8]. From the viewpoint of signal processing, BCI performs real-time detection of the EEG signals associated with special mental activities. This is often referred to as single-trial EEG classification [9] that makes decision on each trial (a single execution of a mental task). Recently there has been rapidly growing interest in related pattern recognition research (see comprehensive reviews in [1], [10]).

Computing discriminative features in single-trial EEG can be very challenging due to EEGs' poor specificity caused by volume conduction effects, brain nonstationarity, and various background noises [11]–[13]. In addition, the characteristics of EEG signals may vary significantly from person to person [14] and among different conditions [15]. In order to solve these problems, spatial filtering has been introduced to explore discriminative spatial characteristics of multichannel EEG signals [9], [16], [17], as long as the mental activities of interest show different spatial patterns. The basic principle is to transform EEG signals acquired from a large and equivocal array of sensors into a small set of components containing discriminative information about task-related brain activities. To present, one of the most popular and effective techniques is the common spatial pattern (CSP) method [9], [18], [19]. As a supervised method, it can produce better features than independent component analysis (ICA) which is unsupervised. This has been demonstrated in [20] using various ICA algorithms, including Infomax, FastICA, and SOBI.

CSP consists of a real-valued linear projection that transforms each EEG time sample into one or multiple specific vectors representing particular spatial patterns. Subsequently it calculates the average powers of each resultant signal in the trial, and the power values will represent the EEG trial in classification. The linear transformation of CSP is usually optimized for each individual subject to account for cross-subject variations, by minimizing the Rayleigh quotient [19] [see (21) and related descriptions]. The quotient can be viewed as the ratio of the average power feature of one class to that of the other, while its minimization can be casted as a generalized eigenvalue problem. It is recognized that CSP is especially useful in sensorimotor rhythm (SMR)-BCI [21] with motor-related mental tasks [19], [22]. Various extensions of CSP have also been proposed in recent years [23]–[28], with the Rayleigh quotient continuing to play an important role.

More recently, two interesting developments of CSP were reported in [28] and [29]. The former proposed a multiclass Bayes error bound estimate with a closed-form expression and then developed a learning algorithm for minimizing the bound estimate in multiclass data. The latter introduced a Bayesian generative model for representing SMR EEG, and devised a variational Bayesian method for learning. It has proved that CSP is a maximum-likelihood estimation of the generative model under certain conditions.

From Bayesian viewpoint, however, the direct relationship between power features and Bayes error of single-trial EEG classification has not yet been addressed rigorously. Existing works (e.g., [28]) often relate CSP to the solution for minimizing an upper bound of Bayes error, i.e., the Bhattacharyya bound [30], in classification of individual EEG time samples. However, that error bound may not be appropriate for representing the Bayes error because of the following reasons. The Bhattacharyya bound would match a tighter bound called the

Manuscript received March 6, 2012; revised December 16, 2012; accepted February 12, 2013. Date of publication March 21, 2013; date of current version May 14, 2013.

The authors are with the Institute for Infocomm Research, Agency for Science, Technology and Research, 138632 Singapore (e-mail: hhzhang@i2r.a-star.edu.sg; hjyang@i2r.a-star.edu.sg; ctguang@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2249087

Chernoff bound only if the two classes have equal-covariance matrices but unequal means (see [30]), whereas rhythmic EEG samples in sensorimotor BCI have zero mean and different covariance matrices between classes [18]. More importantly, the actual problem of concern is not classification of an individual time sample but that of a whole trial comprising a number of time samples.

The objective of this paper is to build a theory that links Bayes learning and spatial filtering for EEG classification. We introduce the maximum entropy principle to probabilistic modeling of EEG power features. We show that the features can be described by gamma distributions. Then we study the Bayes error for gamma random variables and prove that the error is monotonic over the Rayleigh quotient [19] under simple conditions. To validate the theory in practical EEG classification, our experimental study involves three independent EEG data sets consisting of 18 human subjects performing motor imagery BCI tasks. With the analysis of the EEG power features, we validate that the Bayes error is closely correlated with the Rayleigh quotient. We also investigate the possibility to further reduce Rayleigh quotient over CSP and the effect on classification accuracy. Various classifiers, including linear discriminant analysis, linear and nonlinear support vector machines are used in evaluating the classification accuracy.

The rest of this paper is organized as follows. Section II analyzes the Bayes error for EEG power features, and formulates the direct theoretic link between the Bayes error and the Rayleigh quotient. Section III describes the experimental results, followed by discussions in Section IV. Finally, Section V concludes this paper.

II. Bayesian Analysis of EEG Features and Spatial Filtering

A. Gamma Probability Models of Power Features

In this section, we aim to connect EEG power features to gamma probability distributions. To this end, let us consider the primary phenomena of SMR EEG, namely, event-related de-synchronization (ERD) and event-related synchronization (ERS) [9]. They are the attenuation or increase of the rhythmic activity over the sensorimotor cortex generally in the μ (8–13 Hz) (highly overlapping with the alpha band) and β (14–30 Hz) bands. ERD is relatively much more prominent, which has been demonstrated to be inducible by both imagined movements in healthy people or intended movements in paralyzed patients [31].

To address the low signal-to-noise ratio of EEG, spatial filtering can be employed to enhance the ERD power feature especially in multichannel EEG [18], [27], as long as the mental activities of interest show different spatial patterns. The power feature y of a spatially-filtered EEG segment is given by

$$y = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \|\mathbf{w}^T \mathbf{x}(t)\|^2 dt = \mathbf{w}^T \Sigma_x^{(t_1,t_2)} \mathbf{w} \qquad (1)$$

where $\mathbf{x}$ is a vector function of time representing the multichannel time series of EEG. Note that the EEG signal is band-pass filtered beforehand to pick up the ERD/ERS related rhythmic information. The two time variables t_1 and t_2 are the time points defining the time segment, and the vector $\mathbf{w}$ is the spatial filter. Each element in $\mathbf{w}$ is associated with a particular EEG channel. The operator $\|\cdot\|$ denotes the magnitude of a value. The term $\Sigma_x^{(t_1,t_2)}$ represents the correlation (equivalent to covariance if $\mathbf{x}$ is zero-meaned) matrix in the time frame.

Bayesian analysis of the ERD feature requires investigation of its probability distribution in real EEG data. Thus, we compute the feature from EEG samples in the data sets used in our experiment (see Section III-A), using the following method. We apply the widely used CSP algorithm [18] to spatially filter the EEG and then compute the energy feature according to (1). It is known that the principle of CSP is to find directions that maximize variance for one class and simultaneously minimize variance for the other class (we will come back to the principle of CSP at the end of this section).

Fig. 1 plots the histograms of the feature values. Here, we consider the first motor imagery class as the positive class. In each subject, we select the CSP filter for the smallest Rayleigh quotient (see [18] or [21]), and use it to generate the EEG power features. It can be seen that these features are valid for representing the ERD phenomenon, as they have smaller values (attenuated powers) in the positive class than in the negative class.

For probability modeling of these features, Gaussian distributions obviously are out of the question, because the actual distributions are skewed and left-bounded. So the important question is: what probabilistic models may accurately describe the underlying probability functions? It is worthwhile to note that this can be a difficult question because EEG is rather complex, nonstationary, and exhibits large variations across subjects. Hence, it is hard to appropriately build *a priori* knowledge into probabilistic modeling.

To address this issue, we introduce the principle of maximum entropy, which states that one should choose the probability distribution with the largest entropy by default. In other words, this principle aims to minimize the amount of prior information built into the distribution. Importantly, it also agrees with the fact that many physical systems tend to move toward maximal entropy configuration over time.

As shown in [32], under certain conditions, the principle of maximum entropy will lead to a solution described by gamma distributions [see Fig. 2(a) for examples]. In this paper, we generalize the solution by relieving the related condition (see Appendix A), and have the following theorem.

Theorem 1: For a probability distribution of nonnegative random variable that satisfies either of the following two constraints:

I. the expectation of the distributions is known to be m;
II. in addition to I, the expectation of the logarithm of the variable is known to be s.

The probability density $f_x(x)$, which has the maximum entropy is a gamma distribution. Particularly, in case of constraint I, the distribution takes a simple exponential form.

$$f_x(x) = \frac{1}{m} \exp\left(-\frac{x}{m}\right). \qquad (2)$$

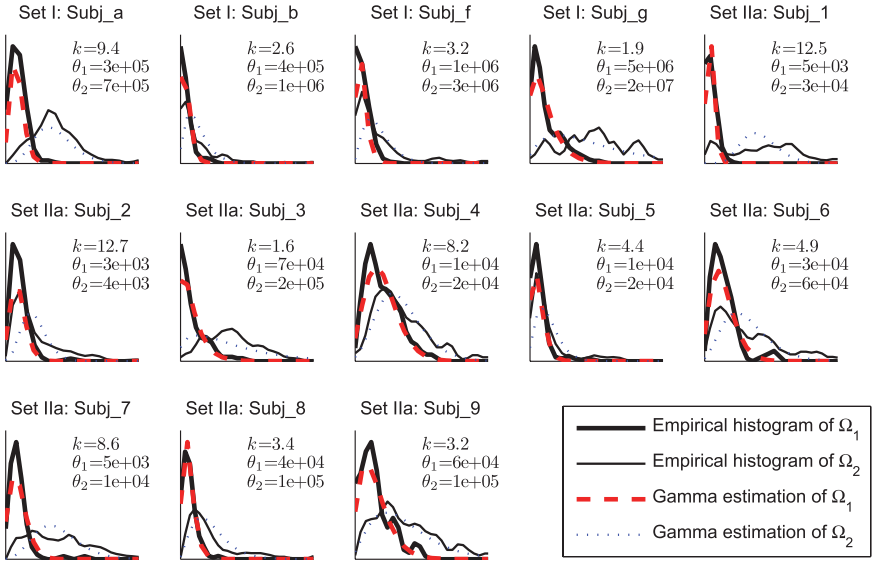

Fig. 1. Empirical histogram and gamma approximations of EEG power feature distributions for each subject in BCI Competition IV Set I and IIa. Each sub-figure plots the histogram of EEG power features [see (1)] together with the gamma approximations. Ω_1 or Ω_2 denotes the positive or the negative class. The horizontal axis indicates the EEG power value, while the vertical axis is related to the estimated probability density. Here, we consider equal shape parameter k between two classes and show the estimated k value as well as the class-specific rate parameter estimate θ_1 and θ_2 (respectively, for Ω_1 and Ω_2). See Section II-A for detailed descriptions. Note that the shape of the distribution is of interest but not the scale of the distributions, so the axis ticks are omitted to improve the clarity of the graphs.

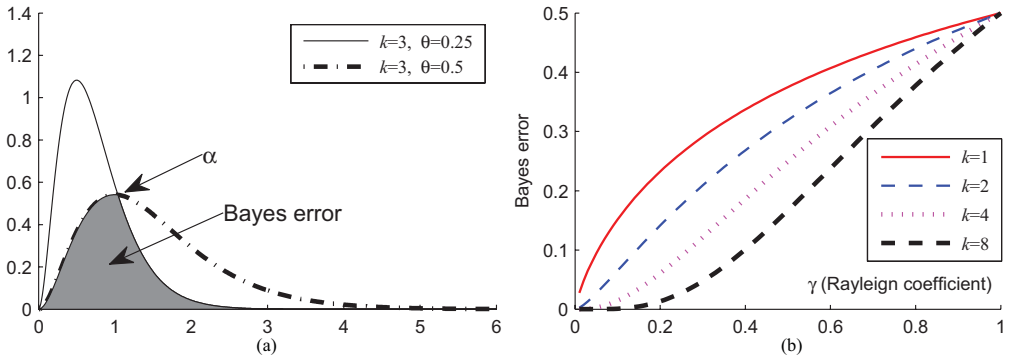

Fig. 2. Bayes error for gamma distributions and its relation to Rayleigh quotient. See Section II for related descriptions.

In case of constraint II, it takes a general gamma form completely determined by both m and s

$$f(x) = \frac{m^{\alpha}}{\Gamma(\alpha)} \exp(-mx) x^{\alpha-1} \tag{3}$$

where α is the solution to the equation $\psi(\alpha) - \log \alpha = s + \log m$. Here, $\psi()$ is the digamma function, and $\Gamma()$ the gamma function (see details of the two functions in Appendix A or in text later).

Let us now discuss the two constraints or conditions. The first constraint basically indicates that all the prior knowledge about the distribution is the mean value (expectation). According to the theorem and its details in Appendix A, if the expectation is given, the best probability distribution for maximum entropy will be an exponential distribution, which is a special case of the gamma distribution family. If additional simple prior knowledge about the expectation of logarithm is added as stated in constraint II, the best probability distribution for maximum entropy will be a general gamma distribution, where the parameters are determined by the expectations (see Appendix A for details).

Briefly, this theorem implies that one may consider gamma probability models for describing the nonnegative-valued EEG power features. To validate this, we examine the estimated gamma distributions in comparison with histograms of real data samples in Fig. 1. It can be seen that there is a good match between the gamma curves and the histograms. Generally, the probability density function of a gamma distribution is determined by two parameters: the scale parameter θ, or its inverse, the rate parameter $\lambda = 1/\theta$ and the shape parameter k.

That is

$$p(y;k,\theta)=\frac{1}{\theta^k}\frac{1}{\Gamma(k)}y^{k-1}e^{-\frac{y}{\theta}};\quad y\geq 0 \text{ and } k,\theta>0 \tag{4}$$

where $\Gamma()$ is a gamma function given by

$$\Gamma(k)=\int_0^\infty t^{k-1}e^{-t}dt. \tag{5}$$

B. Bayes Error for Classification of Gamma Random Variables

Let us consider the EEG classification problem using the Bayes rule [30], on the input data of EEG trials from either of two EEG classes. Without loss of generality, we denote the positive class by Ω_1 and the negative class by Ω_2, and say that the feature (a random variable) is associated with the ERD signal in Ω_1 only. According to the maximum *a posteriori* probability (MAP) rule in Bayes decision, a sample will be classified into the class for which the *a posteriori* probability of the feature value is larger. We would like to stress that all the theoretical analysis can be applied to the ERS signal as well just by considering that the positive class generates a higher EEG power instead of a lower one. Another reason that we consider ERD instead of ERS is that ERD is generally a stronger and more important signal for BCI.

For the ERD feature in EEG, we expect that a stronger ERD (i.e., a smaller power) should indicate a more likely positive class EEG trial than a weaker ERD. Correspondingly in Bayesian classification, the posterior probability functions of the two classes should have only one cross point. Otherwise, if there are two or multiple cross points, the Bayes decision regions will be fragmented and there would be a paradox: a relatively smaller feature value (stronger ERD) may be classified as negative while a relatively weaker ERD signal may be classified as positive.

As will be shown below, a sufficient condition for single cross point between two distributions is that they have the same skewness, i.e., sharing the same shape parameter k.

The two classes of EEG features having equal-k gamma distributions are defined by three Parameters, including k, θ_1, and θ_2 according to (4). Denote the *a priori* probability by $P(\Omega_1)$ or $P(\Omega_2)$. Let $p_{\Omega_1}(\alpha)$ or $p_{\Omega_2}(\alpha)$ be the class-conditional probability density at α. Setting the cross point condition $P(\Omega_1)p_{\Omega_1}(\alpha)=P(\Omega_2)p_{\Omega_2}(\alpha)$ leads to

$$\alpha=\left(\frac{1}{\theta_1}-\frac{1}{\theta_2}\right)^{-1}\left\{\log\left(\frac{P(\Omega_1)}{P(\Omega_2)}\right)-k\log\left(\frac{\theta_1}{\theta_2}\right)\right\}. \tag{6}$$

Therefore, there is one and only one cross point α, which can be easily determined by k, θ_1, and θ_2 together.

Therefore, we assume equal k in the two classes Ω_1 and Ω_2. This assumption is also supported by empirical results illustrated in Fig. 1, where the estimated gamma distribution with equal k provides a close approximation to the histogram of the real EEG data.

Furthermore, assuming equal prior probability and equal shape parameter in the two classes, we can show that the Bayes error is a function of the ratio of class means.

Theorem 2: Let y be a gamma distributed random variable in either EEG class Ω_1 or Ω_2, where the two classes have equal *a priori* probability and the same shape parameter k in probability distribution. The MAP Bayes error ϵ of classifying y will be

$$\epsilon=\frac{1}{2}\left(1+\frac{1}{\Gamma(k)}\int_0^{\frac{k\Gamma}{\Gamma-1}\log\Gamma}t^{k-1}e^{-t}dt-\frac{1}{\Gamma(k)}\int_0^{\frac{k}{\Gamma-1}\log\Gamma}t^{k-1}e^{-t}dt\right) \tag{7}$$

for $\Gamma\leq 1$ where Γ is the ratio given by: $\Gamma=\frac{\theta_1}{\theta_2}$.

Proof: An illustration of the Bayes error is given in Fig. 2(a). With equal prior probability for the two classes, the cross point in (6) becomes

$$\alpha=\frac{k\theta_1\theta_2}{\theta_2-\theta_1}\log\left(\frac{\theta_2}{\theta_1}\right) \tag{8}$$

It is straightforward to show that because of the condition $1\geq\Gamma(=\frac{\theta_1}{\theta_2})$ and

$$\log\left(\frac{p_{\Omega_1}(y)}{p_{\Omega_2}(y)}\right)=k\log\left(\frac{\theta_2}{\theta_1}\right)-\left(\frac{1}{\theta_1}-\frac{1}{\theta_2}\right)y \tag{9}$$

there is always $p_{\Omega_1}(y)\geq p_{\Omega_2}(y)$ when $y\leq\alpha$, and the class-conditional probability density $p_{\Omega_1}(y)<p_{\Omega_2}(y)$ when $y>\alpha$. So we can develop the Bayes error function below

$$\begin{aligned}\epsilon&=P(\Omega_2)\int_0^\alpha p_{\Omega_2}(y)dy+P(\Omega_1)\int_\alpha^{+\infty}p_{\Omega_1}(y)dy \\ &=\frac{1}{2}\left(1+\frac{1}{\Gamma(k)}\phi\left(k,\frac{\alpha}{\theta_2}\right)-\frac{1}{\Gamma(k)}\phi\left(k,\frac{\alpha}{\theta_1}\right)\right)\end{aligned} \tag{10}$$

where ϕ is the lower gamma function

$$\phi(k,x)=\int_0^x t^{k-1}e^{-t}dt. \tag{11}$$

Since the Rayleigh quotient $\Gamma=\frac{\theta_1}{\theta_2}$, we have

$$\frac{\alpha}{\theta_2}=\frac{k\Gamma}{\Gamma-1}\log\Gamma \tag{12}$$

$$\frac{\alpha}{\theta_1}=\frac{k}{\Gamma-1}\log\Gamma \tag{13}$$

and then it is straightforward to write the Bayes error as (7). ■

The quotient Γ is the ratio in scale parameter between the two class distributions. Since the two classes share equal k, the quotient is equivalent to the ratio between the intra-class expectations $\mu_1=k\theta_1$ and $\mu_2=k\theta_2$. Therefore, the condition $\Gamma\leq 1$ is essentially meant for the variable to be a valid ERD feature, since it implies that the expected power value is smaller in the positive class Ω_1: $\mu_1\leq\mu_2$.

We plot some examples of the Bayes functions (7) in Fig. 2(b). All the Bayes error function curves meet at two points: $\{\Gamma=0,\epsilon=0\}$ and $\{\Gamma=1,\epsilon=0.5\}$ (again, we consider the effective range $0\leq\Gamma\leq 1$ only). The former point represents perfect feature variable for zero Bayes error, the latter point represents completely invalid feature variable and thus chance-level classification. Interestingly, the curves also appear to be smooth and monotonic. To prove the monotonicity of the Bayes error over Γ, we have the following theorem, and will discuss its important implications thereafter.

Theorem 3: Let y be a gamma random variable in either EEG class Ω_1 or Ω_2, where the two classes have equal *a priori* probability and a given shape parameter k in probability distribution. The Bayes error ϵ of classifying y into Ω_1 or Ω_2 is a monotonically increasing function of the coefficient Γ for $\Gamma \leq 1$, where Γ is the ratio of the shape parameters between the two classes: $\Gamma = \theta_1/\theta_2$.

Proof: From the expression in (7) and Proof of Theorem 2.2, we can write the derivative of the Bayes error ϵ with respect to the ratio $\Gamma = \theta_2/\theta_1$ as

$$\begin{aligned}\frac{\partial \epsilon}{\partial \Gamma} &= \frac{1}{2\Gamma(k)}\left[\frac{\partial \phi(k, \frac{k\Gamma}{\Gamma-1}\log\Gamma)}{\partial \Gamma} - \frac{\partial \phi(k, \frac{k}{\Gamma-1}\log\Gamma)}{\partial \Gamma}\right] \\ &= \frac{1}{2\Gamma(k)}\left[t^{k-1}e^{-t}\Big|_{t=\frac{k\Gamma}{\Gamma-1}\log\Gamma}\frac{\partial \frac{\Gamma}{\Gamma-1}\log(\Gamma)}{\partial \Gamma}\right. \\ &\qquad \left. -t^{k-1}e^{-t}\Big|_{t=\frac{k}{\Gamma-1}\log\Gamma}\frac{\partial \frac{1}{\Gamma-1}\log(\Gamma)}{\partial \Gamma}\right] \quad (14)\end{aligned}$$

with

$$\frac{\partial \frac{\Gamma}{\Gamma-1}\log(\Gamma)}{\partial \Gamma} = \left(-\frac{\log(\Gamma)}{(\Gamma-1)^2} + \frac{1}{\Gamma-1}\right) \quad (15)$$

and

$$\frac{\partial \frac{1}{\Gamma-1}\log(\Gamma)}{\partial \Gamma} = \left(-\frac{\log(\Gamma)}{(\Gamma-1)^2} + \frac{1}{\Gamma(\Gamma-1)}\right). \quad (16)$$

For simplicity of description, we write

$$b = t^{k-1}e^{-t}\Big|_{t=\frac{k}{\Gamma-1}\log\Gamma}. \quad (17)$$

It is straightforward to further develop the derivative of Bayes error into

$$\frac{\partial \epsilon}{\partial \Gamma} = \frac{b}{2\Gamma(k)}\frac{\log(\Gamma)}{\Gamma(\Gamma-1)}. \quad (18)$$

We can show that

$$\frac{\log(\Gamma)}{\Gamma(\Gamma-1)} > 0, \quad 0 < \Gamma < +\infty \quad (19)$$

since

1) if $\Gamma > 1$, then $\Gamma(\Gamma-1) > 0$, $\log(\Gamma) > 0$, so $\frac{\log(\Gamma)}{\Gamma(\Gamma-1)} > 0$;
2) if $0 < \Gamma < 1$, then $\Gamma(\Gamma-1) < 0$, $\log(\Gamma) < 0$, so $\frac{\log(\Gamma)}{\Gamma(\Gamma-1)} > 0$;
3) if $\Gamma = 1$, then

$$\frac{\log(\Gamma)}{\Gamma(\Gamma-1)}\Big|_{\Gamma=1} = \frac{\lim_{\Gamma\to 1}\frac{d\log(\Gamma)}{d\Gamma}}{\lim_{\Gamma\to 1}\frac{d\Gamma(\Gamma-1)}{d\Gamma}} = 1 > 0. \quad (20)$$

Therefore, the derivative of the Bayes error function (although it is meant for $\Gamma \leq 1$ only) is always positive for every Γ in $\Gamma > 0$. The theorem is thus proved. ■

The theorem indicates that minimum Bayes error can be obtained by minimizing the Γ coefficient, provided that the shape parameter is fixed. Interestingly, this coefficient is equivalent to the Rayleigh quotient [19], which is determined by the covariance matrices of the two EEG classes together with the spatial filter

$$\frac{\mathbf{w}^T\mathbf{E}[\mathbf{R}|\Omega_1]\mathbf{w}}{\mathbf{w}^T\mathbf{E}[\mathbf{R}|\Omega_2]\mathbf{w}} = \frac{\mathbf{E}[\mathbf{w}^T\mathbf{R}\mathbf{w}|\Omega_1]}{\mathbf{E}[\mathbf{w}^T\mathbf{R}\mathbf{w}|\Omega_2]} = \frac{\mathbf{E}[y|\Omega_1]}{\mathbf{E}[y|\Omega_2]} = \frac{\mu_1}{\mu_2} = \Gamma. \quad (21)$$

Thus, we have established the theoretical relationship between the Bayes error and the Rayleigh quotient. The Rayleigh quotient is the basis of the CSP technique that we have introduced earlier. The minimization of the quotient can be casted as a generalized eigenvalue problem, which can be readily solved by standard linear analysis algorithms. Therefore, our theorem gives a theoretical account of the success of the CSP technique for spatial filtering in EEG classification.

III. Experimental Results

A. Materials: EEG Datasets from Sensorimotor BCI

Our numerical study consists of a ten-fold cross-validation based investigation of the proposed method using human EEG data. Particularly, we aim to answer two important questions. First, is the Rayleigh quotient an effective indicator of the Bayes error for a uni-variate feature in real EEG data? Second, can the proposed method generate more separable multivariate features for real applications that often use two or multiple features?

To this end, we use the following three publicly available datasets in our numerical experiment. Since the signal of interest (i.e., ERD) is primarily associated with the μ rhythm [8–13]Hz [9], this paper considers the μ rhythm only.

1) *BCI Competition IV Dataset I:* The dataset [33] consists of both human and artificially generated EEG data, while we consider human EEG data in calibration sessions only. The data were collected from four healthy subjects using a 59-channel EEG device with a sampling rate of 1000 Hz. During data collection, each trial started by displaying a visual cue on a computer screen that prompted the subject to perform the mental task according to the cue. The mean tasks included motor imagination of left hand, right hand, or foot, while each subject pre-selected only two classes and performed a total of 200 trials, equally in the two classes. There was also a 4-s short break after each trial. Here, the numerical study uses the 100-Hz version of the data, and considers only the time interval of [0.5 4]s in each trial.
2) *BCI Competition IV Dataset IIa:* The dataset [34] involves nine subjects performing four-class motor imaginations related to left-hand, right-hand, foot, or tongue, in a cue-based protocol similar to that for the previous dataset. The data consist of 288 trials of EEG collected from 22 channels, with a sampling rate of 250 Hz and a band-pass filter between 0.5 and 100 Hz. The time interval of [0.5 3]s after cue in each trial is considered.
3) *BCI Competition III Dataset IVa:* The data set [35] was recorded from five healthy subjects from nonfeedback BCI sessions. Visual cues of 3.5s duration

indicated the subject to perform which of the following three motor imagery tasks: left hand (L), right hand (R), and right foot (F). Note that each subject chose only two particular tasks. A total of 118 EEG channels were measured at positions of the extended international 10/20-system. Signals were band-pass filtered between 0.05 and 200 Hz and then digitized at 1000 Hz with 16-bit (0.1 uV) accuracy. In this paper, we use the 100-Hz down-sampled data, and chose only the 15 EEG channels in or around the sensorimotor areas.

In the frequency-domain processing, we use $N = 2^{\text{nextpow2}(n_s)}$ points in FFT computation, where nextpow2() is the next higher power of 2, and n_s is the length of the signal. It turns out that there are 21 frequency points in mu band for BCI Competition IV Dataset IIa, 26 points for BCI Competition IV Dataset I, and 13 points for BCI Competition III Dataset IVa.

B. Toward Lower Rayleigh Quotient

For validating the proposed theory, we need to demonstrate that if the Rayleigh quotient can be reduced for each EEG dataset, so is the Bayes error. To this end, we devise a simple extension to the conventional CSP by using a complex-valued solution to reduce the Rayleigh quotient. Unlike in [36] that explores analytic representation of real signals by Hilbert transformation in the time domain, we account for rich phase information in the frequency domain and also put in place a linear-phase spatial filter across different frequency points. We refer to this new filtering technique as *ComplexCSP*. For details, please see Appendix B.

We would like to emphasize that this paper is not meant for developing a new spatial filtering technique. Rather, this new technique only serves the purpose of validating the developed Bayesian spatial filtering theory in practical use. Nevertheless, our experimental results in the following sections demonstrate that the new technique can effectively improve classification accuracy and outperform the method proposed in [36] and other state-of-the-art methods.

C. Rayleigh Quotient Versus Bayes Error

The relationship between the Bayes error and the Rayleigh quotient in real EEG features is fundamental for the Bayesian learning theory. Only if the Rayleigh quotient is closely correlated with Bayes error, it can serve as an alternative metric for designing Bayesian discriminative learning.

We consider two-class data sets only, as the theory is based on two-class Bayes classification while multiclass Bayes error is difficult to compute or estimate. Without loss of generality, let us name one class Ω_1 and the other class Ω_2. We run traditional CSP to extract the ERD feature for Ω_1. The Bayes error is then estimated by finding the minimal empirical error rate: for the given set of uni-variate feature samples, we try each samples' value as a threshold and then select the one, which generates the minimal error rate.

Fig. 3 plots the Rayleigh quotient and the Bayes error computed in each cross-validation fold for each subject, for BCI Competition III Dataset IVa. Statistical analysis gives coefficient of determination $R^2 = 0.95$, correlation coefficient of 0.97 with p-value $= 0$. The result indicates that there is a statistically significant positive correlation between the two measures. We have also computed the mutual information estimate (using the technique reported in [37]). The mutual information [in mean (std) format] is 0.03(0.01) in CSP features and 0.15(0.14) in ComplexCSP features. Statistical analysis shows that the mean of mutual information is larger in ComplexCSP, with p-value of 2.5e-07.

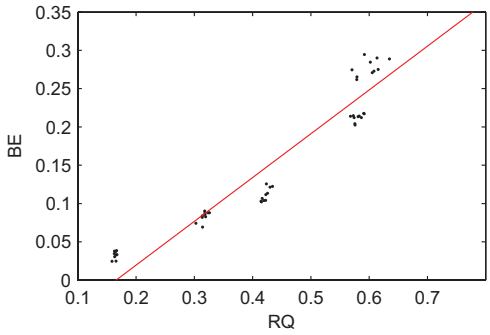

Fig. 3. Correlation between Bayes error (BE: y-axis) and Rayleigh quotient (RQ: x-axis). Each point is generated from a subset (from ten-fold cross-validation) of a subjects' data and the first CSP feature. EEG data are from BCI Competition III Dataset IVa. See Section III-C for detailed description. The red line is a linear regression function for the data points.

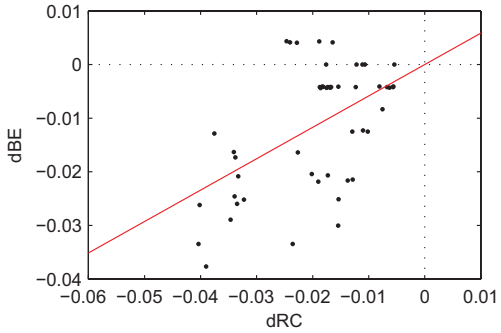

Fig. 4. Difference in Bayes error (dBE: x-axis) versus difference in Rayleigh quotient (dRC: x-axis), from CSP to ComplexCSP. Each point represents a particular subset of a human subjects' EEG data. The x-value denotes the difference in Rayleigh quotient from CSP-generated feature and ComplexCSP-generated feature. EEG data are from BCI Competition III Dataset IVa. See Section III-C for detailed description. The red line is a through-zero-point linear regression function for the data points.

Now we would like to examine the dual questions: is it possible to further reduce the Rayleigh quotient? Will the reduction in the Rayleigh quotient improve classification accuracy? To this end, we run the ComplexCSP algorithm described above and compare the result against that produced by the conventional CSP. Particularly, we compute the differential Bayes error rate and the differential Rayleigh quotient between the two algorithms, in each of the cross-validation folds and for each subject. Fig. 4 plots the correlation between the two differentials. Statistical analysis yields $R^2 = 0.33$, correlation coefficient of 0.57, p-value of 1.5e-5, indicating a correlation with statistical significance.

TABLE I
COMPARISONS OF CLASSIFICATION ACCURACY RATES (%) ON TRAINING DATA

	Subject	LDA (two-class) or L_SVM (multiclass)				G_SVM			
		CSP	mCSP [42]	ACSP [36]	ComplexCSP	CSP	mCSP [42]	ACSP [36]	ComplexCSP
Two-class	Sub_a	91.6(1.7)	82.0(0.8)	92.6(4.4)	**93.0(0.8)**	91.3(1.8)	83.5(0.7)	92.4(4.4)	**92.6(0.7)**
	Sub_b	85.2(2.5)	67.9(4.1)	74.1(3.6)	**92.0(1.1)**	85.2(1.8)	67.7(4.0)	75.3(3.6)	**91.9(0.9)**
	Sub_f	92.6(1.6)	72.1(1.6)	79.4(3.1)	**94.9(0.9)**	93.0(1.2)	72.1(1.2)	78.8(3.8)	**95.0(0.9)**
	Sub_g	92.8(3.1)	76.0(11.9)	81.0(1.5)	**92.9(3.1)**	92.2(3.3)	76.0(12.0)	80.9(1.2)	**92.7(2.9)**
	Sub_aa	80.2(2.5)	77.8(0.7)	80.3(1.9)	**84.5(1.4)**	80.1(2.6)	78.2(0.8)	81.3(1.2)	**84.5(1.5)**
	Sub_al	**96.3(0.4)**	95.8(0.5)	95.4(0.4)	**96.3(0.4)**	96.2(0.5)	96.4(0.5)	95.4(0.4)	**96.6(0.3)**
	Sub_av	75.0(1.5)	69.8(1.6)	70.0(2.2)	**77.0(1.9)**	73.9(1.5)	68.7(2.0)	68.4(1.9)	**76.5(2.0)**
	Sub_aw	90.8(0.9)	84.5(0.7)	84.0(1.4)	**92.5(0.9)**	90.4(1.0)	84.9(0.9)	83.7(1.2)	**92.1(1.0)**
	Sub_ay	93.1(0.5)	89.9(0.9)	91.8(0.7)	**94.4(0.7)**	94.2(0.5)	89.9(0.7)	93.7(0.7)	**94.4(0.7)**
multiclass	Sub_1	71.0(1.1)	57.9(2.5)	54.7(3.1)	**74.3(2.0)**	70.6(0.6)	53.7(2.2)	50.7(2.9)	**74.1(1.3)**
	Sub_2	52.4(1.7)	46.9(7.2)	58.7(2.8)	**62.3(2.6)**	50.4(3.2)	40.3(7.8)	58.1(1.7)	**61.8(1.9)**
	Sub_3	75.0(1.2)	76.3(1.6)	57.3(1.7)	**81.3(1.3)**	73.5(1.1)	75.7(1.7)	55.9(2.4)	**80.0(1.1)**
	Sub_4	53.0(3.2)	41.2(4.5)	42.5(1.7)	**63.5(2.0)**	53.3(2.8)	35.3(5.2)	38.0(2.8)	**62.3(2.3)**
	Sub_5	49.4(1.9)	38.9(2.6)	39.7(1.6)	**57.5(2.2)**	46.6(2.0)	34.8(3.3)	36.6(2.5)	**54.4(2.4)**
	Sub_6	50.9(2.3)	37.3(1.9)	39.7(2.2)	**59.0(2.1)**	47.4(3.3)	34.3(2.9)	37.4(2.4)	**55.3(2.6)**
	Sub_7	65.7(2.8)	56.7(2.9)	57.5(2.0)	**72.2(1.9)**	62.9(5.1)	51.7(3.2)	54.9(1.8)	**71.2(2.5)**
	Sub_8	68.1(3.1)	68.2(1.3)	58.2(2.4)	**77.0(1.7)**	67.2(2.0)	67.0(1.4)	52.4(3.9)	**76.4(1.9)**
	Sub_9	59.7(3.5)	50.9(3.4)	56.0(1.9)	**69.8(2.8)**	58.7(3.0)	48.3(2.9)	53.0(1.4)	**69.3(3.1)**

Notes: See Section III-D for details. The numbers are in mean (std) format. Maximum classification accuracy rates are shown in **bold** style for each classifier and each subject. L_SVM and G_SVM denote linear and nonlinear (using Gaussian kernel) support vector machines, respectively. LDA uses the conventional linear discriminant analysis.

Clearly the ComplexCSP algorithm can effectively reduce Rayleigh quotient over CSP in every data fold. Importantly, the Bayes error is also reduced in 40 out of 50 cases.

D. Classification Accuracy Using Multiple Features

One usually takes advantage of multiple features for discriminative information in practice (e.g., using information captured from different viewing angles leads to high classification performance in recognition of human actions [38]). There are a variety of classifiers that can be used for evaluation of discriminative power of features, such as support vector machines, linear discriminant analysis, and dynamic Bayesian networks [39]. In this paper, we employ support vector machines using the LibSVM toolbox [40] with either linear (hereafter referred to as L_SVM) or Gaussian kernels (hereafter referred to as G_SVM). We also employ the widely used linear discriminant analysis using the Statistical Pattern Recognition Toolbox [41].

Besides the conventional CSP and the ComplexCSP algorithm, we also examined two other state-of-the-art algorithms, namely, the analytic representation-based complex-valued filter [36] (referred to as *ACSP* hereafter) mentioned earlier, and the information theoretic algorithm (referred to as *mCSP* hereafter) that has been demonstrated effective for multiclass EEG classification [42]. We are also thankful to the authors for sharing their MATLAB code of the algorithms.

Particularly for BCI Competition IV Dataset IIa, we need to address multiclass classification if the feature extraction is designed for binary classification only. Here, we employ the so-called one-versus-rest technique. For each mental task class, we create a super negative class by combining all the other classes. The binary-class feature extraction algorithms, here, are each applied to learn a single spatial filter. Accordingly, there are a total of four power features generated from each EEG trial.

Tables I and II summarize the classification accuracy rates in the training data and the test data, respectively. Note that our test shows that LDA and L_SVM perform equally well for two-class data but LDA does not perform as well in the multiclass data (BCI Competition IV Dataset IIa). Therefore, we use L_SVM in multiclass linear classification.

As mentioned earlier, the ComplexCSP algorithm serves as a tool for testing the hypothesis that the Bayes error can be reduced by minimizing the Rayleigh quotient. During training, the ComplexCSP algorithm minimizes the Rayleigh quotient for the training data. It can be seen that the reduced Rayleigh quotient translated to the best classification accuracy, as ComplexCSP outperformed all the other algorithms in every case. This indicates that the dual optimum for Rayleigh quotient and Bayes error can be generalized to multivariate feature vectors as well as multiclass problems.

The comparison in classification accuracy is more complex in the test data. In two-class classification, the ComplexCSP algorithm, by producing the minimum Rayleigh quotient, achieved top accuracy in six out of nine subjects using a linear classifier, and seven out of nine subjects using a nonlinear classifier. In multiclass classification, the ComplexCSP achieved top accuracy in five out of nine subjects using either a linear classifier or a nonlinear classifier.

IV. DISCUSSIONS

In Section II, we have shown both theoretically and practically that the Bayes error is closely correlated with the Rayleigh quotient. This proves the Rayleigh quotient, which is a simple analytical function of the spatial filter vector to learn, as a simple and effective learning objective for Bayesian learning. We would like to stress that the theoretical analysis

TABLE II
COMPARISONS OF CLASSIFICATION ACCURACY RATES ON TEST DATA

	Subject	LDA (two-class) or L_SVM (multiclass)				G_SVM			
		CSP	mCSP	ACSP	ComplexCSP	CSP	mCSP	ACSP	ComplexCSP
Two-class	Sub_a	85.0(5.4)	79.7(6.9)	**90.5(5.7)**	87.8(4.6)	84.1(6.2)	79.5(6.4)	87.1(7.9)	**88.1(5.9)**
	Sub_b	67.8(11.5)	53.4(8.7)	53.6(8.3)	**69.1(10.4)**	66.5(12.2)	54.3(8.8)	51.5(13.2)	**67.7(11.2)**
	Sub_f	84.1(7.9)	56.7(6.5)	58.6(12.1)	**88.1(6.7)**	85.3(7.9)	57.7(6.9)	59.2(11.9)	**87.9(6.6)**
	Sub_g	92.4(4.2)	74.1(16.2)	80.3(10.2)	**92.6(4.1)**	**92.7(4.0)**	72.3(19.6)	79.8(10.2)	91.6(5.7)
	Sub_aa	78.3(5.3)	76.1(8.8)	79.7(6.1)	**81.4(5.7)**	77.6(6.0)	77.6(7.3)	80.5(7.2)	**80.6(5.7)**
	Sub_al	**96.5(4.4)**	95.5(4.5)	95.5(4.1)	96.4(3.3)	96.2(4.2)	96.2(4.5)	95.5(4.1)	**96.3(2.7)**
	Sub_av	73.4(7.9)	66.4(11.9)	66.0(8.4)	**74.6(5.9)**	**72.7(7.9)**	65.9(13.0)	64.8(10.3)	71.7(6.1)
	Sub_aw	89.2(5.8)	83.8(5.4)	83.1(7.7)	**90.6(4.4)**	89.5(3.8)	84.0(5.2)	83.3(7.2)	**91.3(3.3)**
	Sub_ay	**93.5(4.4)**	90.5(5.4)	92.5(5.6)	93.3(4.9)	94.0(3.5)	90.7(4.9)	94.2(5.7)	**94.4(5.1)**
multiclass	Sub_1	62.0(9.7)	48.1(13.7)	48.8(9.8)	**62.8(10.2)**	59.9(7.4)	47.5(11.5)	41.7(10.8)	**61.5(11.6)**
	Sub_2	29.8(3.8)	27.3(9.6)	**53.0(8.1)**	31.9(8.1)	26.8(5.9)	24.4(8.3)	**50.8(6.3)**	32.1(9.2)
	Sub_3	68.8(9.1)	70.6(11.7)	51.1(9.1)	**70.6(9.3)**	68.4(8.0)	**72.0(10.3)**	47.1(7.6)	68.6(10.6)
	Sub_4	**28.5(8.5)**	21.4(6.8)	22.9(7.8)	26.4(8.8)	**28.3(8.9)**	19.0(6.9)	18.5(9.9)	27.1(8.6)
	Sub_5	32.4(9.5)	22.7(8.3)	24.9(8.7)	**33.7(8.2)**	31.7(7.6)	22.4(7.5)	25.3(7.9)	**34.3(7.0)**
	Sub_6	**34.1(12.0)**	32.4(10.4)	30.9(11.9)	33.1(9.9)	31.8(10.6)	25.6(9.8)	30.8(13.4)	**35.3(9.6)**
	Sub_7	50.9(11.4)	**52.3(5.4)**	50.9(10.6)	49.4(8.6)	**51.7(10.6)**	46.4(7.4)	47.7(9.6)	48.0(7.4)
	Sub_8	59.6(9.3)	65.8(8.4)	51.8(9.7)	**66.9(7.8)**	60.0(9.5)	63.4(8.3)	48.0(4.9)	**65.6(8.4)**
	Sub_9	37.8(10.8)	34.2(8.7)	40.6(5.7)	**43.8(10.5)**	35.3(8.6)	35.3(9.6)	37.1(5.1)	**41.6(10.0)**

Notes: Refer to the notes of Table I for descriptions.

is based on gamma models for the features. Future study may examine the Bayes error under other probability distribution families.

Nonetheless, we suggest that gamma models are preferable to normal models or their variants for describing the skewed nonnegative data distributions (see Fig. 1 for examples) of EEG power features. The normal models would not fit into nonnegative data. Log-normal models might be useful, as it is well known [43] that both log-normal and gamma distributions may be used quite effectively in analyzing skewed positive data sets. However, with log-normal distributions, the rationale for using the Rayleigh quotient as an indicator of the Bayes error can be difficult to establish. Despite their resemblance in shape, log-normal model and the gamma model may give significantly different results [44].

We would like to stress again that the ComplexCSP algorithm just serves as a tool to test the hypothesis that the Bayes error can be reduced if one can further lower the Rayleigh quotient. It needs further improvements for better robustness and better optimization before it may be well developed. Nonetheless, in the current work, it favorably helped to prove the hypothesis. Especially for the training data, the ComplexCSP algorithm yielded the top accuracy rates in all the subjects with all the classifiers. On average, it increased the accuracy rate by as much as 10% (in "Sub_4" of BCI Competition IV Dataset IIa).

For the test data, the ComplexCSP still produced more top accuracy rates than any other competitive algorithms. However, the margin was much smaller as compared to that for the training data. For example, the maximum increase in accuracy from CSP to ComplexCSP is now only 6% (in "Sub_9" of BCi Competition IV Dataset IIa). This may be due to the nonstationarity nature of brain signals [45], [46]. Therefore, it is important to design a spatial filtering technique that can be robust against the nonstationarity. One possible way may be to introduce plausible optimization constraints that correspond to certain neuro-physiological principles in the neuronal activities of interest. For example, recent advances in neuro-imaging studies, such as [47] and [48] may provide guides into exploring inter-connections (and thus phase information) between brain areas.

V. CONCLUSION

In this paper, we presented a Bayesian learning theory for spatial filtering in EEG feature extraction and classification. Particularly, we showed that the Bayes error can be formulated as a monotonic function over the Rayleigh quotient, where the quotient is a function determined by the spatial filter and the class covariance matrices. Through analysis using real-world EEG data, we verified the positive correlation between the Bayes error and the quotient. Furthermore, we investigated if classification accuracy can be further improved by reducing Rayleigh quotient for a particular filter. To that end, we tested a complex-valued extension to CSP and demonstrated that if the quotient is reduced, so is the classification error (by up to 10% for training data and up to 6% for test data). Therefore, we provided both theoretical and practical accounts for the Rayleigh quotient to be an effective objective measure in Bayes spatial filter learning.

APPENDIX A

PROOF OF THEOREM 2.1

The maximum entropy density is obtained by maximizing Shannon's entropy measure

$$H(f_x) = -\int f_x(x) \log f_x(x) dx \tag{22}$$

subject to l constraints on generalized moment functions $g_j(x)$'s

$$\int g_j(x) f_x(x) dx = \mu_j, \quad j = 0, \ldots, l \tag{23}$$

where μ_j's are known values. Usually the normalization constraint for a probability density function is set by $j = 0$ with $g_0(x) = 1$ and $\mu_0 = 1$.

The Lagrangian for the maximum entropy density is given by

$$\mathcal{L} = -\int f_x(x) \log f_x(x) dx + \sum_j \lambda_j \left[\int g_j(x) f_x(x) dx - \mu_j \right] \tag{24}$$

where λ_j are Lagrangian multipliers. By using the calculus of variations, the necessary condition for a stationary point is known as [49]

$$\log f_x(x) + \sum_j \lambda_j g_j(x) = 0. \tag{25}$$

Now let us consider a simple constraint that the mean m is known

$$g_1(x) = x \text{ and } \int x f_x(x) dx = m. \tag{26}$$

Then the necessary condition (25) leads to

$$f_x(x) = e^{-\lambda_0 - \lambda_1 x}. \tag{27}$$

From the constraints, it follows that $\lambda_0 = \log m$ and $\lambda_1 = 1/m$. So the probability density function that maximizes the entropy is then given by an exponential function

$$f_x(x) = \frac{1}{m} e^{-\frac{x}{m}} \tag{28}$$

which is essentially a gamma function with shape parameter $k = 1$.

Now let us introduce an additional constraint such that the mean of logarithm is

$$g_2(x) = \log x \text{ and } \int_0^{+\infty} \log x f_x(x) dx = \psi(m) \tag{29}$$

where ψ is the digamma distribution: $\psi(a) = d/da \log \Gamma(m)$. Later we will show that this constraint can be effectively relieved.

Then the necessary condition (25) together with the constraints can be satisfied by

$$\begin{cases} \lambda_0 = \log(\Gamma(m)) \\ \lambda_1 = 1 \\ \lambda_2 = m \end{cases}. \tag{30}$$

Therefore, the probability density is a gamma function with parameters $k = 2$ and $\theta = m$.

While the above example appears to imposes a strong condition (29) where the mean of logarithm is a specific function of the mean value m, we will show in below that any random variable can be transformed to meet such condition without altering the optimality of the solution.

Let the expectation of $g_2(x)$ (29) be s, and the expectation of x be m again. Now we introduce a positive scaling factor α

$$y = \alpha x. \tag{31}$$

The entropy after the scaling is given by

$$\begin{aligned} H(f_y) &= \int_0^{+\infty} \log (f_y(y)) f_y(y) dy \\ &= \int_0^{+\infty} \log \left(\frac{1}{\alpha} f_x(x)\right) \frac{1}{\alpha} f_x(x) d(\alpha x) \\ &= s - \log \alpha. \end{aligned} \tag{32}$$

Importantly, this shows that the entropy of a distribution function after scaling is changed by the amount $-\log \alpha$, which is completely determined by α alone. Hence, the function which maximizes the entropy for $H(f_x)$ will still be the one which, after scaling by α, maximizes $H(f_y)$. As a result, even if the expectation of $\log(f_x)$ does not equal $\psi(m)$ as in (29), the maximum entropy solution will still be a gamma distribution, as long as there is a factor α that makes the integral of $\log(f_y)$ satisfy the constraint in (29).

Now it is necessary to prove that such a factor α does exist. For simplicity and according to the above discussions, we can assume that the variable x is pre-scaled such that the expectation of x is 1 ($m = 1$). Then the expectation of $\log(y)$ will be

$$\begin{aligned} \int_0^{+\infty} \log y p(y) dy &= \int_0^{+\infty} \log(\alpha x) \frac{1}{\alpha} f_x(x) d(\alpha x) \\ &= \log \alpha + s. \end{aligned} \tag{33}$$

So the constraint (refer to (29) but now on variable y instead of x) can be written as $\log \alpha + s = \psi(\alpha)$. To show that there exists a solution to α, we first note that because

$$\log(x) \le x - 1 \text{ for } x > 0 \tag{34}$$

there is

$$s = \int_0^{+\infty} \log(x) f_x(x) dx \tag{35}$$

$$\le \int_0^{+\infty} (x - 1) f_x(x) dx = m - 1 = 0. \tag{36}$$

It is well known that $\psi(\alpha)$ can be written as

$$\psi(\alpha) = \log \alpha - \frac{1}{2\alpha} - \sum_{n=1}^{\infty} \frac{B(2n)}{2n(\alpha^{2n})} \tag{37}$$

where $B(2n)$ is the Bernoulli number. It is easy to see that

$$\lim_{\alpha \to \infty} \psi(\alpha) - \log \alpha = 0. \tag{38}$$

For α approaching 0, we consider a sequence of inverse integer $\alpha_n = 1/n : n = 1, \ldots, \infty$. Due to the continuity of the digamma function, there is

$$\lim_{\alpha \to 0} \psi(\alpha) = \lim_{n \to \infty} \psi\left(\frac{1}{n}\right). \tag{39}$$

According to Gauss's digamma theorem, there is

$$\begin{aligned} \psi\left(\frac{1}{n}\right) = &\log\left(\frac{1}{n}\right) - \xi - \log(2) - \frac{\pi}{2} \cot(\pi/n) \\ &+ 2 \sum_{n'=1}^{[(n-1)/2]} \cos\left(\frac{2\pi n'}{n}\right) \log\left(\sin\left(\frac{n'\pi}{n}\right)\right) \end{aligned} \tag{40}$$

where ξ is the Euler–Mascheroni constant (≈ 0.577). As the (negative) cotangent $-\cot()$ and the $\log(\sin())$ terms both go negative infinity when n approaches ∞, there is

$$\lim_{\alpha \to 0} \psi(\alpha) - \log(\alpha) = -\infty. \tag{41}$$

Since continuous function $\psi(\alpha) - \log\alpha$ takes value from $-\infty$ to 0, there must exist a α, which satisfies $\psi(\alpha) - \log\alpha = s \leq 0$.

The implication of the above development is that for any positive-value random variable, we can scale the data such that the mean and the mean of logarithm is connected by the digamma function, and then a maximum entropy estimate of the underlying distribution is a gamma function.

In summary, we suggest that, according to the maximum entropy principle, a positive random variable can be described by either a gamma distribution with $k = 0$ and $\theta = m$ (thus an exponential distribution) if the constraint on mean value is given, or a gamma distribution if the constraint on the expectation of logarithm value is added. For the latter case, it is straightforward from the above development that the gamma density function for maximum entropy can be written as

$$\begin{aligned} f_x(x) &= \frac{\beta}{\Gamma(\alpha)} e^{-\frac{x}{\beta}} \left(\frac{x}{\beta}\right)^{\alpha-1} \\ &= \frac{\beta^{-\alpha}}{\Gamma(\alpha)} e^{-\frac{x}{\beta}} (x)^{\alpha-1} \end{aligned} \tag{42}$$

where α is the solution to $\psi(\alpha) - \log\alpha = s$ and $\beta = 1/m$ is the scaling coefficient that makes the mean. Since after the scaling the expectation of the logarithm becomes $s - \log\beta$, the above equation becomes the equation in Theorem 1. It is easy to see that α is now the shape parameter k of the gamma distribution, and β becomes the rate parameter θ.

Appendix B: Complex-Valued Common Spatial Pattern (ComplexCSP)

Here, we attempt to design a spatial filter that can produce smaller Rayleigh quotient than CSP. This filter is used as a tool in the study of the proposed theory (see Section III-B). As we know, CSP is the optimum real-valued solution to Rayleigh quotient minimization. However, it is straightforward to show that the quadratic form of the power feature [see (1)] with any real-valued spatial filter is independent on the imaginary part of the covariance matrix Σ_x, because the covariance matrix must be Hermitian (i.e., it equals to its own conjugate transpose). In other words, real-valued approach effectively disregards the imaginary part of the covariance matrix, though the imaginary part may contain discriminative information.

It is, therefore, interesting to explore the imaginary part of the covariance matrix by using complex-valued spatial filters. The imaginary part may be directly derived from the real-valued EEG signal through Hilbert transform like in [36]. Here, we propose an alternative solution, which proves more effective for Bayes learning in our experimental study (Section III-D).

Specifically, we believe that it is important to explore phase information in multichannel EEG signals using frequency-domain representation. Because phase is dependent on the frequency point, the imaginary part of the covariance matrix can vary from one frequency point to another. Therefore, the complex-valued spatial filter must also adapt to this variation.

Let the frequency-domain representation of an EEG segment be a matrix $\mathbf{X} \in \mathbb{R}^{n_c \times n_f}$, with n_c the number of channels

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n_f} \\ \vdots & \ddots & \vdots \\ x_{n_c 1} & \cdots & x_{n_c n_f} \end{bmatrix} \tag{43}$$

where x_{ij} denotes the n_f-point discrete Fourier transform of the i-th channel at frequency $\Omega_j = j - 1/2n_f F_s$, with F_s being the sampling frequency. We would like to stress that the representation $\mathbf{X}$ is complex-valued, and every element x_{ij} can be broken into the real part and the imaginary part.

In other words, the phase of the signal varies on the frequency Ω_j. To explore the phase information, we consider a spatial filter whose phase also varies according to a specific function θ_j for channel j. Thus, the spatial filter for a particular frequency f can be written as

$$\mathbf{w}_f = [w_1(f), \ldots, w_{n_c}(f)]^T \text{ with } w_j(f) = \hat{w}_j e^{i\theta_j(f)}. \tag{44}$$

Here, $\hat{w}_j$ and θ_j represent the magnitude and the phase of a spatial coefficient $w_j(f)$, respectively.

In this paper, we use a linear function for θ_j. The linearity ensures that all frequency components at any frequency point have equal time-delay

$$\theta_j(f) = \vartheta_j f. \tag{45}$$

By constraining the phases across different channels, we reduce the number of free parameters by a factor of n_f. This reduction can be critical in practical optimization computation.

Therefore, the spatial filter $\mathbf{w}_f$ at a frequency point f is defined as a vector-function of three real-valued parameters, namely, the frequency f, the array of coefficients ϑ_j's (45), and the array of the amplitude coefficients $\hat{w}_j$'s [Equation (44)]. For the convenience of description, we denote this function as $\mathbf{u}$: $\mathbf{w}_f = \mathbf{u}(f, \vec{\vartheta}, \vec{\hat{w}})$, where $\vec{\vartheta}$ and $\vec{\hat{w}}$ are the array of ϑ_j's and $\hat{w}_j$'s, respectively. The final EEG feature is the combination of powers over all the frequency components in a selected window $\mathbf{f}_{\text{sel}}$

$$y = \sum_{f \in \mathbf{f}_{\text{sel}}} \mathbf{w}_f^T \mathbf{R}_f \mathbf{w}_f \tag{46}$$

where $\mathbf{R}_f$ is the covariance matrix of EEG at frequency f.

Given a set of training samples, the Rayleigh quotient can be expressed as below, which is still the ratio between two class means

$$\hat{\Gamma} = \frac{\sum_{j=1}^{n_0} \sum_{f \in \mathbf{f}_{\text{sel}}} (\mathbf{u}(f, \vec{\vartheta}, \vec{\hat{w}}))^T \mathbf{R}_{f,j,1} \mathbf{u}(f, \vec{\vartheta}, \vec{\hat{w}})}{\sum_{j=1}^{n_1} \sum_{f \in \mathbf{f}_{\text{sel}}} (\mathbf{u}(f, \vec{\vartheta}, \vec{\hat{w}}))^T \mathbf{R}_{f,j,2} \mathbf{u}(f, \vec{\vartheta}, \vec{\hat{w}})} \tag{47}$$

where $\mathbf{R}_{f,j,1}$ and $\mathbf{R}_{f,j,2}$ are the complex-valued covariance matrices of the j-th EEG trial at frequency f in Ω_1 and Ω_2, respectively.

The objective of learning is essentially to minimize the empirical Rayleigh quotient

$$\{\vec{\theta}, \vec{\hat{w}}\}_{\text{opt}} = \underset{\vec{\theta}, \vec{\hat{w}}}{\operatorname{argmin}} \hat{\Gamma}. \tag{48}$$

In this paper, we tentatively use the MATLAB optimization toolbox to search for a resolution to (48), and the initialization is done by the conventional CSP solution. Specifically, we run the *fminunc* function with (48) as the objective function, and for the function we have options, including TolFun set to 1e-4 and MaxIter set to 200. In this paper, we do not perform deliberate engineering or tuning of the optimization tool. Future work may look into more advanced techniques, such as swarm particle [50] that may not be prone to premature convergence [51].

REFERENCES

[1] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, 2007.

[2] J. R. Wolpaw, N. Birbaumer, D. J. MacFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interface for communication and control," *Clin. Neurophys.*, vol. 113, pp. 767–791, Mar. 2002.

[3] J. R. Wolpaw and D. J. MacFarland, "Multichannel EEG-based brain-computer communication," *Electroencephalogr. Clin. Neurophys.*, vol. 90, no. 6, pp. 444–449, 1994.

[4] J. J. Vidal, "Toward direct brain-computer communication," *Annu. Rev. Biophys. Bioeng.*, vol. 2, pp. 157–180, Jun. 1973.

[5] N. Birbaumer, N. Ghanayim N, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, pp. 297–298, Mar. 1999.

[6] N. Birbaumer and L. G. Cohen, "Brain-computer interfaces: Communication and restoration of movement in paralysis," *J. Physiol.*, vol. 579, pp. 621–636, Mar. 2007.

[7] A. Nijholt and D. Tan, "Brain-computer interfacing for intelligent systems," *IEEE Intell. Syst.*, vol. 23, no. 3, pp. 72–79, May–Jun. 2008.

[8] K. K. Ang, G. Guan, S. G. Chua, B. T. Ang, W. K. Kuah, C. Wang, K. S. Phua, Z. Y. Chin, and H. Zhang, "A large clinical study on the ability of stroke patients in using EEG-based motor imagery brain-computer interface," *Clin. Electroencephalogr. Neurosci.*, vol. 42, no. 4, pp. 253–259, 2011.

[9] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimnation between imagination of right and left hand movement," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 6, pp. 642–651, 1997.

[10] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, pp. R32–R57, Mar. 2007.

[11] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch, "EEG coherency I: Statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 5, pp. 499–515, 1997.

[12] P. L. Nunez, R. B. Silberstein, Z. Shi, M. R. Carpenter, R. Srinivasan, D. M. Tucker, S. M. Doran, P. J. Cadusch, and R. S. Wijesinghe, "EEG coherency II: Experimental comparisons of multiple measures," *Clin. Neurophysiol.*, vol. 110, no. 3, pp. 469–486, 1999.

[13] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "EMG contamination of EEG: Spectral and topographical characteristics," *Clin. Neurophysiol.*, vol. 114, no. 9, pp. 1580–1593, 2003.

[14] J. Muller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filtering of single trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, no. 5, pp. 787–798, 1999.

[15] F. P. de Lange, R. C. Helmich, and I. Toni, "Posture influences motor imagery: An fMRI study," *NeuroImage*, vol. 33, no. 2, pp. 609–617, 2006.

[16] L. Qin, L. Ding, and B. He, "Motor imagery classification by means of source analysis for brain-computer interface applications," *J. Neural Eng.*, vol. 1, no. 3, pp. 135–141, 2004.

[17] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1209–1219, Apr. 2009.

[18] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[19] 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, no. 1, pp. 41–56, Jan. 2008.

[20] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, "Separability of four-class motor imagery data using independent components analysis," *J. Neural Eng.*, vol. 3, pp. 208–216, Jun. 2006.

[21] S. Halder, D. Agorastos, R. Veit, E. M. Hammer, S. Lee, B. Varkuti, M. Bogdan, W. Rosenstiel, N. Birbaumer, and A. Kubler, "Neural mechanisms of brain-computer interface control," *Neuroimage*, vol. 55, pp. 1779–1990, Jan. 2011.

[22] B. Blankertz, F. Losch, M. Krauledat, G. Dornhege, G. Curio, and K.-R. Müller, "The Berlin brain-computer interface: Accurate performance from first-session in BCI-naïve subjects," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 10, pp. 2452–2462, Oct. 2008.

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

[24] 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, Nov. 2006.

[25] W. Wu, X. R. Gao, B. Hong, and S. K. 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, Jun. 2008.

[26] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. Int. Joint Conf. Neural Netw.*, 2008, pp. 2391–2398.

[27] H. Zhang, Z. Y. Chin, 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, Jan. 2011.

[28] W. Zheng and Z. Lin, "Optimizing multiclass spatio-spectral filters via bayes error estimation for EEG classification," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA; MIT Press, 2009, pp. 2268–2276.

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

[30] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, USA: Academic, 1990.

[31] A. Kübler, F. Nijboer, J. Mellinger, T. M. Vaughan, H. Pawelzik, G. Schalk, D. J. McFarland, N. Birbaumer, and J. R. Wolpaw, "Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface," *Neurology*, vol. 64, pp. 1775–1777, May 2005.

[32] S. Y. Park and A. K. Bera, "Maximum entropy autoregressive conditional heteroskedasticity model," *J. Econ.*, vol. 150, pp. 219–230, Jan. 2009.

[33] *BCI Competition IV*. (2009) [Online]. Available: http://www.bbci.de/competition/

[34] C. Brunner, M. Naeem, R. Leeb, B. Graimann, and G. Pfurtscheller, "Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis," *Pattern Recognit. Lett.*, vol. 28, no. 8, pp. 957–964, 2007.

[35] 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, Jun. 2004.

[36] O. Falzon, K. P. Camilleri, and J. Muscat, "Complex-valued spatial filters for task discrimination," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2010, pp. 4707–4710.

[37] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognit.*, vol. 37, no. 5, pp. 857–874, 2004.

[38] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 412–424, Mar. 2012.

[39] D. Bouchaffra, "Mapping dynamic bayesian networks to *alpha*-shapes: Application to human faces identification across ages," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1229–1241, Aug. 2012.

[40] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[41] *Statistical Pattern Recognition Toolbox*. (2009) [Online]. Available: http://cmp.felk.cvut.cz/cmp/cmp_software.html

[42] 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, Aug. 2008.

[43] N. Johnson, S. S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. New York, USA: Wiley, 1995.
[44] B. L. Wiens, "When log-normal and gamma models give different results: A case study," *Amer. Stat.*, vol. 53, no. 2, pp. 89–93, 1999.
[45] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, "Nonstationary nature of the brain activity as revealed by EEG/MEG: Methodological, practical and conceptual challenges," *Signal Process.*, vol. 85, no. 11, pp. 2190–2212, 2005.
[46] 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, Feb. 2010.
[47] E. Formaggio, S. F. Storti, R. Cerini, A. Fiaschi, and P. Manganotti, "Brain oscillatory activity during motor imagery in EEG-fMRI coregistration," *Magn. Resonance Imag.*, vol. 28, no. 10, pp. 1403–1412, 2010.
[48] M. Grosse-Wentrup, "Understanding brain connectivity patterns during motor imagery for brain-computer interfacing," in *Advances in Neural Information Processing Systems*, vol. 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Cambridge, MA, USA: MIT Press, 2009, pp. 561–568.
[49] A. Zellner and R. A. Highfield, "Calculation of maximum entropy distributions and approximation of marginal posterior distributions," *J. Econ.*, vol. 37, no. 2, pp. 195–209, 1988.
[50] M. Clerc and J. Kennedy, "The particle swarm—Explosion, stability, and convergence in a multidimensional complex space," *IEEE Trans. Evol. Comput.*, vol. 6, no. 1, pp. 58–73, Feb. 2002.
[51] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, USA: Springer-Verlag, 1999.

Haihong Zhang (M'07) received the B.E degree in electronic engineering from the Hefei University of Technology, Hefei, China, and the M.E degree in electronics and systems from the University of Science and Technology of China, Hefei, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2005.

He joined the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, in 2004, where he is currently the Head of the Neural Signal and Sensing Laboratory. His current research interests include pattern recognition, brain-computer interface, EEG signal processing, and physiological sensing.

Huijuan Yang (A'11–M'11) received the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2008.

She worked in industry for several years before she joined Crimson Logic Pte Ltd, Singapore, in 2001, where she has developed novel algorithms and systems for electronic documents control. She was a Research Associate and Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, from November 2002 to July 2010, where she has been working on digital image processing, such as binary image/document processing and barcode image processing, computer vision and biometrics. She has been with the Signal Processing Department, Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, since August 2010, where she is currently with the Neural and Biomedical Technology Department. Her current research interests include developing adaptative learning algorithms and systems to decode the brain signals for rehabilitation purposes.

Cuntai Guan (S'91–M'97–SM'03) received the Ph.D. degree in electrical and electronic engineering from Southeast University, Nanjing, China, in 1993.

He was with Southeast University, from 1993 to 1994, where he was engaged in speech vocoder, speech recognition, and text-to-speech. In 1995, he was a Visiting Scientist with the Centre de Recherche en Informatique de Nancy/Centre National de la Recherche Scientique Institut National de Recherche en Informatique et en Automatique, Paris, France, where he was involved in keyword spotting. From 1996 to 1997, he was with the City University of Hong Kong, Kowloon, Hong Kong, where he was engaged in developing robust speech recognition under noisy environment. From 1997 to 1999, he was with the Kent Ridge Digital Laboratories, Singapore, where he was involved in multilingual, large vocabulary, continuous speech recognition. He was a Research Manager and the R&D Director for five years in industries, focusing on the development of spoken dialogue technologies. In 2003, he established the Brain-computer Interface Group, Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, where he is currently a Principal Scientist and the Head of the Neural and Biomedical Technology Department. His current research interests include brain-computer interface, neural signal processing, machine learning, pattern classification, and statistical signal processing, with applications to assistive device, rehabilitation, and health monitoring.