Multi-class Common Spatial Patterns and Information Theoretic Feature Extraction

Moritz Grosse-Wentrup, *Student Member, IEEE,* and Martin Buss, *Member, IEEE*

***Abstract*—We address two shortcomings of the Common Spatial Patterns (CSP) algorithm for spatial filtering in the context of Brain-Computer Interfaces (BCIs) based on EEG/MEG: First, the question of optimality of CSP in terms of the minimal achievable classification error remains unsolved. Second, CSP has been initially proposed for two-class paradigms. Extensions to multi-class paradigms have been suggested, but are based on heuristics. We address these shortcomings in the framework of Information Theoretic Feature Extraction (ITFE). We show that for two-class paradigms CSP maximizes an approximation of mutual information of extracted EEG/MEG components and class labels. This establishes a link between CSP and the minimal classification error. For multi-class paradigms, we point out that CSP by joint approximate diagonalization (JAD) is equivalent to Independent Component Analysis (ICA), and provide a method to choose those independent components (ICs) that approximately maximize mutual information of ICs and class labels. This eliminates the need for heuristics in multi-class CSP, and allows incorporating prior class probabilities. The proposed method is applied to the dataset IIIa of the third BCI competition, and is shown to increase the mean classification accuracy by 23.4% in comparison to multi-class CSP.**

I. Introduction

NON-INVASIVE Brain-Computer Interfaces (BCIs) are devices that enable subjects to communicate without using the peripheral nervous system (see [24] for a review). This can be realized by measuring the electric or magnetic field generated by the central nervous system by EEG or MEG, and using these signals to infer the intention of the user of the BCI. One of the main problems in this context is the low signal-to-noise-ratio (SNR) of the recorded EEG/MEG data. This has motivated research on spatial filters that are designed to extract those components of the EEG/MEG data that provide most information on the intention of the BCI-user.

One algorithm that is very frequently used for this purpose is the Common Spatial Patterns (CSP) algorithm. CSP was first proposed in the context of EEG/MEG analysis in [12], and introduced to the BCI community in [18]. Given EEG/MEG data of two different classes, e.g., motor imagery of the left and right hand, the CSP algorithm computes spatial filters that maximize the ratio of the variance of the data conditioned on one class and the variance of the data conditioned on the other class. In this way, spatial filters can be designed that extract those components of the EEG/MEG data that differ maximally (in terms of the variance) between conditions. Such spatial filters are especially suited for BCIs utilizing motor imagery paradigms, in which the intention of the user is typically inferred from frequency specific changes in variance of EEG/MEG components. Excellent classification results have been reported using CSP for pre-processing in non-invasive BCIs based on motor imagery (e.g., in one of the winning entries of the BCI competition 2003 [2]), and improvement of the CSP algorithm, especially its extension to the spectral domain, is an active area of research (cf. [6], [13], [22] and the references therein).

In this article, we address two shortcomings of the CSP algorithm. The first is that at present there is no established connection between the CSP algorithm and the minimal classification error of a BCI. While from an intuitive point of view the optimization problem solved by CSP, i.e., maximization of the ratio of the variance of the extracted EEG/MEG components between conditions, seems sensible, it is an open question whether this approach is actually optimal in terms of the minimal achievable classification error. The second shortcoming is that CSP has been designed for two-class BCIs. While extensions to multi-class paradigms have been proposed and have been shown to deliver good experimental results [6], these extensions are largely based on heuristics. More specifically, the extensions of CSP to multiple classes proposed in [6] are based on a two-step procedure: computation of a set of potential spatial filters and selection of a subset of these filters. The selection of a subset of spatial filters is based on heuristics which are evaluated experimentally. While convincing classification accuracies are reported in [6], it would be desirable to establish a theoretical framework for selecting a subset of spatial filters that is optimal in terms of either the minimum or the expected classification error. In this article, we address these two shortcomings in the framework of Information Theoretic Feature Extraction (ITFE). The principle of ITFE is to extract those components of the EEG/MEG data that maximize mutual information of extracted components and class labels. Under some assumptions, that we argue are justified in the context of non-invasive BCIs based on motor imagery paradigms, we prove that two-class CSP is optimal in terms of maximizing an approximation of mutual information of extracted EEG/MEG components and class labels. Since mutual information establishes a lower and an upper bound on the minimal classification error, this provides an answer to the first shortcoming of CSP: while in general CSP can not be claimed to be optimal in terms of the minimal achievable classification error, it is optimal in terms of maximizing an approximation of mutual information of class labels

M. Grosse-Wentrup and Martin Buss are with the Institute of Automatic Control Engineering (LSR), Technische Universität München, 80290 München, Germany e-mail: moritzgw@ieee.org, m.buss@ieee.org.

and extracted EEG/MEG components. To resolve the second shortcoming, we consider an extension of CSP to multi-class paradigms based on joint approximate diagonalization (JAD) of several EEG/MEG covariance matrices conditioned on class labels that has been shown to perform well in practice [6]. We point out that multi-class CSP by JAD is equivalent to Independent Component Analysis (ICA), and provide a method to choose those independent components (ICs) of the EEG/MEG data that maximize an approximation of mutual information of ICs and class labels. This eliminates the need for heuristics in chosing subsets of spatial filters and provides a solid theoretical foundation for spatial filtering in the context of non-invasive BCIs with multiple classes. Furthermore, it allows incorporating prior probabilities of classes. We apply this procedure to experimental EEG data from a four-class motor imagery paradigm provided by the Laboratory of Brain-Computer Interfaces at the Technische Universität Graz for the third BCI competition, and show that it leads to an average increase in classification accuracy of 23.4% in comparison multi-class CSP as proposed in [6].

The rest of this article is structured as follows. In Section II, we describe the CSP algorithm for two-class problems and present some previously proposed extensions to multi-class paradigms. In Section III, we introduce the framework of ITFE in the context of BCIs based on motor imagery paradigms. This framework is then used in section III-C to prove the optimality of two-class CSP in terms of (an approximation of) maximum mutual information of class labels and extracted EEG/MEG components. In Section III-D, we show how ITFE can be extended to multi-class paradigms. After presenting some experimental results in Section IV, we conclude this article in Section V with a discussion of the results.

II. Common Spatial Patterns

We begin by stating the variables and assumptions used throughout the article. We consider the random variable $\boldsymbol{x} \in \mathbb{R}^N$ to represent the EEG/MEG data, recorded at N electrodes, from which we wish to infer the intention of the BCI-user $c \in \mathcal{C} = \{c_1, \ldots, c_M\}$. We denote the class probability by $P(c_i),\ i = 1, \ldots, M$, and assume that the EEG/MEG data conditioned on any class follows a Gaussian distribution with zero mean, i.e., $p(\boldsymbol{x}|c_i) = \mathcal{N}(\mathbf{0}, R_{\boldsymbol{x}|c_i}),\ i = 1, \ldots, M$. This is no limitation in the context considered here for the following reasons. First, we will only consider linear transformation of $\boldsymbol{x}$, and hence any mean can be first subtracted and added again in the end. Second, BCIs based on motor imagery paradigms typically infer the intention of the user from changes in the variance of the EEG/MEG data in specific frequency bands across conditions. As long as no information contained in higher moments of $\boldsymbol{x}$ is being used for inference, no information is lost by assuming $p(\boldsymbol{x}|c)$ to follow a Gaussian distribution. We then wish to find a linear transformation $W \in \mathbb{R}^{N \times L}$ with $L << N$, such that for finite training data using the dimension reduced $\hat{\boldsymbol{x}} = W^{\mathrm{T}}\boldsymbol{x}$ for inferring the intention of the BCI-user leads to an increased classification accuracy in comparison to using $\boldsymbol{x}$.

A. Two-class Common Spatial Patterns

In this section, we assume a two-class paradigm, i.e., $\mathcal{C} = \{c_1, c_2\}$. The CSP algorithm then solves the optimization problem [16]

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{R}^N}{\operatorname{argmax}} \left\{ \frac{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}} \right\}, \tag{1}$$

with $R_{\boldsymbol{x}|c_1}, R_{\boldsymbol{x}|c_2}$ the covariance matrices of $\boldsymbol{x}$ given c_1, c_2 respectively. Since (1) is in the form of the well-known Rayleigh quotient, solutions to (1) are given by eigenvectors of the generalized eigenvalue problem

$$R_{\boldsymbol{x}|c_1} \boldsymbol{w} = \lambda R_{\boldsymbol{x}|c_2} \boldsymbol{w}. \tag{2}$$

The eigenvectors of (2) thus correspond to the desired spatial filters. Furthermore, for a given eigenvector $\boldsymbol{w}^*$ the corresponding eigenvalue determines the value of the cost function:

$$\lambda^* = \frac{\boldsymbol{w}^{*\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}^*}{\boldsymbol{w}^{*\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}^*}. \tag{3}$$

The eigenvalues thus are a measure for the quality of the obtained spatial filters, i.e., the eigenvalue associated with a spatial filter expresses the ratio of the variance between conditions of the component of the EEG/MEG data extracted by the spatial filter. Pre-processing is then usually done by combining the L eigenvectors of (2) with the smallest/largest eigenvalues to form $W \in \mathbb{R}^{N \times L}$ and computing $\hat{\boldsymbol{x}} = W^{\mathrm{T}}\boldsymbol{x}$.

B. Multi-class Common Spatial Patterns

Extending CSP to multi-class paradigms is either done by performing two-class CSP on different combinations of classes (e.g., by computing CSPs for all combinations of classes or by computing CSP for one class versus all other classes), or by joint approximate diagonalization (JAD) (see [6] and the references therein). Since the first approach is conceptually identical to CSP for two-class paradigms, we will focus here on CSP by JAD.

Given EEG/MEG data from M different classes, the goal of CSP by JAD is to find a transformation $W \in \mathbb{R}^{N \times N}$ that diagonalizes the covariance matrices $R_{\boldsymbol{x}|c_i}$, i.e.,

$$W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = D_{c_i},\ i = 1, \ldots, M, \tag{4}$$

with $D_{c_i} \in \mathbb{R}^{N \times N}$ diagonal matrices. There are several approaches to this problem (discussed in [25]), the details of which are not of interest here. The idea of using JAD for multi-class CSP lies in the fact that CSP for two classes can be understood as diagonalizing two covariance matrices. More precisely, if the eigenvectors of the generalized eigenvalue problem (2) are combined in a matrix W, then $W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = D_{c_i},\ i = 1, \ldots, 2$. It then seems plausible to extend CSP to multi-class paradigms by finding a transformation W that approximately diagonalizes multiple covariance matrices. A total of L columns of the obtained matrix W are then taken as the desired spatial filters.

There are, however, two caveats. First, this approach is motivated heuristically and lacks a firm theoretical foundation. Second, it remains unclear which columns of W provide

the optimal spatial filters. Or, as it is put in [6], *as opposed to the two-class problem, there is no canonical way to choose the relevant CSP patterns for multi-class CSP*. In [6], the following heuristic is proposed to choose the L optimal spatial filters: Given a matrix W obtained by JAD, compute the eigenvalues of all covariance matrices, i.e., compute $\boldsymbol{\lambda}_i = \text{diag}\{W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W\},\ i = 1, \ldots, M$. Then map all $j = 1, \ldots, N$ eigenvalues of each class $i = 1, \ldots, M$ to $\lambda_{i,j} = \max\{\lambda_{i,j}, 1/(1+(M-1)^2 \lambda_{i,j}/(1-\lambda_{i,j}))\}$, and select the L/M eigenvectors with the largest transformed eigenvalues of each class as spatial filters. If one eigenvector is selected more than once, replace it by the eigenvector with the next highest transformed eigenvalue.

We will point out in this article that multi-class CSP by JAD is equivalent to ICA, and provide a method to choose those columns of W that are optimal in terms of maximizing an approximation of mutual information of class labels and extracted EEG/MEG components. We thereby provide a theoretical foundation for multi-class CSP by JAD, and eliminate the need for heuristics in choosing spatial filters.

III. Information Theoretic Feature Extraction

In this section, we introduce the framework of Information Theoretic Feature Extraction (ITFE) for pre-processing. ITFE has recently received considerable attention in the machine learning community, mostly in a non-parametric setting (cf. [9], [17], [23]). The general idea of ITFE is the following. Let $\boldsymbol{x} \in \mathcal{X}$ be a random variable, e.g., the observable EEG/MEG data, from which $c \in \mathcal{C}$, e.g., the intention of the BCI-user, is to be inferred. The goal of ITFE is to find a transformation $f^* : \mathcal{X} \mapsto \hat{\mathcal{X}}$ that maps the original feature space $\mathcal{X}$ into a discrete set $\hat{\mathcal{X}}$ while preserving information on the class labels c in $\boldsymbol{x}$:

$$f^* = \underset{f \in \mathcal{F}}{\text{argmax}} \{I(c, f(\boldsymbol{x}))\}, \tag{5}$$

with $I(c, f(\boldsymbol{x}))$ the mutual information of c and $f(\boldsymbol{x})$ (cf. [5]), and $\mathcal{F}$ some function space. This approach is based on two inequalities that provide upper and lower bounds on the minimal achievable classification error in terms of the mutual information. The first of these two inequalities, called Fano's inequality (cf. [5]), establishes a lower bound on the minium error probability in estimating c from $f(\boldsymbol{x})$ for any classifier $g : \hat{\mathcal{X}} \mapsto \mathcal{C}$:

$$\begin{aligned} P_e &:= \underset{g \in \mathcal{G}}{\text{argmin}} \{\Pr\{c \neq g(f(\boldsymbol{x}))\}\} \geq \frac{H(c|f(\boldsymbol{x})) - 1}{\log |\mathcal{C}|} \\ &= \frac{H(c) - I(c, f(\boldsymbol{x})) - 1}{\log |\mathcal{C}|}, \end{aligned} \tag{6}$$

with $H(.)$ the Shannon entropy, $|\mathcal{C}|$ the number of elements in $\mathcal{C}$, and $\mathcal{G}$ the space of all classifiers. The second inequality, presented in [8], provides an upper bound on the minimum error probability:

$$P_e \leq 1 - 2^{I(c, f(\boldsymbol{x})) - H(c)}. \tag{7}$$

Together these two bounds imply that maximizing mutual information of c and $f(\boldsymbol{x})$ minimizes the minimal error probability, and indeed that $P_e = 0$ iff $I(c, f(\boldsymbol{x})) = H(c)$. The relation of mutual information and minimal error probability for $\mathcal{C} = \hat{\mathcal{X}} = \{c_1, \ldots, c_M\},\ P(c_i) = 1/M,\ i = 1, \ldots, M$ and $M = 4$ is illustrated in Fig. 1.

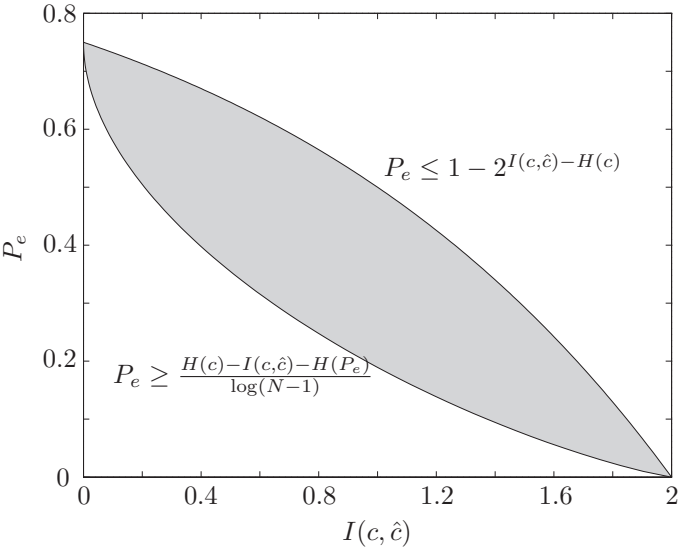

Fig. 1. Admissible combinations (shaded area) of mutual information and minimal error probability for $\hat{c} = f(\boldsymbol{x})$, $\mathcal{C} = \hat{\mathcal{X}} = \{c_1, \ldots, c_M\},\ P(c_i) = 1/M,\ i = 1, \ldots, M$, and $M = 4$.

Note that the bounds in (6) and (7) have been derived using discrete and not differential entropy. As such, they only apply to discrete feature spaces $\hat{\mathcal{X}}$. The bounds can be extended to continuous feature spaces by considering a specific quantization scheme, and thus qualitatively also apply to continuous feature spaces. However, when working in continuous feature spaces (6) and (7) can not be used to quantitatively predict bounds on the minimum classification error in terms of the mutual information.

A. ITFE and Non-invasive BCIs

In the context of non-invasive BCIs, we wish to find a dimension-reduced representation of the EEG/MEG data that maximizes mutual information of class labels and extracted EEG/MEG components. We thus have $\mathcal{X} = \mathbb{R}^N$, and choose $\hat{\mathcal{X}} = \mathbb{R}$ for the dimension-reduced feature space. This has got two implications. First, we only wish to extract one EEG/MEG component at a time. While we could also choose to extract several components simultaneously, this is equivalent to extracting components sequentially in the setting considered here, as we will show in Sections III-C and III-D. Second, since in this context $\hat{\mathcal{X}} = \mathbb{R}$ is not a discrete set, (6) and (7) only apply if additionally a suitable quanitzation scheme for $\hat{\mathcal{X}}$ is specified. This quantization scheme, however, has no qualitative influence, and is thus disregarded. We furthermore limit ourselves to linear transformations. Equation (5) thus simplifies to

$$\boldsymbol{w}^* = \underset{w \in \mathbb{R}^N}{\text{argmax}} \left\{ I(c, \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}) \right\}. \tag{8}$$

Note that this implies that we wish to compute the mutual information of a discrete and a continuous variable. To make this expression well defined, we again need to assume a

quantization that discretizes the continuous variable $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$. This quantization scheme, however, has negligible influence on the mutual information, since the entropy of a n-bit quantization of a continuous random variable is approximately the entropy of the continuous variable plus n [5]. Since the entropy enters twice with different sign into the computation of mutual information, the terms due to the quantization cancel out. We will thus disregard the quantization scheme in the sequel and work with differential entropy.

To the best of our knowledge, no analytic expression for $I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$ exists given our assumptions. Hence, we will first derive an analytic approximation of the mutual information $I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$. We will then find a solution to (8) based on this approximation for two-class paradigms, and finally discuss the extension to multi-class paradigms.

B. Approximation of Mutual Information

First note that the mutual information of c and $\hat{x} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ can be written as

$$\begin{aligned} I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) &= H(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) - H(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}|c) = H(\hat{x}) - H(\hat{x}|c) \\ &= H(\hat{x}) - \sum_{i=1}^{M} P(c_i) H(\hat{x}|c_i). \end{aligned} \tag{9}$$

Since differential entropy is not scale invariant, we assume $\sigma_{\hat{x}}^2 = 1$. This is no loss of generality, since $\boldsymbol{w}$ can always be scaled to meet this assumption. Now note that we assumed $p(\boldsymbol{x}|c) = \mathcal{N}(\boldsymbol{0}, R_{\boldsymbol{x}|c})$. Since $\hat{x}$ is a linear combination of the elements of $\boldsymbol{x}$ it also follows a (now one-dimensional) conditional Gaussian distribution with zero mean, i.e., $p(\hat{x}|c) = \mathcal{N}(0, \sigma_{\hat{x}|c}^2)$. We can thus express the entropy of $\hat{x}$ given class c_i as

$$H(\hat{x}|c_i) = \log\sqrt{2\pi e \sigma_{\hat{x}|c_i}^2} = \log\sqrt{2\pi e \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}}. \tag{10}$$

The marginal distribution $p(\hat{x})$, however, does not follow a Gaussian distribution since

$$p(\hat{x}) = \sum_{i=1}^{M} P(c_i) p(\hat{x}|c_i) = \sum_{i=1}^{M} P(c_i) \mathcal{N}(0, \sigma_{\hat{x}|c_i}), \tag{11}$$

which is a sum of M Gaussian distributions and thus not itself Gaussian. To the best of our knowledge there is no analytical solution to the entropy of a sum of Gaussian distributions, and thus no closed form solution of $H(\hat{x})$. We can, however, approximate $H(\hat{x})$ in the following manner. First, define the negentropy of $\hat{x}$ as

$$J(\hat{x}) := H_g(\hat{x}) - H(\hat{x}), \tag{12}$$

with $H_g(\hat{x})$ the entropy of a Gaussian random variable with the same variance as $\hat{x}$. The negentropy of $\hat{x}$ can be approximated as

$$J(\hat{x}) \approx \frac{1}{12}\kappa_3(\hat{x})^2 + \frac{1}{48}\kappa_4(\hat{x})^2, \tag{13}$$

with the third- and fourth-order cumulants $\kappa_3(\hat{x}) = E\{\hat{x}^3\}$ and $\kappa_4(\hat{x}) = E\{\hat{x}^4\} - 3$ [4]. Since $p(\hat{x})$ is a sum of Gaussian distributions with zero mean it is symmetric, and hence $\kappa_3(\hat{x}) = 0$. Furthermore, $\kappa_4(\hat{x}) = 3\sum_{i=1}^{M} P(c_i)\left(\sigma_{\hat{x}|c_i}^4 - 1\right)$ since the fourth moment of a Gaussian distribution with zero mean and unit variance equals three and $\kappa_4(\alpha x) = \alpha^4 \kappa_4(x)$ (see any textbook on advanced statistics). Combining (12) and (13) we thus have

$$H(\hat{x}) \approx \log\sqrt{2\pi e} - \frac{3}{16}\left(\sum_{i=1}^{M} P(c_i)\left(\sigma_{\hat{x}|c_i}^4 - 1\right)\right)^2. \tag{14}$$

Combining (9), (10) and (14) we obtain an estimate of the mutual information of c and $\hat{x}$ as

$$\begin{aligned} I(c, \hat{x}) \approx\ & -\sum_{i=1}^{M} P(c_i) \log\sqrt{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}} \\ & -\frac{3}{16}\left(\sum_{i=1}^{M} P(c_i)\left((\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w})^2 - 1\right)\right)^2 \end{aligned} \tag{15}$$

It then remains to investigate the accuracy of this approximation of mutual information. The only approximation used in deriving (15) is the approximation of negentropy in (13). This approximation is based on an Edgeworth expansion up to order four of the true probability density function (11) about its best Gaussian approximation. As such, (15) is exact if $p(\hat{x})$ is Gaussian distributed, and the quality of the approximation deteriorates with deviation of $p(\hat{x})$ from Gaussianity.

To quantitatively evaluate the accuracy of the approximation of mutual information, the true mutual information in (9) was computed by numerical integration (using recursive adaptive Lobatto quadrature as implemented in Matlab©) for $\mathcal{C} = \{c_1, c_2\}$ and $\sigma_{\hat{x}|c_1} \in]0, 1]$. Note that this covers the whole range of $\sigma_{\hat{x}|c_i},\ i \in \{1, 2\}$ due to symmetry of (9) with respect to $\sigma_{\hat{x}|c_i}$ and the assumption of unit variance of $\hat{x}$. The error of the approximation of mutual information in (15) was then evaluated for different prior class probabilities by subtracting the numerically computed true mutual information from the approximation of mutual information. The resulting error (in per cent of the true mutual information) is shown in Fig. 2. Note that $\sigma_{\hat{x}|c_1} = 1$ implies $\sigma_{\hat{x}|c_2} = 1$ and hence $p(\hat{x}) = \mathcal{N}(0, 1)$. As expected, the error between the approximated and true mutual information is zero for $\sigma_{\hat{x}|c_1} = 1$ and small for $\sigma_{\hat{x}|c_1}$ close to one. In fact, the error of the approximation is below one per cent for $\sigma_{\hat{x}|c_1} \in [0.84, 1]$. As long as $\sigma_{\hat{x}|c_1} > 0.36$ the error stays below ten per cent. However, for even smaller values of $\sigma_{\hat{x}|c_1}$ the error grows large, limiting the usefulness of the approximation. Qualitatively, this behavior of the approximation is independent of the number of classes, i.e., if $p(\hat{x})$ is close to Gaussianity a small error can be expected also for $M > 2$. Quantitatively, the goodness of the approximation varies as a function of the number of classes. The validity of the approximation in (15) for multiple classes will be experimentally validated in Section IV.

The applicability of the approximation of mutual information in the context of non-invasive BCIs thus depends on by how much EEG/MEG sources that provide information on the intention of the user of a BCI deviate from Gaussianity, i.e., how much their variances vary across conditions. In general, such sources can be expected to be rather close to Gaussianity, and thus the approximation to be accurate, for the simple reason that inferring the intention of the user of a BCI is a hard

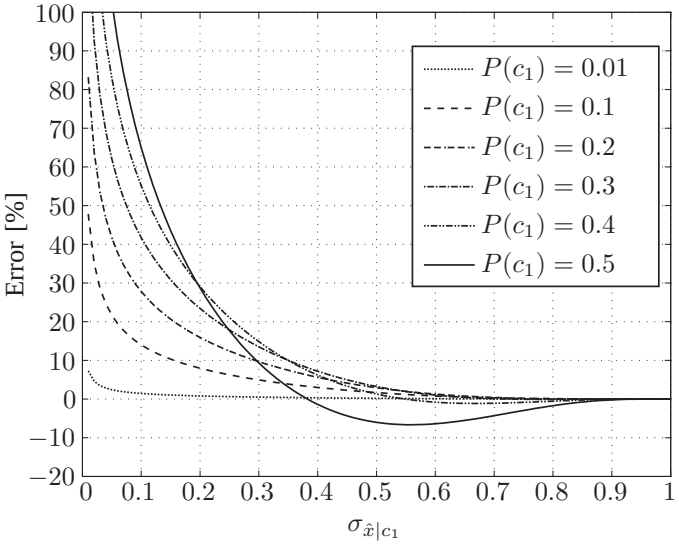

Fig. 2. Error of the approximation of mutual information (15) in per cent for $\mathcal{C} = \{c_1, c_2\}$ as a function of $\sigma_{\hat{x}|c_1}$ for different prior class probabilities.

task. If variances of EEG/MEG sources providing information on the intention of the user would vary significantly across conditions, inferring the intention of the user of a BCI could be expected to be substantially easier than it is the case. This claim will be experimentally validated in Section IV.

C. Two-class ITFE and Optimality of Two-class CSP

We will now discuss solutions to (8) based on the above approximation of mutual information for two-class paradigms, i.e., we again assume $\mathcal{C} = \{c_1, c_2\}$. Equation (15) then reduces to

$$\begin{aligned} I(c, \hat{x}) \approx & -P(c_1) \log \sqrt{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}} \\ & -P(c_2) \log \sqrt{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}} - \frac{3}{16} \Big(P(c_1) \left(\sigma^4_{\hat{x}|c_1} - 1 \right) \\ & + P(c_2) \left(\sigma^4_{\hat{x}|c_2} - 1 \right) \Big)^2 . \end{aligned} \tag{16}$$

We will from here on refer to this expression as mutual information, keeping in mind that it is in fact an approximation thereof. Taking the derivative of (16) with respect to $\boldsymbol{w}$ then yields

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}} I(c, \hat{x}) = & -\frac{P(c_1)}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w}} R_{\boldsymbol{x}|c_1} \boldsymbol{w} - \frac{P(c_2)}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w}} R_{\boldsymbol{x}|c_2} \boldsymbol{w} \\ & - \frac{3}{2} \left(P(c_1) (\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w})^2 + P(c_2) (\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w})^2 - 1 \right) \\ & \cdot \left(P(c_1) \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_1} \boldsymbol{w} R_{\boldsymbol{x}|c_1} \boldsymbol{w} + P(c_2) \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_2} \boldsymbol{w} R_{\boldsymbol{x}|c_2} \boldsymbol{w} \right) \end{aligned} \tag{17}$$

Letting

$$\begin{aligned} \alpha_i &:= -\frac{P(c_i)}{\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}}, \\ \beta_i &:= -\frac{3}{2} \left(\sum_{j=1}^{2} P(c_j) (\boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_j} \boldsymbol{w})^2 - 1 \right) \boldsymbol{w}^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}, \end{aligned}$$

and setting (17) to zero results in

$$(\alpha_1 + \beta_1) R_{\boldsymbol{x}|c_1} \boldsymbol{w} + (\alpha_2 + \beta_2) R_{\boldsymbol{x}|c_2} \boldsymbol{w} = 0. \tag{18}$$

Rearranging and letting $\lambda := -\frac{\alpha_2 + \beta_2}{\alpha_1 + \beta_1}$ then yields

$$R_{\boldsymbol{x}|c_1} \boldsymbol{w} = \lambda R_{\boldsymbol{x}|c_2} \boldsymbol{w}. \tag{19}$$

In the case of two-class paradigms and the stated assumptions, solutions to (8) are thus given by the eigenvectors of the generalized eigenvalue problem (19). Comparing the solutions obtained by ITFE (19) and CSP (2) shows that for two-class paradigms both methods yield identical spatial filters. Furthermore, if equal class probabilities are assumed, i.e., $P(c_1) = P(c_2) = 1/2$, and the obtained spatial filters are ranked in terms of the ratio of the variance between conditions (CSP) and in terms of mutual information (ITFE) the ordering is the same. This can be seen by the following argument. For CSP, spatial filters are ranked according to

$$\begin{aligned} f(\lambda^*) &:= \max \left\{ \lambda^*, \frac{1}{\lambda^*} \right\} = \max \left\{ \frac{\sigma^2_{\hat{x}|c_1}}{\sigma^2_{\hat{x}|c_2}}, \frac{\sigma^2_{\hat{x}|c_2}}{\sigma^2_{\hat{x}|c_1}} \right\} \\ &= \max \left\{ \frac{\sigma^2_{\hat{x}|c_1}}{2 - \sigma^2_{\hat{x}|c_1}}, \frac{2 - \sigma^2_{\hat{x}|c_1}}{\sigma^2_{\hat{x}|c_1}} \right\}, \end{aligned} \tag{20}$$

with the third equality due to the assumption of equal class probabilities and unit variance of $\hat{x}$. For $\sigma^2_{\hat{x}|c_1} \in]0, 2[$ this is a convex function that attains its minimum at $\sigma^2_{\hat{x}|c_1} = 1$ and is symmetric about $\sigma^2_{\hat{x}|c_1} = 1$. As it is easy to check, the same holds true for the approximation of mutual information in (15), which is used to rank spatial filters in ITFE (note that $\sigma^2_{\hat{x}|c_1} = \boldsymbol{w}^{\mathrm{T}} R_{\hat{x}|c_1} \boldsymbol{w}$). For equal class probabilities, both functions used for ranking spatial filters thus depend only on $\sigma^2_{\hat{x}|c_1}$, are convex, symmetric about $\sigma^2_{\hat{x}|c_1} = 1$, and attain their minimum at $\sigma^2_{\hat{x}|c_1} = 1$. Now consider two spatial filters $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ with associated eigenvalues λ_1 and λ_2. If $f(\lambda_1) > f(\lambda_2)$ then either $(\sigma^2_{\hat{x}|c_1})_{(\lambda_1)} > (\sigma^2_{\hat{x}|c_1})_{(\lambda_2)}$ and $(\sigma^2_{\hat{x}|c_1})_{(\lambda_1)} > 1$, or $(\sigma^2_{\hat{x}|c_1})_{(\lambda_1)} < (\sigma^2_{\hat{x}|c_1})_{(\lambda_2)}$ and $(\sigma^2_{\hat{x}|c_1})_{(\lambda_1)} < 1$. Since the approximation of mutual information is convex, attains its minimum at $\sigma^2_{\hat{x}|c_1} = 1$, and is symmetric about $\sigma^2_{\hat{x}|c_1} = 1$, it follows that also $I(c, \boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{x}) > I(c, \boldsymbol{w}_2^{\mathrm{T}} \boldsymbol{x})$. Consequently, the ordering of spatial filters ranked by CSP and ITFE is the same.

Summarizing the results of this section, we have shown that for equal class probabilities, conditionally Gaussian distributed EEG/MEG data, and linear transformations pre-processing by CSP and ITFE leads to the same spatial filters. Under the given assumptions, two-class CSP thus maximizes an approximation of mutual information of extracted EEG/MEG components and class labels.

It should be pointed out that $p(\hat{x})$ is completely described by the class conditional variances. As such, an approximation of mutual information using Edgeworth expansion of arbitrary order will also be a function of the class conditional variances only. Hence, it should be possible to extend the above argument to approximations of mutual information of arbitrary order, thereby proving optimality of two-class CSP in terms of true mutual information of class labels and extracted EEG/MEG components (under the assumption of conditionally Gaussian distributed sources). However, a rigorous proof of this conjecture is beyond the scope of this work.

D. Multi-class Information Theoretic Feature Extraction

We will now discuss possible solutions of (8) for multi-class paradigms, i.e., for $\mathcal{C} = 1, \ldots, M$. In principle, taking the derivative of (15) with respect to $\boldsymbol{w}$ and setting it to zero gives an implicit solution for the spatial filters that correspond to local extrema of (15). However, due to the presence of multiple covariance matrices $\partial I(c, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})/\partial \boldsymbol{w} = 0$ can not be formulated as a generalized eigenvalue problem anymore. Furthermore, to the best of our knowledge, no analytic solution to this expression exists. This leaves the possibility of deriving a gradient descent rule for finding a solution to (8). While this is a straightforward procedure, (8) does not constitute a convex optimization problem. Consequently, gradient descent is not an efficient approach for finding all local extrema of (15).

Due to these difficulties we consider a different approach. First note that the problem of finding optimal linear spatial filters can be understood as a subspace identification problem. Assume there are $L < N$ EEG/MEG sources within the brain providing information on the intention of the user of a BCI. These L sources span a L-dimensional subspace of the data, which we denote as the signal subspace. The space orthogonal to the signal subspace, spanned by the sources that do not provide information on the intention of the BCI-user, is denoted as the noise subspace. The goal of spatial filtering in the context of non-invasive BCIs is to extract the signal subspace. The actual procedure of extracting the signal subspace can be decomposed into two steps. The first step is to find a transformation of the data space such that the signal and the subspace become orthogonal in the new basis, and the second step is to identify the subset of the new basis vectors that span the signal subspace. This procedure is equivalent to first finding a set of potential spatial filters, and then identifying those spatial filters that extract sources which provide information on the intention of the user of the BCI, i.e, sources that span the signal subspace.

As discussed above, the approximation of mutual information in (15) is not well suited for computing a set of potential spatial filters. However, once a transformation of the data space has been obtained in which the signal and noise subspace are orthogonal, (15) can be employed to identify those basis vectors that span the signal subspace. In multi-class CSP as presented in Section II-B, JAD of class-conditional data covariances matrices is used for computing a set of potential spatial filters. In this Section, we will point out that this is an implementation of Independent Component Analysis (ICA) (cf. [15]). We will then show that under certain conditions ICA is capable of separating the signal and noise subspace. In multi-class CSP, the identification of the sources spanning the signal subspace is then carried out by means of the heuristic presented in Section II-B. Here, it will be shown that the identification of the sources spanning the signal subspace can be done by means of the derived approximation of mutual information. This eliminates the need for heuristics in identifying the signal subspace, and provides a solid theoretical foundation for spatial filtering in the context of non-invasive BCIs with multi-class paradigms. The complete procedure, i.e., performing ICA to obtain a suitable transformation of the data space and using the derived approximation of mutual information to identify the signal subspace, will be termed multi-class ITFE.

We begin by considering a linear model for the EEG/MEG data, i.e.,

$$\boldsymbol{x} = A\boldsymbol{s}. \tag{21}$$

Here, $\boldsymbol{s} \in \mathbb{R}^N$ is a random vector with zero mean representing the original EEG/MEG current sources inside the cortex, and $A \in \mathbb{R}^{N \times N}$ is a full-rank mixing matrix with each column $\boldsymbol{a}_j,\ j = 1, \ldots, N$ describing the projection strength of source s_j to each of the N EEG/MEG electrodes. We furthermore assume $p(\boldsymbol{s}) = \prod_{j=1}^{N} p(s_j)$, i.e., we assume the elements of $\boldsymbol{s}$ to be mutually statistically independent. This is the standard instantaneous mixing model assumed in ICA, which has been shown to be a good working assumption for EEG/MEG data (cf. [11] and the references therein). Finally, we assume that there are only L sources that provide information on the intention of the BCI-user. Without loss of generality, we assume these to be the first L sources, i.e., $I(c, s_i) = 0,\ i = L+1, \ldots, N$.

We will now show how for this model the original sources $\boldsymbol{s}$ can be reconstructed only from observations of $\boldsymbol{x}$ by JAD of the class-conditional covariance matrices of $\boldsymbol{x}$. First note that the covariance matrix of $\boldsymbol{x}$ given condition c_i is given by

$$R_{\boldsymbol{x}|c_i} = A R_{\boldsymbol{s}|c_i} A^{\mathrm{T}}, \tag{22}$$

with $R_{\boldsymbol{s}|c_i}$ the covariance matrix of $\boldsymbol{s}$ given condition c_i. If we now perform JAD, it is obvious that $W^{\mathrm{T}} = A^{-1}$ is a solution of the JAD procedure that diagonalizes all covariance matrices:

$$W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = R_{\boldsymbol{s}|c_i} = D_{c_i} \tag{23}$$

for $i = 1, \ldots, M$. Note that $R_{\boldsymbol{s}|c_i} = D_{c_i}$ are diagonal matrices because of the mutual independence of the elements of $\boldsymbol{s}$. In this case we have that

$$\hat{\boldsymbol{x}} = W^{\mathrm{T}}\boldsymbol{x} = W^{\mathrm{T}} A \boldsymbol{s} = \boldsymbol{s}, \tag{24}$$

and the obtained spatial filtering matrix W applied to the EEG/MEG data results in estimates of the underlying independent components (ICs) of the observed data. It remains to be established if, or under which conditions, $W^{\mathrm{T}} = A^{-1}$ is the only matrix that jointly diagonalizes all covariance matrices. This question of uniqueness has been addressed for orthogonal mixing matrices A (or for sphered data) in [1], and for arbitrary mixing matrices in [20]. It turns out that in the context considered here a necessary and sufficient condition for $W^{\mathrm{T}} = A^{-1}$ to be the unique joint diagonalizer (up to scaling and permutations) of $R_{\boldsymbol{x}|c_i},\ i = 1, \ldots, M$, is that the matrix

$$S := \begin{bmatrix} \sigma^2_{s_1|c_1} & \cdots & \sigma^2_{s_N|c_1} \\ \vdots & \ddots & \vdots \\ \sigma^2_{s_1|c_M} & \cdots & \sigma^2_{s_N|c_M} \end{bmatrix} \tag{25}$$

has no pair of proportional columns, i.e, that for no pair of ICs the variances covary across conditions. Under these conditions any JAD procedure that converges, i.e., that jointly diagonalizes all covariance matrices, returns a matrix W that,

if applied to the observed EEG/MEG data, returns (scaled and permuted) estimates of the underlying ICs according to (24). While it is not possible to ensure a-priori that the variances of no pair of ICs covary across conditions, we consider this to be highly unlikely. Consequently, JAD of the EEG/MEG covariance matrices conditioned on the class labels can be considered an implementation of ICA.

It then remains to be shown that the unmixing matrix W indeed separates the signal and the noise subspace, and a procedure has to be derived that identifies those columns of W that extract the signal subspace. Now note that if the ICA model (21) and the uniqueness condition hold a matrix W obtained by JAD that diagonalizes all EEG/MEG covariance matrices conditioned on class labels implies that

$$\begin{aligned} I(c,\boldsymbol{x}) &= I(c,W^{\mathrm{T}}\boldsymbol{x}) = I(c,\boldsymbol{s}) = \sum_{i=1}^{N} I(c,s_i) \\ &= \sum_{i=1}^{L} I(c,s_i) = \sum_{i=1}^{L} I(c,\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}). \end{aligned} \tag{26}$$

with $\boldsymbol{w}_i$ the i^{th} column of W. Here, the first equality follows from the fact that mutual information is invariant under invertible transformations [5], the second equality follows from (24), the third equality follows from the mutual independence of the elements of $\boldsymbol{s}$, and the fourth equality from our assumption that only the first L sources provide information on c. Hence, all information in $\boldsymbol{x}$ on c is contained in the first L ICs, i.e., the signal subspace is spanned by the first L ICs. This establishes that under the stated assumptions ICA does indeed separate the signal and the noise subspace. In practice, the L spatial filters that extract the signal subspace are then chosen as those L columns of W with the highest mutual information $I(c,\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x})$. This term can be easily evaluated, and thus the optimal spatial filters identified, according to the approximation of mutual information (15) derived in Section III-B.

To summarize the results of this section, we have pointed out that the problem of finding a set of optimal linear spatial filters can be interpreted as a subspace identification problem, with the signal subspace defined as the space spanned by all sources that provide information on the intention of the user of the BCI. We have further shown that JAD of the EEG/MEG covariance matrices conditioned on class labels is an implementation of ICA, which is capable of separating the signal and the noise subspace under the stated assumptions and thus provides a suitable set of potential spatial filters. We then showed how the derived approximation of mutual information can be used to identify those spatial filters that provide most information on the intention of the user of the BCI. We have thereby eliminated the need for heuristics in and provided a sound theoretical basis for spatial filtering in the context of non-invasive BCIs with multi-class paradigms. Finally, multi-class ITFE, as derived here, allows incorporating unequal class probabilities by choosing those spatial filters that maximize mutual information in (15). For convenience, the complete procedure of multi-class ITFE is summarized in Fig. 3.

IV. Experimental Results

We will now present experimental results from a four-class motor imagery paradigm supporting the results of the previous section. The purpose of this section is to compare pre-processing by multi-class ITFE with multi-class CSP, i.e., comparing the effect of choosing spatial filters that maximize mutual information versus choosing spatial filters according to the heuristic presented in Section II-B.

The data we use was recorded in the Laboratory of Brain-Computer Interfaces at the Technische Universität Graz for the third BCI Competition (data set IIIa), and is available at *http://ida.first.fraunhofer.de/ projects/bci/competition_iii/*. A detailed description of the recording procedure can be found in [3]. Three subjects (k3b, k6b, and l1b) were asked to perform motor imagery of the left/right hand, one foot, or tongue. Each trial lasted for seven seconds, with the motor imagery performed during the last four seconds of each trial. During the experiment EEG was recorded at 60 channels, using the left mastoid as reference and the right mastoid as ground. The sampling rate was 250 Hz, and the data was filtered between 1 and 50 Hz with a notchfilter on. For subjects k6b and l1b a total of 60 trials per condition were recorded, and for subject k3b 90 trials per condition were recorded. Four trials of subject k6b had to be discarded due to missing data. Otherwise no trials were rejected and no artifact correction was performed.

For each subject the following evaluation procedure was performed. First, all data was filtered with a fifth-order butterworth filter with cut-off frequencies 5 and 35 Hz. Then, the four seconds of each trial in which motor imagery was performed were extracted. Afterwards, the data was randomly partitioned into a training and a test set. The size of the training set was varied between 10 and 50 trials in steps of ten trials for subjects k6b and l1b, and between 10 and 80 trials for subject k3b. The covariance matrices of all four conditions were computed using only data of the training set. JAD was performed on the obtained covariance matrices using the algorithm presented in [25], and the L optimal spatial filters were chosen according to a) the heuristic presented in Section II-B (multi-class CSP), b) the procedure described in Fig. 3 (multi-class ITFE), and c) multi-class ITFE with evaluation of the mutual information of class labels and extracted EEG components by numerical integration as described in Section III-B. Note that while procedure c) is feasible due to the knowledge of $p(\hat{x})$ in (11), it is undesirable from a practical point of view due to increased computational complexity. For multi-class ITFE equal class probabilities were assumed. Note that the choice of L is a problem of model identification that is beyond the scope of this article. We arbitrarily chose $L = 8$. The spatial filters obtained by procedures a) - c) were then applied to the training- and test data sets. This resulted in eight-dimensional signals for each trial of the test and training data set. Features were then computed by extracting 15 frequency bands of 2 Hz width ranging from 5 to 35 Hz using a fifth-order butterworth filter, and computing the sample variance in each frequency band for each of the extracted EEG/MEG components. This resulted in a 120-dimensional feature vector for each trial. The feature vectors of

Input: Covariance matrices $R_{\boldsymbol{x}|c_i},\ i = 1, \ldots, M$

1) Perform joint approximate diagonalization s.t. $W^{\mathrm{T}} R_{\boldsymbol{x}|c_i} W = D_{c_i},\ i = 1, \ldots, M$ (e.g., with the FFDiag-algorithm [25]).
2) For each column $\boldsymbol{w}_j,\ j = 1, \ldots, N$, of W scale $\boldsymbol{w}_j$ s.t. $\boldsymbol{w}_j^{\mathrm{T}} R_{\boldsymbol{x}} \boldsymbol{w}_j = 1$ and estimate mutual information according to

$$I(c, \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}) \approx -\sum_{i=1}^{M} P(c_i) \log \sqrt{\boldsymbol{w}_j^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}_j} - \frac{3}{16} \left(\sum_{i=1}^{M} P(c_i) \left((\boldsymbol{w}_j^{\mathrm{T}} R_{\boldsymbol{x}|c_i} \boldsymbol{w}_j)^2 - 1 \right) \right)^2 .$$

3) Choose the L columns of W with highest mutual information.

Output: Pre-processing matrix $W \in \mathbb{R}^{N \times L}$

Fig. 3. Multi-class Information Theoretic Feature Extraction

the training set were then used to train four logistic regression classifiers with L1-regularization, since this classifier is known to perform well in the presence of many irrelevant features [14]. Each classifier was trained on one versus all other conditions, with a regularization parameter chosen manually as 0.1. To infer the class label of trials in the test data set the continuous output of each classifier was computed for all trials. The output of each logistic regression classifier ranges from zero to one, representing the probability of a certain class. Then, the class label attached to each trial was chosen as the index of the classifier with maximum output for that trial. For each partitioning of the data in a test- and training set this procedure was repeated 20 times.

The resulting classification accuracies for all subjects and evaluation procedures a) and b) are shown in Fig. 4, with the thin horizontal line indicating chance level. Results of evaluation procedure c) are not shown, since on average these differed from procedure b) by only 0.4%. This experimentally validates the accuracy of the derived approximation of mutual information (15) in the context of non-invasive BCIs. While the classification accuracies vary significantly across subjects, it is evident that multi-class ITFE outperforms multi-class CSP by far, with a mean increase in classification accuracy of 23.4%. This increase is especially significant for subject l1b, for which multi-class CSP performs only slightly above chance. With spatial filters chosen according to multi-class ITFE, subject k3b even achieves classification accuracies of about 95%.

It should be pointed out that the classification accuracies achieved here do not, with the exception of subject k3b, compare favorably with the best entries to the BCI competition III for the same data set [19]. We attribute this to the fact that while the algorithms submitted to the third BCI competition were extensively tuned, there are several parameters in the procedure presented here that were determined arbitrarily. For example, it is well known that computing spatial filters in narrow frequency bands, tuned according to the most reactive frequency bands for each subject, significantly improves classification accuracy as opposed to selecting a rather broad frequency band as done here. Furthermore, the number of spatial filters retained was chosen arbitrarily as eight for all subjects and training sets, and the regularization parameter of the classification procedure was also determined manually and constant for all subjects. All of these parameters could be tuned using methods such as cross-validation on the training set to achieve higher classification accuracies. This, however, is not the point of this study. We chose a rather simple classification procedure to emphasize the importance of choosing the optimal spatial filters: while the total set of spatial filters is identical for multi-class CSP and multi-class ITFE, choosing a subset of filters that maximize mutual information according to the procedure of multi-class ITFE summarized in Fig. 3, as opposed to the procedure proposed in [6], leads to a significant increase in classification accuracy.

V. Conclusions

In this article, we investigated the Common Spatial Patterns algorithm for spatial filtering in non-invasive Brain-Computer Interfaces in the framework of Information Theoretic Feature Extraction. We showed that for two-class paradigms CSP maximizes (an approximation) of mutual information of extracted EEG/MEG components and class labels. This provides a previously unknown link between CSP and the minimal achievable error probability of a BCI for a given data set. In the context of multi-class paradigms, we pointed out that finding a set of optimal linear spatial filters can be understood as a subspace identification problem. We further pointed out that multi-class CSP solves this problem by ICA and subsequent identification of the most suitable spatial filters by means of a heuristic. We could eliminate the need for this heuristic by showing that ICA is capable of separating the signal and noise subspace, and providing a procedure for choosing a subset of spatial filters that maximize (an approximation) of mutual information of class labels and extracted EEG/MEG components. This procedure, termed multi-class ITFE, was shown to outperform multi-class CSP by on average 23.4% on the data set IIIa of the third BCI competition. Furthermore, the framework of multi-class ITFE allows incorporating prior class probabilities into the feature extraction process. Finally, the accuracy of the employed approximation of mutual information was validated experimentally in the context of non-invasive BCIs by also considering a numerical integration of mutual information in multi-class ITFE. Since the obtained classification results did on average not differ by more than 0.4% between using the analytic approximation and the numerical integration of mutual information, this supports the claim, made at the end of Section III-B, that EEG/MEG sources providing information on the intention of the user do not deviate much from Gaussianity.

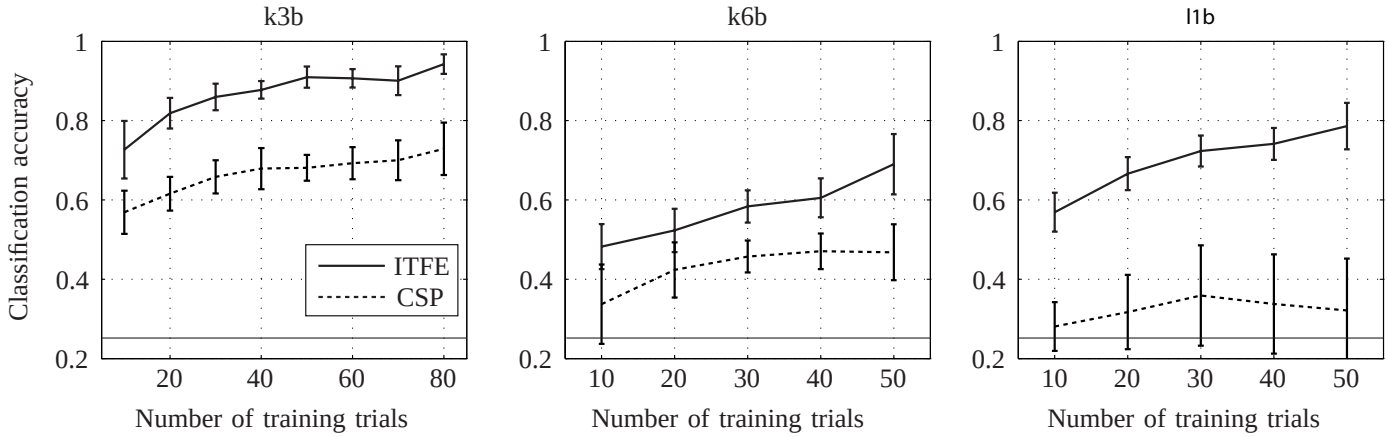

Fig. 4. Classification accuracies of subjects k3b, k6b, and l1b as a function of the number of training trials for multi-class ITFE and multi-class CSP. The thin horizontal line indicates chance level.

The derived approximation of mutual information can hence be considered sufficiently accurate in the context considered here.

As already pointed out in Section III, most recent studies on ITFE consider a non-parametric setting. The primary advantage of non-parametric approaches is their generality, i.e., no restrictions on the probability distribution of the observed data have to be imposed. However, non-parametric approaches are computationally intensive and often require substantial amounts of training data. This is in contrast to the parametric approach considered in this work. By incorporating informed restrictions on the distribution of the EEG/MEG sources a computationally simple and effective feature extraction algorithm could be derived. Furthermore, please note again that in the context considered here the assumption of conditionally Gaussian distributed EEG/MEG sources is no limitation. As a long as only variance changes are used as features for classification, no information that could improve the classification procedure is lost by already discarding higher-order information in the feature extraction process. Notwithstanding this argument, it would indeed be very interesting to investigate whether information on the intention of the user of a BCI is encoded in higher moments of EEG/MEG sources.

In summary, spatial filtering for non-invasive BCIs has evolved to a point where even for multi-class paradigms high classification accuracies have become possible. While this constitutes an important step away from toy problems to real-world BCI applications, there are still several problems to be addressed. One significant problem, that has not been addressed here and from which all supervised spatial filtering algorithms such as multi-class CSP and multi-class ITFE suffer, is overfitting. If strong artifacts are present in the recorded data, these algorithms tend to train on the artifacts instead on pattern changes in the EEG/MEG data intentionally induced by the user of the BCI. For two-class paradigms, unsupervised approaches to spatial filtering such as beamforming have already been shown to achieve classification accuracies comparable to CSP while being robust against artifacts [10]. It remains to be seen if supervised spatial filtering algorithms can be rendered more robust to such artifacts, e.g., by regularization [7] or by logistic regression [21], or if unsupervised approaches also provide viable alternatives for multi-class paradigms.

REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. IEEE Transactions on Signal Processing, 45(2):434–444, 1997.

[2] G. Blanchard and B. Blankertz. BCI competition 2003 - data set IIa: Spatial patterns of self-controlled brain rythm modulations. IEEE Transactions on Biomedical Engineering, 51(6):1062–1066, 2004.

[3] B. Blankertz, K.R. Mueller, D. Krusienski, G. Schalk, J.R. Wolpaw, A. Schloegl, G. Pfurtscheller, J.R. Millan, M. Schroeder, and N. Birbaumer. The BCI competition III: Validating alternative approaches to actual BCI problems. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(2):153–159, 2006.

[4] P. Comon. Independent component analysis, a new concept? Signal Processing, 36:287–314, 1994.

[5] T.M. Cover and J.A. Thomas. Elements of Information Theory. Wiley, 2006.

[6] G. Dornhege, B. Blankertz, G. Curio, and K.R. Mueller. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. IEEE Transactions on Biomedical Engineering, 51(6):993–1002, 2004.

[7] J. Farquhar, N.J. Hill, T.N. Lal, and B. Schoelkopf. Regularised CSP for sensor selection in BCI. In Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course, pages 14–15. Verlag der Technischen Universitaet Graz, Graz, 2006.

[8] M. Feder and N. Merhav. Relations between entropy and error-probability. IEEE Transactions on Information Theory, 40(1):259–266, 1994.

[9] K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. Journal of Machine Learning Research, 5:73–99, 2004.

[10] M. Grosse-Wentrup, K. Gramann, and M. Buss. Adaptive spatial filters with predefined region of interest for EEG based brain-computer-interfaces. In B. Schoelkopf, J. Platt, and T. Hoffman, editors, Advances in Neural Information Processing Systems 19, pages 537–544. MIT Press, Cambridge, MA, 2007.

[11] A. Hyvaerinen, J. Karhunen, and E. Oja. Independent Component Analysis. John Wiley & Sons, 2001.

[12] Z.J. Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. Electroencephalography and Clinical Neurophysiology, 79:440–447, 1991.

[13] S. Lemm, B. Blankertz, G. Curio, and K.R. Mueller. Spatio-spectral filters for improving the classification of single trial EEG. IEEE Transactions on Biomedical Engineering, 52(9):1541–1548, 2005.

[14] A.Y. Ng. Feature selection, L_1 vs.L_2 regularization, and rotational invariance. In Carla E. Brodley, editor, Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada, July 4-8. ACM, 2004.

[15] L. Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. Journal of Machine Learning Research, 4:1261–1269, 2003.

[16] L.C. Parra, C.D. Spence, A.D. Gerson, and P. Sajda. Recipes for linear analysis of EEG. Neuroimage, 28:326–341, 2005.

[17] J.C. Principe, D. Xu, Q. Zhao, and J.W. Fisher III. Learning from examples with information theoretic criteria. Journal of VLSI Signal Processing, 26(1-2):61–77, 2000.

[18] H. Ramoser, J. Mueller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Transactions on Rehabilitation Engineering, 8(4):441–446, 2000.

[19] A. Schloegl. Results of the BCI competition 2005 for data set IIIa and IIIb. Technical report, Institute for Human-Computer Interfaces - BCI Lab, University of Technology Graz, Austria, 2005. available at http://www.dpmi.tu-graz.ac.at/~schloegl/publications/TR_BCI2005_III.pdf.

[20] J.M.F. ten Berge. On uniqueness in CANDECOMP/PARAFAC. Psychometrika, 67(3):399–409, 2002.

[21] R. Tomioka, K. Aihara, and K.R. Mueller. Logistic regression for single trial EEG classification. In B. Schoelkopf, J. Platt, and T. Hoffman, editors, Advances in Neural Information Processing Systems 19, pages 1377–1384. MIT Press, Cambridge, MA, 2007.

[22] R. Tomioka, G. Dornhege, G. Nolte, K. Aihara1, and K.R. Mueller. Optimizing spectral filters for single trial EEG classification. In Lecture Notes in Computer Science, pages 414–423. Springer, Berlin/Heidelberg, 2006.

[23] K. Torkkola. Feature extraction by non parametric mutual information maximization. Journal of Machine Learning Research, 3:1415–1438, 2003.

[24] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain-computer interfaces for communication and control. Clinical Neurophysiology, 113(6):767–791, 2002.

[25] A. Ziehe, P. Laskov, G. Nolte, and K.R. Mueller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. Journal of Machine Learning Research, 5:777–800, 2004.