Brain electrical activity discriminant analysis using Reproducing Kernel Hilbert spaces

Cristian Alejandro Torres Valencia

Department of Engineering

Universidad Tecnológica de Pereira

This dissertation is submitted for the degree of

Doctor of Engineering

Automatics Research Group

2020

I would like to dedicate this thesis to everyone that was close to me in these years . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 60,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Cristian Alejandro Torres Valencia
2020

Acknowledgements

Thanks to all of you who have been part of my life during this difficult times.

Thanks to my advisor and colaborators.

I would like to acknowledge my source of funding given by Convocatoria 647 - 2014 from Administrative Department of Science, Technology and Innovation of Colombia (COLCIENCIAS).

Abstract

A deep an adequate understanding of the human brain functions has been an objective for interdisciplinar teams of scientists. Different types of technological acquisition methodologies, allow to capture some particular data that is related with brain activity. Commonly, the more used strategies are related with the brain electrical activity, where reflected neuronal interactions are reflected in the scalp and obtained via electrode arrays as time series. The processing of this type of brain electrical activity (BEA) data, poses some challenges that should be addressed carefully due their intrinsic properties. BEA in known to have a non-stationaty behavior and a high degree of variability dependenig of the stimulus or responses that are being adressed.

The main idea behind BEA precessing is to extract relevant information from the data that could be easily vinculated with especific conditions i.e. affective states, cognitive status and motor responses. From the first stage of preprocessing, feature extraction, classification and interpretation of the results, there is a high number of works with different proposals to reach the goal. However, there is an evident lack of consistency and in most of the cases, a low generalization capability among the state-of-art.

In this thesis, we propose a BEA processing scheme based on enhancing some properties of multichannel data and using space projections by means of reproducing kernel Hilbert spaces (RKHS). The proposal is divided into three different frameworks, going from the most straightforward one, with a feature extraction stage on the frequency domain by means of the power spectral densitiy from different brain rythms and a classification stage with a relevant feature selection and support vector machine (SVM) decision function. The second framework is based on spatial dependencies analysis from the multichannel data follow by relevant information selection and classification using RKHS. Finaly, a framework based on spatio-temporal dependencies using both discriminative an generative models is implemented.

The introductory chapter gives an overview about BEA, particularly about the multichannel data that is acquired by Electroencephalography. The remaining chapters depicts each one of the frameworks mentioned before, that gives the final scheme for BEA analysis

using RKHS. Each chapter contains the results of the frameowrk tested on two publicily available datasets, one for emotional stimulus and the other for motir imagery experiments. A final chapter of considerations and remarks from each framework is included also.

Contents

List of Figures

List of Tables

Abreviations

BEA	Brain Electrical Activity
CNS	Central Nervous System
EPSP	Excitatory Postsynaptic Potential
IPSP	Inhibitory Postsynaptic Potential
EEG	Electroencephalography
MEG	Magnetoencephalography
MRI	Magnetic Resonance Imaging
CT	Computed Axial Tomography
RKHS	Reproducing Kernel Hilbert Space
BCI	Brain Computer Interfaces
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
KNN	K-th Nearest Neighbor
NN	Neural Networks
SVM	Support Vector Machines
HMM	Hidden Markov Models
MSE	Mean Squared Error
SNR	Signal to Noise Ratio
ICA	Independent Component Analysis
PAA	Period Amplitude Analysis
DFA	Detrended Fluctuation Analysis
PSD	Power Spectral Density
DBN	Deep Belief Networks
SEN	Spectral Entropy
STFT	Short Time Fourier Transform
CKA	Centered Kernel Alignment
FC	Functional Connectivity

PDF	Probability Density Function
CRP	Cross Recurrence Plots
RFE	Recursive Feature Elimination
MFE	Margin-Maximizing Feature Elimination
RBF	Radial Basis Function
PCA	Principal Component Analysis
GSR	Galvanic Skin Response
HR	Heart Rate
CA	Classification Accuracy
Nf	Number of features
COH	Coherence
TGMA	Time Generalized Measure of Association
MI	Mutual Information
FCV	Functional Connectivity Variability
LMS	Least Mean Square
KLMS	Kernel Least Mean Square
QKLMS	Quantized-Kernel Least Mean Square
GP	Gaussian Process
MOGP	Multi-output Gaussian Process
MOSM	Multi-output Spectral mixture
NLL	Negative Log-Likelihood
ALRA	Adaptive learning-based relevance analysis

Chapter 1

Introduction

The brain is by far the most complex organ in the human body and has been the object of continuous research. Different studies have been mainly focused on the possible mapping of electrical activity in specific brain regions to physiological and cognitive processes [41][52]. Moreover, the understanding and modeling of brain functions are directed by the analysis of brain electrical activity (BEA) that can be acquired with brain imaging techniques. Some methodologies such as the Electroencephalogram (EEG), Magnetoencephalogram (MEG), Magnetic Resonance Imaging (MRI), functional MRI (fMRI), Computed axial tomography (CT), among others, have been developed to improve the amount of data that can be studied via the acquisition of BEA [54].

However, there are some advantages and issues related to each specific methodology. For example, the MRI and fMRI schemes are known to have an invasive acquisition protocol, a considerable high cost, and also for the MRI a low temporal resolution [41]. Moreover, the CT acquisition scheme involves the subject exposition to x-rays that could be leading to undesired second effects; also, the related cost is very high and very low temporal resolution [76]. Finally, MEG and EEG schemes pose some advantages in comparison to the methodologies above, their high temporal resolution, lower cost, and lesser invasive protocol are properties that make these technologies more selected for the study of BEA [6]. Nevertheless, the MEG technology has a higher cost, is less portable and needs some specialized locations to perform the recordings acquisition, so, the variety of cognitive and physiological scenarios that can be tested is restricted [76]. Then, in general terms, the preferred scheme for BEA analysis is the EEG. Although there are evident advantages of using EEG, the information decoding for representation and recognition tasks poses a set of challenges that should be solved to design a reliable methodology for BEA processing [66].

The modeling of the acquired data at the original domain space is not a straightforward

problem to solve. The BEA data acquired from EEG is known to have a high no-stationary behavior, as a result of the high quantity of processes that takes place inside the human brain to perform a particular task [52][6]. The EEG temporal variations that are obtained under different conditions or stimulus has qualities of a random process and their statistical moments are known to be unstable [52]. Some quasi-stationary approximations have been used by assuming Gaussianity in windowed versions of the data; however, those conditions are only valid for resting mind state and not for physical or mental activity [41][6]. Additionally, EEG recordings correspond to conducted electrical activity to the brain cortex from all the possible brain sources, and their variation across different brain regions has a behavior that is not easily model [87]. This activity is known to have a low spatial resolution due to the high amount of potential sources but a limited number of electrodes [77]. The mixing property of the BEA into EEG recordings poses a challenge on the selection of a strategy that allows representing the data efficiently; there is a high quantity of information that is repeated among EEG channels and has to be carefully addressed [6][76].

Classical processing methodologies that deal with non-stationary behavior are related to short time representations, frequential decomposition, and spectral-domain analysis [6] [102]. However, these strategies treat each channel as an independent process, so the processed information could be repeated across extracted patterns and also resulting in high dimensionality [66][6][70]. Additionally, decomposition of each channel into frequency bands (alpha, beta, theta, delta, and gamma) is known to allow the specialist to detect some known patterns that are related to specific BEA [76]. Finally, the aforementioned decomposition derives in a higher amount of signals to process, and more influence of noise for some specific frequencies with eye and face muscles artifacts [41][76]. So, although some relevant patterns could be found in this data representations [37], these techniques result in data that contains redundant information in their extracted latent variables that affect the performance of further classification stages [52]. Besides producing redundant information, the independent analysis of EEG channels results in a large number of features deriving in the well known “curse of dimensionality” issue. This issue is related to the problem where there is no sufficient ammount of data to properly train a predictive model with than high number of features [31]. The problem of high dimensionality data has been approached by means of strategies that develop the selection or elimination of features, but some of those strategies compromises directly the results interpretability [31]; and, due the wide spectrum of possible representations, finding an optimum feature set for classification purposes is not an easy task to solve [47].

On the other hand, due to the low spatial resolution of the EEG, some recent method-

ologies for BEA representations include the assessment of spatial dependencies between channels [57]. These methods are categorized as brain connectivity analysis, and their use is motivated by addressing brain functioning as a complex system with relationships between their regions and not as independent processes [27]. The use of brain connectivity measures is motivated by brain function principles, in which segregated regions of the brain are integrated to develop a specific task to produce a specialized information processing [79]. Different schemes that allow to include spatial dependencies in BEA are related to the computation of structural, functional, and effective connectivity which reflects the particular properties of the brain functioning [27]. The structural connectivity characterizes the biological paths that the information follows inside the brain, but it only can be measured with complex, expensive, and time-consuming acquisition technologies [57]. The effective connectivity that measures the effects that some brain activity in one region has over another segregated region. However, as well as structural connectivity, it only can be measured from complex acquisition techniques such as diffusion tensor imaging, Magnetic Resonance (MRI), functional MRI, which are discarded for the development of this work [27]. Finally, the functional connectivity characterizes the continuous dependencies between segregated regions of the brain for a specialized information processing [98][57]. There is a wide range of measures that can be applied to code these interdependences, but their performance seems to bee condition-dependent [27], and there is no general consensus on which measure should be applied in every single case [57]. Also, though algorithms based on FC seem to be promising, the variability of the inter-channel dependencies and the selection of the FC measure still pose an open issue. Besides, the assessment success highly depends on the subject at hand that is related to the particular form in which the brain of each person works. Additionally, the BEA known variability found on the recordings of subjects under similar conditions (developing a similar task or under similar stimulus) [68][79] affects the generalization capacity of the classifiers that use data from different representations schemes. Those differences in the BEA across subjects possess a hard challenge in the task of extracting relevant patterns that can be transversally used, even if the process is performed either in the original domain of the EEG data or the versions derived from other representations [58][1].

Finally, the merely spatial dependencies analysis of the EEG data is not enough to satisfy completely processing objectives [19]. It has been stated from some recent studies, that dynamic analysis of these dependencies will give further interpretability of EEG data and allows a more in-depth analysis of the brain functioning phenomena [17][68]. One of the problems of developing discriminative methodologies over BEA is the lack of tempo-

ral references or tags of the recordings in relevant data segments [72]. The data tagging process from the specialists could be very time consuming and is often performed in some individual studies. Moreover, most of the publicly available datasets of BEA within an EEG acquisition scheme are not tagged by a specialist in terms of temporal references; just in some cases where the experiments have short periods of realization i.e., motor imagery, some useful tags can be included within the discriminative processing [37][52][18]. However, in general terms, the analysis has to be performed in long-duration recordings without any temporal references that prevent the characterization from being guided over specific and relevant data. So, a strategy that is capable of analyzing the dynamical changes in the spatial dependences forms BEA data could be the alternative to this lack of temporal references [68][74]. A complete analysis of the changes in the relationships between channels will reduce the need for previous data segments selection and allows a global interpretation of the connectivity along the experiment [98].

From all the depicted problems about BEA recognition, the development of this dissertation is focused on answering the following question: **How to develop a methodology for representation and classification of multichannel BEA data with non-stationary and low spatial resolution constraints by spatio-temporal dependencies analysis and RKHS feature mapping aiming to improve classification accuracy and results interpretability?**

1.1 Justification

The complete set of physiological and cognitive functions of the human body have their origin in the brain, so the brain functions reflected in its electrical activity has a high degree of relevance across different medical, neurological and engineering applications. Although there are some schemes of measuring the BEA from the currents generated by electrical impulses within the brain, some of them are invasive, long time duration and expensive such as Magnetic Resonance Imaging (MRI) or Computerized axial tomography (CT) making them less used. Unlike CT or MRI, the EEG poses a less invasive/expensive scheme with the possibility of real-time analysis [66][76].

Measuring and quantifying the BEA can be useful in some diagnosis and treatment processes of the nervous system. The detection of epileptic sources, mental disorders, emotion recognition and recently, the inclusion of brain-computer interfaces (BCI) are some of the possible applications derived from systems that efficiently process BEA. From all the possible application scenarios, we can find affective computing, detection and treatment of

neurological pathologies, analysis of cognitive task performance and there is a particular BCI application based on the recognition of motor tasks. The processing of BEA for motor tasks could be used in the improvement of life quality for people with body disabilities using it as a component of systems of monitoring and surveillance or the construction of smart prosthesis that can be moved using BEA [6].

Besides the possible applications, there are many different schemes to characterize and quantify the BEA, namely, temporal analysis and frequential analysis among others. Some methodologies belonging to each scheme have proved during years a reliable but not optimal representation of BEA [66][11]. Mainly, from the biological analysis of the brain functioning, the characterization of BEA from brain connectivity measures have been introduced. The human brain is composed of different regions that are dedicated to the development of specific tasks. Those regions can present anatomical-wise o functional-wise connections that can be triggered on to develop a required neural activity. So basically, the brain functions are summed as the integration of segregated regions that perform a determined process [12, 27, 35, 57, 103]. Then, a quantification of those relationships that occurs inside the human brain could bring relevant information in the characterization and recognition of different BEA states. Some measures such as the correlation, cross-correlation, coherence and some measures related to phase synchronization and with information theory have proved to quantify statistical interdependences capacity that describes functional brain relationships [27][57].

On the other hand, when it comes to machine learning algorithms applied to BEA processing, different models have been developed following representations aforementioned. Thus, schemes based on linear and quadratic discriminant analysis (LDA,QDA), nearest neighbors classifiers (KNN), neural networks (NN), support vector machines (SVM) and hidden Markov models (HMM) have all been applied to EEG signals classification tasks. Even though the recognition task is proved to perform under a variety of conditions of EEG acquisition, the majority of works are focused on the classification of motor tasks (under the "motor imagery" scheme), neurological pathologies detection, or in the field of affective computing where the conditions are related to emotional states. Nevertheless, the discriminative power of those schemes is still far from optimal for this type of EEG data representations. The inclusion in recent years of the theory around the Reproducing Kernel Hilbert spaces (RKHS) has produced an improvement in the generalization capacity of discriminative schemes using a feature space of representation that can be infinite-dimensional [67]. By using the RKHS, a high variety of problems could be effectively addressed and performance-improved in pattern recognition problems [52][79].

From the high interest generated around the study of BEA, there can be found a considerable number of works that serve as a platform to develop further methodologies within this scope. From the bare acquisition of EEG multichannel data, to further recognition stages, there are a variety of methodologies that allow solving in some degree the proposed problem. The particular BEA condition that tends to be studied must be induced for the EEG data acquisition. This induction process is carried out by stimulating the subjects with a specific task that triggers on a physiological or cognitive response. With this schemes of stimulation, several databases have been constructed and posted publicly available for the development of works on BEA processing [66]. Similarly, within the neuroengineering field, the use of multichannel data in processing tasks have preferred EEG acquisition technology over other schemes aforementioned. Processing stages of brain source recognition, characterization of BEA during sleep or motor tasks and the final pattern recognition from those conditions, motivate the development of new methodologies that can efficiently process the high quantity of information that is extracted [14].

1.2 Literature Review

Within the EEG data processing field for recognition of BEA patterns, there are several works in the state-of-art that solved a particular problem within a wide spectrum of possible applications. Areas such as affective computing [4], development of brain computer interfaces (BCI) [54], motor tasks recognition from BEA [56] and brain source localization [88] among others, tries to find specific BEA patterns related to a particular condition. From the high number of EEG processing methodologies, a categorization and grouping of the most relevant approaches is presented within this section, going from the initial preprocessing step to the final BEA recognition.

First, on the analysis of strategies for artifact removal from EEG recordings, there is not an only one accepted scheme that works efficiently in every scenario. There are several measures such as mean squared error (MSE), the signal to noise ratio (SNR), the signal to artifact ratio (SAR) among others, that allow quantifying the efficiency of the method for artifact removal in simulated and real EEG signals [84]. Despite ICA-based algorithms are preferred over other algorithms in several applications, the performance can decrease depending on the measurement artifacts and the type of EEG signals. Nevertheless, the preprocessing stage is not the core of this proposal, since the main objective is related to information extraction and interpretation for BEA recognition. Therefore, a second step related to EEG data representation is needed for further processing stages. There are some

categories in which the methodologies for data representation could be grouped [52]. At first, works that search for informative patterns within the temporal domain are grouped as temporal representations, where instantaneous statistical measures are employed by assuming each EEG channel as an independent random process. Statistical moments such as mean, variance, kurtosis, among others can be computed in combination with other temporal representations such as zero crossing and the period amplitude analysis (PAA) to represent the EEG data in temporal domain [52] [41]. Similarly, the Hjorth parameters [58], based on the signal variance of derivatives and the Detrended fluctuation analysis (DFA) [48] have been applied to representation of BEA within the temporal domain. Although these techniques have been employed efficiently into classic signal processing applications, the particularities of EEG signals does not allow a high performance from them. The non-stationarity behavior of the data, which is not adequately modeled from instantaneous statistics, and the decoding capacity of the multichannel EEG data within this scheme is considerably low.

Secondly, a spectral representation of BEA data can be depicted, where the analysis of the EEG is performed in different frequency bands that are well known as the EEG fundamental rhythms, namely, alpha, theta, beta, delta and gamma [66]. Then, measures such as the power spectral density (PSD) can be computed over each frequency bands to obtain some BEA descriptors of Alzheimer patients [89], features for motor imagery classification tasks [64] and within emotion elicitation experiments [37]. Similarly, with the spectral representation. Other features such as the spectral entropy (SEN) [97] have also been used in the discrimination of BEA during motor imagery tasks. Additionally, the use of digital filters based on auto-regressive (AR) models [100], have been included as parametric methodologies for spectral representations of EEG data in cognitive states recognition. However, the high dimensionality of data affects the performance of this methodologies, also, the signals remain treated as independent random processes and quasi-stationary criterion must be assumed. Additionally, the optimization of parameters in the filter based models adds another variability degree that reduces the flexibility of these models [89][52]. In third place, we can find time-frequency or short time representations, where an initial stage of signal windowing followed by frequency bands decomposition is applied. Works based in this scheme include the application of the short time Fourier Transform (STFT) [70][72] for emotion recognition, Wavelet transform [100] for motor imagery tasks and the use of waveforms dictionaries and the Matching pursuit strategy [14] for cognitive states classification. However, the decoding capacity of these strategies and the interpretability of characterized data is still low, furthermore, the analysis is performed for each channel independently, so the amount of resulting data is considerable high [82].

With these high dimensional spaces of representation, strategies that propose the selection of relevant information have been applied to EEG extracted data. Works like [99] tries to overcome the emotion recognition problem by a strategy known as Empirical mode decomposition. This allows to divide each channel into various frequency scales signals which are then analyzed by sample entropy (SE). The resulting data corresponds to selected information characterized by the (SE), with lower dimensionality compared with the original data. However this selection was performed only over two channels from the EEG array. Similarly, the use of differential entropy (DE) with the correlation coefficient as a tool for selecting relevant features is proposed in [44]. Furthermore, in [8] a extraction of several spectral domain features, time domain features, and statistical measures from EEG is performed. Then, a scheme for feature selection employing the minimum-redundancy-maximum-relevancy is applied, to evaluate the correlation between each feature to the corresponding classification variable. Moreover, methodologies based on wrapper methods such as feature elimination, have been applied in the context of emotion recognition form BEA. Works like [34] and [16] tried a SVM-based method for feature elimination while searching for patterns related with region activation under emotional stimulus. The more relevant outcome from all these works is that an adequate feature selection stage will provide an improved dataset that can be more efficiently classified despite any specific application.

Temporal and spectral features have been used in the classification of BEA patterns by using classical machine learning approaches, for example, the use of linear discriminant analysis (LDA) from spectral features have been developed in works like [104][13][55][64]. Moreover, works based on Deep belief networks (DBN) have been applied to the recognition of emotions using BEA data from fundamental rhythms representation [101]. Other works using neural networks and deep learning approaches includes [102][101][75] and ANOVA-based statistical analysis in works like [89]. Acceptable results in terms of classification accuracy have been achieved through these methods, however, generalization problems associated with training tests overfitting and low interpretability qualities have been reported[13]. On the other hand, a clustering methodology have been also applied to EEG in order to recognize EEG segments and the comprehensive grouping of segments that presents similar behavior of BEA [33]. This clustering method is affected directly for the data dimensionality in the representation space and the each cluster center initialization.

In addition to the short time models for EEG representation, in recent years the analysis of brain connectivity has provided a tool for BEA processing more related to the physiologic processes that happen inside the brain [27]. Functional connectivity (FC) have been widely used from EEG signals as an attempt to include some spatial information due to the consid-

erably low spatial resolution [66][57]. One of the properties of FC analysis is that the signals are no longer treated as independent processes but as part of a network that shares statistical dependences between each pair of channels [27]. In this case, the objective is to find relevant patterns of spatial dependences that allows to recognize particular BEA and giving a deep insight regarding the interpretability of the extracted patterns [57]. Again, this schemes of EEG representation have been used in different applications from Alzheimer and dementia patterns classification to emotional and motor task recognition [27][94]. However, there are a considerable number of FC measures among different works in the state-of-art, and the selection of an optimal measure across different scenarios is not yet defined. Among them, we can find measures such as the correlation, cross-correlation, some other with time/frequency analysis such as the coherence and from the information theory framework measures such as the entropy or the mutual information [57][103]. For example in [5], a study of FC for stress states detection is proposed, the use of coherence and cross mutual information measures with an evaluation of EEG asymmetry via significance probability maps (SPM) produced some remarks about the variations in connectivity patterns related to stressful conditions. Likewise, in [94] a clustering strategy is proposed for FC patterns under conditions of BEA in subjects with dementia or Alzheimer with the Phase lag index (PLI) measure and the use of a minimum spanning trees (MSP) clustering. Emotional classification algorithms have also been developed using FC measures. In [42], the correlation index, coherence, and the phase synchronization index were applied to emotion recognition of positive and negative stimulus. A selection of connectivity features was performed used ANOVA, and then a quadratic discriminant analysis (DQA) classifier implemented in the recognition process.

However, the inclusion of feature spaces based on transformations via Kernel functions as the basis of the well known Reproducing Kernel Hilbert spaces (RKHS) framework, allows formulating the classification procedure from spectral features in BEA in a more efficient way [2]. RKHS are indeed the basis of a widely used support vector machine (SVM) classifier that has been applied to EEG classification with spectral features [13][72][70][100][64]. In [82] a feature elimination scheme based on SVM over EEG data was proposed to find a optimal subset of features to recognize emotional states, the results were higher than other works that have tested similar methodologies over an specific dataset. Later, the combination of FC measures with the RKHS framework to detect spatiotemporal variations in the EEG signals have been proposed. In general terms, the use of SVM is widely accepted across different applications. In [95], three connectivity measures with an SVM classifier were used to determine discrete emotional states from Parkinson's disease patients, finding important remarks on poor classification performance for the PD patients

as clear evidence of the disease influence in connectivity patterns. A similar study on the differentiation of healthy and affected with the major depressive disorder (MMD) subject using FC measures is presented in [53]. The results evidenced a higher performance of the methodology when using SVM as a classifier in comparison with logistic regression and a Naive Bayesian classifier, giving a reliable methodology for depression diagnosis from FC. Another study on FC with applications to emotion recognition is presented in [79], where FC measures were used in combination with a relevance analysis based on centered kernel alignment (CKA) for dimensional emotion classification using SVM. FC has also been applied to studies on motor imagery for BCI development. In [98] an analysis of the fronto-parietal connectivity patterns was developed with the inclusion of fMRI information to evaluate the performance of rehabilitation abilities. Finally, the evaluation of cognitive states has also been addressed using functional connectivity in [67]. However, one of the problems associated with the use of FC into the RKHS framework is the ambiguity in the measure selection that allows a high generalization power in different scenarios under several conditions [68]. Likewise, the variability between subjects is also a problem associated, not only to the connectivity analysis but also for all the other schemes of BEA representation [27].

Although there is an evident improvement on the interpretability of the methodologies that implement FC in comparison to the short time analysis, is evident that the results obtained are far from optimal. So, the dynamic analysis of spatial dependences from FC is proposed to improve the performance of the methodologies of BEA recognition [17]. An initial attempt to include a sort of temporal analysis of FC was presented in [18], where a network functionality-based strategy is designed to quantify the variations of the connectivity networks in resting state BEA. Then, the use of Hidden Markov moddel (HMM) is used to estimate the transition probabilities from graph network measures via FC. Some limitations as the selection of the number of hidden states are reported within this application. Another attempt to introduce a dynamical analysis of FC is proposed in [68] for cognitive states recognition. In this case, the time generalized measure of association (TGMA) within each pair of channels is modeled as a random variable within segmented data windows by a probability density function (PDF) using Parzen estimation with a Gaussian Kernel. For the classification of BEA, the PDFs are compared for consecutive windows with the Cauchy-Schwartz divergence, from the results it can be seen that some relationships between regions could be observed in particular BEA patterns across time variations.

A scheme that summarizes the set of possible EEG data representations and schemes of classification is presented in Figure 1.1.

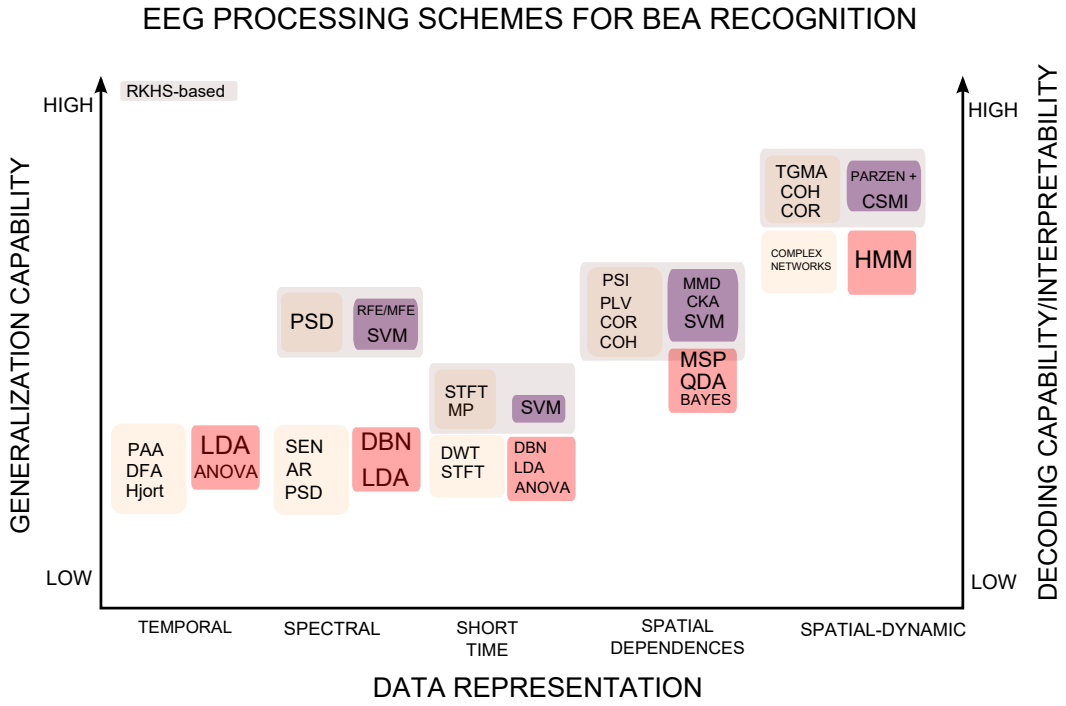


Fig. 1.1 EEG representations methodologies, differentiation between RKHS-based methodologies and by the type of data representation

1.3 Aims

1.3.1 General aim

To develop a representation scheme using RKHS for brain electrical activity discriminant pattern recognition to code relevant information in terms of classification and interpretability of data, related to spatio-temporal dependencies.

1.3.2 Specific aims

1. To develop a methodology of short time representation to code the non-stationarity of BEA data in classification tasks with discriminant models in RKHS.

2. To develop a methodology of data representation within RKHS that allows to identify the relevant spatial dependencies considering subject variability in BEA classification tasks.

3. To propose a methodology of data representation based on RKHS that allows to identify relevant spatio-temporal dependencies to code the non-stationarity and the inter-

dependence between channels in BEA classifcation tasks.

1.4 Outline of the Thesis

From the depicted problems and proposed aims of this thesis, we address the contributions of the research fields involved.

From the perspective of neuroscience, this thesis tries to improve the brain functioning comprehension by means of result interpretability in BEA discriminative scenarios. We explore classical short-time techniques for characterization of EEG-BEA data like the ones proposed in [37] [89], going through more recent techniques such as connectiviy analysis [27]. The variety of possible conditions of BEA that could be analyzed is considerable large, so in this case we are applying the developet methodologies on motor imagery and emotional databases. Those conditions allows an adequate validation of the proposed framework. Finally a new scheme for spectral content analysis is implemented using a particular kernel proposed in [61] within a probabilistic framework.

On the other hand from the machine learning perspective, the contributions of this thesis are mainly related with the selection of relevant information for discriminative purposes. An strategy for RKHS-based selection of relevant features from short-time is proposed using the methodology in [31]. We also proposed an information theory scheme for adaptive learning known as Quantized-Kernel least mean square (QKLMS) for relevant saptio-temporal connectiviy data selection. And finally the design of a probabilistic framework for spatio-temporal analysis using multi-output Gaussian processes. All these methodologies attempt to discriminate the BEA data from both processed and raw data not only assigning a label, but also allowing to assess the possibly associated brain regions involved into the depicted condition. In detail, the remainder of this thesis is strcutured as follows:

- Chapter 2 defines some basic knowledge and fundamental concepts for the rest of the thesis. Also, a general description of the proposed framework following the aims definition.

- Chapter 3 discusses the basis of BEA analysis from EEG data, and a decsription of classical methodologies for its processing. Additionally, the proposed framework of short-time analysis within RKHS is introduced regarding the selection of relevant information for discriminative purposes.

- Chapter 4 introduces the basis of spatial dependencies analysis of BEA data, including the formal definition of few connectiviy measures. Then, the proposed framework for

data selection by implementing centered kernel alignment (CKA) [21] within RKHS is described.

- Chapter 5 presents the basis of the proposed spatio-temporal framework, involving an adaptive information theory algorithm for relevant temporal selection of spatial analysis of BEA [67]. The spatial dependencies analysis uses the proposed time generalized measure of association proposed in [68]. Then, a scheme of spatiotemporal analysis involves the use of a recent proposed kernel [61] within a porbabilistic framework of multiple-output Gaussian processes MOGP [7].
- Finally, Chapter 6 summarizes the main contributions of this thesis and discusses the possible future works.

1.5 Associated Publications and Software

- A journal publication including one of the spatio-temporal methodologies proposed in chapter 5 [83].
- A journal publication including the short-time representations depicted in chapter 2 [82].
- A conference publication including the relevant selection of spatial dependencies depicted in chapter 3 [79].
- A conference publication including the proposed framework for spatio-temporal analysis using an adaptive QKLMS depicted in chapter 4 [80].
- A journal publication (in revision) including the proposed framework within the probabilistic MOSM-GP for discriminative BEA.
- Additional publications of related works of the research field of this thesis during the time of the Ph.D. development includes: Two conferences on works about source localization from EEG [77][81]. A work on deep brain stimulation simulation is included in a Journal [92].

Chapter 2

Fundamentals

This chapter contains the basis of the brain electrical activity (BEA) and reproducing kernel Hilbert spaces (RKHS) needed along the remainder of this thesis.

2.1 Brain Electrical Activity

From the $17-$th week of prenatal development, the neural activity of the human brain begins. It is believed that from this early stage of human life, the status of the whole body is represented by its electrical activity. Then, the capacity of represent this high amount of information, is only possible by processing multiple electrical signals originated in several neural cells. The central nervous system (CNS) is composed of nerve cells and glia cells, which are located between neurons and are responsible of stimulus responses and information transmision. Each nerve cell consists of axon, dendrites and cell bodies, being the latter the responsible for most of the nerve cell metabolism (protein synthesis). Dendrites are conected either to the axons or dendrites of other cells, and through these conections each nerve is associated to aproximately other 10000 nerves by receiving impulses.

The synaptic currents transferred between the junctions of axons and dendrites are the core of the CNS activities. During synapses, the membrane potential changes and travels along the fibres producing inhibitory or excitatory synapses in other neuron cells. Excitatory postsynaptic potential (EPSP) and inhibitory postsynaptic potential (IPSP) are the results of action potentials traveling along the fibres that reaches or not a certain threshold on the postsynaptic neuron respectively. Figure 2.1 shows the changes in the membrane potential and how the current flow during synaptic activation. The portion of these currents that flows through the extracellular space is responsible of field potentials generation. The AP's of most nerves last between 5 and 10 ms with a conduction velocitiy around 1 and

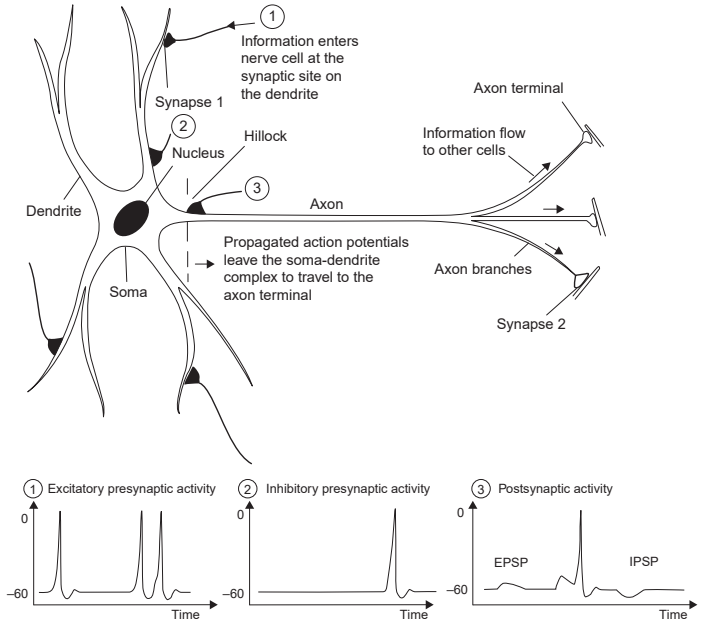


Fig. 2.1 Action potentials in the excitatory and inhibitory presynaptic fibre, lead to EPSP and IPSP in the postsynaptic neuron

$100m/s$, and potential amplitudes approximately -60 to 10 mV. Those potentials are initiated by many different stimulus and targeted for acquisition following an adequate scheme of amplification.

Different technologies allows to measure indirectly the brain electrical activity with few schemes of acquisition. The costs of each technology, the possibility of avoid invasive procedures, and restricted study environments are the main properties that allows to select one scheme over the others. Moreover, the particularities of Electroencephalography (EEG) have been preferred in several research fields that needed a portable, non-invasive, scheme for BEA studies. Figure 2.2 shows a comparison chart using the spatial resolution, time resolution and cost as the parameters for the adequate selection. To the remain of this work, the selected scheme for BEA processing would be the EEG, giving us the desired time resolution needed into the framework tha we want to develop.

2.2 Electroencephalograpy (EEG)

The electroencephalograph technology for BEA acquisitions consists in a set of electrodes placed in the subject head. An EEG signal contains the measurement of flowing currents

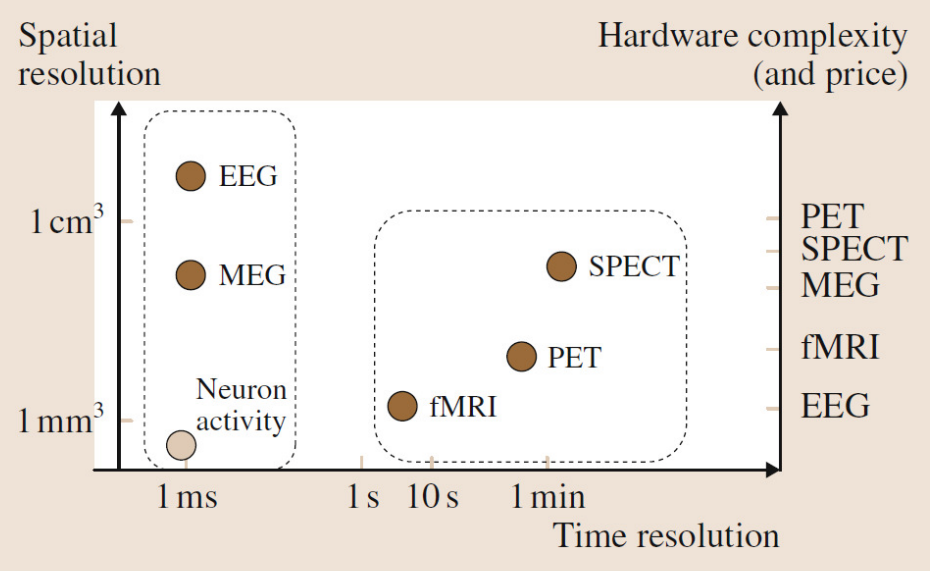


Fig. 2.2 BEA acquisition technologies

during synaptic excitation of many neural cells in the brain cortex. Morevoer, these flowing currents generate a magnetic field measurable by magnetoencephalographic (MEG) devices and a secondary electric field over the scalp that can be measured by means of an EEG array. The summed postsynaptic potentials described earlier, create electric dipoles between the soma and apical dendrites which branch from neurons. Then, for an adequate acquisition of these currents without an invasive scheme, the EEG uses a set of electrodes that are placed in the scalp following an specific distribution. However, some appreciations regarding the layers that compose the human head should be considered. Each layer has different properties as thickness and current resistivity i.e. the skull has a tickness of $0.3-0.7$ cm with a ressitivity of $10-25k\Omega$. Furthermore, the cortex is a thin layer that covers the brain with $0.1-0.3$ cm and an in vivo resisitivity of $50-150\Omega$.

Due the diferences in the electrical properties of each layer, the data obtained in EEG recordings are a non-linear sum of the brain sources. Consequently, the information in each EEG channel could be interpreted as mixed neural activations from several regions. Besides, EEG recordings are also affected by noise generated either within the brain or over the scalp. These particular conditions mean that only large populations of active neurons can generate enough potential to be acquired by the scalp electrodes, so an amplification stage is needed for EEG visualization.

The EEG acquisition scheme has some particularities that promotes its use across the neuroscience field. The possibility of high time resolution by means of the allowed sampling frequency from the electrodes, a portable and low cost technology compared with other

neuroimage techniques.

Let $\chi = \{\boldsymbol{X}_n \in \mathbb{R}^{C\times T}, l_n \in [l_{min}, l_{max}]\}_{n=1}^{N}$ be the BEA data from a particular condition $n = 1,2,\cdots N$ with $\boldsymbol{X}_n$ a set of EEG signals from C channels at T time samples, and l_n the condition label in the range from l_{min} to l_{max}. However, the adjustment of a model $\mathscr{F} : \mathbb{R}^{C\times T} \rightarrow [l_{min}, l_{max}]$ that allows to estimate l_n from $\boldsymbol{X}_n$ is a challenging task. The problems associated with EEG data processing within the original data space such as, non-stationarity, redundant data, variability inter-conditions and inter-subjects reduce the generalization capability of an specific model.

2.3 Reproducing kernel Hilbert spaces

A space $\mathbb{H}$ with an inner product, could be considered as an pre-Hilbert space that has an ortonormal base $\{\boldsymbol{\chi}\}_{k=1}^{\infty}$. If $\mathbb{H}$ is the largest space of vectors for which an infinite set $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ is a basis. Then, vectors not necesarily lying in the original inner product space represented in the form $\boldsymbol{x} = \sum_{k=1}^{\infty} a_k \boldsymbol{x}_k$, are said to be spanned by the basis $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ with a_k representation coefficients. Now, if we define two new vectors $\boldsymbol{y}_n = \sum_{k=1}^{n} a_k \boldsymbol{x}_k$ and $\boldsymbol{y}_m = \sum_{k=1}^{m} a_k \boldsymbol{x}_k$, with $n > m$, the Euclidean distance between vector $\boldsymbol{y}_n$ and $\boldsymbol{y}_m$ could be expressed as:

$$\|\boldsymbol{y}_n - \boldsymbol{y}_m\|^2 = \left\| \sum_{k=1}^{n} a_k \boldsymbol{x}_k - \sum_{k=1}^{m} a_k \boldsymbol{x}_k \right\|^2 \tag{2.1}$$

$$= \left\| \sum_{k=m+1}^{n} a_k \boldsymbol{x}_k \right\| \tag{2.2}$$

$$= \sum_{k=m+1}^{n} a_k^2, \tag{2.3}$$

by the orthonormality condition. Then, the following is needed to make the definition of $\boldsymbol{x}$ meaningful:

1. $\sum_{k=m+1}^{n} a_k \rightarrow 0$ as both $n, m \rightarrow \infty$
2. $\sum_{k=1}^{m} a_k < \infty$

Moreover, a sequence of vectors $\{\boldsymbol{y}_k\}_{k=1}^{\infty}$ is a Cauchy sequence. Then, from the basis $\{\boldsymbol{x}_{k=1}^{\infty}\}$ a vector $\boldsymbol{x}$ can be expanded if, and only if, $\boldsymbol{x}$ is a linear combination of the basis vectors and the $\{a_k\}_{k=1}\infty$ associated coefficients are square summable. Consequently, the space $\mathbb{H}$ is more complete than the starting inner product space.

Definition 1: *An inner product space $\mathbb{H}$ is complete if every Cauchy sequence of vectors taken from the space $\mathbb{H}$ converges to a limit in $\mathbb{H}$; a complete inner product space is called a Hilbert space*

Now, a kernel is a continous, symetric, and positive definite function $\kappa : \mathbb{U} \times \mathbb{U} \rightarrow \mathbb{R}$, with $\mathbb{U}$ the input domain. Then, for a vectorial space $\mathbb{H}$ of real-valued functions of $\boldsymbol{u}$, that are generated from the kernel $\kappa(\boldsymbol{u}, \cdot)$ if there are two functions:

$$h = \sum_{i=1}^{l} a_i \kappa(\boldsymbol{c}_i, \cdot) \tag{2.4}$$

$$g = \sum_{j=1}^{m} b_j \kappa(\tilde{\boldsymbol{c}}_j, \cdot), \tag{2.5}$$

where a_i and b_j are the expansion coefficients and both $\boldsymbol{c}_i$ and $\tilde{\boldsymbol{c}}_j \in \mathbb{U}$ for all i and j. Then, the bilinear form is defined as:

$$\langle h, g \rangle = \sum_{i=1}^{l} \sum_{j=1}^{m} a_i \kappa(\boldsymbol{c}_i, \tilde{\boldsymbol{c}}_j), \tag{2.6}$$

that satisfies the symmetry property, the scaling and distributive property, and the squared norm property. These properties makes the bilinear term $\langle h, g \rangle$ is indeed an inner product. There is one adittional property that follows directly. Specifically, seting $g = \kappa(\boldsymbol{u}, \cdot)$, we obtain:

$$\langle h, \kappa(\boldsymbol{u}, \cdot) \rangle = \sum_{i=1}^{l} a_i \kappa(\boldsymbol{c}_i, \boldsymbol{u}) \tag{2.7}$$

$$= h(\boldsymbol{u}) \tag{2.8}$$

This is the *Reproducing property*. The kernel $\kappa(\boldsymbol{u}, \boldsymbol{u}')$ represents a function of the two vectors $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{U}$ is called a reproducing kernel of the vector space $\mathbb{H}$ if it satisfies:

1. For every $\boldsymbol{u} \in \mathbb{U}$, $\kappa(\boldsymbol{u}, \boldsymbol{u}')$ as a function of the vector $\boldsymbol{u}'$ belongs to $\mathbb{H}$.

2. It satisfies the reproducing property.

If the inner product space $\mathbb{H}$, where the reproducing kernel is defined is also complete, then is called a reproducing kernel Hilbert space (RKHS). The Mercer theorem expresses

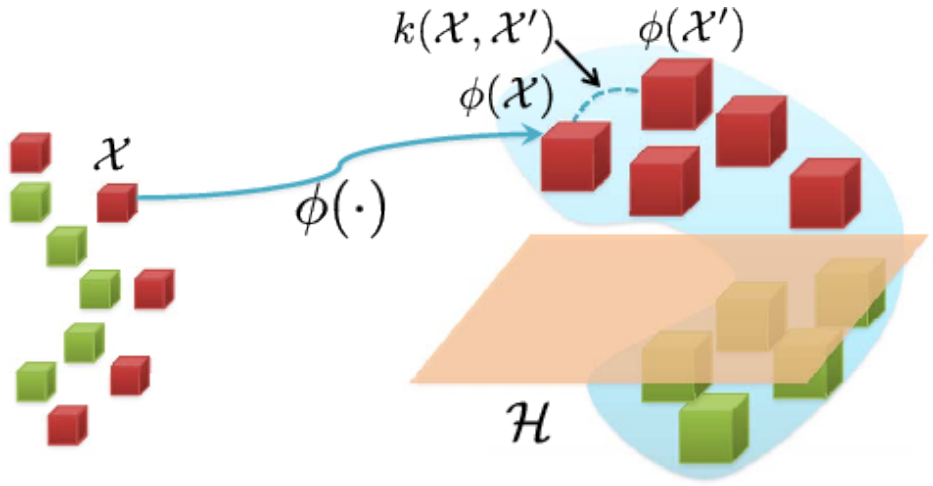


Fig. 2.3 Feature projection from original input space to an RKHS via a kernel function $\kappa(\cdot,\cdot)$

the analytic power of RKHS by stating that any reproducing kernel $\kappa(\boldsymbol{u},\boldsymbol{u}')$ can be expanded as follows:

$$\kappa(\boldsymbol{u},\boldsymbol{u}') = \sum_{i=1}^{\infty} \zeta_i \phi_i(\boldsymbol{u}) \phi_i(\boldsymbol{u}'), \tag{2.9}$$

where ζ_i and ϕ_i are the non-negative eigenvalues and eigenfunctions, respectivley. Therefore, a mapping $\boldsymbol{\varphi}$ could be defined as:

$$\boldsymbol{\varphi} : \mathbb{U} \rightarrow \mathbb{F} \tag{2.10}$$

$$\boldsymbol{\varphi}(\boldsymbol{u}) = \left[\sqrt{\zeta_1}\phi_1(\boldsymbol{u}), \sqrt{\zeta_2}\phi_2(\boldsymbol{u}), \ldots\right] \tag{2.11}$$

The dimensionality of $\mathbb{F}$ is determined by the number of positive eigenvalues, which are infinite in the Gaussian kernel case. In general, $\boldsymbol{\varphi}$ is treated as the feature mapping and $\boldsymbol{\varphi}(\boldsymbol{u})$ is the transformed feature vector lying into the inner product space $\mathbb{F}$, with the implicit outcome that:

$$\boldsymbol{\varphi}(\boldsymbol{u})^T \boldsymbol{\varphi}(\boldsymbol{u}') = \kappa(\boldsymbol{u},\boldsymbol{u}') \tag{2.12}$$

Consequently, it can be stated that $\mathbb{F}$ is the same RKHS induced by the kernel, identifying $\boldsymbol{\varphi}(\boldsymbol{u}) = \kappa(\boldsymbol{u},\cdot)$ which are the bases of the two spaces. In Figure 2.3 it can be seen the process of mapping features into RKHS from the kernel function over the features in the original input space.

2.4 General framework overview

From the presented details about the BEA data and the machine learning framework that is defined in the aims of this thesis, a general overview of the proposed methodology that will be detailed in the following chapters is included in this section. Figure 2.4 presents an scheme of the general framework of BEA processing under the RKHS scope.

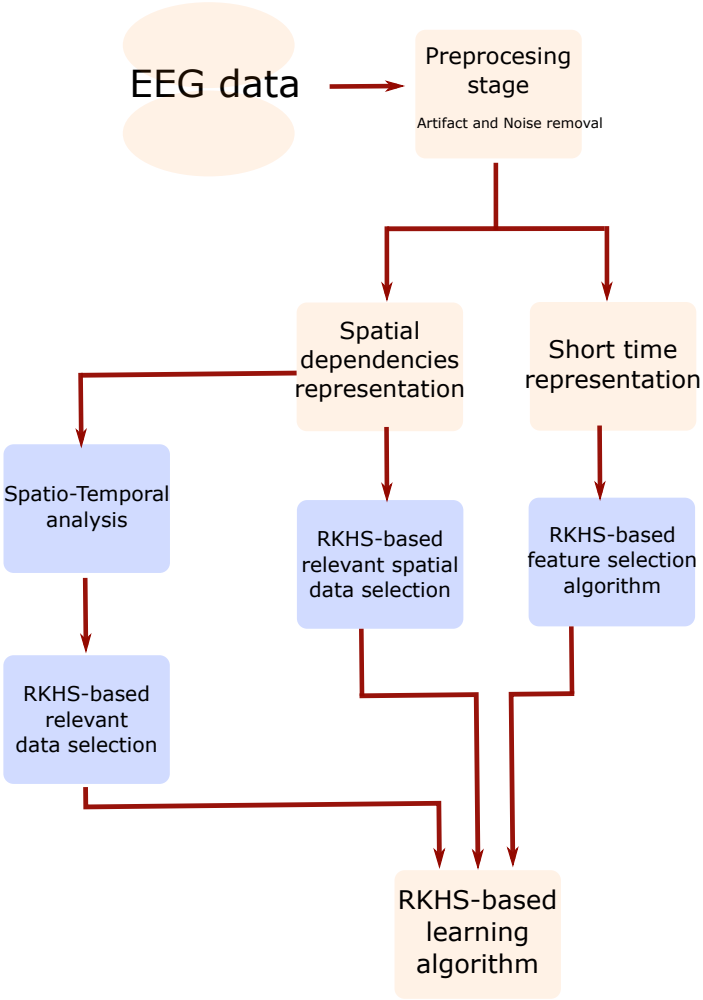


Fig. 2.4 Proposed general framework for the disxcriminant analysis of BEA using RKHS

The first stage is related to the data that is going to be processed, this data is obtained from publicly available databases. The conditions included in the databases are from emotional and motor imagery experiments with an adequate acquisition of EEG data. *The more detailed description of the databases employed in this thesis for testing the methodologies could be found in Appendix B*

Then, a second stage is related to data pre-processing, with some adjustment of signals

in some cases where a de-noising procedure may be needed. After this stage, the framework is divided into three paths, each one corresponding to different BEA processing techniques that at the end converges into a discriminative outcome. The first path of BEA processing is reported in the Chapter 3 and it is related with BEA processing in the short time representation of the signals. The individual extraction of features from EEG channels produces the complete dataset that should be discriminated. The second path on the scheme is related with the spatial dependencies processing of BEA data and it is included in Chapter 4. This particular framework is related with the connectivity analysis of EEG data with the selection of relevant measures and the final RKHS-wise classification. Finally the third path on the scheme is related with the proposal of saptio-temporal BEA analysis and its presented in Chapter 5. In this case, the connectivity measures are adaptively selected to perform an improved condition classification with more interpretability of the results from the selection itself.

Chapter 3

Short-Time processing

The initial proposals for BEA processing techniques, worked by performing a representation of EEG data assuming each channel as a independent random process. From each one of this random processes, several features could be extracted searching for patterns that allows to discriminate BEA. The classical approaches includes time domain and frequency domain processing of each channel. The former techniques includes Hjort parameters [58], detrended fluctuaction analysis (DFA) [48] among others. Although these techniques have been employed efficiently in typical signal processing applications, the particularities of EEG signals does not allow a high performance from them. Moreover, the number of features could increase considerably regarding the number of channels and frequential decompositions. Also, the non-stationarity nature of the EEG recordings, some artifacts and noise affects directly the analysis on the temporal domain.

To avoid the EEG processing in the original temporal domain, a space transformation is also available. Spectral representation of BEA, allows the analysis of the EEG in different frequency bands, namely, alpha, theta, beta, delta, and gamma. Those are known as the fundamental rhythms of BEA, and allows the neurological specialists to search for visual patterns in those frequency bands that are related with BEA in different scenarios [66]. Each cognitive state, neurological condition or disease, or physiological responses, produces different neural activation that can be differentiated in the frequency domain. In this case, the decomposition is used to extract more precise information that helps the discriminative procedure.

In the scope of emotion recognition works, some of the more relevant works are presented in [37] and [72]. Not only for the publication of a very complete affective dataset including physiological responses, but also for proposing a multimodal scheme by using the EEG in combination with other responses of the human body. Regarding the EEG pro-

cessing, each channel was processed independently, computing the power spectral density (PSD) from different frequency bands related with the fundamental rythms. Then the set of features computed from each channel was concatenated in one feature vector with some other descriptors of additional signals. The scheme of classification from this features was the Support vector machine with a division of the emotional experiments from their labels in the arousal and valence dimension. This dataset (DEAP) has ben used for several works of emotion recognition not only from EEG but in this type of multimodal scheme.

The particularities of the Alzheimer disease with a proved increasing in activity in the theta band, allows to propose a strategy of BEA modeling [89]. In this case, the analysis of the PSD of the theta band is used to quantify the decreasing of the activity in particular recordings of Alzheimer patients. The dataset is acquired form people already diagnosed with the disease and with respect to a control group. The results from this framework using an autoregresive method for the PSD estimation allows to conclude that even using a statistical test for classification of both groups, this data is a reliable representation of the BEA for particular scenarios. On the other hand, frequency domain features have also been studied for developing works related with motor imagery. In [64] a set of features

One of the drawbacks of this scheme is the high dimensionality of the resulting feature vector. Additionally, this large set of features computed from different channels independently could include redundant information that can be captured from different electrodes in the array. For the second case, the particularities of the Alzheimer disease restricted the generalization capability of this framework. Again, the number of resulting features could also lead to high computational complexity and redundant information from the analysis of channels independently. In this work a parallel approach using the coherence between the theta bands of each channel is proposed giving good results in terms of discriminative capacity. This last result starts to give some ide that to improve the performance of the BEA study, some quantification of the relationships between channels must be included [90].

Temporal and spectral features have been used in the classification of BEA patterns by classical machine learning approaches, for example, the use of linear discriminant analysis (LDA) from spectral features have been developed in works like [104][55][64]. Moreover, works based on Deep belief networks (DBN) have been applied to the recognition of emotions using BEA data from fundamental rhythms representation [101]. Acceptable results regarding classification accuracy have been achieved through these methods; however, generalization problems associated with training tests overfitting and low interpretability qualities have been reported[13].

3.0.1 Short-time representation

The preferred strategy for computing spectral features form EEG data is the PSD. It allows to analyze the power distribution of each channel over a specific frequency range. The information captured is related with the overall frequency content of specific neural activity and is expressed as Fourier transform of attained EEG signal as equation 3.1 shows [10].

$$\boldsymbol{X}(\omega) = \frac{1}{2\pi N} \| \sum_{n=1}^{N} \boldsymbol{x}(n) e^{-j\omega n} \|^2 \tag{3.1}$$

In order to deal with some of the EEG issues related with non-stationarity behavior, a initial stage of data windowing is performed, and also different strategies of BEA feature extraction. In this case, a windowing function $\boldsymbol{W} : \mathbb{R}^{C\times T} \rightarrow \mathbb{R}^{C\times t\times w}$ is set to produce a segmented version of BEA where some assumptions of quasi-stationarity can be done. Then, $\hat{\chi} = \boldsymbol{W}(\boldsymbol{X}_n, \theta) = \{\hat{\boldsymbol{X}}_{n,k} \in \mathbb{R}^{C\times t}; l_n\} \forall n = 1, \cdots, N; k = 1, \cdots, W$, where each k window has t time samples, and θ parameters associated with selected window and overlapping producing the final number of t and W.

We can find time-frequency or short time representations, where an initial stage of signal windowing followed by frequency bands decomposition is applied. Works based in this scheme include the application of the short time Fourier Transform (STFT) [70][72] for emotion recognition, Wavelet transform [100] for motor imagery tasks and the use of waveforms dictionaries and the Matching pursuit strategy [14] for cognitive states classification. However, the decoding capacity of these strategies and the interpretability of characterized data is still low. Furthermore, the analysis is performed for each channel independently, so the amount of resulting information is considerable high [82].

3.1 Short-time proposed Methodology

With the idea to extract information from the EEG data using short-time spectral representations, the proposed framework in this stage is presented in figure 3.1. From the EEG recordings and addittional biosignals available in DEAP and MAHNOB databases, a linear analysis for feature extraction is developed following certain statistical measures and frequency analysis as stated before. This analysis is commonly used to extract information about the physiological signals [37] as a result from various studies in which it has been concluded that some of these features have a direct relationship with certain BEA conditions [71]. The features extracted from the EEG and peripheral signals are presented in Table 3.1.

Based on the work presented by G. Valenza in 2012 [85], a set of non-linear features are

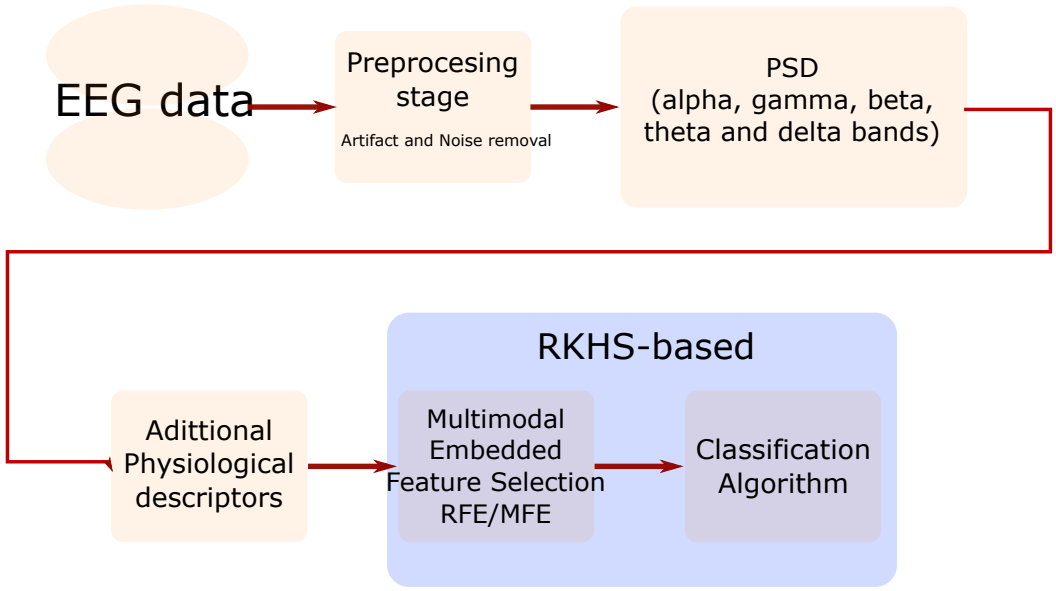


Fig. 3.1 Proposed Short-time representation framework.

also included in the scheme of multimodal emotion classification. From the results obtained with the use of the non-linear features, an improvement in the performance is reported in terms of percentages of accuracy. These nonlinear features are based on a methodology called Recurrence Plots and are extracted from the GSR, respiratory pattern and the heart rate signals. In recent years, this nonlinear analysis has been applied in several works related with the affective computing, including bipolar patients for emotional response analysis in [86], [29]. In these works, the nonlinear analysis shows higher performance in the recognition of affective states than the classical time-frequency analysis of the signals [85].

Recurrence plots are based on a technique for the analysis of complex dynamic systems called embedding procedure [85], where a set of vectors X_i is constructed from the time series representing the behavior of the system. The evolution of the system can be represented by the projection of the vectors on a path between a multidimensional space that is commonly known as a phase space or phase state [85]. Eckmann in [24] introduced a tool that can be used to display states recurrence X_i in phase space. This tool called Recurrence Plots (RP) allows to investigate the $m-$dimensional phase space trajectory from a $2-$dimensional representation of their recurrences [50]. Following RP calculation, a Recurrence Quantification Analysis (RQA) proposed in [96], is computed to quantify the number and length of recurrences from a dynamic system, represented by its state space trajectory. Then, from the RQA some features are extracted [85]. For further details of the RP technique and the RQA analysis see [85] and [24].

The embedding of a time series $x_t = (x_1, x_2, \cdot, x_N)$ is constructed by creating a set of vectors X_i such that:

$$X_i = \left[x_i, x_{i+\Delta}, x_{i+2\Delta}, \ldots, x_{i+(m-1)\Delta}\right], \tag{3.2}$$

where Δ is a delay in the number of incremental samples and m is the number of incremental samples (dimension) of the X_i. The evolution of the system can be represented by the projection of the vectors X_i on a path between a multidimensional space that is commonly known as a phase space or phase state [85]. Eckmann in [24] introduced a tool that can be used to display states recurrence X_i in phase space, because usually these phase spaces have no dimension that can be displayed. However, this tool called Recurrence Plots (RP) allows to investigate the $m-$dimensional phase space trajectory from a $2-$dimensional representation of their recurrences [50].

When performing the embedding of a time series, it must be chosen the dimension m of X_i and the delays Δ such that each vector X_i represents values that reveal the topological relationship between subsequent points in the time series. The number of incremental samples in the embedded vector is usually chosen so long to cover the dominant frequency of the time series, but m should not be so long that the first and last value in each division of time became unrelated.

When a state at time i recurred at time j, the element (i, j) of the square matrix $N \times N$ is set at 1, 0 in the opposite case [85]. This representation is called RP and can be expressed mathematically as follows:

$$R_{i,j} = \Theta\left(\varepsilon_i - \left\|x_i - x_j\right\|\right), \tag{3.3}$$

where $x_i \in {}^m, \quad i, j = 1, \ldots, N$, N is the number of incremental states considered x_i, ε_i is a threshold distance , $\|\cdot\|$ the norm and $\Theta(\cdot)$ the Heaviside function which is defined as

$$H(z) = \begin{cases} 1, & si \quad z \geq 0 \\ 0, & si \quad z < 0 \end{cases} \tag{3.4}$$

The quantification f the RP's is based on the evaluation of the diagonal lines to estimate chaos-order transitions or the vertical lines to estimate chaos-chaos transitions.

Following RP calculation, a Recurrence Quantification Analysis (RQA) proposed in

[96] is computed to quantify the number and length of recurrences from a dynamic system, represented by its state space trajectory. Then, from the RQA some features are extracted [85].

- Recurrence Rate (RR)
- Determinism (DET)
- Laminarity (LAM)
- Trapping Time (TT)
- Ratio (R)
- Averaged diagonal line length (L)
- Entropy (ENT)
- Longest diagonal line (L_{max})

Table 3.1 Features from EEG, peripheral and video signals [37]

Signal	*Extracted Features*
GSR	Average skin resistance, average of derivative, average of derivative for negative values only, proportion of negative samples in the derivative vs. all samples, number of local minima in the GSR signal, average rising time of the GSR signal, 10 spectral power in the $[0-2.4]$Hz bands, zero crossing rate of Skin conductance slow response (SCSR) $[0-0.2]$ Hz, zero crossing rate of Skin conductance very slow response (SCVSR) $[0-0.08]$Hz, SCSR and SCVSR mean of peaks magnitude.
Skin Temperature	Average, average of its derivative, spectral power in the bands ($[0-0.1]$Hz, $[0.1-0.2]$Hz).
Respiration pattern	Average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, 10 spectral power in the bands from 0 to 2.4Hz.
Blood volume pressure	Average and standard deviation of HR, HRV, and inter beat intervals, energy ratio between the frequency bands $[0.04-0.15]$Hz and $[0.15-0.5]$Hz, spectral power in the bands ($[0.1-0.2]$Hz, $[0.2-0.3]$Hz, $[0.3-0.4]$Hz), low frequency $[0.01-0.08]$Hz, medium frequency $[0.08-0.15]$Hz and high frequency $[0.15-0.5]$Hz components of HRV power spectrum.
EEG	theta, slow alpha, alpha, beta, and gamma Spectral power for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.
EMG and EOG	Eye blinking rate, energy of the signal, mean and variance of the signal.
Video	Mean shape (shape of the face in all the frames from each video).

3.1.1 Feature selection and classification

The embedded methods for feature selection and classification stage using SVM's RFE and MFE are discussed in this section. A brief introduction of the basis of SVM's is presented first, followed by the explanation of how the RFE and MFE methods involve the SVM training into the feature selection task.

Support Vector Machines

Support vector machines are the state-of-art machine-learning algorithm. The SVM methodology propose that the inputs from the observations **x**, could be mapped into a higher dimensional space, where a class separation hyperplane could be computed [20]. The computed function then is used to assign a label on the output y [20]. To find an optimum hyperplane that effectively separates the different classes of the data inputs, a small amount of the observations that lies on the edge of separation called support vectors (SV) is used [20].

Let $\mathbf{w_0} \cdot \mathbf{z} + b_0 = 0$, be an optimal hyperplane in feature space. The weights $\mathbf{w}_0$ for the optimal hyperplane can be written as a linear combination of support vectors [20] $\mathbf{w}_0 = \sum_{SV} \alpha_i \mathbf{z}_i$. The optimal hyperplane $\mathbf{w}_0 \cdot \mathbf{x} + b_0 = 0$ is the unique one capable of separate correctly the training data with a maximal margin. It determines the direction $\mathbf{w}/|\mathbf{w}|$, where the distance between the projections of the training vectors of different classes is maximal. If the training data are separable an SVM is a maximum margin classifier. A peculiarity of the SVM 's is that the weights w_i are functions of the support vectors.

The optimal hyperplane (w_0, b_0) are the arguments that maximize the distance in (5.12) and it is constructed from the support vectors[20]. Vectors x_i for which $y_i(w \cdot x_i + b) = 1$ will be tagged as support vectors. The vector w_0 that determines the optimal hyperplane can be written as a linear combination of training vectors[20]:

$$\mathbf{w}_0 = \sum_{i=1}^{l} y_i \alpha_i^0 \mathbf{x}_i, \tag{3.5}$$

where $\alpha_i^0 \geqslant 0$. Since $\alpha > 0$ only for support vectors, the expression (3.5) represents a compact form of writing $\mathbf{w}_0$[20]. To solve the problem of finding the optimal hyperplane (SVM training stage), a constrained optimization problem of maximizing the distance for a given weight vector can be determined by the Lagrangian multiplier method [91]. The SVM training then consist in the implementation of the quadratic problem in 3.6, minimizing over

α_k subject to 3.7 [31].

$$J = \left(1/2\right) \sum_{hk} y_h y_k \alpha_h \alpha_k \left(x_h \cdot x_k + \lambda \delta_{hk}\right) - \sum_k \alpha_k, \tag{3.6}$$

$$0 \leqslant \alpha_k \leqslant C \quad and \quad \sum_k \alpha_k y_k = 0, \tag{3.7}$$

where $x_k \cdot x_h$ denotes the scalar product, y_k corresponds to the class label, δ_{hk} is the Kronecker symbol and α and C are positive constants (soft margin parameters) that ensure convergence even when the problem is non-linearly separable [31].

Recursive Feature Elimination (RFE)

Evaluating how one feature contributes to the separation between classes can produce a feature ranking. One of the possible uses for the feature ranking is the design of a classifier on a pre-selected subset of features [31]. Each feature that is correlated with the separation of interest is by itself a class separator. The entries that are associated with larger weights, have a greater influence on the classification decision, therefore if a classifier has a good performance, those entries with the highest weights are the more relevant characteristics [31]. This feature ranking could be obtained during the SVM training stage.

In classification problems the ideal target function is the expected value of the error, this is the error rate calculated on a infinity number of examples, whereas in the training stage this ideal objective function is replaced by a cost function **J** estimated only for training patterns. Given this, in [31], the authors introduced the idea of calculating the change in the cost function $\mathbf{DJ}(i)$ from removing a single feature or equivalently, from making the weight $\mathbf{w}_i$ zero. Using the change in the cost function when a feature is removed, a feature ranking could be constructed in order to discard the features with the least ranking value. Nevertheless a good criterion of feature ranking is not necessarily a good criterion for selecting a subset of them. To use the ranking criterion in order to eliminate features, an iterative procedure called Recursive Feature Elimination (RFE)[31] was proposed. The procedure follows as:

1. Train the classifier (optimizing the weights $\mathbf{w}_i$ with respect to **J**)
2. Compute the ranking criterion for all the features $\mathbf{DJ}(i)$.
3. Remove the feature with the smallest ranking criterion.

$$\mathbf{J} = \left(1/2\right) \alpha^T \mathbf{H} \alpha - \alpha^T 1, \tag{3.8}$$

where $\mathbf{H}$ is the matrix with elements $y_h y_k \mathbf{K}(x_h, x_k)$, $\mathbf{K}$ is a kernel function that measures the similarity between x_h and x_k, and 1 is an l dimensional vector of ones. To calculate the change in the cost function due to the elimination of the component i, the $\alpha's$ remains unmodified and the $\mathbf{H}$ matrix is recalculated. This corresponds to calculating $K(x_h(i), x_k(-i))$, giving the $\mathbf{H}(-i)$ matrix, where the notation $(-i)$ means that the component i has been removed [31]. The resulting ranking coefficient is:

$$DJ(i) = \left(1/2\right) \alpha^T \mathbf{H} \alpha - \left(1/2\right) \alpha^T \mathbf{H}(-i) \alpha. \tag{3.9}$$

The input with the smallest difference $\mathbf{DJ}(i)$ is eliminated. The elimination of the input with the smallest difference is repeated iteratively producing the Recursive Feature Elimination (RFE) method. The change in the $\mathbf{H}$ matrix must be calculated only for the support vectors [31]. Following the basis of the RFE algorithm, the index m^* of the first feature to remove is $\arg \min_{m \in \{1,\ldots m\}} |\omega_m|$, and generally in the iteration i the same rule of selection is applied to the $M - i$ remaining features.

Margin-maximizing Feature Elimination

The use of the weights from the trained classifier proposed in RFE algorithm has no consideration of the maximal margin of separation between classes of the SVM. RFE is equivalent to the elimination by maximization of the margin if the following equation is always satisfied [3]:

$$\max_m \min_n \frac{y_n f(x_n) - y_n x_{n,m} w_m}{\sqrt{\|w\|^2 - w_m^2}} = \min_n \frac{y_n f(x_n) - y_n x_{n,m^*} w_{m^*}}{\sqrt{\|w\|^2 - w_{m^*}^2}}, \tag{3.10}$$

where $\mathbf{x}$ are the input examples, $\mathbf{y}$ the corresponding outputs. w is associated to the $DJ(i)$ vector computed by RFE, so w_m corresponds to the ranking coefficient of the m feature. In order to consider the margin of separation computed from the SV's, a recursive algorithm based in SVM's called Margin-maximizing Feature Elimination (MFE) was proposed in [3]. The authors argue that experimentally they have shown that RFE is not in agreement with

margin maximization. RFE is focused on minimally reducing the squared weight vector 2-norm 3.5, ignoring the margin constraints [3]. The authors in [3] also demonstrate that for the kernel case that the assumption of RFE that the squared weight vector 2-norm is strictly decreasing as features are eliminated is not valid for all the kernels. MFE then propose a feature elimination method based on the recursion over the kernels. For example for the polynomial kernel, $K(u,v) = \exp\left(\left\|u-v\right\|^2 \Big/ \left(\gamma^2\right)\right)$ and denoting $H_{k,n}^{i,m} = H\left(s_k^{i,m}, x_n^{i,m}\right)$ in iteration i the recursion [3]:

$$\mathbf{H}_{k,n}^{i,m} = H_{k,n}^{i-1,m_{i-1}} - \left\|s_{k,m} - x_{n,m}\right\|^2 \Big/ \left(\gamma^2\right) \forall k, \forall n \tag{3.11}$$

where s_k corresponds to the support vector k. This recursively calculated kernels are used to evaluate the discriminant function

$$f(x) = \sum_{k \in S} \lambda_{s_k} y_{s_k} K(s_k, x) + b \tag{3.12}$$

and the weight vector norm through

$$\|\omega\|^2 = \sum_{k \in S} \sum_{l \in S} \lambda_{s_k} y_{s_k} \lambda_{s_l} y_{s_l} K(s_k, s_l) \tag{3.13}$$

building a MFE-kernel algorithm. The MFE method at each iteration i eliminates the feature m_{MFE} that preserve the maximum positive margin for the training set from the following equation [3]:

$$(m_{MFE}, n_{MFE}) = \arg \max_{m \in S = \left\{m' | g_{n'}^{i,m'} > 0, \forall n'\right\}} \min_n \frac{g_n^{i,m}}{\|w\|^{i,m}}, \tag{3.14}$$

where notation $q^{i,m}$ corresponds to quantity q at feature elimination step i upon elimination of feature m and $g_n^{i,m} = y_n b + \sum_{m=1}^{M} \delta_n^m$ with M being the set of eliminated features and $\delta_n^m = y_n x_{n,m} w_m$.

3.1.2 Methodological details

For the development of the routines for non-linear feature extraction and the schemes of feature selection based on discriminant methods (SVM's), two toolboxes were used. For the RQA non-linear analysis, the Cross Recurrence Plot Toolbox (CRP) [49] is used. The Pattern Recognition Toolbox (PRTools) is used for the routines that allow the training and testing of the SVM's [23]. The implementation of the SVM algorithm is from the PRTOOLS toolbox, and the kernel used for the RFE implementation was the Radial Basis Function (RBF) kernel. An own implementation of the RBF kernel is made for further computing of RFE-SVM and MFE-SVM and combined with PRTOOLS in the "user kernel" mode. The PRTOOLS toolbox uses quadratic programming from the MATLAB Optimization toolbox. The regularization parameter for the SVM, and the parameter of the RBF kernel are estimated using cross-valiadtion over a grid of values for both parameters.

To assess the functionality of the discriminant selection algorithms, a validation stage is developed in which we compare the results produced by the algorithms of selection on different data sets against a classical feature space dimensionality reduction scheme such as Principal Component Analysis (PCA) [36]. The PRTools toolbox includes an implementation of the PCA algorithm. PCA uses an orthonormal transformation ton convert possibly correlated variables into linearly uncorrelated variables called principal components. The first principal component should have the largest possible variance, and the succeeding components should have the highest variance. Each component is subject to the constraint that has to be orthogonal with the preceding components [36]. With these reduced spaces, a SVM with a radial basis function is trained to determine the respective classification accuracy.

The signals from the database are processed to obtain the set of features from the linear analysis in Table 3.1 proposed in [37], and the features from the non-linear analysis RQA for the GSR, HR, and respiration pattern [85]. Using the labels from the database for the arousal and valence dimensions, several datasets are generated for different classification problems. Sets $D1$ and $D2$ correspond to biclass problems for both dimensions with levels from $1-5$ to $6-9$. In the arousal dimension the range cover the classes of active and passive. For the valence dimension the range cover the classes of pleasant and unpleasant. For the MAHNOB database, datasets are extracted in a equivalent form as the $D1-D2$ datasets. In this case the datasets are named as $M1-M2$ for the two spaces of classification. For this database we only take into account the linear features. A summary of the different sets of data generated is presented in Table 3.2. For a further quantification analysis, an exact description of the position of each feature into the feature space is presented in Table

Table 3.2 Datasets

Dataset	*Description*
D1	DEAP database, Arousal dimension, biclass problem, patterns with values $1-5$ and $6-9$
D2	DEAP database, Valence dimension, biclass problem, patterns with values $1-5$ and $6-9$
M1	MAHNOB database, Arousal dimension, biclass problem, patterns with values $1-5$ and $6-9$
M2	MAHNOB database, Valence dimension, biclass problem, patterns with values $1-5$ and $6-9$

Table 3.3 Feature indexes for the two databases

DEAP		MAHNOB	
Signal	*Index*	*Signal*	*Index*
GSR	1 : 11	GSR	1 : 11
GSR - RQA	12 : 19	Temp	12 : 15
Temp	20 : 23	Resp	16 : 19
Resp	24 : 27	HR1	20 : 29
Resp - RQA	28 : 35	HR2	30 : 39
HR	36 : 45	HR3	40 : 49
HR - RQA	46 : 53	EEG	50 : 273
EEG	54 : 276		
EOG y EMG	277 : 288		
Video	289 : 323		

3.3. This information allows a clear understanding of the features that are selected after each feature elimination step, when the algorithms of feature selection are applied.

Based on the different data sets extracted, several experiments are performed to evaluate the performance of RFE and MFE. the RFE-SVM and MFE-SVM algorithms were set to eliminate one feature at each iteration in order to avoid possible elimination of correlated features when a bunch of features is eliminated. A modification of the RFE algorithm proposed in [93] is implemented in order to test possible wrong feature removal from the datasets. This RFE-SVM-CBR implementation is used when RFE-SVM is set to eliminate a bunch of features in each iteration. The size of the final subset of features was computed as the 5% of the original features set size. This is 16 features for the DEAP database and 14

features for the MAHNOB database.

The classification accuracy (CA) was computed after each iteration of the selection algorithms. Crossvalidation is performed by using 80% of the patterns for training and 20% of the patterns for test. The procedure is repeated 10 times in order to have an statistical validation of the test. The F1-Score is a measure of the test accuracy. F1-score is defined as the harmonic mean of precision and recall as the following equation shows:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The F1-score was computed at each iteration of the feature selection algorithms. The vectors with the indexes of the selected features and the CA after each iteration are stored for further analysis. An statistical analysis based on the equal median test is applied to the classification results for each method. The equal median test allows to determine which method has a higher classification accuracy, and if it is statistically different against the other feature selection algorithms [62]. A final test for the RFE and MFE algorithms is made by applying the selection algorithms to EEG features only. The resulting features from the EEG are combined with the other signals for a new selection test.

3.2 Results

This section presents the results for the different feature selection experiments with RFE and MFE over each dataset, computing the F1-Score and the accuracy rate from every feature elimination method performed and presented as the mean an the standard deviation of the 10 realizations of each experiment.

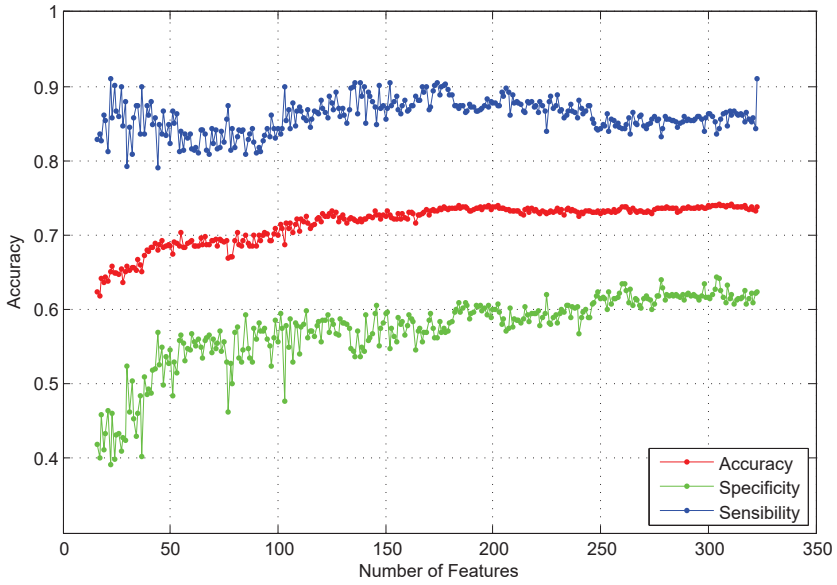
3.2.1 Results for $D1$ dataset

For the $D1$ dataset the Figures 3.2(a) and 3.2(b) show the behavior of the accuracy rate against several feature eliminations using the RFE and MFE algorithms. For the RFE method, the accuracy begins around 72% for the complete set of features, and only starts to decrease below 65% when the size of the subset corresponds to less than the 20% of the original feature space dimension. In the MFE case, the classification accuracy for subsets of 10% of the original size, reaches a value of around 70%. The value for the sensibility increases for smaller subsets of features when RFE is applied, in the case of MFE the value

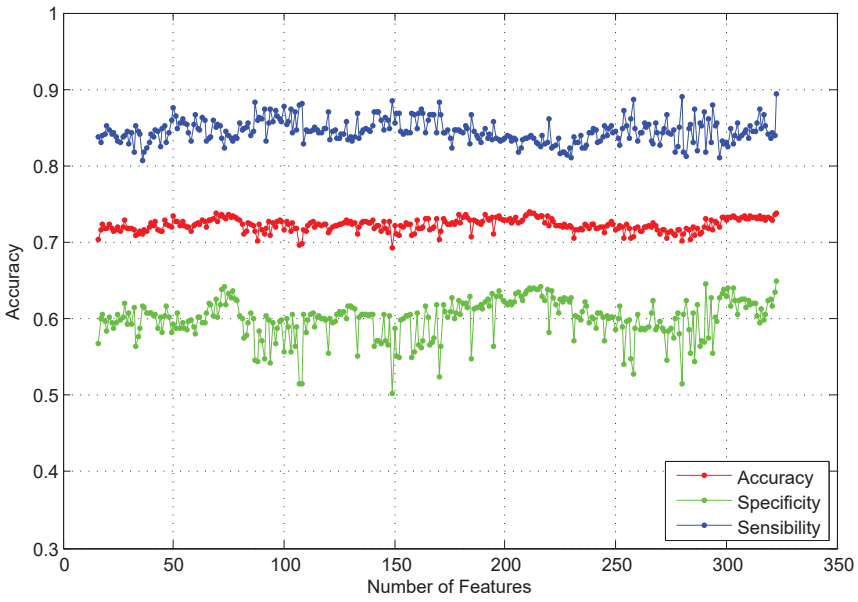
of sensibility maintain similar levels for smaller subsets. The specificity tend to decrease at each feature elimination for RFE and for MFE the value of specificity shows variations with each elimination but maintain a similar value compared to the initial value at the original size of features.

Analysis of the features selected in the different subsets is presented in Figures 3.3(a) and 3.3(b). These figures show the percentages of occurrence of each feature in different subsets. For example, when the algorithm is analyzing a subset of features of size 129, and in the ten repetitions of the experiment a particular feature x_h appeared five times, we assign a percentage of 0.5 to the occurrence of that feature. The percentage is represented in a color scale (red for 1 and blue for 0). These histograms allow us to analyze which signals are more relevant at the recognition step, of the different affective states. Notice that the histograms can be understood from two points of view. The first point of view is that given a particular feature x_h, we can see in what percentage that feature appeared when the algorithm analyzed subsets of different size. The feature corresponds to one of the features extracted from one of the EEG channels. For this single feature, it can be seen how the inclusion of the feature varies for different subsets selected. From the complete size of the set of features (323 features) to 246 features, this particular feature is selected always for all the realizations of the experiment. When the size of the selected subset reaches 169 features, the index of selection is around 0.7 following the color scale. The index of selection continues decreasing as the size of the set of selected features also decreases. When the size of the set is 93, the index of selection is around 0.2, and finally when the elimination of features reaches the smallest size, this feature was not selected in any realization. The second point of view for analyzing the histograms is that given a fixed size of features selected, S, we can see in what percentage each of the available features was included when performing the ten repetitions. While features $1-5$ from GSR show percentages of occurrence of 1, features $6-10$ from GSR have percentages of occurrence from 0.1 to 0.4. All temperature features, all respiratory features, and most of the HR features have an index of selection of 1. While some of the EEG features were completely discarded at this stage, others have percentage of occurrence from 0.3 to 0.5.

From the index distribution of the features in Table 3.3, it can be noticed that the features that are discarded in early iterations, when the RFE algorithm is used, are the features from the EEG signal. Some features from the physiological signals are retained despite several feature eliminations. In the case of the MFE algorithm, the features initially discarded are from the physiological signals, while a higher number of features from the EEG are selected in the final subsets.

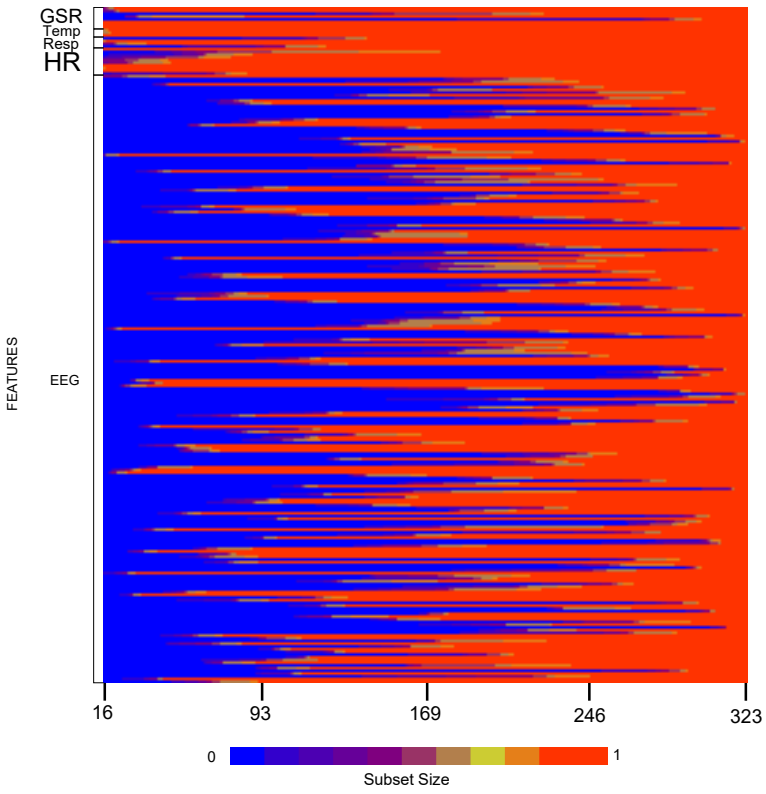


(a) RFE

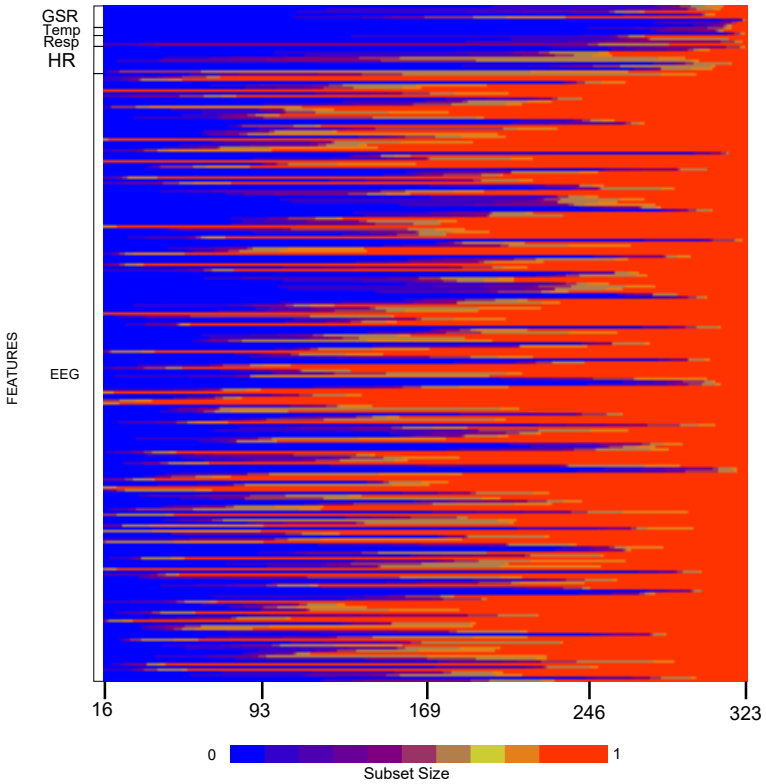


(b) MFE

Fig. 3.2 Classification accuracy for different sized feature subsets in $D1$ dataset, selected from RFE and MFE methodologies



(a) RFE selected features



(b) MFE selected features

Fig. 3.3 Feature apparition within different sized subsets in $D1$ dataset

3.2.2 Results for $D2$ dataset

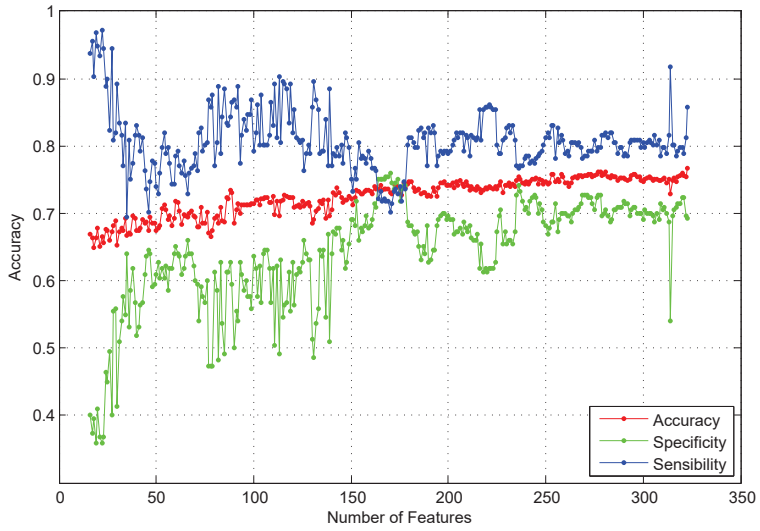
Following the same analysis for the $D2$ dataset, corresponding to the biclass problem in the valence dimension, in Figures 3.4(a) and 3.4(b) the results show a similar behavior compared to the results from $D1$ dataset. An initial accuracy of 73% is reached using the total set of features in both methodologies, and this percentage remains around 72% for subsets of less than 30 features when the RFE algorithm is used, also the level of the sensibility show a considerable improvement for the final subset but the specificity level decrease for those smaller subsets of features . In the MFE selected subsets, the initial classification accuracy is maintained even for the final subset as Figure 3.4(b) shows, also the levels of sensibility and specificity maintain similar values for all the selected subsets.

The corresponding percentage of occurrence for the features in different subsets for the $D2$ dataset are presented in Figures 3.5(a) and 3.5(b). Features from all the physiological signals are selected in smaller subsets using RFE with more influence from all the physiological signals. Some features from the EEG are also selected. With the MFE algorithm the features retained are mostly selected from the EEG as in the $D1$ dataset.

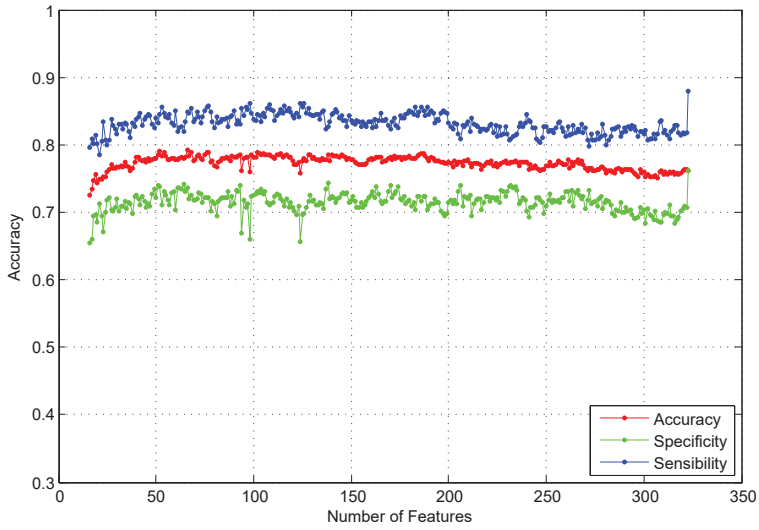
3.2.3 Results for $M1$ dataset

For the biclass problem in the arousal space using dataset $M1$, the two feature selection methods are employed in a similar manner as in previous experiments. Figure 3.7 shows the variations in the accuracy rate as several features are eliminated via RFE and MFE. For RFE, Figure 3.6(a) shows an initial success rate of 69% with the complete set of 273 features, decreasing below rates of 65% for subsets of less than 50 features. For the selection using MFE, the initial success rate with all the features is around 69% and this percentage is maintained despite the different feature eliminations with some slight increase for reduced subsets of less than 50 features with accuracy rates close to 60%.

Occurrence histograms for the selected features in the different $M1$ subsets are presented in Figure 3.7. It can be observed that the features that were not discarded by the RFE algorithm comes from the HR, GSR and EEG, Figure 3.7(a). When the MFE algorithm is used the features are selected predominantly from the EEG with few coming from the Temperature signal and the respiratory pattern, see Figure 3.7(b).



(a) RFE

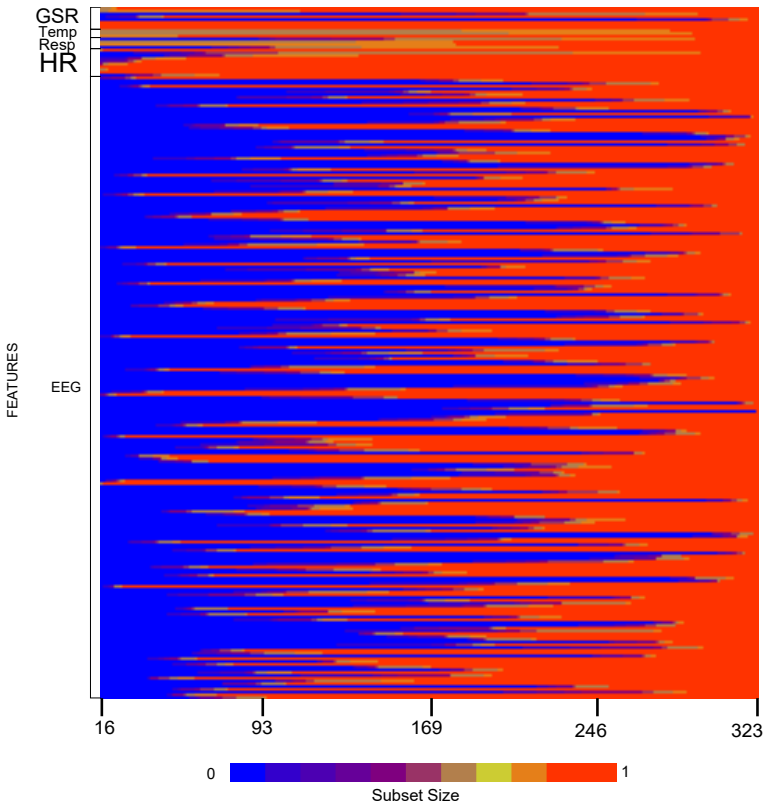


(b) MFE

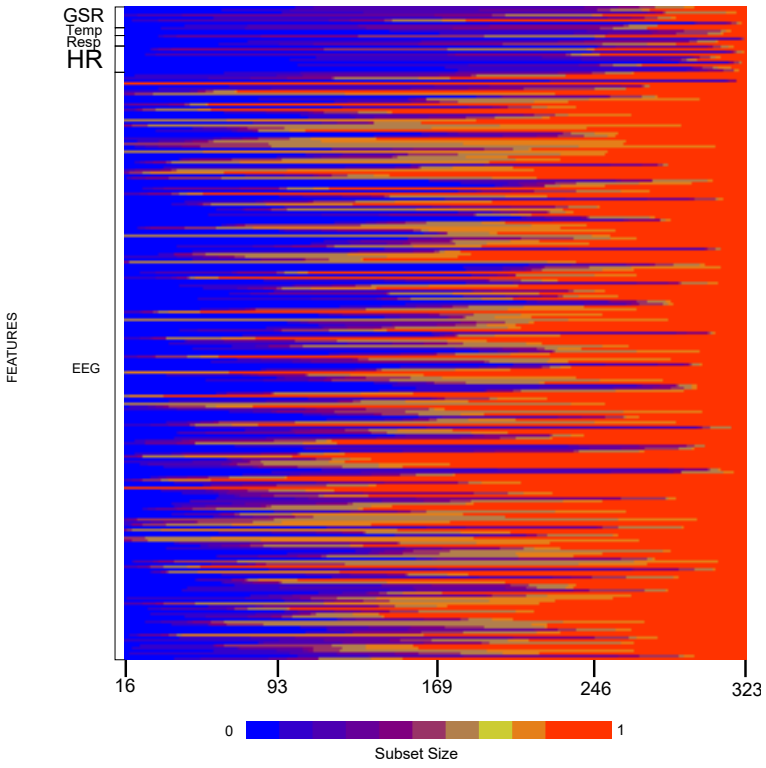
Fig. 3.4 Classification accuracy for different sized feature subsets in $D2$ dataset, selected from RFE and MFE methodologies

3.2.4 Results for $M2$ dataset

For the valence dimension using $M2$ dataset, both selection algorithms show a similar behavior. From the initial percentage of 65% using the total set of features, several feature eliminations are applied and the accuracy is maintained around the initial percentage. The

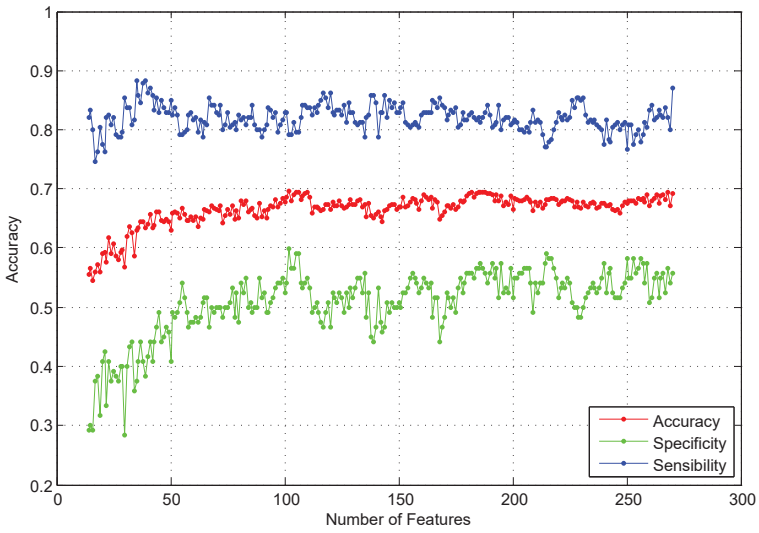


(a) RFE selected features

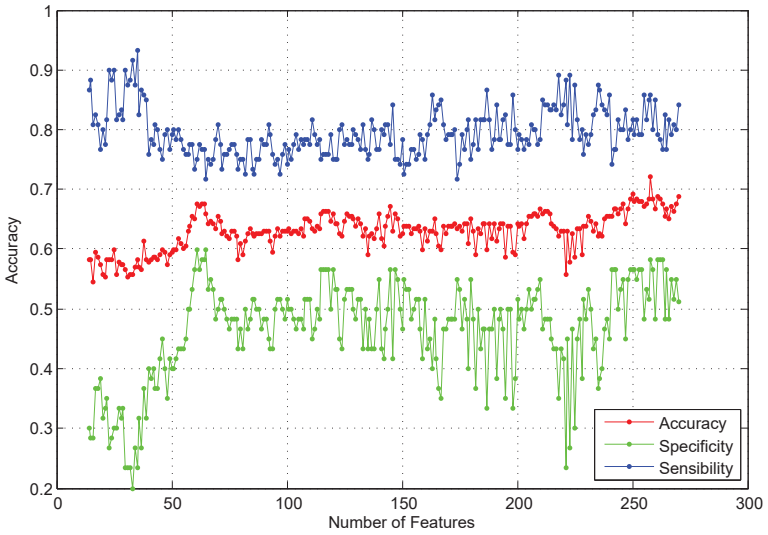


(b) MFE selected features

Fig. 3.5 Feature apparition within different sized subsets from *D*2 dataset



(a) RFE

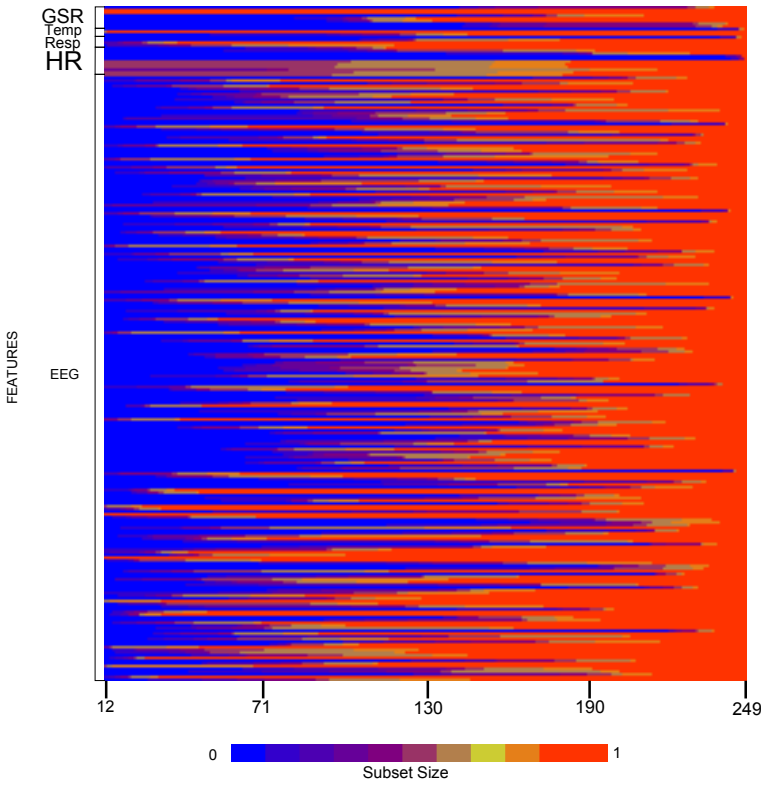


(b) MFE

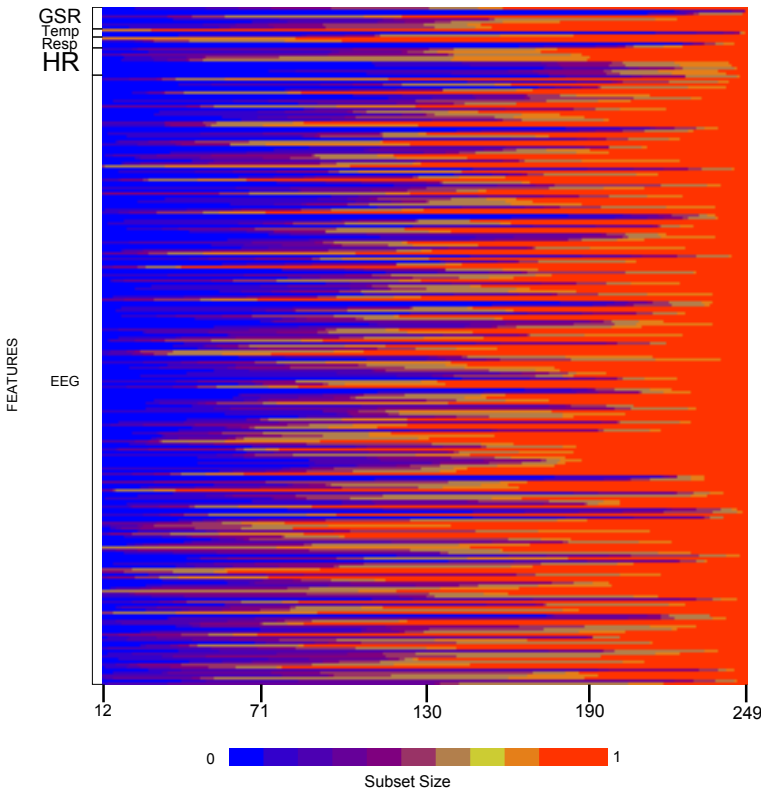
Fig. 3.6 Classification accuracy for different sized feature subsets in $M1$ dataset, selected from RFE and MFE methodologies

accuracy declines for smaller subsets to percentages around 60%, as shown in Figures 3.8(a) and 3.8(b).

From the analysis of the percentage of occurrence, the most selected features come from the HR and the respiratory signal when the RFE algorithm is applied, see Figure 3.9(a).

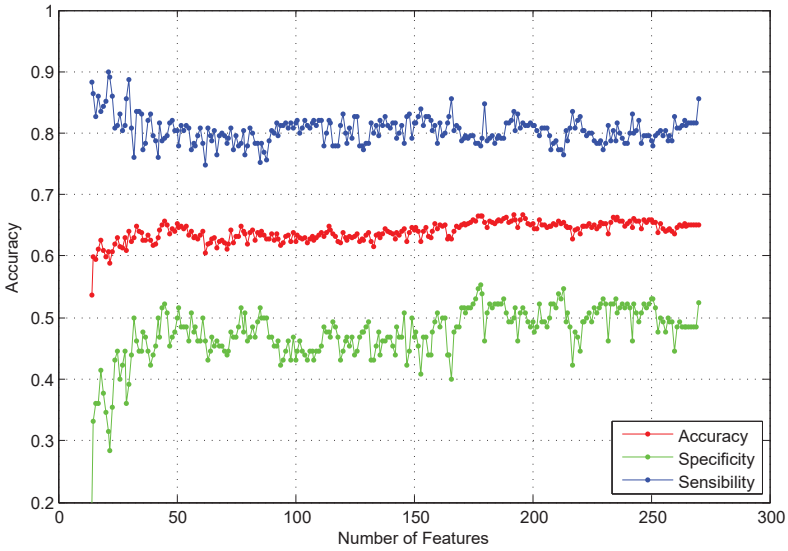


(a) RFE selected features

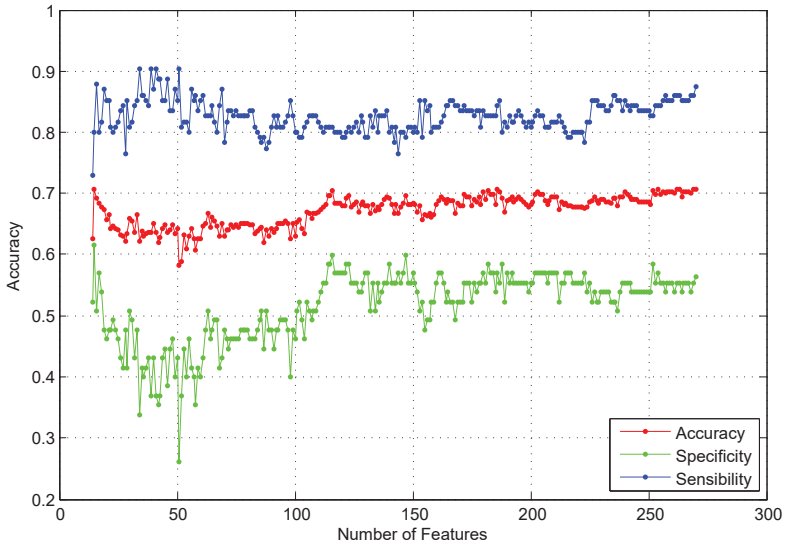


(b) MFE selected features

Fig. 3.7 Feature apparition within different sized subsets in $M1$ dataset



(a) RFE

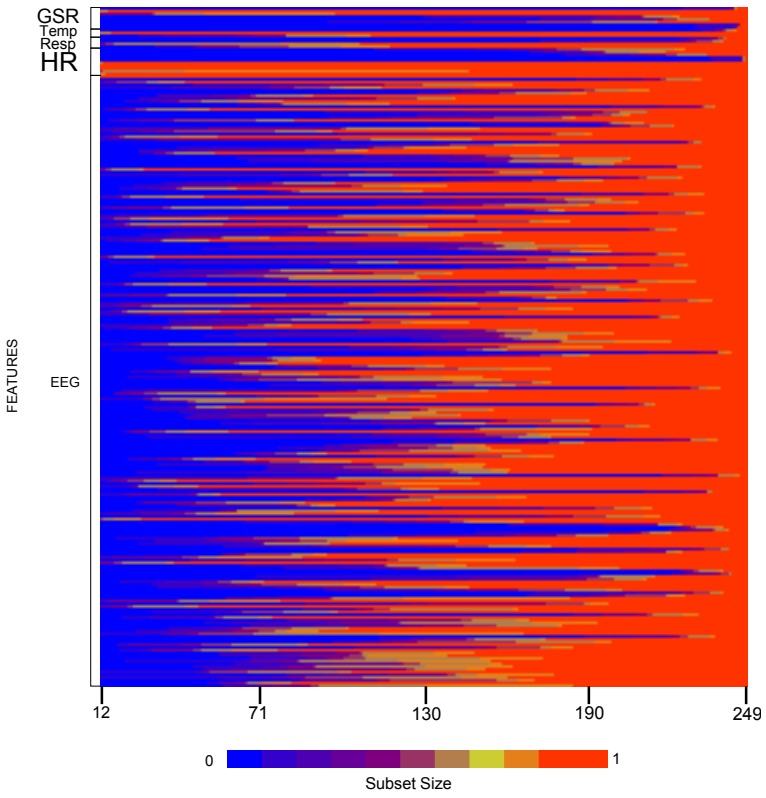


(b) MFE

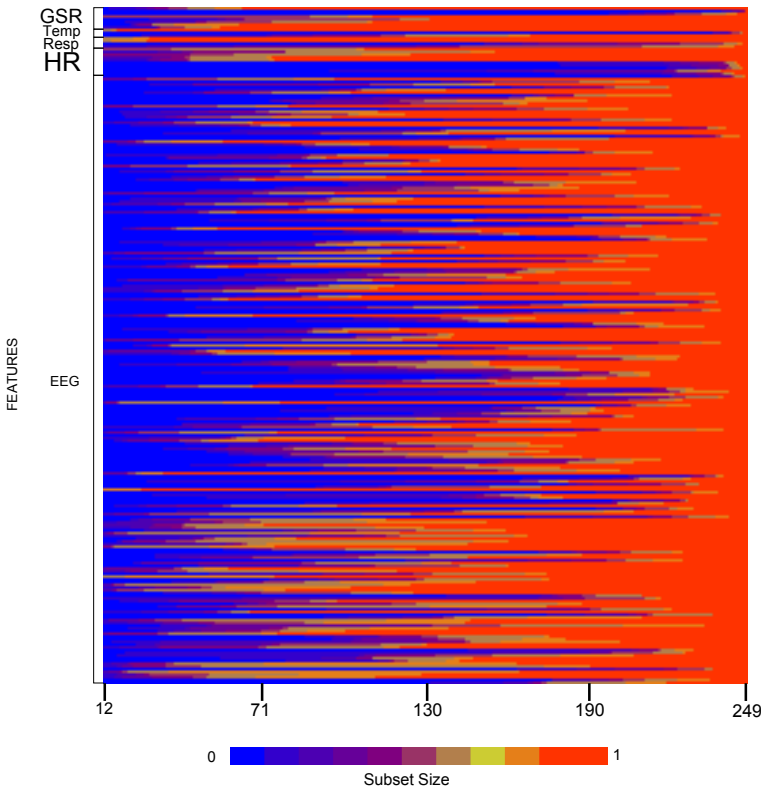
Fig. 3.8 Classification accuracy for different sized feature subsets in $M2$ dataset, selected from RFE and MFE methodologies

When MFE is used, the selected features in the smaller subsets come from the EEG signal, respiratory pattern and the temperature signal as well, see Figure 3.9(b).

A compilation of the classification results from the experiments of feature selection is



(a) RFE selected features



(b) MFE selected features

Fig. 3.9 Feature apparition within different sized subsets in $M2$ dataset

Table 3.4 Classification Accuracy and F1-Score for RFE and MFE over different datasets

		Metric		
Dataset	Algorithm	CA (average) [%]	CA (max) [%]/ (Nf)	F1-Score (average) [%]
D1	RFE	72.09 ± 2.40	74.82 (115)	72.58 ± 3.02
	MFE	72.63 ± 0.82	74.31 (124)	72.65 ± 0.92
D2	RFE	72.35 ± 1.42	74.15 (103)	71.06 ± 1.47
	MFE	73.69 ± 0.80	75.17 (167)	72.07 ± 0.74
M1	RFE	67.01 ± 3.57	69.59 (176)	79.25 ± 1.93
	MFE	64.20 ± 0.87	67.71 (240)	77.92 ± 1.16
M2	RFE	62.58 ± 1.32	66.09 (52)	75.41 ± 1.08
	MFE	63.38 ± 2.20	65.17 (53)	65.17 ± 1.32

Table 3.5 Summary of most relevant features

Method	Arousal		Valence	
	Index	Signal	Index	Signal
RFE - DEAP	21, 23	Temp	21, 23	Temp
	26	Resp	24	Resp
	37	HR	37, 41, 44	HR
	52	HR - RQA		
RFE - MAHNOB	2, 3, 10	GSR	2	GSR
	$41 - 46$	HR	41, 42, 43, 45, 46	HR
MFE - DEAP	6	GSR	46	HR-RQA
	46	HR-RQA	93, 170	EEG
	65, 74, 80, 156, 217	EEG		
MFE - MAHNOB	11	GSR	9	GSR
	58, 91	EEG	178	EEG

presented in Table 3.4. The classification accuracy (CA) and the F1-Score are computed in every iteration of the RFE and MFE algorithms for each dataset following the elimination of the less relevant features. The table contains the CA average, maximum CA with the number of features (Nf) where the maximum CA was reached, and the F1-Score average for each experiment. The equal median test is performed over all the results from the two selection algorithms. For all the feature subsets in each experiment, MFE brings better classification results than RFE. The statistical significance analysis based on the equal median test is applied for each dataset selected in each iteration using the RFE and MFE methods. The test allows to determine which method selected the dataset which provides a higher classification accuracy (CA). Over the DEAP database, the equal median test shows that MFE has higher performance than RFE in the selection of 12 datasets in the valence dimension and 22 datasets in the arousal dimension. For the MAHNOB database, the MFE algorithm obtained a higher performance than RFE in the selection of 26 datasets in the arousal dimension and 7 datasets in the valence dimension.

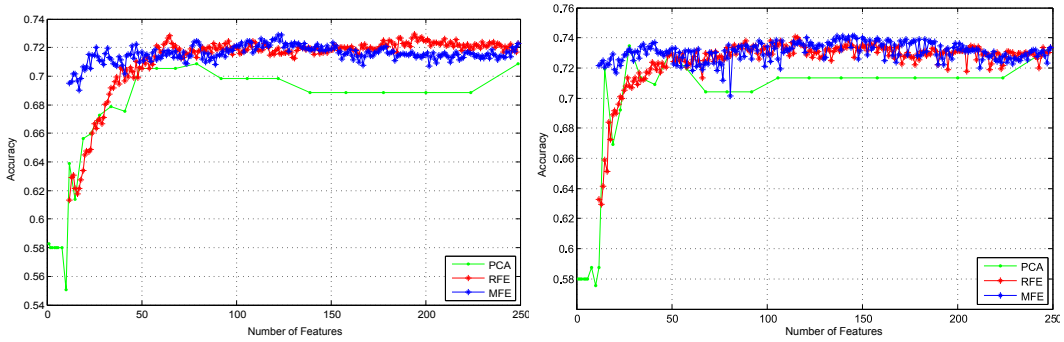
In Table 3.5, we present a summary of the most selected features in the smallest subset finally obtained by RFE and MFE. We only included the features with a percentage of occurrence higher than 0.7, both for the arousal and valence dimensions.

As the summary on Table 3.5 shows, the RFE algorithm selects peripheral signal features

in most of the experiments. Meanwhile, the MFE algorithm gives more relevance to EEG features when choosing the optimal subset. These results are similar in both datasets.

3.2.5 Feature selection validation

A validation stage for the discriminant feature selection methods is developed. A classical dimensional reduction algorithm is used and the classification accuracy from the reduced subset is obtained for several biclass experiments in each database. In Figure 3.10 it can be observed the results for the principal component analysis (PCA) method in comparison against RFE and MFE for the $D1$ and $D2$ datasets. As it can be seen from the results, the performance of the discriminant feature selection algorithms generally have better classification accuracy as the feature space is reduced in comparison to PCA. Note that the performance of PCA in some cases is equal or slightly improves the precision on the classification for some subspaces in comparison with RFE. The selection scheme by MFE has clearly better results in terms of accuracy in all cases, see Figure 3.10.



(a) RFE, MFE and PCA over D1 dataset (b) RFE, MFE and PCA over D2 dataset

Fig. 3.10 Comparison between RFE and MFE against PCA

3.2.6 Additional Results

Several additional tests for feature selection were performed in order to asses other relevant aspects of the emotion recognition problem.

- Since the size of the features extracted from all the EEG signals is more than four times the size of the features from the peripheral signals, a feature selection step using RFE and MFE is previously performed only over the EEG features. By doing this, we reduced the number of EEG features from 219 to 50 most relevant features, this

is, to a similar size in comparison to the size of the peripheral features (50 for EEG and 49 for the other signals). We then formed a new feature set of 99 features, and we again perform RFE and MFE over this new set. The results obtained for this selection scheme show that MFE continues selecting predominantly EEG features in the final subset, even when the pre-selection step is performed, achieving similar levels of classification accuracy, around 72%. RFE keeps selecting heart rate signal features predominantly, reaching classification accuracies around 70%, for the smallest subsets of features selected. These results allow us to conclude that even in the original test, when the EEG features outnumbered the other signal features, the RFE and MFE algorithms selected optimal subsets with consistency. Accuracy levels also have similar values in both scenarios: when the number of EEG features is greater than the number of peripheral features, and when the number of EEG features is similar to the number of peripheral features.

- The analysis of correlation proposed in [93] is also performed in an additional scheme. The original selection test using RFE were made by eliminating one feature at each iteration. In this scheme the algorithm is set to remove a bunch of features in each iteration to perform the RFE-SVM-CBR that allows the inclusion of possible misseliminated features into the selected set based on a correlation analysis. The results from this test over the DEAP and MAHNOB database did not show any inclusion of features possibly removed again into the selected subset.

3.3 Discussion

With all the experiments performed over the different databases using the discriminant feature selection algorithms, the results are consistent with the theory of the recursive feature elimination, where at each iteration a feature (or set of them) is discarded on the basis that it has less relevance for the separation of classes. RFE removes features in each iteration while the classification accuracy remains around the value obtained with the original set of features in both spaces, arousal and valence. Experiments as the one presented in Figure 3.2(a), show that the classification accuracy for $D1$ dataset, after several eliminations, maintains a constant level. This applies even for subsets with sizes less than half of the original set. This behavior is the same for both databases MAHNOB and DEAP.

Based on the results obtained from the experiments, using the selection algorithm MFE, it can be observed that bunches of features are eliminated retaining a percentage of accuracy

of the same magnitude as the total set of features. MFE improves the performance compared to the RFE method in some cases according the statistical analysis. Even with subsets of less than 15% of the original feature space size, the classification accuracy is close to the value obtained with the whole set of features. The results are similar in all experiments for different classification spaces in biclass problems for both databases. The values of sensibility and specificity that has high relevance in medical studies gives an additional insight about the relevance of the study. In general terms, the sensibility of the classifier improves when less features are used, that is more realizations of the principal class recognized adequately. This behavior has an important relevance for using a small set of features for an initial detection of the emotional state.

The CA and F1-Score metrics presented in Table 3.4 show a compilation of the classification results. It can be seen that in most of the cases the CA is higher in the MFE experiment and also the F1-Score shows that the test accuracy using several subsets of features is also higher in most of the cases for the MFE algorithm. Nevertheless the two selection algorithms prove to effectively reduce the dimension of the feature space without affecting the CA dramatically. Also a statistical test confirms that MFE is superior in the CA metric than RFE, as it was presented in section 3.2.

Results from the selected features in the different datasets reveals important information of the signals that are more relevant for the emotion recognition problem. For the RFE selection algorithm, it can be seen that the most selected features come from signals as the heart rate and the respiratory pattern for both arousal and valence dimensions for the DEAP database. For the study conducted on the MAHNOB database, the trend in the set of selected features is the greater inclusion of features from the heart rate HR. A similar analysis for the MFE algorithm shows from the occurrence histograms that the features from the EEG are selected in both arousal and valence dimensions for the final subset.

The EEG signals and the heart rate signal are the signals from where most of the features were selected. This result would be expected in the sense that the brain and the heart (as part of the central nervous system) are the organs that react more rapidly to an external stimulus [40]. Results show that both RFE and MFE are able to pick on this fact, even when both methods do not select exactly the same features. On the other hand, as it was pointed out in section 2.3.3, while MFE removes features taking into account the margin that separates both classes, and attempts to maximize that margin, RFE only looks for reducing the squared weight vector associated to the SVM, ignoring the margin criterion. For testing the relevance of some signals and features individually, another scheme of selection and classification must be implemented.

Additionally, the experiments for the validation of the discriminant feature selection algorithms showed that RFE and MFE outperforms, in the majority of the experiments, the results obtained with PCA. The results from the MAHNOB database are comparable with the works presented in the state of art for this database. Some of the classification results are even higher than those reported in [38] and [39] that work in a similar framework. In [38], the authors used RFE to select features extracted from the EEG signals, and the video signal. For those features selected, the highest F1-scores were 67.1% for arousal, and 71.5% for valence. In our experiments, we obtained an averaged F1-score of 79.25 for arousal, and an averaged accuracy of 75.41 for valence, both results using RFE. In [39], the authors only analyzed the EEG features. The selection algorithm is based on a sequential search of the best feature subset by an inclusion/exclusion features scheme. The best results reported in this work were 65.1% in the arousal dimension, and 63.0% in the valence dimension [39]. Our results show that adding the features from the peripheral signals to the EEG features gives an increment in the F1-score metric in both dimensions, with similar classification accuracy results. Also, the scheme of feature selection from the MFE algorithm outperforms the results obtained with RFE for the smaller subsets of selected features.

The discriminant feature selection methods performed successfully in all the experiments by removing several features without affecting the accuracy rate in the classification task. Several experiments show that the emotion classification in the Arousal/Valence space using the multimodal approach could be improve with an adequate selection feature stage. Nevertheless, any conclusions obtained in terms of the features selected are given in terms of the specific classifier used in this paper. Further studies are needed to assess the performance of other classifiers with the same sets of features selected by our RFE-SVM and MFE-SVM implementations.

For the biclass experiments, the MFE algorithm presented higher classification accuracy than the RFE algorithm for reduced feature subsets in most of the experiments. From the evidence of the results, the more relevant features for emotion classification due to the selection of the features in the smaller subsets are the EEG for both methods. The features from the EEG signal seem to be more relevant in the selection with MFE and the different physiological signals were selected in smaller subsets of features when RFE was applied.

Since there are few works that have made an effort in the feature selection for MER, this work has demonstrated that MER with a stage of feature selection with an embedded methodology based on SVM's could be adapted to this field. Future work could be heading to include audiovisual information and multiclass problems that allows the differentiation of more ranges of Valence and Arousal.

Chapter 4

Spatial dependencies analysis for BEA

4.1 Brain Connectivity

The modern neuroscience has focused on understanding how the brain functions are associated with communication between regions containing a large number of neuronal elements [9]. The formation of distributed networks of the neural population in particular brain regions and the connectivity patterns between them are related to brain processes such as behavior and cognition [73]. Since EEG data is well known by their poor spatial resolution, the analysis of the network connectivities allows adding some spatial information into the processing that will lead to higher performance in BEA recognition [27]. The determination of brain region relationships in behavioral or cognitive responses will improve our understanding of the brain dynamics and the subjacent processes that are related to every conscious state [73].

Three possible interactions can occur between brain cells, connectivity at structural, functional and effective levels. The first one is determined directly by the brain anatomical distribution and the relationships between regions formed by nervous tracts [27]. The second type called functional connectivity is the relationship between isolated brain regions that occurs to perform a specialized information processing. Finally, the effective connectivity is related to the concept of causality and depicts the neural activity produced in some region due to previous activations that happened in other brain regions that can be structurally isolated [27][35]. From the three mentioned connectivity concepts, the functional one attracts a lot of interest from the scientist that are likely to be related to the processes that govern the BEA, integration, and segregation.

From the windowed data in $\hat{\chi}$, a set of connectivity measures that operate between each pair of channels at each window could be defined to find the interdependence between

them. Let $\boldsymbol{u}_n, \boldsymbol{v}_n \in \hat{\boldsymbol{X}}_{n,k}$ be a pair of EEG channels and a connectivity quantification function $\Phi : \mathbb{R}^t \times \mathbb{R}^t \to \mathbb{R}$ that allows to compute statistical dependences between channels u, v. Different Φ functions measures some conditions of dependence between channels related to functional connectivity (FC), the most commonly used are the Pearson correlation coefficient, coherence, phase index synchronization and measures based on information theory such as the entropy and mutual information [12][103].

Functional connectivity measures

Come common FC measures have been used in different works on BEA processing, some of them are presented in this section [27].

Correlation index The linear correlation $\phi_{COR}(\boldsymbol{u}, \boldsymbol{v}) \in [-1, 1]$ between $\boldsymbol{u}$ and $\boldsymbol{v}$ in the time domain is computed by the Pearson's correlation coefficient as:

$$\phi_{COR}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{\sigma_u \sigma_v} \sum_{l=1}^{t} (u_l - \bar{u})(v_l - \bar{u}), \tag{4.1}$$

where $\sigma_u, \sigma_v \in \mathbb{R}^+$ and $\bar{u}, \bar{v} \in \mathbb{R}$ are the standard deviation and the mean values of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively.

Coherence-(COH). The linear time-invariant relationship between $\boldsymbol{u}$ and $\boldsymbol{v}$ at frequency range $[f_{\min}, f_{\max}]$ is calculated trough the coherence measure as:

$$\phi_{COH}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{f_{\max} - f_{\min}} \sum_{f=f_{\min}}^{f_{\max}} \frac{|\zeta_{uv}(f)|^2}{\zeta_{uu}(f)\, \zeta_{vv}(f)}, \tag{4.2}$$

where $\xi_{COH}(\boldsymbol{u}, \boldsymbol{v}) \in [0, 1]$, $\zeta_{uv}(f) \in \mathbb{C}$ is the cross-spectrum of $\boldsymbol{u}$ and $\boldsymbol{v}$, and $\zeta_{uu}(f)$, $\zeta_{vv}(f) \in \mathbb{C}$ are the power spectrum of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively.

Mutual Information-(MI). The MI between $\boldsymbol{u}$ and $\boldsymbol{v}$ allows revealing the uncertainty amount of one time series by observing the other. So, high-order correlations can be computed utilizing probability density estimators as follows:

$$\phi_{MI}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{l=1}^{L} \hat{p}(u_l, v_l) \log\left(\frac{\hat{p}(u_l, v_l)}{\hat{p}(u_l)\hat{p}(v_l)}\right), \tag{4.3}$$

where $\hat{p}(u_l, v_l) \in [0, 1]$ is an estimation of the joint probability density function and $\hat{p}(u_l), \hat{p}(v_l) \in [0, 1]$ are the marginal density function approximations of u_l and v_l,.

Time generalized measure of association (TGMA) Since the correlation captures second order statistics only, the use of generalized measure of association (GMA) is motivated to capture nonlinear structure in data without the cost of free parameter selection. A measure of association computes how dependent are larger values from one random variable (RV) to larger values of a second RV. In the case of GMA, the dependences of a realization $\boldsymbol{u_i}$ from one RV to a realization $\boldsymbol{u_j}$ to the corresponding realizations $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ of the second RV, are measured via a rank vector r_i. Here, the dependence is quantified via a distance metric (i.e. Euclidean distance) in the input spaces of $\boldsymbol{u}$ and $\boldsymbol{v}$, namely, $\delta_u(\cdot,\cdot)$ and $\delta_v(\cdot,\cdot)$ [26]. To estimate the GMA, the computation of the rank $r_i; 1 \leq i \leq t$ of realization $\boldsymbol{v}_{j^*}$ in terms of δ_v, with j^* the index of the closest realization u_j to u_i in terms of δ_u.

$$j^* = arg\min_{j \neq i} \delta_u(\boldsymbol{u}_i, \boldsymbol{u}_j), \tag{4.4}$$

The rank r_i could be considered itself as a RV R, and the distribution of R will quantify the dependence between $\boldsymbol{u}$ and $\boldsymbol{v}$. Then the GMA could be defined mathematically as [26]:

$$\phi_{GMA}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{t-1} \sum_{r=1}^{t-1} (t-r) P(R=r), \tag{4.5}$$

where $P(R=r) = \#\{i : r_i = r\}/t$ and represents the empirical probability of the rank variable. GMA assumes values between $[0.5, 1]$ and its parameter free scheme posses an advantage over other connectivity measures. A possible issue related to the GMA is that the nearest neighbor of a given point will be the nearest in time, and this is not the desired dependence to quantify. TGMA then proposes a modification of the GMA algorithm by decreasing the effect of temporal structure in the input time series [68]. This is performed via a window restriction based on the autocorrelation function (ACF) of each time series $\boldsymbol{\zeta_u}$ and $\boldsymbol{\zeta_v}$. From the connectivity functions applied over the windowed $\hat{\boldsymbol{X}}$ data, the new space of BEA representation is denoted by $\boldsymbol{\Phi} = \{\boldsymbol{U}_{n,k} \in \mathbb{R}^{C \times C}, l_n \in [l_{min}, l_{max}]\} \forall n = 1, \cdots, N; k = 1, \cdots, W$ with N conditions and W windows, and matrix $\boldsymbol{U}_{n,k}$ is generated as equation (4.6) shows [67].

$$\boldsymbol{U}_{n,k}(i,j) = \Phi(\boldsymbol{u}, \boldsymbol{v}) \forall i, j = 1, 2, \cdots, C \tag{4.6}$$

The aforementioned data $\boldsymbol{U}_{n,k}$ is known to be symmetrical since each measure computed is equal between two channels $\Phi(\boldsymbol{u}, \boldsymbol{v}) = \Phi(\boldsymbol{v}, \boldsymbol{u})$.

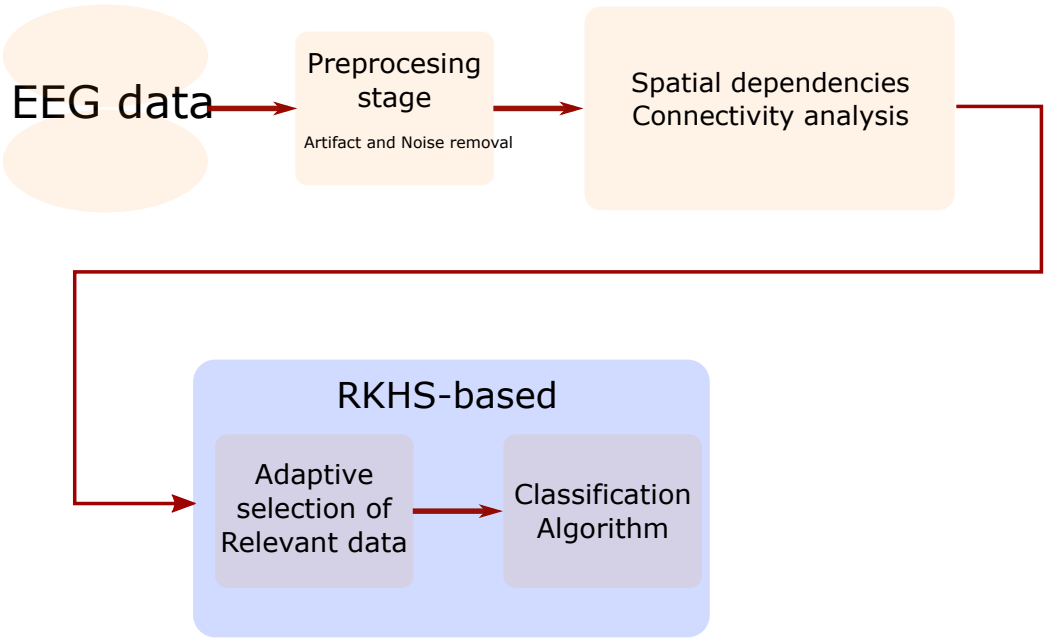


Fig. 4.1 Proposed framework for spatial dependencies analysis within RKHS.

4.2 Connectivity data discrimination

The proposed framework in this section is presented in figure 4.1. The main idea is to use spatial dependencies analysis with a selectio of relevance information.

From the data in $\mathbf{\Phi}$ different schemes of classification can be proposed. But, for the best of our knowledge, the majority of methodologies rely on statical analysis of this type of data. In practice, a class label $\Gamma(l_n)$ must be assigned for discriminative purposes, with:

$$\Gamma(l_n) = \begin{cases} 1; l_n > \xi \\ -1; l_n <= \xi \end{cases} \tag{4.7}$$

where ξ a threshold that allows to divide the data into two classes and $\Gamma(l_n) \in \{-1, +1\}$. Then, from the elements at each $\boldsymbol{U}_{n,k} = a^k_{uv}$, with $u, v = 1, \cdots, C$, the mean and variance of each provided measure along segments are stored in matrices $\mathbf{\Delta}_n \in \mathbb{R}^{C \times C}$ and $\mathbf{\Omega}_n \in \mathbb{R}^{C \times C}$, holding elements:

$$\Delta_{uv} = \frac{1}{W} \sum_{k=1}^{W} a^k_{uv}, \tag{4.8}$$

$$\Omega_{uv} = \frac{1}{W} \sum_{k=1}^{W} \left(a^k_{uv} - \Delta_{uv} \right)^2. \tag{4.9}$$

This analysis derive in the called FC variability (FCV) that allows a general analysis of the change in connectivity across the complete trial of BEA acquisition. Finally, the feature vector $\boldsymbol{y} \in \mathbb{R}^{C(C-1)}$, is built after vector concatenation of $\boldsymbol{\Delta}_n$ and $\boldsymbol{\Omega}_n$ matrices (Δ_{uv}=Δ_{vu} and Ω_{uv}=Ω_{vu}).

4.2.1 Relevance analysis of extracted FCV

Given a provided EEG set, a feature matrix $\boldsymbol{Y}_m \in \mathbb{R}^{N \times C(C-1)}$ can be obtained from Eqs. (4.8) and (4.9) by extracting FCV patterns based on the m-th measure, i.e., COR, COH, and MI. So, to highlight the most relevant connectivity measure regarding the set (subject) at hand, here, we employ a supervised kernel-based relevance analysis to take advantage of the available joint information, associating FCV variations to a given emotion dimension value. Namely, the FCV similarities among EEG trials $\boldsymbol{y}_n, \boldsymbol{y}_{n'} \in \boldsymbol{Y}_m$ are coded by estimating a Gaussian kernel matrix $\boldsymbol{K}_m \in \mathbb{R}^{N \times N}$ on $\boldsymbol{Y}_m$, as follows:

$$k_{nn'} = \exp\left(-\|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|/2\sigma^2\right), \tag{4.10}$$

where $n, n' \in N$ and $\sigma \in \mathbb{R}^+$ is termed the kernel bandwidth. Further, on the emotion dimension space, we also estimate a kernel matrix $\boldsymbol{L} \in \mathbb{R}^{N \times N}$ as follows:

$$l_{nn'} = \delta\left(b_n - b_{n'}\right), \tag{4.11}$$

where $\delta(\cdot)$ is the delta function. It is worth noting that each defined kernel reflects a different notion of similarity (FCV vs. labels). Therefore, we must still evaluate how well the kernel-based similarity matrix $\boldsymbol{K}_m$ matches with the target matrix $\boldsymbol{L}$. To this end, a Centered Kernel Alignment (CKA) functional is used to appraise such a match as the inner product of both kernels to estimate the dependence $\mu_m \in [0,1]$ between the jointly sampled data as follows [21]:

$$\mu_m = \frac{\langle \bar{\boldsymbol{K}}_m, \bar{\boldsymbol{L}} \rangle_{\mathrm{F}}}{\sqrt{\langle \bar{\boldsymbol{K}}_m, \bar{\boldsymbol{K}}_m \rangle_{\mathrm{F}} \langle \bar{\boldsymbol{L}}, \bar{\boldsymbol{L}} \rangle_{\mathrm{F}}}}, \tag{4.12}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ is the matrix-based Frobenius inner product. $\bar{\boldsymbol{K}}$ stands for the centered kernel matrix $\bar{\boldsymbol{K}}$=$\tilde{\boldsymbol{I}}\boldsymbol{K}\tilde{\boldsymbol{I}}$, $\tilde{\boldsymbol{I}}$=$\boldsymbol{I}$–$\mathbf{1}^\top \mathbf{1}/N$, $\boldsymbol{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, and $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector. In this sense, μ_m weights allow ranking the relevance of an FCV, that is, the higher μ_m value the better the m-th FCV representation regarding the emotion labels. So, the highstest weigth value is employed to select the most relevant FCV (RFCV) for a given EEG set.

4.2.2 Testing dataset and preprocessing

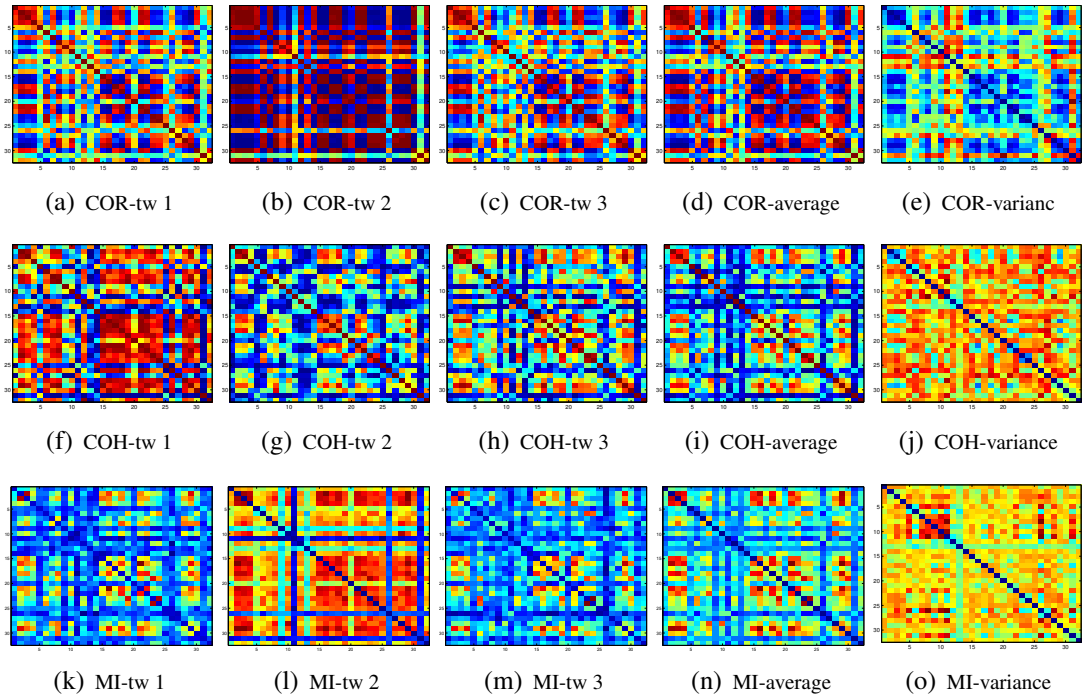
The well-known *Database for Emotion Assessment using Physiological Data* (DEAP) is used to test the introduced FCV approach. The DEAP is publicly available and contains physiological recordings from 40 emotion elicitation experiments of 32 subjects. Each subject was requested to watch a one minute portion of a video that induces a particular emotion, then, an auto-tagging system captured the arousal, valence, dominance, and liking level of each video within the range 1 to 9. The collected data includes the following signals: EEG, electrooculogram, galvanic skin response, temperature, among others. The EEG data were acquired using a 32 channel biosemi configuration at 128 Hz and filtered by an artifact removal stage [37].

4.2.3 FCV training

The proposed FCV approach is tested as feature extraction tool for emotion assessment. Thus, each DEAP subject dataset is configured as a biclass problem for both arousal and valence dimensions. The first class corresponds to arousal/valence levels between 1 and 5. Meanwhile, the second one holds levels between 5 and 9. Furthermore, a window of 9 seconds with 25% overlapping is employed to compute the inter-channel dependencies based on FCV. The fixed window size aims to highlight channel dependencies under alpha, beta, gamma, and theta rhythms along time. Likewise, the configuration of the frequencies bands for the coherence measure are related to the aforementioned rhythms (f_{min}=4Hz and f_{max}=47Hz). Here, the FC measures are computed using the HERMES MatLab toolbox [57]. Subsequently, the FCV-COR, FCV-COH, FCV-MI, and RFCV are computed as in sections 4.2 and 4.2.1, yielding a feature extraction matrix $\boldsymbol{Y}\in\mathbb{R}^{N\times P}$ with N=40 emotion elicitation videos and P=992 features for each considered representation. Finally, the discrimination between emotion classes is carried out based on a k-nearest neighbor classifier under Gaussian similarity criteria. A nested 10-fold cross-validation strategy is used to test the system performance, where the number of nearest neighbors of the applied classifier is fixed as the one reaching the best accuracy within the following testing range $\{1,3,5,7,9,11\}$.

4.3 Results

The FC scheme detailed in section 4.2 allows the visualization of the variability in the connectivity patterns between EEG channels. Figure 4.2 shows an example of some time win-



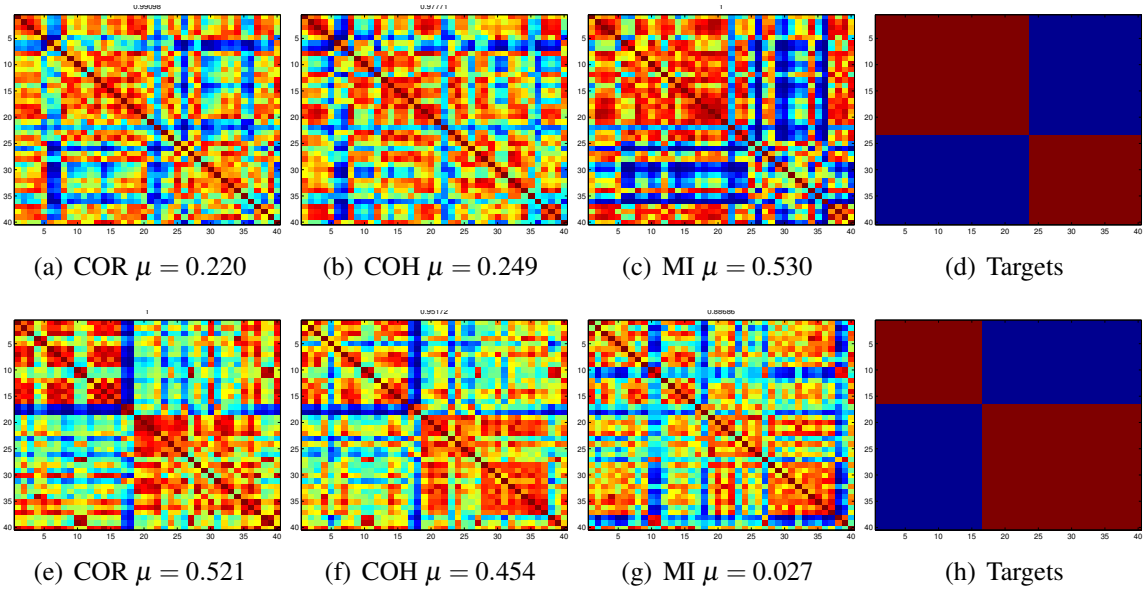
(a) COR-tw 1 (b) COR-tw 2 (c) COR-tw 3 (d) COR-average (e) COR-varianc

(f) COH-tw 1 (g) COH-tw 2 (h) COH-tw 3 (i) COH-average (j) COH-variance

(k) MI-tw 1 (l) MI-tw 2 (m) MI-tw 3 (n) MI-average (o) MI-variance

Fig. 4.2 FC measures for the 32 EEG array in different time window (TW). Top row - COR measure, middle row - COH measure and bottom row - MI measure. Columns $1-3$ from left to right corresponds to each measure in different (non-subsequently) TW. Column 4 is the average and column 5 the variance for all the time windows.

dows from the three measures over the subject 13 in a experiment with arousal and valence ratings of 8.09 and 6.15 respectively. It can be seen in Figure 4.2 the variations in the dependences of channels from the EEG array for few time windows. As seen, the relationships on different channels from the EEG array varies in time, and some strong interdependences could be found according each FC measure. For this particular subject/experiment, the COR measure exhibit a strong interdependences between the majority of channels with a small degree of variability among all the time windows (figures from 4.2(a) to 4.2(c)). On the other hand, for the COH (figures from 4.2(f) to 4.2(h)) and MI measures (figures from 4.2(k) to 4.2(m)), there is a higher degree of variability among time windows. The discussed variability for each measure is consequently summed up in the average and variance figures (columns $4-5$ from figure 4.2). The average FC allows to observe the channels with strong interdependences as well as the channels with weak interdependences in the whole experiment. Likewise, the FC variance shows the channels interdependence variability across the experiment, with a higher degree of variability for the majority of channels in the COH (figure 4.2(j)) and the MI (figure 4.2(o)) measures.



(a) COR $\mu = 0.220$ (b) COH $\mu = 0.249$ (c) MI $\mu = 0.530$ (d) Targets

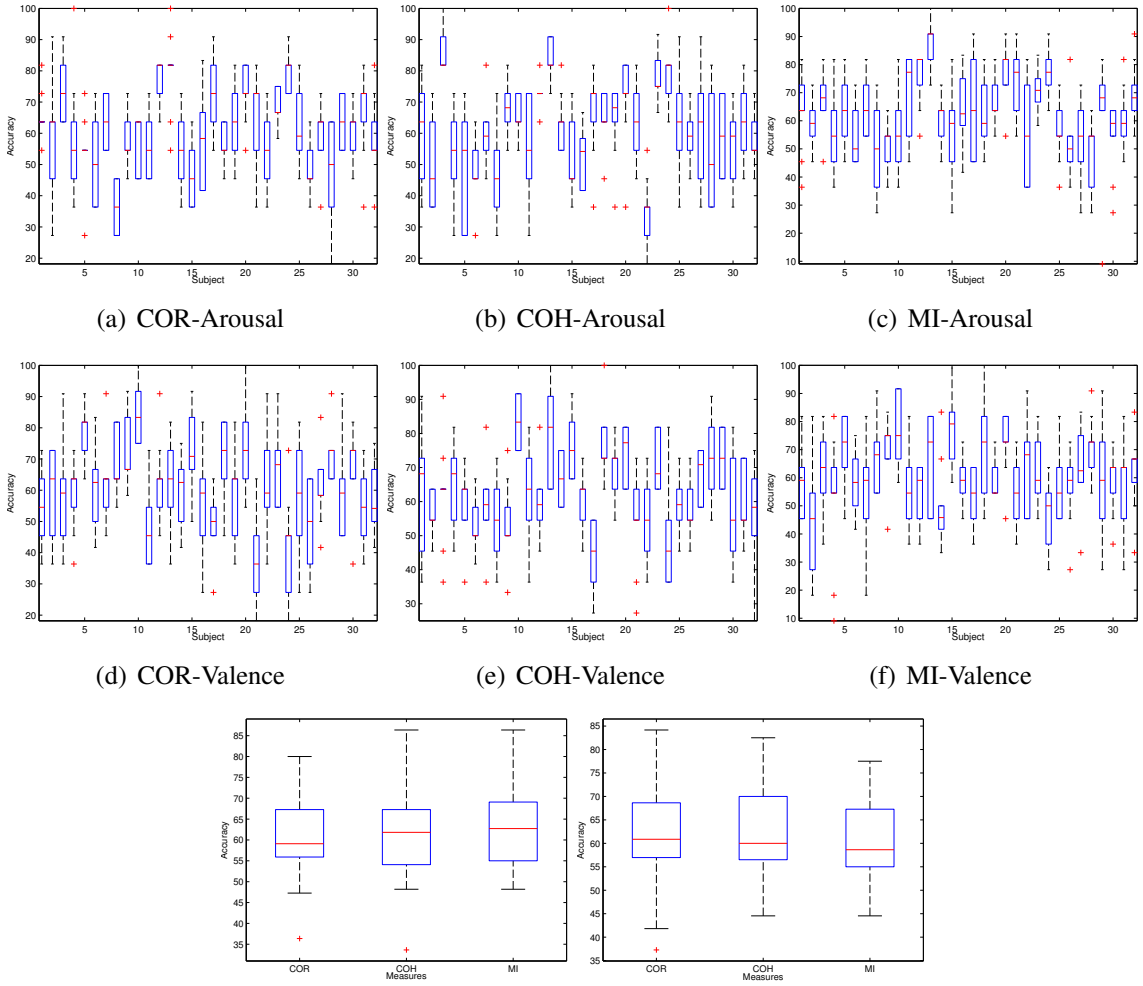
(e) COR $\mu = 0.521$ (f) COH $\mu = 0.454$ (g) MI $\mu = 0.027$ (h) Targets

Fig. 4.3 Gaussian kernel transformation applied to the three FC connectivity measures and the targets matrix for two subjects 13 and 18

On the other hand, in figure 4.3, the FCV-based representation and emotion label similarities for each considered measure (see section 4.2.1) can be analyzed. In this particular case, the FCV corresponds to the subjects 13 and 18 and the set of 40 emotion elicitation experiments. We can infer by visual inspection that exist a higher similarity between the FCV-MI approach and the target matrix for the subject 13 (figures from 4.3(a) to 4.3(d)), which is also coded by the computation of the weights μ in the RFCV representation. In the other case, for the subject 18 there is a higher relation in the FCV-COR with the targets representation than for the FCV-COH and the FCV-MI (figures from 4.3(e) to 4.3(h)). For both cases the RFCV allows to code the measure that seems to present the highest correlation with the targets.

FCV is used for classification purposes as stated in the experimental setup. A graphical description of those results can be found in figure 4.4, where the classification accuracy (CA) for each subject and each dimension are presented. Figures 4.4(a), 4.4(b), and 4.4(c) show CA for the 32 subjects in arousal dimension using the FCV-COR, FCV-COH and FCV-MI representation respectively. Likewise, figures 4.4(d), 4.4(e), and 4.4(f) present the CA for all subjects in the valence dimension. From the figures it can be noticed the differences in CA among subjects that evidences the subject-dependency of the FC measures. Also, a summary for each FCV measure is included in figure 4.4(g) for the arousal dimension and figure 4.4(h) for the valence dimension. From those figures, small differences in the CA

when the FCV scheme is applied could be noticed and there is no evidence of one of the FCV schemes to present a superior performance in comparison to the others.



(a) COR-Arousal (b) COH-Arousal (c) MI-Arousal

(d) COR-Valence (e) COH-Valence (f) MI-Valence

(g) Arousal (h) Valence

Fig. 4.4 Boxplots of classification accuracy (CA) per subject in each FC measure. Top row - arousal, middle row - valence and the bottom row, average CA for both dimensions

Finally, a summary of the results of CA for all the subjects is presented in Table 4.1 for the FCV and RFCV schemes. In this table the results of the proposed methodology are compared against state-of-art works that have been developed in a similar framework using EEG data and the same database (DEAP). It can be seen that for all the works using the DEAP dataset, there is still room for improvement, since the higher results are around 67.00%. Our RFCV approach proves to obtain the higher CA for the valence dimension with 65.73% and the second higher CA for arousal dimension with 66.00%.

Table 4.1 Mean emotion classification results [%] for all considered DEAP subjects.

Reference	Approach	Arousal	Valence
Koelstra et.al. [37]	Linear features, power spectral - SVM	62.00	57.50
Soleymani et.al. [70]	Power spectral - SVM	50.00	62.00
Gupta et.al. [30]	Power spectral - HJORT - SVM	60.00	60.00
Castellanos et.al. [59]	MSP - ROI signal - SVM	58.6	55.76
Daimi et.al. [22]	Wavelet Packet - SVM	67.00	65.00
This proposal	FCV-COR - KNN	61.93	63.35
This proposal	FCV-COH - KNN	63.45	61.62
This proposal	FCV-MI - KNN	62.48	60.78
This proposal	**RFCV - KNN**	66.00	65.73

4.4 Discussion

We introduced a novel FC representation approach for feature extraction to enhance automatic emotion assessment from EEG data. To this end, the proposed strategy incorporates three well-known FC measures: coherence, correlation, and mutual information, to code the temporal variability of EEG inter-channel dependencies. Moreover, a supervised kernel-based relevance analysis based on CKA is used to evaluate the significance of each FC variability regarding the considered measures. Our approach learns both important temporal inter-channel variations and relevant FC measures to deal with inter-subject dependency in emotion classification. Validation of the proposed feature extraction, termed RFCV, is carried out in a public dataset (DEAP). Attained results demonstrate that RFCV is a reliable methodology for emotion assessment in comparison to the state-of-art works. As future work authors plan to couple RFCV with a space state strategy to deal appropriately with the intrinsic EEG nonstationarity. Besides, information theory measures could be employed to reveal connectivity variations instead of variability-based criteria.

Chapter 5

Spatio-temporal dependencies analysis

Instead of selecting the relevant spatial data or coding its variability, a more complex framework include the dynamical variations of them. This scheme is known as spatial-dynamic BEA analysis and can be developed by the implementation of information-theory metrics. Some works have developed methodologies that include the analysis of the dynamics of spatial dependences by including additional BEA data from fMRI such as [17]. An initial attempt to include a sort of temporal analysis of functional connectivity (FC) was presented in [18], where a network functionality-based strategy is designed to quantify the variations of the connectivity networks in resting state BEA. Then, the use of Hidden Markov moddel (HMM) is used to estimate the transition probabilities from graph network measures via FC. Some limitations as the selection of the number of hidden states are reported within this application. Another attempt to introduce a dynamical analysis of FC is proposed in [68] for cognitive states recognition. In this case, the time generalized measure of association (TGMA) within each pair of channels is modeled as a random variable within segmented data windows by a probability density function (PDF) using Parzen estimation with a Gaussian Kernel. For the classification of BEA, the PDFs are compared for consecutive windows with the Cauchy-Schwartz divergence, from the results it can be seen that some relationships between regions could be observed in particular BEA patterns across time variations. Figure 5.1 shows the proposed framework related with the analysis of spatio-temporal information on BEA data.

5.1 Relevance Analysis based on kernel adaptive filtering

BEA data processing. For a particular subject, let $\boldsymbol{\chi} = \{\boldsymbol{X}_n \in \mathbb{R}^{C\times T}; \boldsymbol{y}_n \in \mathbb{R}^T; l_n \in [l_{min}, l_{max}]\}_{n=1}^N$ be and EEG set data with C channels at T time instants, where $\boldsymbol{X}_n$ holds the $n-$th EEG trial

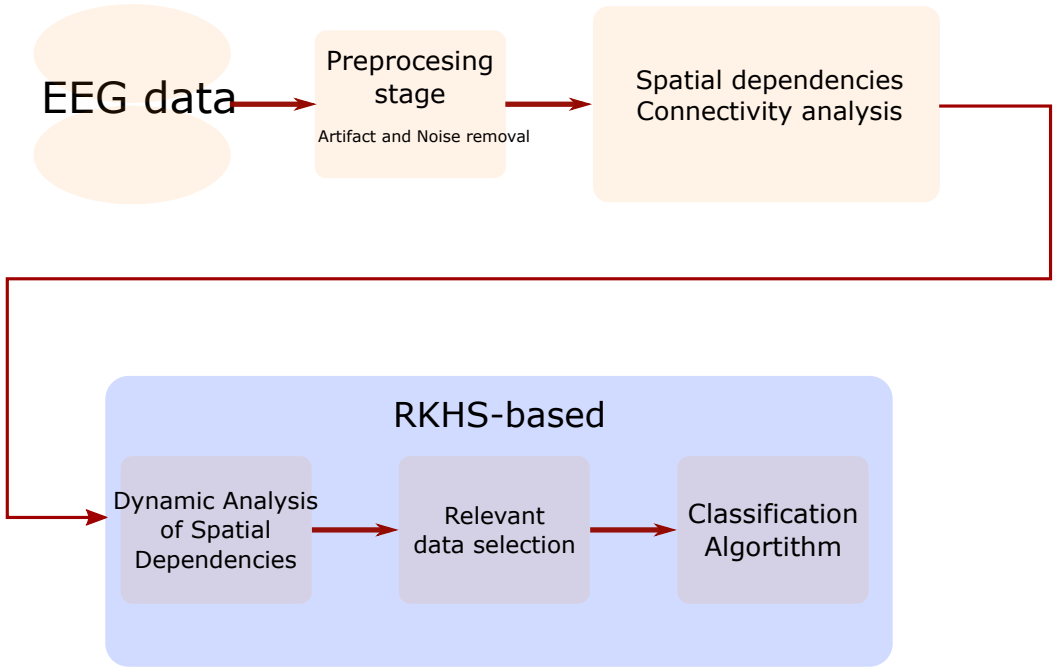


Fig. 5.1 Proposed methodology for the spatio-temporal analysis of BEA using RKHS

with emotional level l_n and physiological response $\boldsymbol{y}_n$. Then, a windowing function is applied to segment the data into W windows holding t time samples per window, yielding $\mathscr{Z}_n = \{\bar{\boldsymbol{X}}_{n,k} \in \mathbb{R}^{C\times t}; \bar{\boldsymbol{y}}_{n,k} \in \mathbb{R}^t\}$, with $n = \{1,\ldots,N\}$ and $k = \{1,\cdots,W\}$. Now, a feature extraction stage is performed over $\mathscr{Z}_n$ based on a given connectivity measure to obtain the representation set $\mathscr{U}_n = \{\boldsymbol{U}_{n,k} \in \mathbb{R}^{C\times C}; \boldsymbol{d}_{n,k} \in \mathbb{R}\}$. In particular, $\boldsymbol{U}_{n,k} = \phi_e(\bar{\boldsymbol{X}}_{n,k}) \in \mathbb{R}^{C\times C}$ codes the connectivity quantification and $\boldsymbol{d}_{n,k} = \phi_p(\bar{\boldsymbol{y}}_{n,k}) \in \mathbb{R}^W$ characterizes the physiological response, both within the k-th window. Here, we employ the correlation index and the time generalized measure of association (TGMA) as connectivity measures to capture the channel dependencies [67]. A detailed explanation of the correlation index and the TGMA could be found in [27] and [32], respectively. For simplicity and taking into account the symmetric property of the correlation and the TGMA measures, the feature space is rearranged via vector concatenation to build the set $\Omega_n = \{\boldsymbol{U}_{n,k} \in \mathbb{R}^{(C*(C-1)/2)}; \boldsymbol{d}_{n,k} \in \mathbb{R}\}$.

5.2 Information Theory

Any machine learning algorithm could be seen as the process of adjusting an specific mathematical model to particular data. Depending of the selected model, a set of parameters should be adjusted and optimized according the available data. In a supervised scheme, the

data comes as a set of $\{\boldsymbol{u}(i), d(i)\}$ with $\boldsymbol{u}(i) \in \mathbb{R}^{M \times N}$ and $d(i) \in \mathbb{R}^N$ for M data dimensions in N samples.

From this point of view, if the initial available data has a high number of N samples, the parameters of the model could be adjusted following a classic machine learning scheme of optimization. But in another scenario, if the data comes sequentially sample by sample, the adjustment of the model parameters should be done in an adaptive way. This implies that if a new sample of data $\{\boldsymbol{u}(i), d(i)\}$ is available, the parameters of the model are updated according some rules that are related to an specific cost function. Then the model produces a hypothesis $h(i)$ that depends not only on the previous hypothesis $h(i-1)$ but also on the new data sample $\{\boldsymbol{u}(i), d(i)\}$. In general terms, the adaptive algorithm should generate the hypothesis from the set of $\{\boldsymbol{u}(0), d(0), \boldsymbol{u}(1), d(1), \cdots \boldsymbol{u}(i-1), d(i-1)\}$ and the current one, but some assumptions have to be included in order to optimize the memory usage and the computation complexity.

5.2.1 Linear Adaptive Filters

This type of algorithms allows a sequential learning that enables a "filter" or mapping to adjust its free parameters automatically in response to statistical variations in the environment in which the filter operates. Commonly an error-based cost function is implemented in order to perform the automatically updating task. A general scheme of this type of filters could be depicted in Figure 5.2. In this scheme, it can be seen how the error-based rule depends on the $\boldsymbol{w}(i-1)$ weights corresponding to the adjustable parameters of the model at time sample $(i-1)$. With an incoming input $\boldsymbol{u}(i)$ applied to the filter at time i, a response $y(i)$ is produced and compared to the external label of the data sample $d(i)$ that generates an error signal $e(i)$. The error signal gives the amount $\Delta\boldsymbol{w}(i)$ of an incremental quantity to produce the adjustment in the weight vector $\boldsymbol{w}(i-1)$. So, the new weights will be $\boldsymbol{w}(i) = \boldsymbol{w}(i-1) + \Delta\boldsymbol{w}(i)$. This process is repeated until the filter reaches a condition or there is no more input data to learn.

5.2.2 Least-Mean-Square algorithm

The simplest an commonly used filter for adaptive learning is the Least Mean Squares (LMS). The LMS algorithm tries to minimize the instant cost function:

$$J(i) = \frac{1}{2}e^2(i); e(i) = d(i) - \boldsymbol{w}(i-1)^T\boldsymbol{u}(i)$$

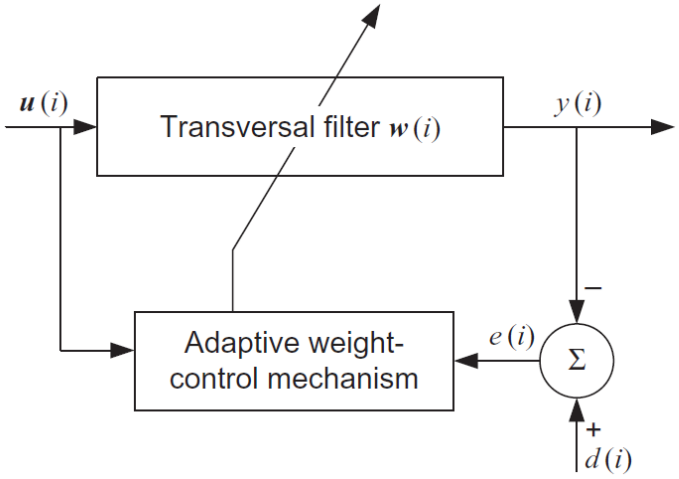


Fig. 5.2 General scheme of linear filter algorithms

Then, the minimization is as follows:

$$\frac{\delta}{\delta \boldsymbol{w}(i-1)} J(i) = -e(i)\boldsymbol{u}(i), \tag{5.1}$$

and using the gradient descent algorithm the adjustment $\Delta \boldsymbol{w}(i)$ applied to the algorithm at time i is:

$$\Delta \boldsymbol{w}(i) = \eta e(i)\boldsymbol{u}(i), \tag{5.2}$$

with η the step parameter or learning rate. And the updated weight estimate is computed as:

$$\boldsymbol{w}(i) = \boldsymbol{w}(i-1) + \eta e(i)\boldsymbol{u}(i). \tag{5.3}$$

See appendix D.0.1 for the detailed algorithm of the LMS. This algorithm could perform very efficiently besides its simplicity if the η parameter is well chosen. For best performance the value of η should be small but from a practical perspective it derives in slow convergence.

5.2.3 Kernel Least Mean Squares (KLMS)

The use of kernel functions over data and the derived mapping into a reproducing kernel Hilbert space (RKHS) is included in this case. The RKHS properties allows the derivation of linear adaptive algorithms and to obtain non-linear filters in the input space. A kernel

induced mapping is employed to transform the input $\boldsymbol{u}(i)$ into a higher dimensional feature space $\mathbb{F}$ as $\boldsymbol{\phi}(\boldsymbol{u}(i))$. In this higher dimensional space, the prediction $\boldsymbol{\omega}^T\boldsymbol{\phi}(\boldsymbol{u})$ is much powerful than $\boldsymbol{w}^T\boldsymbol{u}$ because the difference in dimensionality of $\boldsymbol{u}$ and $\boldsymbol{\phi}(\boldsymbol{u})$. So, the LMS algorithm over the new data $\{\boldsymbol{\phi}(i), d(i)\}$ representation will be:

$$\boldsymbol{\omega}(0) = 0 \tag{5.4}$$

$$J(\boldsymbol{\omega}) = \sum_{i=1}^{N}(d(i) - \boldsymbol{\omega}(i-1)^T\boldsymbol{\phi}(i))^2 \tag{5.5}$$

$$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta e(i)\boldsymbol{\phi}(i), \tag{5.6}$$

with $J(\boldsymbol{\omega})$ the cost function to minimize and $\boldsymbol{\omega}(i)$ denoting the estimate of weights at iteration i in $\mathbb{F}$. However, since ϕ is only implicitly known an alternative way of computation the weight updating step is needed. That is, by the kernel trick the computation of the filter output in the input space by kernel evaluations.

$$\boldsymbol{\omega}(i) = \eta \sum_{j=1}^{i} e(j)\boldsymbol{\phi}(j) \tag{5.7}$$

$$\boldsymbol{\omega}(i)^T\boldsymbol{\phi}(\boldsymbol{u}') = \eta \sum_{j=1}^{i} e(j)\kappa(\boldsymbol{u}(j), \boldsymbol{u}') \tag{5.8}$$

See appendix D.0.2 for the detailed algorithm of the KLMS. In this case $f_{i-1}\boldsymbol{\omega}(i)^T\boldsymbol{\phi}(\boldsymbol{u}')$ is the output of the filter and $C(i) = \boldsymbol{u}(i)$ corresponds to each center and the quantity $\boldsymbol{a}(i)$ the corresponding coefficient. In the particular case when the $\kappa(\cdot,\cdot)$ is a Gaussian kernel, the σ parameter, known as the kernel bandwidth has to be selected as input of the KLMS algorithm. The specification of σ could be done via crossvalidation, nearest neighbor, penalizing functions among others schemes.

5.2.4 Quantized Kernel Least Mean Squares (QKLMS)

One of the possible drawbacks of the KLMS algorithm is that the size of the network increases according the number of training data related to the stored centers. In order to reduce the memory cost of the network, a novelty criterion could be included in the KLMS algorithm to store only the relevant centers form the data $\boldsymbol{u}$. Suppose the present dictionary is $C(i) = \{\boldsymbol{c}_j\}_j = 1^m$, a decision over a new input $\{\boldsymbol{u}(i+1), d(i+1)\}$ to be added into the

dictionary as a new center depends on a distance criterion:

$$D(C,\boldsymbol{u}(i+1)) = min_{\boldsymbol{c}j\in C(i)}\|\boldsymbol{u}(i+1)-\boldsymbol{c}(j)\| \tag{5.9}$$

If $D(C,\boldsymbol{u}(i+1)) < \delta$ with δ a predefined threshold the new input $\boldsymbol{u}(i+1)$ will be not added into the dictionary. In this case the cost function is computed only over the set of stored centers.

$$J(\boldsymbol{\omega}) = \sum_{i=1}^{N}(d(i) - \sum_{j=1}^{\|C(i-1)\|}\boldsymbol{a}(j)\kappa(C(j),\boldsymbol{u}(i))^2) \tag{5.10}$$

See appendix D.0.3 for the detailed algorithm of the QKLMS.

5.2.5 Physiological responses estimation and BEA data selection

With the aim of estimating $\boldsymbol{d}_{n,k}$ from Ω_n, a kernel induced mapping $\kappa(\cdot,\cdot)$ is employed to transform $\boldsymbol{U}_{n,k}$ into a higher dimensional feature space $\mathbb{F}$ through the nonlinear mapping function $\boldsymbol{\psi} : \mathbb{R}^{C*(C-1)/2} \rightarrow \mathbb{F}$. Then, an adaptive learning approach is applied, termed the kernel least mean squares (KLMS), which minimizes the following cost function [46]:

$$\min_{\boldsymbol{w}} \sum_{k=1}^{W}(\boldsymbol{d}_{n,k} - \boldsymbol{\omega}(k-1)^\top \boldsymbol{\psi}(\boldsymbol{U}_{n,k}))^2, \tag{5.11}$$

where an updating scheme can be inferred as $\boldsymbol{\omega}(k) = \eta\sum_{j=1}^{i} e(j)\boldsymbol{\psi}(j)$, allowing to efficiently estimate the desired output as $\eta\sum_{j=1}^{k} e(j)\kappa(\boldsymbol{U}_{n,j},\boldsymbol{u}')$, being $\eta \in [0,1]$ a learning rate parameter and $e(j) = \boldsymbol{d}_{n,j} - \boldsymbol{\omega}(j-1)^\top\boldsymbol{\psi}(\boldsymbol{U}_{n,j})$. Now, to extract relevant BEA patterns, the quantized-KLMS (QKLMS) algorithm is used via the following novelty criterion:

$$D_k = \min_{\boldsymbol{c}(j)\in\boldsymbol{C}_k} \|\boldsymbol{U}_{n,k+1} - \boldsymbol{c}(j).\|_2 \tag{5.12}$$

where $\boldsymbol{c}(j) \in \mathbb{R}^{(C*(C-1)/2)}$ is the j-th codeword in the codebook $\boldsymbol{C}_k$. If $D_k < \delta$, ($\delta \in \mathbb{R}^+$) then the new input $\boldsymbol{U}_{n,k+1}$ will be not added into the dictionary, otherwise, $\boldsymbol{U}_{n,k+1}$ is added as a new codeword of $\boldsymbol{C}_k$. So, not only the estimation of the physiological response is performed but also the most relevant centroids will be stored. Then, at this stage only the selected centroids as well as the QKLMS coefficients correspond to relevant BEA data.

The models in the input space from adaptive filtering algorithms are often linearly (i.e Linear Mean squares (LMS)). LMS performs the adjustment of a $w \in \mathbb{R}^L$ set of weights that in combination with an input $\boldsymbol{U}(k)$, produces an output $y(k) = \boldsymbol{w}^\top \boldsymbol{U}(k)$ that is compared to the reference output $d(k)$ via an error function. Then, a cost function is used to update the weights values. Although this type of algorithms achieve efficient results, the linearity of its scheme underperforms in scenarios with high non-linear relationship between the inputs and the outputs. In those cases, the Reproducing Kernel Hilbert spaces (RKHS) allows to use a similar scheme with a non-linear mapping in a higher dimensional space where linear dependences could be find.

For the initial step of BEA data processing, a partition of the time series into 2 seconds segments is performed, resulting in $W = 19$ possible windows [67]. A second step corresponds to the brain connectivity computation, using the correlation and TGMA measures that are applied to the EEG data. Then, the physiological estimation from the adaptive filtering is performed using two different signals, the GSR, and the BVP as the desired outputs. The data is divided in two classes from the $\boldsymbol{l}$ labels by splitting the data into high and low ratings in each dimension (arousal and valence), with $1-3$ for the low class and $7-9$ for the high-class [82]. In this stage, the subjects with non-balanced classes are discarded for classification consistency. Now, for the relevance analysis using QKLMS, a validation procedure is employed by using the 80% of the data for training and the left 20% for testing purposes, a mean square error is computed in the testing set at each condition. Some parameters needed in the ALRA and the classification methodologies are settled, the Kernel bandwidth $\sigma = 10$, the novelty criterion $\delta_1 = \sqrt{\frac{1}{2}\sigma}$ [46] and the learning rate $\eta = 0.9$ are fixed for the QKLMS algorithm. Likewise, in the classification stage, a strategy of cross-validation with 80% for training and 20% for testing within 10 fold repetitions is performed. The value of σ_{svm} is selected by heuristic search into a grid of predefined values within the cross-validation scheme. For further reference, the $\kappa(\cdot,\cdot)$ Kernel function used in the QKLMS algorithm implementation and for classification purposes is the Gaussian Kernel.

BEA classification

From the connectivity measures and the application of the QKLMS strategy, discrimination of the BEA patterns of each condition could be developed. Under a supervised learning scheme, the l_n labels for all the conditions are used to partition the data into biclass problems, but only for those subjects with balanced classes. As a matching criterion, both datasets, $\mathscr{Z}$ containing the connectivity data and Ω holding the relevant data, are employed in the following experiments.

– Classification of emotional states by using relevant data. The number of *Nc* selected centroids varies at each condition, and the data is ordered by weighting each selected window by the corresponding QKLMS expansion coefficients.

– Classification using connectivity data. The mean and variance for each $\mathscr{Z}_n$ along the whole time windows is stored in the matrices $\blacksquare \in \mathbb{R}^{C\times C}$ and $\blacksquare \in \mathbb{R}^{C\times C}$. Next, the feature vector coding the variability of the connectivity measures is built after vector concatenation of $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$ resulting in the set $\{\boldsymbol{X}_C \in \mathbb{R}^{N\times P}; \boldsymbol{l}^\top \in \mathbb{R}^N\}$.

5.3 Gaussian Processes

5.3.1 Cross-spectral estimation from kernel mixtures

The communication between neuron cells is the basis of every neuronal processing task. The electrical impulses result in every possible cognitive or physiological condition, such as behavior, sensation, thoughts, and emotions [6]. Due to equally measuring normal and abnormal BEA, EEG is considered a well-suited neuroimaging technique for diagnosis, treatment, and clinical procedures across several neurological pathologies [65]. Equation (5.13) presents the mathematical representation of the BEA from an EEG of C channels holding T time instants [78].

$$\boldsymbol{\chi} = \{\boldsymbol{t} \in \mathbb{R}^T, \boldsymbol{X} \in \mathbb{R}^{C\times T}\}, \tag{5.13}$$

with $\boldsymbol{t}$ as the time sample positions of the recordings, and $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^C$ holding the brain electrical responses measured by the EEG array at channel i. To quantify the spectral content between channels, the introduced cross-spectrum estimation relies on specific covariance functions. The Cramer's theorem states that a family of integrable functions $\{\kappa_{ij}(\tau)\}_{i,j=1}^C$ are the covariance functions of a weakly-stationary stochastic process if and only if they admit the following representation:

$$\kappa_{ij}(\tau) = \int_{\mathbb{R}^n} e^{\imath\omega^\top\tau} S_{ij}(\omega) d\omega, \tag{5.14}$$

being $\imath$ the imaginary unit, each $S_{ij}(\omega)$ an integrate complex-valued function $S_{ij} : \mathbb{R} \to \mathbb{C}$ that is also positive definite, and i, j the indices of two EEG channels. This relationship

between covariance functions κ_{ij} in the time domain with argument $\tau \in \mathbb{R}$ and their corresponding spectral density S_{ij} with arguments ω in the Fourier domain allows designing a desired spectral density and obtaining a covariance function [61].

Now, a family $\boldsymbol{S} = \{S_{ij}\}_{i,j=1}^{C} \in \mathbb{R}^{C\times C}$ of positive-definite complex-valued functions can be used as cross-spectral densities for multi-output data [61]. These functions are designed by including specific parameters that allow physical interpretation of the obtained covariance kernel regarding the input data. Moreover, complex-valued and positive-definite matrices can be decomposed in the form $\boldsymbol{S}(\omega) = \boldsymbol{R}^H(\omega)\boldsymbol{R}(\omega)$ where $\boldsymbol{R}(\omega) \in \mathbb{R}^{Q\times C}$, Q represents the rank of decomposition, and $(\cdot)^H$ denotes the Hermitian operator. Since Fourier transforms and multiplications of squared exponential (SE) functions are also SE, the auto-covariance function $R_i(\omega)$ of i-th channel is modeled as the complex-valued SE in Equation (5.15).

$$R_i(\omega) = w_i \exp\left(-\frac{1}{4}\frac{(\omega-\mu_i)^2}{\sigma_i^2}\right)\exp\left(-\imath(\theta_i\omega+\phi_i)\right), \tag{5.15}$$

with $w_i, \phi_i, \mu_i, \theta_i \in \mathbb{R}$ and $\sigma_i \in \mathbb{R}^+$. With such a choice of functions, the cross-spectral density between channels i and j is given by Equation (5.16) with covariance $\sigma_{ij} \in \mathbb{R}^+$, mean $\mu_{ij} \in \mathbb{R}$, magnitude $w_{ij} \in \mathbb{R}$, delay $\theta_{ij} \in \mathbb{R}$, and phase $\phi_{ij} \in \mathbb{R}$.

$$S_{ij}(\omega) = w_{ij}\exp\left(-\frac{1}{2}\frac{(\omega-\mu_{ij})^2}{\sigma_{ij}} + \imath(\theta_{ij}\omega+\phi_{ij})\right) \tag{5.16}$$

Finally, in order to guarantee that the model is restricted to real-valued stochastic processes, the spectral density is reassigned to become symmetric with respect to ω by $S_{ij}(\omega) \to \frac{1}{2}(S_{ij}(\omega) + S_{ij}(-\omega))$. Then, the inverse Fourier transform of the resulting cross-spectral density becomes the corresponding temporal domain real-valued kernel; the kernel and the symmetric version of the spectral density are presented in Equations (5.17) and (5.18), respectively.

$$\kappa_{ij}(\tau) = \alpha_{ij} \exp\left(-\frac{1}{2}(\tau+\theta_{ij})^2\sigma_{ij}\right) cos\left((\tau+\theta_{ij})\mu_{ij}+\phi_{ij}\right) \tag{5.17}$$

$$\begin{aligned} S_{ij}(\omega) = &\frac{w_{ij}}{2}\exp\left(-\frac{1}{2}\frac{(\omega-\mu_{ij})^2}{\sigma_{ij}}+\imath(\theta_{ij}\omega+\phi_{ij})\right) \\ &+\frac{w_{ij}}{2}\exp\left(-\frac{1}{2}\frac{(\omega+\mu_{ij})^2}{\sigma_{ij}}+\imath(-\theta_{ij}\omega+\phi_{ij})\right) \end{aligned} \tag{5.18}$$

where the term $\alpha_{ij} = w_{ij}\sqrt{2\pi}|\sigma_{ij}|^{1/2}$ absorbs the constant resulting from the inverse Fourier transform. Equation (5.17) allows computing the real-valued autocovariances $(i = j)$ and cross-covariances $(i \neq j)$ with negatively and positively correlated channels through the magnitude parameter $\alpha_{ij} \in \mathbb{R}$; delayed channels through the $\theta_{ij} \neq 0$ delay parameter, and channels out-of-phase through the $\phi_{ij} \neq 0$ phase parameter. Moreover, increasing the rank of decomposition Q corresponds to considering more components in the multiple-output spectral mixture (MOSM) kernel as shown in equations (5.19) and (5.20).

$$\kappa_{ij}(\tau) = \sum_{q=1}^{Q} \alpha_{ij}^{(q)} exp\left(-\frac{1}{2}\frac{(\tau+\theta_{ij}^{(q)})^2}{\sigma_{ij}^{(q)}}\right) cos\left((\tau+\theta_{ij}^{(q)})\mu_{ij}^{(q)}+\phi_{ij}^{(q)}\right) \tag{5.19}$$

$$\begin{aligned} S_{ij}(\omega) = &\sum_{q=1}^{Q}\frac{w_{ij}^{(q)}}{2}\exp\left(-\frac{1}{2}\frac{(\omega-\mu_{ij}^{(q)})^2}{\sigma_{ij}^{(q)}}+\imath(\theta_{ij}^{(q)}\omega+\phi_{ij}^{(q)})\right) \\ &+\sum_{q=1}^{Q}\frac{w_{ij}^{(q)}}{2}\exp\left(-\frac{1}{2}\frac{(\omega+\mu_{ij}^{(q)})^2}{\sigma_{ij}^{(q)}}+\imath(-\theta_{ij}^{(q)}\omega+\phi_{ij}^{(q)})\right) \end{aligned} \tag{5.20}$$

denoting the superindex (q) the $q-$th spectral component. Then, MOSM effectively computes autocovariance and cross-covariances through the spectral-mixture of positive-definite kernels from the Fourier transform of spectral functions $S_{ij}(\omega)$. In practice, the adjustment of the cross-spectrum parameters should be performed on the evidence of the EEG data.

5.3.2 Multi-output spectral mixture Gaussian Process

Given an EEG trial, the Gaussian Process (GP) probabilistic framework computes the mixture parameters by maximizing the data likelihood as the cost function. A Gaussian Process

(GP) is a real-valued stochastic process $(f(t))$ over a input set $\boldsymbol{t}$, such that for any finite subset of inputs $t \in \{1,\dots,T\}$, the random variables $f(t)$ are jointly Gaussian [7]. Additionally, the GP is uniquely determined by its mean function $m(\boldsymbol{t}) := E_t(f(t))$, typically assumed $m(\boldsymbol{t}) = 0$ and its covariance function $\kappa(\boldsymbol{t},\boldsymbol{t}) := cov(f(t), f(t')) \in \mathbb{R}^{T\times T}$ known as the kernel.

Then, the multivariate extension of GPs is derived by assembling C different scalar-valued stochastic processes, one for each EEG channel. Any finite collection of values across all such processes are jointly Gaussian, termed multiple-output Gaussian Process (MOGP). This extension results in a vector-valued process $\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{m},\boldsymbol{K})$, where $\boldsymbol{m}(\boldsymbol{t}) \in \mathbb{R}^{TC}$ is a concatenated vector from the mean vectors associated to the outputs and $\boldsymbol{K} \in \mathbb{R}^{TC\times TC}$ a block partitioned matrix of the form [7]:

$$\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{K}(\boldsymbol{X}_1,\boldsymbol{X}_1) & \cdots & \boldsymbol{K}(\boldsymbol{X}_1,\boldsymbol{X}_C) \\ \boldsymbol{K}(\boldsymbol{X}_2,\boldsymbol{X}_1) & \cdots & \boldsymbol{K}(\boldsymbol{X}_2,\boldsymbol{X}_C) \\ \cdots & \cdots & \cdots \\ \boldsymbol{K}(\boldsymbol{X}_C,\boldsymbol{X}_1) & \cdots & \boldsymbol{K}(\boldsymbol{X}_C,\boldsymbol{X}_C) \end{bmatrix}, \tag{5.21}$$

where each block $\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_j)$ is a $T \times T$ matrix denoting the covariance between output channels i, j. Furthermore, a multivariate kernel $\boldsymbol{K}(t,t')$ is stationary if $\boldsymbol{K}(t,t') = \boldsymbol{K}(t-t')$. In this case, the kernel becomes $\kappa_{ij}(t,t') = \kappa_{ij}(\tau)$ if substituting $\tau = t - t'$. Therefore, we use the MOSM kernel in Equation (5.19) as the covariance function to be implemented within the MOGP. By defining such a process as the MOSM kernel, the model adjustment to the data is performed by maximizing the data log-probability. Since the observations in the multiouput case are jointly Gaussian, they are concatenated into the vector $\boldsymbol{y} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top \cdots, \boldsymbol{x}_C^\top]^\top \in \mathbb{R}^{CT}$ the channel observed value. Then, the negative log-likelihood (NLL) can be expressed as in Equation (5.22).

$$-\log p(\boldsymbol{y}|\boldsymbol{t},\boldsymbol{\Theta}) = \frac{CT}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{K}| + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y}, \tag{5.22}$$

with $\Theta = \{w_i^{(q)}, \mu_i^{(q)}, \sigma_i^{(q)}, \theta_i^{(q)}, \phi_i^{(q)}, \sigma_i^2\}_{i=1,q=1}^{C,Q}$ holding the complete set of parameters. As a result, minimization of NLL with respect to Θ designs an spectral kernel quantifying the EEG channel relationships at automatically tuned frequency bands.

5.3.3 Discriminative scheme using MOSM-GP

Let a set of N labeled BEA trials $\{\boldsymbol{\chi}_n, l_n\}_{n=1}^{N}$, each of them belonging either class A or B, that is, $l_n \in \{A, B\}$. In the case of DEAP dataset, classes correspond to low and high valence, while for MI dataset left and right hand movement imagination are considered. As stated in section 5.3.2, a single MOSM-GP models each BEA trial as the stochastic process $\boldsymbol{f}_n^{(l)}$ resulting in N_A and N_B MOSM-GPs for classes A and B, respectively. Furthermore, on the evidence of a new BEA trial $\boldsymbol{X}_*$, the marginal likelihood for each learned MOSM-GP is computed as:

$$p(\boldsymbol{f}_n^{(l)}(\boldsymbol{X}_*)|\boldsymbol{X}_n, \boldsymbol{f}_n^{(l)}, \boldsymbol{X}_*) = \mathcal{N}(\boldsymbol{f}_n^{(l)}(\boldsymbol{X}_*), \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*)), \tag{5.23}$$

By evaluating the marginal likelihoods on all training MOSM-GPs, the new BEA trial label is estimated as follows:

$$l_* = \underset{l \in \{A,B\}}{\arg\max}\, E\left\{ p(\boldsymbol{f}_n^{(l)}(\boldsymbol{X}_*)|\boldsymbol{X}, \boldsymbol{f}_n^{(l)}, \boldsymbol{X}_*) : l_n = l \right\}, \tag{5.24}$$

where $E\{\cdot : l_n = l\}$ denotes the expectation operator over training trials belonging to class l. Figure 5.3 illustrates the proposed discriminative MOSM-GP framework, termed DMOSM-GP.

5.3.4 Methodological details

Before the model training stage, a channel selection is carried out to reduce the training computational cost. For the MI dataset, channels are selected based on the evidence that body movement triggers neural activity in the opposite brain hemisphere within the sensorimotor area. Regarding the DEAP dataset, channels related to brain regions more likely to participate in affective states are considered. Figure 5.4 depicts the subset of selected channels for both EEG datasets. Regarding the DMOSM-GP free parameter, the rank of decomposition is chosen from a grid search within the range $Q \in \{1, \ldots, 10\}$ to minimize the mean absolute error (MAE) of the model prediction against the original EEG data. The GPflow framework is employed for the model definition [51], and the kernel function is optimized via the minimization of NLL cost function using the autograd library. Finally, for the statistical significance assessment of the classification performance, an 10-fold cross-

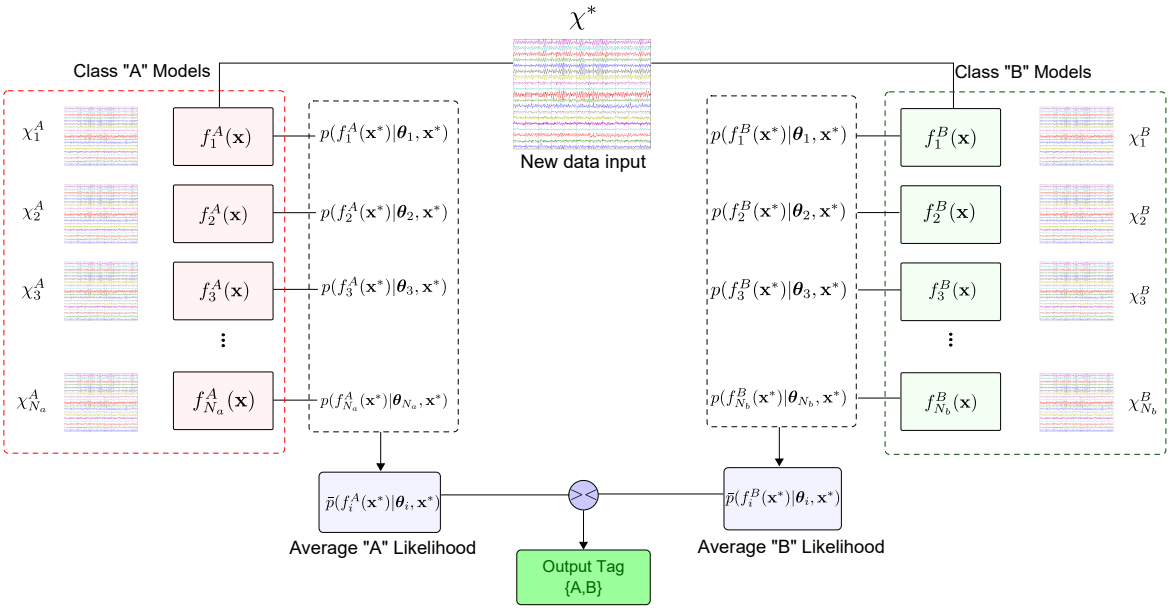


Fig. 5.3 Proposed Framework for discriminative MOSM

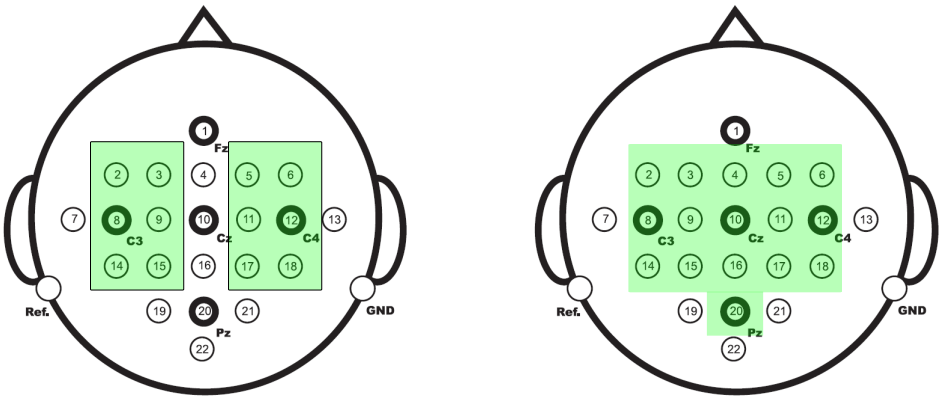
validation scheme was applied.

5.4 Results

The results regarding both approaches for spatio-temporal analysis of BEA are presented.

5.4.1 Results of QKLMS-adaptive for BEA data

Examples of the physiological estimation results obtained by the QKLMS algorithm using BEA are discussed first. Estimation of the GSR by the correlation measure for subject 12 under a particular condition with valence rating of 7 is presented in figure 5.5. Likewise, an estimation of the BVP using the TGMA measure for subject 2 with an arousal rating of 2 is presented in figure 5.6. The top row on both figures presents the filter output against the reference value at each one of the 19 windows. At this point, the mean square error is computed for the complete test set of realizations for each subject. For the valence dimension, the error obtained is around 9.34 ± 2.11 and 4.30 ± 1.82 for GSR and BVP respectively when the correlation measure is used; and approximately 5.22 ± 1.85 and 7.15 ± 2.15 for GSR and BVP respectively when TGMA is used instead. On the other hand, for the Arousal dimension, the mean error is around 10.00 ± 2.44 for GSR and 8.74 ± 2.32 for BVP whit the correlation index and finally a mean error of 6.45 ± 1.87 for GSR and 8.85 ± 2.23 for



(a) DEAP dataset (b) MI dataset

Fig. 5.4 Positioning scheme of electrodes for acquisition in both datasets. Green boxes highlight the selected channels for testing the proposed methodology.

BVP with the TGMA measure.

Likewise, at the middle and low rows of both figures, an example of the relevance analysis from the connectivity measures is presented. The displayed data corresponding to the relevant windows selected from QKLMS on a scheme of 32 electrodes from the EEG array displayed into a head model, with the lines representing the existent connectivity between each pair of channels that satisfies a defined threshold. Also, since the number of selected centroids $\mathscr{X}$ varies across the different conditions and subjects, the displayed results contains only the selected data for that particular experiment.

Classification of BEA

The comparative results from the proposed approach in comparison to other works in the state-of-art that uses the DEAP database, are presented in Table 5.1. As can be seen in Table 5.1, there is an improvement on the classification results when the proposed ALRA approach is employed. The number of selected subjects in each dimension that satisfies the balanced class experiments are 10 subjects for the valence dimension and 9 subjects for the arousal dimension. The highest results were obtained by the ALRA approach when employing the GSR signal with $76.62\% \pm 5.81$ for the arousal dimension, and by the ALRA when employing the BVP response with $76.21\% \pm 4.06$ for the valence dimension.

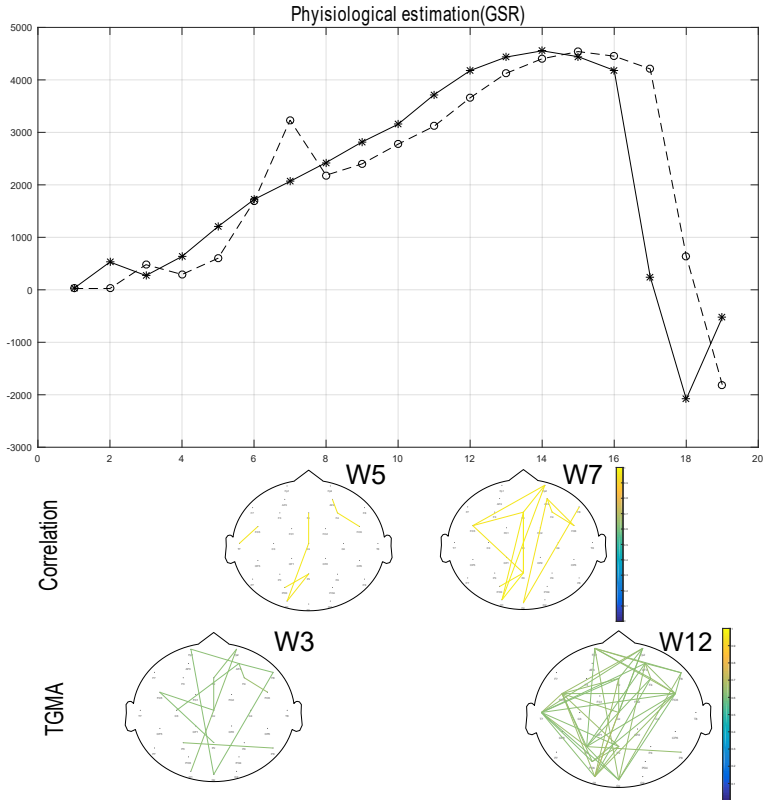


Fig. 5.5 Estimation of GSR physiological response and relevance selection of connectivity patterns. Top row, reference output (solid line) and filter output estimation (dashed line) for the 19 windows. Middle row, selected centroids at $\{5, 7\}$ windows for correlation and bottom row the selected centroids at $\{3, 12\}$ windows for TGMA.

5.4.2 Results of DMOSM-GP for BEA data

Parameter tuning and spectral modeling

To tune the rank of decomposition Q, we evaluate the MOSM-GP performance for modeling the selected output channels from the data posterior distribution at specific temporal locations. Moreover, the mean absolute error (MAE) quantifies the difference between the target and predicted outputs as a function of the rank. Figure 5.7 presents the mean MAE across the GP outputs against the number of spectral components defined for the MOSM kernel. As a first insight, the MAE values evidence that the MOSM-GP effectively reconstruct EEG recordings. Nonetheless, the error increases over six spectral components, due to a large number of kernel parameters to be tuned, which in turn increases the computational complexity without providing relevant information to the probabilistic model. On the contrary, a single component lacks the complexity to account for the brain activity changes. Therefore, a rank of decomposition between three and five benefits the most the MAE performance,

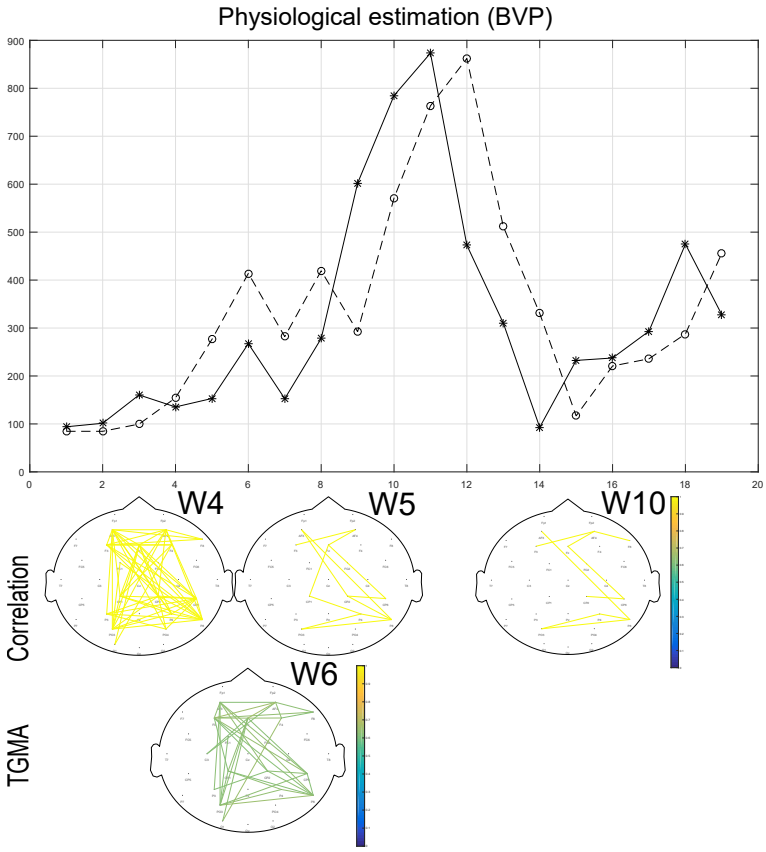


Fig. 5.6 Estimation of BVP physiological response and relevance selection of connectivity patterns. Top row, reference output (solid line) and filter output estimation (dashed line) for the 19 windows. Middle row, selected centroids at $\{4,5,10\}$ windows for correlation and bottom row the selected centroids at $\{6\}$ window for TGMA

implying a balance between model complexity and generalization capability. Consequently, for the remaining of the work, we selected three spectral mixtures as the optimal Q for testing the MOSM-GP scheme. For the purpose of visualization, Figure 5.8 exemplifies the MOSM-GP output for channels FP1 and AF3 using $Q = 3$. As seen, the posterior MOSM distribution suitably models EEG data at all time locations with bounded deviations.

An analysis of the spectral information quantified by the MOSM-GP is carried out using Equation (5.18) that decomposes the spectral content shared by two channels into Q terms. Figure 5.9 plots the component-wise spectral distribution between channels $FC2$ and $FC6$ in a trial from left (Figure 5.9(a)) and right hand movement (Figure 5.9(b)). Attained spectra prove that each component automatically fits a particular frequency band. Moreover, the component magnitude $\alpha_{ij}^{(q)}$ highlights the dominant component from each trial as the most discriminative frequency band. As expected, such frequencies lay around 15Hz to 35Hz for MI experiments, being associated with an activation of alpha and beta rhythms.

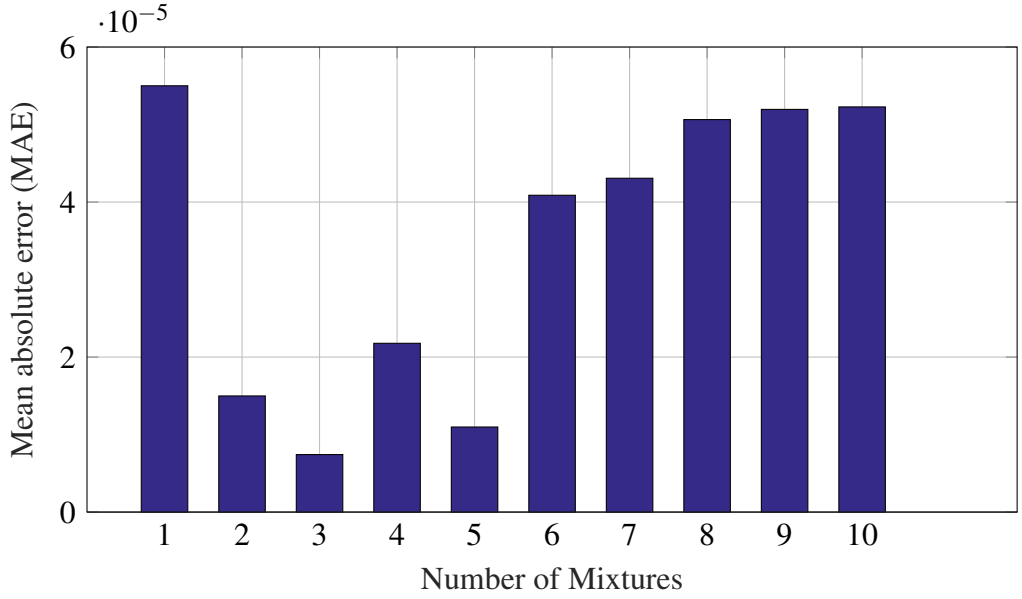
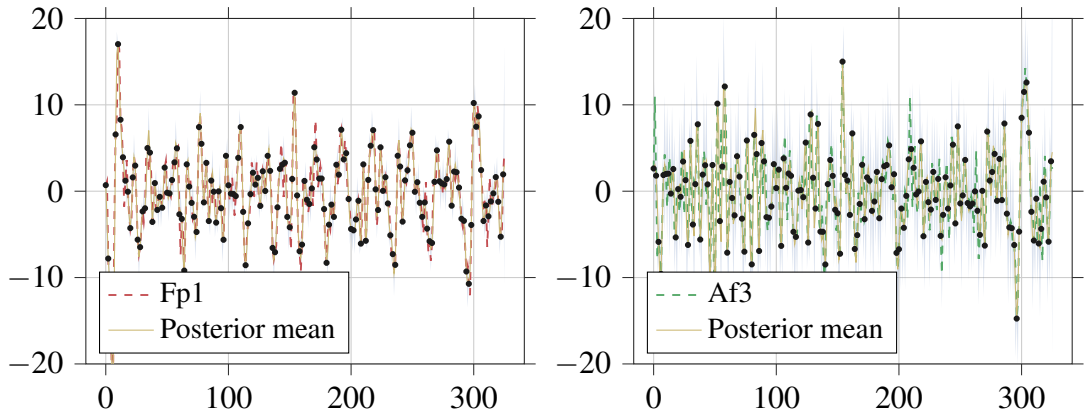


Fig. 5.7 MAE values for the data reconstruction of EEG channels after training a MOSM-GP compared to the original recordings.



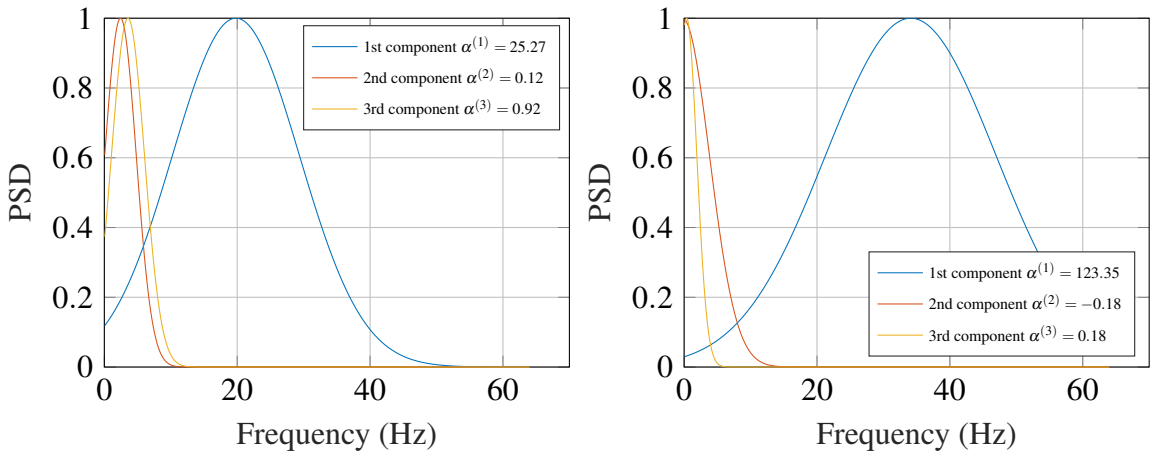
(a) Channel FP1 (b) Channel AF3

Fig. 5.8 Channels FP1 and AF3 from EEG recordings adjusted by an MOSM with three spectral mixtures and its corresponding model posterior along time.

Table 5.1 Mean emotion classification results [%] for selected DEAP subjects, only mean value for comparison consistency against reported works in [59]

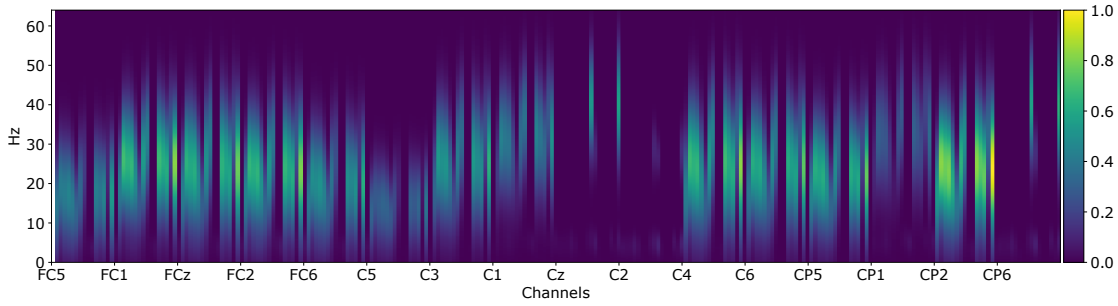
Reference	Approach	Arousal	Valence
Koelstra et.al. [37]	Linear features, power spectral - SVM	62.00	57.50
Soleymani et.al. [70]	Power spectral - SVM	50.00	62.00
Gupta et.al. [30]	Power spectral - HJORT - SVM	60.00	60.00
Castellanos et.al. [59]	MSP - ROI signal - SVM	58.6	55.76
Daimi et.al. [22]	Wavelet Packet - SVM	67.00	65.00
Torres et.al. [78]	RFCV - KNN	66.00	65.73
This work	SVM - correlation	65.31	69.63
This work	SVM - TGMA	65.31	69.99
This work	**ALRA - GSR**	**76.62**	74.38
This work	**ALRA - BVP**	75.19	**76.21**

Latter, Equation (5.20) is used to compute the cross-spectrum visualizing the computed covariances by the MOSM kernel for every channel pair. Since there is valuable information on the analysis of the cross-covariances between channel pairs, a complete trial visualization of the quantified spectral information is presented in Figures 5.10(a) and 5.10(b) for left and right hand movement trials, respectively. Each one of the horizontal axis sections, corresponds to a particular channel i and its MOSM PSD against all the channels. The vertical axis represents the frequency bin at which the connectivity is assessed. Then, lighter green colors are related with strong spectral densities shared between i, j channels, while darker blue colors are related with lower interdependencies. Despite most of the strong spectral density is constrained to the $[15, 40]$Hz band, there is clear evidence that not all the channels are synchronized at this frequency when performing MI activity. Particularly for Figure 5.10(a), there are channels sections that seems to be uncorrelated in the complete spectrum for this particular task. For example, channels Cz, $C2$, and $CP6$ seems to have low frequency dependencies shared with the rest of the EEG array. On the other hand, channel $CP2$ presents strong connections with most of the channels. Furthermore, for the opposite MI task, there exist variations on the captured frequency relationships among the EEG array. $C2$ and $C4$ result as the most highly correlated the other channels when performing the MI task.

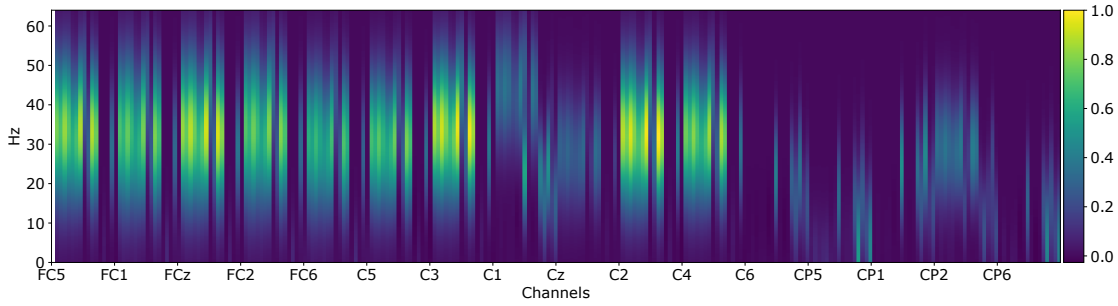


(a) Left hand MI (b) Right hand MI

Fig. 5.9 Power spectral density $S_{ij}(\omega)$ for a given pair of channels ($FC2$ - $FC6$) from MI database for two opposite conditions. Top row for left hand movement condition and low row for right hand movement condition. Figures 5.9(a) and 5.9(b) are the spectral content per spectral component.



(a) Left hand condition MI



(b) Right hand condition MI

Fig. 5.10 Quantified power spectral interdependencies for two particular conditions of the motor imagery dataset.

Discriminative MOSM-GP

For the DMOSM-GP framework, each EEG trial is trimmed into two-seconds lasting time series to find the desired spectral relationships. Table 5.2 presents the absolute value of the average likelihood as the class dependency measure, depicted in equation (5.24). The first column corresponds to the test condition level regarding the emotional content (valence for this experiment), columns two and three are the values obtained for the model testing, and fourth is the valid tag of the corresponding test signal. Finally, the fifth column is the predicted tag from the magnitude of the mean test likelihood. Similarly, in Table 5.3, the results for the DMOSM-GP strategy over the BCI database are presented. The first column corresponds to the movement associated with the experiment. Columns two and three are the average likelihoods obtained by testing the new input against each class's models. The fourth column is associated with the resulting tag. Each row in Tables 5.2 and 5.3 correspond to a particular trial, for emotional or motor imagery experiments.

From this test on complete datasets, it can be evidenced that the prediction regarding the likelihood of the trained model with the test signals is promising. The accuracy of the test data is around 73.3% for the DEAP database and about 87.33 for the BCI database. Specifically, for comparison purposes against works using the DEAP dataset, a selection of 9 subjects is performed. This selection is based on the evidence of an uncertainty test performed in [59], where the authors concluded that the subject itself did not adequately tag some experiments of the DEAP database. The scheme of implicit tagging used in this particular database, allows the subjects to rate their affective result, so on, some of the acquired signals seem not correctly related to the emotional tests. The results reported in table 5.4 shows the accuracy resulting from the cross-validated DMOSM-GP scheme. The subject IDs reported are the same in the database, and some high accuracy (state-of-art comparable) were achieved around 78%, that is the case of subject 18.

On the other hand, for the BCI database, the complete set of 9 subjects is employed, and Table 5.5 reports the accuracy of classification within 5 folds of DMOSM-GP. Some higher accuracy was obtained for the MI database in the subject-dependent experiments around 87% for subject ID 2. In general terms, the results associated with the BCI database are higher than the DEAP database, and it can be related directly with the condition that is tested. In the case of emotional experiments, there is a high degree of subjectivity in the elicitation of the states. In contrast, in the case of motor imagery, the conditions are somehow more consistently replicated.

Table 5.2 MOSM-GP Test for BEA analysis of emotional conditions of two classes. The mean absolute value of the likelihood from the test signals against the trained models is included - DEAP database

Level	Low Models	High Models	True	Predicted
9.0	138.6293	143.0221	A	A
9.0	10.8743	14.7149	A	A
9.0	103.9478	100.7107	A	B
9.0	299.0507	304.0222	A	A
8.81	299.7143	305.0394	A	A
8.56	363.5983	369.2760	A	A
9.0	348.1999	353.7941	A	A
9.0	38.8484	42.9119	A	A
1.0	193.7571	191.0369	B	B
1.0	145.3716	142.2964	B	B
1.0	400.7168	406.6640	B	A
1.0	36.0769	32.4378	B	B
1.1	48.8835	53.0558	B	A
1.0	132.5064	136.9884	B	A
1.08	83.4116	80.2945	B	B
Total Accuracy $73.35 \pm 5.64\%$				

Table 5.3 MOSM-GP Test for BEA analysis of emotional conditions of two classes. The mean absolute value of the likelihood from the test signals against the trained models is included - BCI database

Side	Left Models	Right Models	Predicted
L	163.8711	159.3530	L
L	181.6294	176.9928	L
L	29.7624	35.1477	R
L	183.1083	179.0200	L
L	128.3101	123.8268	L
L	24.0204	20.2636	L
L	142.6790	138.4443	L
L	11.2064	7.0390	L
R	59.3072	63.8494	R
R	111.6419	95.6175	L
R	191.7138	196.6392	R
R	193.1533	196.7049	R
R	35.7346	39.8927	R
R	194.2656	198.3379	R
R	54.5305	59.0000	R
Total Accuracy 87.33 ± 3.68			

Table 5.4 Discriminative test for subjects from the DEAP database using the DMOSM-GP scheme, the test are developed within the valence emotional dimension

Subject ID	Accuracy	Subject ID	Accuracy
17	0.65 ± 4.32	2	0.70 ± 6.19
18	0.78 ± 8.12	16	0.50 ± 5.28
22	0.60 ± 3.49	19	0.65 ± 5.66
29	0.65 ± 4.28	31	0.50 ± 6.77
30	0.70 ± 5.28	**Total**	0.64 ± 0.09

Table 5.5 Discriminative test for subjects from the Motor Imagery database using the DMOSM-GP scheme

Subject ID	Accuracy	Subject ID	Accuracy
1	0.82 ± 3.45	6	0.85 ± 3.18
2	0.87 ± 4.16	7	0.80 ± 3.73
3	0.85 ± 2.56	8	0.82 ± 4.21
4	0.77 ± 4.56	9	0.79 ± 2.66
5	0.82 ± 4.03	**Total**	0.88 ± 0.14

Results comparison

The results obtained from the DMOSM-GP strategy are compared with some state of art results in terms of classification accuracy. However, since the proposed methodology uses generative models for discriminative purposes, it has to be stated that the comparison should be made on a few additional items than the classification accuracy. As can be seen in Tables 5.6 and 5.7, the average classification results obtained by the DMOSM-GP strategy are competitive among the state-of-art works. It has to be noticed that this work does not implement a preprocessing or feature extraction stage, but uses the acquired data to train the MOSM-GP's and perform the discriminative task.

5.5 Discussion and Concluding remarks

To the best of our knowledge, the estimation of physiological responses by emotional stimulus from BEA data was not developed before, although the emotion recognition was performed by using descriptors of this data. The use of adaptive filtering learning seems adequate for BEA data analysis since this data is particular for each subject and the methodologies of processing must be adapted to these particularities. The QKLMS algorithm allows the estimation of the physiological value from the connectivity measures. The amounts of

Table 5.6 Mean emotion classification results [%] for selected DEAP subjects, only mean value for comparison consistency against reported works in [59]

Reference	**Approach**	**Valence**
Koelstra et al. [37]	PSD - SVM	57.50
Soleymani et al.[70]	PSD - SVM	62.00
Gupta et al.[30]	PSD - HJORT - SVM	60.00
Padilla-Buritica et al.[59]	MSP - SVM	55.76
Daimi and Saha[22]	Wavelet - SVM	65.00
Torres-Valencia et al.[78]	RFCV - KNN	65.73
Pan et.al. [60]	PSD - LORSAL	63.29
Pan et.al. [60]	DE - LORSAL	77.17
This work	DMOSM-GP	64.35

Table 5.7 Mean emotion classification results [%] for MI dataset, only mean value for comparison consistency against reported works

Reference	**Approach**	**Valence**
Qureshi et al[63]	ICA - ELM	94.29
Li et al.[43]	CSP - SVM	78.78
Liang et al.[45]	PDC - MEMD	70.22
Elastuy and Eldawlatly[25]	DBN	73.44
Gómez et al.[28]	CSP - SVM	81.41
This work	DMOSM-GP	88.71

the estimation error enable to conclude that this framework can be used with a high degree of confidence as evidenced in the results. Also, an additional essential outcome is the implicit selection of some centroids via the novelty criterion, which allows extracting the connectivity data holding relevant information for classification purposes on the absence of labeling in most of the BEA. An improvement in the classification accuracy is observed when using the relevant selected data in comparison to the case when the whole connectivity data is used. Also, higher values of classification accuracy were obtained in comparison with the state-of-art works as detailed in Table 5.1. As future work, a scheme of automatic tagging of BEA could be derived from this strategy as a remarkable outcome of the proposed methodology. Finally, the adjustment of the parameters in each stage could be improved since the heuristic search may result in non-optimal parameter selection.

Brain information processing is a complex task that is not yet entirely mapped and understood. Despite the previous knowledge about brain regions interactions in motor or emotional means, works that allow improving the interpretability of results in different BEA scenarios would lead to the development of more precise frameworks for analysis, diagnosis, and treatment of mental pathologies, among other tasks. In this work, we proposed a framework for discriminate BEA using raw data with the support of a spectral kernel that identifies relationships between channels on behalf of a probabilistic methodology of multi-output Gaussian process. One of the essential remarks of this framework is the capability of learning EEG connectivity patterns by estimating raw data spectral components without a feature extraction stage. Further, this proposal of generative models working directly with EEG data has the advantage of adjusting the model relying on the optimization of kernel hyperparameters just from the channels information in a data-driven framework.

The results presented in section 5.4 evidenced that introducing the MOSM kernel to MOGPs becomes a reliable tool for BEA modeling, due to the spectral designing properties. It is well known that EEG channels share complex frequency relationships that can be exploited using the design of a covariance function in that particular domain. Regarding this property, the posterior distribution over the data measured by the MAE in Figure 5.7 shows an adequate adjustment of the model on the original data. It also allows us to conclude that the inclusion of more spectral components into the covariance function benefits the model adjusting to the data.

Moreover, the identification of the spectral relationships between channels, performed by the MOSM kernel, allows gaining a better understanding of the latent functional connectivity between brain regions. As Figure 5.10 evidenced, the MOSM-GP identifies rep-

resentative frequency bands for the cognitive process. The positive linear relationships are grouped among channels from the same hemisphere, with strong specific dependencies between channels like $P3-P7$, and $Fc6-Af4$. The negative relationships are explained by lower PSD amplitudes at Cz, $C2$ and $CP6$ for the left, and $C6, CP1, CP6$ for the right-hand MI task. All these interactions quantified by the MOSM kernel in terms of higher or lower values of the PSD can be directly related with the activations of neural cells in different regions of the brain related with emotions (amygdala, hippocampus, and frontal cortex) or related with motor activities (primary motor cortex and posterior parietal cortex). Nevertheless, further studies must be completed from the evidence of these relationships to an accurate source reconstruction before determining the specific brain region of the acquired neural activity.

Finally, the discriminative results regarding the emotional and motor imagery conditions conclude that probabilistic models can be efficiently employed as a classification tool for EEG data. In this case, the probability distribution of tested data against the trained models directly becomes a classification algorithm by following a direct comparison of the mean likelihood value between the models from two classes. Despite lacking a feature extraction stage, the proposed DMOSM-GP produces discriminative information from the data. Moreover, the total accuracy of the subject-dependent and condition-dependent tests is comparable with state-of-art works as Tables 5.6 and 5.7 illustrate.

One of the drawbacks of this framework is the computational complexity of training a considerable number of MOSM-GP models. Also, increasing the number of MOSM kernel mixtures and the size of BEA data will derive into an exponential growth of the training time. Further improvements of this methodology will be directed towards using lighter versions of probabilistic models such as the sparse GPs aiming at solving more difficult supervised learning tasks from EEG data as multilabel classification or regression.

Chapter 6

Conclusions and Future Work

This thesis considered different approaches for brain electrical activity analysis. The objective of improving interpretability and generalization capability of the results was achieved by the results that allows to explain some mental conditions on the evidence of processed EEG data. Moreover, the proposed frameworks works independently and could be used within several scenarios regarding the origin and experiments of the data. This chapter summarizes the contributions and research work done in the thesis, besides some future research lines are presented.

6.1 Conclusions

- Chapter 1 provided the fundamental description of the mathematical basis of RKHS and the data context that allows the interpretation of the results.
- Chapter 3 introduced the solution of the high dimensionality problem of EEG. We stated that selecting relevant features within a channel-independent feature extraction scheme, improved considerably the generalization capacity, while reducing the model complexity. We implemented two schemes for EEG data feature elimination and test it in emotional experiments. Several features were previously extracted using short time representations and successfully reduced the number of features while maintaining the classification accuracy of an Support vector machine.
- Chapter 4 presented a spatial dependencies analysis for EEG data. The advantage of using connectivity measures is to improve the low spatial resolution of the data. We tested some connectivity measures and proposed an automatic selection scheme of relevant data for classification purposes. The kernel centered alignment was extended

for using it within the spatial domain of connectivity data. The results show that the selected measure performs well in terms of classification accuracy for detecting emotional states.

- Finally, in chapter 5 includes the more important proposals within this thesis. The use of adaptive filters for selecting relevant temporal marks of connectivity data allows to identify more deeply the interactions between brain regions for specific conditions. We extended the QKLMS algorithm for spatial dependencies analysis for EEG data including and additional physiological response. This methodology allows to reduce the computational complexity of the classifier while giving important evidence of spatio-temporal data that improves our brain understanding. Additionally, we proposed a discriminative scheme while using a probabilistic model such as Gaussian Processes for computing frquency dependencies between channels. The use of a spectral-designed kernel allows to analyze the interdependencies between channels into a transformed domain representation.

6.2 Future Work

Some interesting paths for future work involve approaches to solving current limitations of the presented methods and further extensions of the developed methodologies. From the results obtained, some improvement could be focused in the extension of the methodologies to multiclass scenarios. Also, additional information theory metrics could be involved into the already built frameworks that could improve the selection of relevant data. Regarding the proposed framework that uses multiple-output GP's, a sparse approach could substitute the full covariance matrix that was implemented. This change allows to reduce the computational cost of the training stage that is quite expensive for this framework.

Other future work could be related with the information contained in Appendix C, were a exploration of the Cauchy-Schwartz quadratic mutual information (CSQMI) could be employed to measure differences between connectivity data in subsequent temporal windows. These detected transitions might be indicators of BEA in different brain regions related with a particular condition. Then, a quantification of these differences could be evidenced via a p-value test.

References

[1] Berdakh Abibullaev. Learning suite of kernel feature spaces enhances smr-based eeg-bci classification. In *Brain-Computer Interface (BCI), 2017 5th International Winter Conference on*, pages 55–59. IEEE, 2017.

[2] Berdakh Abibullaev. Learning suite of kernel feature spaces enhances smr-based eeg-bci classification. In *Brain-Computer Interface (BCI), 2017 5th International Winter Conference on*, pages 55–59. IEEE, 2017.

[3] Yaman Aksu, David J. Miller, George Kesidis, and Qing X. Yang. Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions. *Trans. Neur. Netw.*, 21(5):701–717, May 2010. ISSN 1045-9227. doi: 10.1109/TNN.2010.2041069. URL `http://dx.doi.org/10.1109/TNN.2010.2041069`.

[4] S. M. Alarcao and M. J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017. ISSN 1949-3045. doi: 10.1109/TAFFC.2017.2714671.

[5] JF Alonso, S Romero, MR Ballester, RM Antonijoan, and MA Mañanas. Stress assessment based on eeg univariate features and functional connectivity measures. *Physiological measurement*, 36(7):1351, 2015.

[6] Turky Alotaiby, Fathi E. Abd El-Samie, Saleh A. Alshebeili, and Ishtiaq Ahmad. A review of channel selection algorithms for eeg signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015(1):66, Aug 2015. ISSN 1687-6180. doi: 10.1186/s13634-015-0251-9. URL `https://doi.org/10.1186/s13634-015-0251-9`.

[7] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3): 195–266, 2012.

[8] John Atkinson and Daniel Campos. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers. *Expert Systems with Applications*, 47:35 – 41, 2016. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2015.10.049. URL `http://www.sciencedirect.com/science/article/pii/S0957417415007538`.

[9] Andrea Avena-Koenigsberger, Bratislav Misic, and Olaf Sporns. Communication dynamics in complex brain networks. *Nature Reviews Neuroscience*, 19(1):17, 2018.

[10] Dipali Bansal and Rashima Mahajan. Chapter 2 - eeg-based brain-computer interfacing (bci). In Dipali Bansal and Rashima Mahajan, editors, *EEG-Based Brain-Computer Interfaces*, pages 21 – 71. Academic Press, 2019. ISBN 978-0-12-814687-3. doi: https://doi.org/10.1016/B978-0-12-814687-3.00002-8. URL `http://www.sciencedirect.com/science/article/pii/B9780128146873000028`.

[11] Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering*, 4(2):R32, 2007.

[12] André M Bastos and Jan-Mathijs Schoffelen. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9:175, 2016.

[13] Aayush Bhardwaj, Ankit Gupta, Pallav Jain, Asha Rani, and Jyoti Yadav. Classification of human emotions from eeg signals using svm and lda classifiers. In *Signal Processing and Integrated Networks (SPIN), 2015 2nd International Conference on*, pages 180–185. IEEE, 2015.

[14] Austin J Brockmeier and José C Príncipe. Learning recurrent waveforms within eegs. *IEEE Transactions on Biomedical Engineering*, 63(1):43–54, 2016.

[15] C Brunner, R Leeb, G Müller-Putz, A Schlögl, and G Pfurtscheller. Bci competition 2008–graz data set a. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 16, 2008.

[16] Jiahui Cai, Wei Chen, and Zhong Yin. Multiple transferable recursive feature elimination technique for emotion recognition based on eeg signals. *Symmetry*, 11(5):683, 2019.

[17] Catie Chang and Gary H Glover. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

[18] Sharon Chiang, Alberto Cassese, Michele Guindani, Marina Vannucci, Hsiang J Yeh, Zulfi Haneef, and John M Stern. Time-dependence of graph theory metrics in functional connectivity analysis. *NeuroImage*, 125:601–615, 2016.

[19] Sharon Chiang, Alberto Cassese, Michele Guindani, Marina Vannucci, Hsiang J Yeh, Zulfi Haneef, and John M Stern. Time-dependence of graph theory metrics in functional connectivity analysis. *NeuroImage*, 125:601–615, 2016.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20 (3):273–297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL `http://dx.doi.org/10.1023/A:1022627411411`.

[21] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

[22] Syed Naser Daimi and Goutam Saha. Classification of emotions induced by music videos and correlation with participants' rating. *Expert Systems with Applications*, 41(13):6057 – 6065, 2014. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2014.03.050. URL http://www.sciencedirect.com/science/article/pii/S0957417414001882.

[23] R. P. W. Duin. Prtools version 3.0: A matlab toolbox for pattern recognition. In *Proc. of SPIE*, page 1331, 2000.

[24] J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *EPL (Europhysics Letters)*, 4(9):973, 1987.

[25] Basem Elasuty and Seif Eldawlatly. Dynamic bayesian networks for eeg motor imagery feature extraction. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 170–173. IEEE, 2015.

[26] B. Fadlallah, S. Seth, A. Keil, and J. Principe. Quantifying cognitive state from eeg using dependence measures. *IEEE Transactions on Biomedical Engineering*, 59(10): 2773–2781, Oct 2012. ISSN 0018-9294. doi: 10.1109/TBME.2012.2210283.

[27] Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1 (1):13–36, 2011.

[28] V. Gómez, A. Álvarez, P. Herrera, G. Castellanos, and A. Orozco. Short time eeg connectivity features to support interpretability of mi discrimination. In Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 699–706, Cham, 2019. Springer International Publishing. ISBN 978-3-030-13469-3.

[29] A. Greco, G. Valenza, A. Lanata, G. Rota, and E.P. Scilingo. Electrodermal activity in bipolar patients during affective elicitation. *Biomedical and Health Informatics, IEEE Journal of*, 18(6):1865–1873, Nov 2014. ISSN 2168-2194. doi: 10.1109/JBHI.2014.2300940.

[30] Rishabh Gupta, Khalil ur Rehman Laghari, and Tiago H. Falk. Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization. *Neurocomput.*, 174(PB):875–884, January 2016. ISSN 0925-2312. doi: 10.1016/j.neucom.2015.09.085. URL http://dx.doi.org/10.1016/j.neucom.2015.09.085.

[31] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3): 389–422, March 2002. ISSN 0885-6125. doi: 10.1023/A:1012487302797. URL http://dx.doi.org/10.1023/A:1012487302797.

[32] Bilal H Fadlallah, Austin Brockmeier, Sohan Seth, Lin Li, Andreas Keil, and Jose Principe. An association framework to analyze dependence structure in time series. 2012:6176–9, 08 2012.

[33] Behzad Hejrati, Abdolhossein Fathi, and Fardin Abdali-Mohammadi. Efficient lossless multi-channel eeg compression based on channel clustering. *Biomedical Signal Processing and Control*, 31:295–300, 2017.

[34] A.R. Hidalgo-Muñoz, M.M. López, I.M. Santos, A.T. Pereira, M. Vázquez-Marrufo, A. Galvao-Carmona, and A.M. Tomé. Application of svm-rfe on eeg signals for detecting the most relevant scalp regions linked to affective valence processing. *Expert Systems with Applications*, 40(6):2102 – 2108, 2013. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2012.10.013. URL http://www.sciencedirect.com/science/article/pii/S095741741201127X.

[35] Barry Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19(2):466 – 470, 2003. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/S1053-8119(03)00112-5. URL //www.sciencedirect.com/science/article/pii/S1053811903001125.

[36] Ian Jolliffe. *Principal Component Analysis*. John Wiley and Sons, Ltd, 2005. ISBN 9780470013199. doi: 10.1002/0470013192.bsa501. URL http://dx.doi.org/10.1002/0470013192.bsa501.

[37] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. ISSN 1949-3045. doi: http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2011.15.

[38] Sander Koelstra and Ioannis Patras. Fusion of facial expressions and eeg for implicit affective tagging. *Image Vision Comput.*, 31(2):164–174, 2013.

[39] Jukka Kortelainen and Tapio Seppanen. Eeg-based recognition of video-induced emotions: Selecting subject-independent feature set. *Conf Proc IEEE Eng Med Biol Soc*, 2013:4287–4290, 2013. ISSN 1557-170X. URL http://www.biomedsearch.com/nih/EEG-based-recognition-video-induced/24110680.html.

[40] Sylvia D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3):394 – 421, 2010. ISSN 0301-0511. doi: http://dx.doi.org/10.1016/j.biopsycho.2010.03.010. URL http://www.sciencedirect.com/science/article/pii/S0301051110000827. The biopsychology of emotion: Current theoretical and empirical perspectives.

[41] M Rajya Lakshmi, Dr TV Prasad, and Dr V Chandra Prakash. Survey on eeg signal processing methods. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(1), 2014.

[42] You-Yun Lee and Shulan Hsieh. Classifying different emotional states by means of eeg-based functional connectivity patterns. *PloS one*, 9(4):e95415, 2014.

[43] Duan Li, Hongxin Zhang, Muhammad Saad Khan, and Fang Mi. A self-adaptive frequency selection common spatial pattern and least squares twin support vector machine for motor imagery electroencephalography recognition. *Biomedical Signal Processing and Control*, 41:222–232, 2018.

[44] H. Li, C. Qing, X. Xu, and T. Zhang. A novel de-pccm feature for eeg-based emotion recognition. In *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 389–393, 2017.

[45] Shuang Liang, Kup-Sze Choi, Jing Qin, Qiong Wang, Wai-Man Pang, and Pheng-Ann Heng. Discrimination of motor imagery tasks via information flow pattern of brain connectivity. *Technology and Health Care*, 24(s2):S795–S801, 2016.

[46] Weifeng Liu, Jose C. Principe, and Simon Haykin. *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley Publishing, 1st edition, 2010. ISBN 0470447532, 9780470447536.

[47] Félix García López, Miguel García Torres, Belén Melián, Batista José, A. Moreno Pérez, and J. Marcos Moreno-vega. Solving feature subset selection problem by a parallel scatter search. *European Journal of Operational Research*, 2005, 2006.

[48] Akash Kumar Maity, Ruchira Pratihar, Anubrato Mitra, Subham Dey, Vishal Agrawal, Shankha Sanyal, Archi Banerjee, Ranjan Sengupta, and Dipak Ghosh. Multifractal detrended fluctuation analysis of alpha and theta eeg rhythms with musical stimuli. *Chaos, Solitons & Fractals*, 81:52–67, 2015.

[49] Norbert Marwan. Cross recurrence plot toolbox for matlab, 2011.

[50] Norbert Marwan, M. Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237 – 329, 2007. ISSN 0370-1573. doi: 10.1016/j.physrep.2006.11.001.

[51] Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo Le'on-Villagr'a, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, apr 2017.

[52] Shayan Motamedi-Fakhr, Mohamed Moshrefi-Torbati, Martyn Hill, Catherine M Hill, and Paul R White. Signal processing techniques applied to human sleep eeg signals—a review. *Biomedical Signal Processing and Control*, 10:21–33, 2014.

[53] Wajid Mumtaz, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, and Aamir Saeed Malik. A machine learning framework involving eeg-based functional connectivity to diagnose major depressive disorder (mdd). *Medical & biological engineering & computing*, pages 1–14, 2017.

[54] A Myrden and T Chau. Towards psychologically adaptive brain–computer interfaces. *Journal of Neural Engineering*, 13(6):066022, 2016. URL `http://stacks.iop.org/1741-2552/13/i=6/a=066022`.

[55] Emanuel Neto, Felix Biessmann, Harald Aurlien, Helge Nordby, and Tom Eichele. Regularized linear discriminant analysis of eeg features in dementia patients. *Frontiers in aging neuroscience*, 8:273, 2016.

[56] Luis F. Nicolas-Alonso, Rebeca Corralejo, Javier Gomez-Pilar, Daniel Álvarez, and Roberto Hornero. Adaptive semi-supervised classification to reduce intersession non-stationarity in multiclass motor imagery-based brain–computer interfaces. *Neurocomputing*, 159:186 – 196, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.02.005. URL `http://www.sciencedirect.com/science/article/pii/S0925231215001496`.

[57] Guiomar Niso, Ricardo Bruña, Ernesto Pereda, Ricardo Gutiérrez, Ricardo Bajo, Fernando Maestú, and Francisco del Pozo. Hermes: Towards an integrated toolbox to characterize functional and effective brain connectivity. *Neuroinformatics*, 11(4): 405–434, 2013. ISSN 1559-0089. doi: 10.1007/s12021-013-9186-1. URL `http://dx.doi.org/10.1007/s12021-013-9186-1`.

[58] Seung-Hyeon Oh, Yu-Ri Lee, and Hyoung-Nam Kim. A novel eeg feature extraction method using hjorth parameter. *International Journal of Electronics and Electrical Engineering*, 2(2):106–110, 2014.

[59] Jorge I. Padilla-Buritica, Juan D. Martinez-Vargas, and German Castellanos-Dominguez. Emotion discrimination using spatially compact regions of interest extracted from imaging eeg activity. *Frontiers in Computational Neuroscience*, 10:55, 2016. ISSN 1662-5188. doi: 10.3389/fncom.2016.00055. URL `http://journal.frontiersin.org/article/10.3389/fncom.2016.00055`.

[60] Chao Pan, Cheng Shi, Honglang Mu, Jie Li, and Xinbo Gao. Eeg-based emotion recognition using logistic regression with gaussian kernel and laplacian prior and investigation of critical frequency bands. *Applied Sciences*, 10(5):1619, Feb 2020. ISSN 2076-3417. doi: 10.3390/app10051619. URL `http://dx.doi.org/10.3390/app10051619`.

[61] Gabriel Parra and Felipe Tobar. Spectral mixture kernels for multi-output gaussian processes. In *Advances in Neural Information Processing Systems*, pages 6681–6690, 2017.

[62] Joaquín Pizarro, Elisa Guerrero, and Pedro L. Galindo. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48(174):155 – 173, 2002. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/S0925-2312(01)00653-1. URL `http://www.sciencedirect.com/science/article/pii/S0925231201006531`.

[63] M. N. I. Qureshi, D. Cho, and B. Lee. Eeg classification for motor imagery bci using phase-only features extracted by independent component analysis. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2097–2100, July 2017. doi: 10.1109/EMBC.2017.8037267.

[64] Oluwarotimi Williams Samuel, Yanjuan Geng, Xiangxin Li, and Guanglin Li. Towards efficient decoding of multiple classes of motor imagery limb movements based on eeg spectral and time domain descriptors. *Journal of medical systems*, 41(12):194, 2017.

[65] Saeid Sanei. *Adaptive processing of brain signals*. John Wiley & Sons, 2013.

[66] Saeid Sanei. *Adaptive processing of brain signals*. John Wiley & Sons, 2013.

[67] C. S. Silva, M. K. Hazrati, A. Keil, and J. C. Principe. Quantification of neural functional connectivity during an active avoidance task. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 708–711, Aug 2016. doi: 10.1109/EMBC.2016.7590800.

[68] Catia S Silva, Jose C Principe, and Andreas Keil. A novel methodology to quantify dense eeg in cognitive tasks. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2442–2446. IEEE, 2017.

[69] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3 (1):42–55, 2012. ISSN 1949-3045. doi: http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2011.25.

[70] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, April 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.37.

[71] Mohammad Soleymani, Guillaume Chanel, Joep J. M. Kierkels, and Thierry Pun. Affective characterization of movie scenes based on content analysis and physiological changes. *Int. J. Semantic Computing*, 2:235–254, 2008.

[72] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[73] Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual review of psychology*, 67:613–640, 2016.

[74] Vairavan Srinivasan, Chikkannan Eswaran, and Natarajan Sriraam. Approximate entropy-based epileptic eeg detection using artificial neural networks. *IEEE Transactions on information Technology in Biomedicine*, 11(3):288–295, 2007.

[75] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of neuroscience methods*, 274:141–145, 2016.

[76] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[77] C Torres-Valencia, J Hernandez-Muriel, W Gonzalez-Vanegas, A Alvarez-Meza, A Orozco, and M Alvarez. Non-parametric source reconstruction via kernel temporal enhancement for eeg data. In *Iberoamerican Congress on Pattern Recognition*, pages 443–450. Springer, 2016.

[78] C. Torres-Valencia, A. Alvarez-Meza, and A. Orozco-Gutierrez. *Emotion Assessment Based on Functional Connectivity Variability and Relevance Analysis*, pages 353–362. Springer International Publishing, Cham, 2017.

[79] C Torres-Valencia, A Alvarez-Meza, and A Orozco-Gutierrez. Emotion assessment based on functional connectivity variability and relevance analysis. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 353–362. Springer, 2017.

[80] C. Torres-Valencia, A. Alvarez-Meza, and A. Orozco-Gutierrez. Emotion assessment using adaptive learning-based relevance analysis. In Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny, editors, *Image Analysis and Recognition*, pages 193–200, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93000-8.

[81] C. A. Torres-Valencia, M. C. J. Santamaria, and M. A. Alvarez. Kernel temporal enhancement approach for loreta source reconstruction using eeg data. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4527–4530, 2016.

[82] Cristian Torres-Valencia, Mauricio Álvarez-López, and Álvaro Orozco-Gutiérrez. Svm-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, pages 1–15, 2016. ISSN 1783-8738. doi: 10.1007/s12193-016-0222-y. URL http://dx.doi.org/10.1007/s12193-016-0222-y.

[83] Cristian Torres-Valencia, Álvaro Orozco, David Cárdenas-Peña, Andrés Álvarez Meza, and Mauricio Álvarez. A discriminative multi-output gaussian processes scheme for brain electrical activity analysis. *Applied Sciences*, 10(19):6765, Sep 2020. ISSN 2076-3417. doi: 10.3390/app10196765. URL http://dx.doi.org/10.3390/app10196765.

[84] Jose Antonio Urigüen and Begoña Garcia-Zapirain. Eeg artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, 12(3):031001, 2015.

[85] G. Valenza, A. Lanata, and E.P. Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *Affective Computing, IEEE Transactions on*, 3(2): 237–249, 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.30.

[86] G. Valenza, L. Citi, C. Gentili, A. Lanata, E.P. Scilingo, and R. Barbieri. Characterization of depressive states in bipolar patients using wearable textile technology and instantaneous heart rate variability assessment. *Biomedical and Health Informatics, IEEE Journal of*, 19(1):263–274, Jan 2015. ISSN 2168-2194. doi: 10.1109/JBHI.2014.2307584.

[87] Johannes Vorwerk, Jae-Hyun Cho, Stefan Rampp, Hajo Hamer, Thomas R Knösche, and Carsten H Wolters. A guideline for head volume conductor modeling in eeg and meg. *NeuroImage*, 100:590–607, 2014.

[88] Johannes Vorwerk, Jae-Hyun Cho, Stefan Rampp, Hajo Hamer, Thomas R. Knösche, and Carsten H. Wolters. A guideline for head volume conductor modeling in eeg and meg. *NeuroImage*, 100:590 – 607, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.06.040. URL http://www.sciencedirect.com/science/article/pii/S1053811914005175.

[89] Ruofan Wang, Jiang Wang, Haitao Yu, Xile Wei, Chen Yang, and Bin Deng. Power spectral density and coherence analysis of alzheimer's eeg. *Cognitive neurodynamics*, 9(3):291–304, 2015.

[90] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[91] J. Weston and I Guyon. Feature selection for kernel machines using stationary weight approximation. in preparation., 2000.

[92] Pablo Alvarado y Cristian Torres Valencia y Álvaro Orozco Gutiérrez y Mauricio Álvarez López y Genaro Daza Santacoloma y Hans Carmona Vilada. Modelado y comportamiento de la simulación de propagación eléctrica durante la estimulación cerebral profunda. *DYNA*, 83(198):49–58, 2016. ISSN 2346-2183. doi: 10.15446/dyna.v83n198.51766. URL `https://revistas.unal.edu.co/index.php/dyna/article/view/51766`.

[93] Ke Yan and David Zhang. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, 212: 353 – 363, 2015. ISSN 0925-4005. doi: http://dx.doi.org/10.1016/j.snb.2015.02.025. URL `http://www.sciencedirect.com/science/article/pii/S0925400515001872`.

[94] Meichen Yu, Alida A Gouw, Arjan Hillebrand, Betty M Tijms, Cornelis Jan Stam, Elisabeth CW van Straaten, and Yolande AL Pijnenburg. Different functional connectivity and network topology in behavioral variant of frontotemporal dementia and alzheimer's disease: an eeg study. *Neurobiology of aging*, 42:150–162, 2016.

[95] R Yuvaraj, M Murugappan, U Rajendra Acharya, Hojjat Adeli, Norlinah Mohamed Ibrahim, and Edgar Mesquita. Brain functional connectivity patterns for emotional state classification in parkinson's disease patients without dementia. *Behavioural brain research*, 298:248–260, 2016.

[96] Joseph P. Zbilut and Charles L. Webber. Recurrence quantification analysis. In *Wiley Encyclopedia of Biomedical Engineering*. John Wiley and Sons, Inc., 2006. ISBN 9780471740360.

[97] Aihua Zhang, Bin Yang, and Ling Huang. Feature extraction of eeg signals using power spectral entropy. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, volume 2, pages 435–439. IEEE, 2008.

[98] Tao Zhang, Tiejun Liu, Fali Li, Mengchen Li, Dongbo Liu, Rui Zhang, Hui He, Peiyang Li, Jinnan Gong, Cheng Luo, et al. Structural and functional correlates of motor imagery bci performance: Insights from the patterns of fronto-parietal attention network. *Neuroimage*, 134:475–485, 2016.

[99] Yong Zhang, Xiaomin Ji, and Suhua Zhang. An approach to eeg-based emotion recognition using combined feature extraction method. *Neuroscience Letters*, 633: 152 – 157, 2016. ISSN 0304-3940. doi: https://doi.org/10.1016/j.neulet.2016.09.037. URL `http://www.sciencedirect.com/science/article/pii/S0304394016307200`.

[100] Yong Zhang, Bo Liu, Xiaomin Ji, and Dan Huang. Classification of eeg signals based on autoregressive model and wavelet packet decomposition. *Neural Processing Letters*, 45(2):365–378, 2017.

[101] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[102] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[103] Dongli Zhou, Wesley K Thompson, and Greg Siegle. Matlab toolbox for functional connectivity. *Neuroimage*, 47(4):1590–1607, 2009.

[104] Rob Zink, Borbála Hunyadi, Sabine Van Huffel, and Maarten De Vos. Alpha and low-beta oscillatory patterns extracted with canonical polyadicdecomposition relate to lda classifier performance in real-life mobile. In *Proc. 8th International Workshop on Biosignal Interpretation*, pages 1–4, 2016.

Appendix A

Publications

A.1 Published Papers

- Cristian Torres-Valencia, Mauricio Álvarez-López, and Álvaro Orozco-Gutiérrez. Svm-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, pages 1–15, 2016. ISSN 1783-8738. doi: 10.1007/s12193-016-0222-y. URL http://dx.doi.org/10.1007/s12193-016-0222-y

- C Torres-Valencia, J Hernandez-Muriel, W Gonzalez-Vanegas, A Alvarez-Meza, A Orozco, and M Alvarez. Non-parametric source reconstruction via kernel temporal enhancement for eeg data. In *Iberoamerican Congress on Pattern Recognition*, pages 443–450. Springer, 2016

- C. A. Torres-Valencia, M. C. J. Santamaria, and M. A. Alvarez. Kernel temporal enhancement approach for loreta source reconstruction using eeg data. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4527–4530, 2016

- C Torres-Valencia, A Alvarez-Meza, and A Orozco-Gutierrez. Emotion assessment based on functional connectivity variability and relevance analysis. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 353–362. Springer, 2017

- C. Torres-Valencia, A. Alvarez-Meza, and A. Orozco-Gutierrez. Emotion assessment using adaptive learning-based relevance analysis. In Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny, editors, *Image Analysis and Recognition*, pages 193–200, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93000-8

- Cristian Torres-Valencia, Álvaro Orozco, David Cárdenas-Peña, Andrés Álvarez Meza, and Mauricio Álvarez. A discriminative multi-output gaussian processes scheme for brain electrical activity analysis. *Applied Sciences*, 10(19):6765, Sep 2020. ISSN 2076-3417. doi: 10.3390/app10196765. URL `http://dx.doi.org/10.3390/app10196765`

Appendix B

Databases

Two databases of emotion elicitation experiments with EEG acquisition, and one motor imagery database are the sources of the data to testing the different methodologies of this thesis.

B.1 DEAP database

A dataset for emotion analysis using EEG, physiological and video signals (DEAP) is a publicly available datase of emotion elicitation experiments. Published in 2012, was introduced to the scientific community as a database for multimodal dataset from different biological sources such as EEG, galvanic skin response, respiratory pattern, temperature, electrooculography and electromiography. The $20-10$ EEG configuration is defined and 32 channels acquired with one minute recordings. The signals are sampled at 128 Hz and stored in a file as a three dimensional array such as channel x stimulus x sample. The acquisition experiments follows the scheme of emotion elicitation, where 40 videos are displayed to each on fo the 32 subjects, and after watching it, they give its rating for the video using a scale in different emotion dimensions such as, valence, arousal, dominance and liking. These scales allows to determine the induced emotion with a wider range of possible emotional outcomes [37].

B.2 MAHNOB-HCI

This dataset has a similar scheme than the DEAP, with emotion elicitation experiments applied to 30 participants. The main idea of this dataset was to provide a multimodal (includ-

ing EEG) database for affective computing with the tagging being made for the participants while watching short duration videos. 32 EEG channels among other biosignals and facial video were acquired during the execution of the experiments. Again, the ratings of the experienced emotion were given into scales of different dimensions such as arousal, valence and dominance. The primery objective of the dataset was to provide the scientific community adequate data for the development of human-computer interfaces (HCI) [69].

B.3 Motor Imagery BCI competition 2008

The third dataset corresponds to motor imagery experiments as part of a challenge in the development of brain computer interfaces in 2008. This dataset contains experiments from 9 subjects performing four specific tasks involving movements of hands and feet. A set of 22 EEG channels were acquired and sampled at 250Hz, then the signals were band-pass filtered between 0.5Hz and 100Hz. Each subject performed a couple of experiment sessions, each consisting in 6 runs and 48 trials per run of the different tasks, with a duration of 6 seconds per trial. From this set of experiments, the movements corresponding to right hand an left hand were selected, the tag *R* were assigned to the trials for right hand and *L* for the trials of left hand [15].

Appendix C

Additional scheme for spatio-temporal analysis of BEA

C.1 Cauchy-Schwartz Quadratic Mutual Information

In [68] a attempt of spatiotemporal quantification of functional connectivity is presented for cognitive states clustering. This is the determination of people with similar cognitive states while developing a visual-motor-driven task. In this processing stage, the idea is to treat each column in $\boldsymbol{U}_{n,k}$ as an independent random variable (RV) and its probability density function (PDF) is estimated. This scheme results in a set of C PDFs, one for each channel. Computing each PDF by their observations is performed using the Parzen estimator without including the element $\boldsymbol{U}_{n,k}(i,i)$ from the RV resulting in a total number of $C-1$ points.

For the spatio-temporal variations quantification, the use of the Cauchy-Schwartz Quadratic Mutual Information (CS-QMI) is employed. The CS-QMI is an estimator of continuous RVs that can be evaluated in a RKHS [68]. Let $U_i, U_j \in \boldsymbol{U}_{n,k}$ represent two RVs from the connectivity data. Then, the CS-QMI between them is defined as the Cauchy-Schwartz divergence between the joint distribution of U_i and U_j an the product of the marginals as equation (C.1) shows.

$$I_{CS}(U_i,U_j) = D_{CS}(f_U|(x_i,xj), f_{U_i}(u_i)f_{U_j}(u_j)), \tag{C.1}$$

with

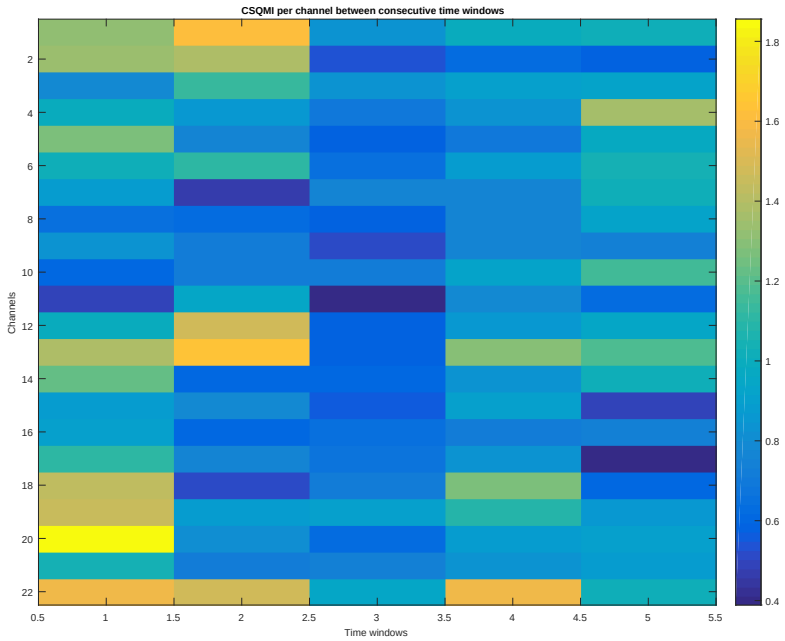


Fig. C.1 Example of CSQMI computed over an experiment of Motor Imagery

$$D_{CS}(f,g) = -log\frac{(\int f(u)g(u)du)^2}{\int f^2(u)dx \int g^2(u)du} \tag{C.2}$$

$$= log\int f^2(u)du + log\int g^2(u)du - 2log\int f(u)g(u)du \tag{C.3}$$

$$= 2\hat{H}_2(fg) - \hat{H}_2(f) - \hat{H}_2(g), \tag{C.4}$$

with $\hat{H}_2(f)$ as the Renyi's quadratic entropy estimator of the PDF f. Using a Gaussian kernel, $\hat{H}_2(f)$ can be written as:

$$\hat{H}_2(f) = -log\int_{-\infty}^{\infty}\left(\frac{1}{N}\sum_{i=1}^{N}G_\sigma(u-u_i)\right)^2 du \tag{C.5}$$

$$= -log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}G_{\sigma\sqrt{2}}(u_j-x_i)\right) \tag{C.6}$$

At this step, the estimation of all $\boldsymbol{U}_{n,k}(i)\forall i = \{1,2,\cdots,C-1\}$ as $f(\boldsymbol{U}_{n,k}(i))$ with a Gaus-

sian kernel and a quantification of the temporal dependencies as $Q_{i,k} = I_{CS}(f(\boldsymbol{U}_{n,k}(i)), g(\boldsymbol{U}_{n,k+1}(i)))$. After this processing we get the following data set:

$$\mathscr{Q}_n = \{\boldsymbol{Q}_n \in \mathbb{R}^{C\times k}; l_n \in [l_{min}, l_{max}]\}, \tag{C.7}$$

Then, this data can be used for discriminative purposes.

C.1.1 P-value computing for CSQMI matrices

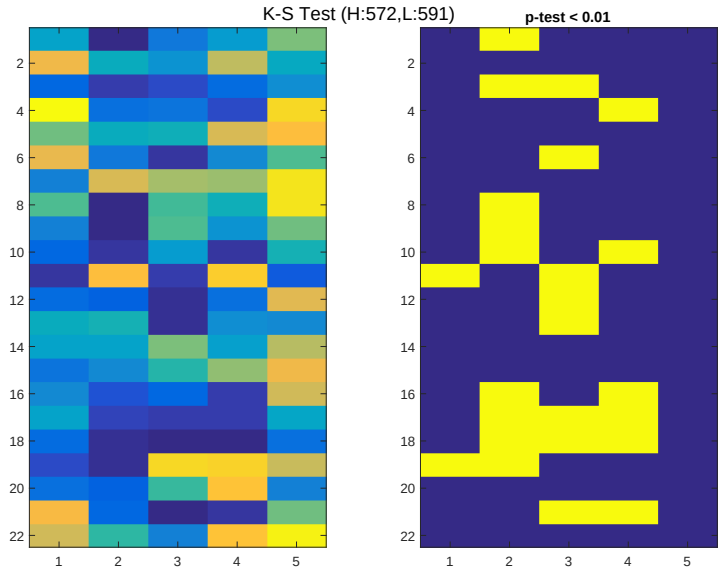


Fig. C.2 Kolmogorv-Smirnov Test for CSQMI matrix array of classes $+$ and $-$ in motor imagery experiments

We propose the use of CSQMI matrices for discriminative finding of BEA patterns. This is, computing the p-value as an indicator of the differences between a couple of classes. Such a indicator, should compute the mutual information and the change in subsequent temporal windows. Let $\chi^+ \in \mathbb{R}^{C\times T\times N}$ be a set of CSQMI matrices of N examples from class $+$; and let $\chi^- \in \mathbb{R}^{C\times T\times M}$ the set of M CSQMI matrices corresponding to $-$ class. We implemented the two sample Kolmogorov-Smirnov for each channel within each temporal window transition. This is a non-parametric hypothesis test that evaluates the cumulative

Fig. C.3 False discovery rate for Kolmogorv-Smirnov Test in motor imagery experiments

distribution function differences from the two vectors of each set.

$$KS = \max_{x} \left(|\hat{F}_{+}(x) - \hat{F}_{-}(x) \right) \tag{C.8}$$

The computed values allow to analyze which channels in which time intervals are representative for containing discriminative information between the two defined classes. The null hypothesis rejection or small p-values are related to these times/channels in which the data distribution seems to come from two different functions.

C.1.2 False detection Rate

Now, to finally quantify the more relevant channels and time transitions regarding the connectivity matrices from both evaluated classes, we propose the False discovery rate (FDR) use. This measure represents the probability of wrongly rejecting the null hypothesis from a set of p-values.

$$\rho = \mathscr{G}(KS_t) \forall t \tag{C.9}$$

In this case, the smaller FDR ρ correspond to those channels which has a lower probability of rejecting the null hypothesis wrongly. This es again related with the case in which is assumed that CSQMI data comes from different probability distributions of classes $+$ and $-$. For this particular experiment, channels $1,3,9,10,13,14y18$ seems to contain information that discriminate the BEA from two classes in different temporal transitions.

Appendix D

Information theory algorithms

D.0.1 Least mean square (LMS)

Data: $\boldsymbol{u}(i), d(i)$
Result: $\boldsymbol{w}$
Initialization
$\boldsymbol{w}(0) = 0$
Computation
while $\{\boldsymbol{u}(i), d(i)\}$ *available* **do**
  $e(i) = d(i) - \boldsymbol{w}^T(i-1)\boldsymbol{u}(i)$
  $\boldsymbol{w}(i) = \boldsymbol{w}(i-1) + \eta e(i)\boldsymbol{u}(i)$
end

D.0.2 Kernel least mean square (KLMS)

Initialization
$\boldsymbol{a}_1(1) = \eta d(1), C(1) = \{\boldsymbol{u}(1)\}, f_1 = \boldsymbol{a}_1 \kappa(\boldsymbol{u}(1), \cdot)$
Computation
while $\{\boldsymbol{u}(i), d(i)\}$ *available* **do**
  Compute the Output
  $f_{i-1}(\boldsymbol{u}(i)) = \sum_{j=1}^{i} \boldsymbol{a}_j(i-1)\kappa(\boldsymbol{u}(i), \boldsymbol{u}(j))$
  Compute the error
  $e(i) = d(i) - f_{i-1}(\boldsymbol{u}(i))$
  Store the new center and the respective coefficient
  $C(i) = \{C(i-1), \boldsymbol{u}(i)\}$
  $\boldsymbol{a}(i) = \eta e(i)$
end

D.0.3 Quantized kernel least mean square (QKLMS)

Initialization
$\boldsymbol{a}_1(1) = \eta d(1), C(1) = \{\boldsymbol{u}(1)\}, f_1 = \boldsymbol{a}_1 \kappa(C(1), \boldsymbol{u}(1))$
Computation
while $\{\boldsymbol{u}(i), d(i)\}$ *available* **do**
  Compute the Output
  $f_{i-1}(\boldsymbol{u}(i)) = \sum_{j=1}^{i} \boldsymbol{a}_j(i-1)\kappa(\boldsymbol{u}(i), \boldsymbol{u}(j))$
  Compute the error
  $e(i) = d(i) - f_{i-1}(\boldsymbol{u}(i))$
  Compute the distance
  $D(C, \boldsymbol{u}(i+1)) = min_{\boldsymbol{c}j \in C(i)} \|\boldsymbol{u}(i+1) - \boldsymbol{c}(j)\|$
  if $D < \delta$ **then**
    $C(i) = C(i-1)$
    Updating the center with the input $\boldsymbol{u}(i)$ *and the coefficient*
    $\boldsymbol{a}_{j*}(i) = \boldsymbol{a}_{j*}(i-1) + \eta e(i)$
    $j* = arg\min_{1 \le j \le |C(i-1)|} \|\boldsymbol{u}(i+1) - \boldsymbol{c}(j)\|$
  else
    Adding new center to the dictionary
    $C(i) = \{C(i-1), \boldsymbol{u}(i)\}$
    $\boldsymbol{a}(i) = \eta e(i)$
  end
end