Analysis and Classification of EEG Signals using Probabilistic Models for Brain Computer Interfaces

THÈSE N^o 3547 (2006)

presentée à la Faculté sciences et techniques de l'ingénieur

École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de docteur ès sciences

par

Silvia Chiappa

Laurea Degree in Mathematics,

University of Bologna, Italy

acceptée sur proposition du jury:

Prof. José del R. Millán, directeur de thèse

Dr. Samy Bengio, co-directeur de thèse

Dr. David Barber, rapporteur

Dr. Sara Gonzalez Andino, rapporteur

Prof. Klaus-Robert Müller, rapporteur

Lausanne, EPFL

2006

Riassunto

Questa tesi esplora l'utilizzo di modelli probabilistici a variabili nascoste per l'analisi e la classificazione dei segnali elettroencefalografici (EEG) usati in sistemi Brain Computer Interface (BCI).

La prima parte della tesi esplora l'utilizzo di modelli probabilistici per la classificazione. Iniziamo con l'analizzare la differenza tra modelli generativi e discriminativi. Allo scopo di tenere in considerazione la natura temporale del segnale EEG, utilizziamo due modelli dinamici: il modello generativo hidden Markov model e il modello discriminativo input-output hidden Markov model. Per quest'ultimo modello, introduciamo un nuovo algoritmo di apprendimento che è di particolare beneficio per il tipo di sequenze EEG utilizzate. Analizziamo inoltre il vantaggio nell'utilizzare questi modelli dinamici verso i loro equivalenti statici.

In seguito, analizziamo l'introduzione di informazione più specifica circa la struttura del segnale EEG. In particolare, un'assunzione comune nell'ambito di ricerca relativa al segnale EEG è il fatto che il segnale sia generato da una trasformazione lineare di sorgenti indipendenti nel cervello e altre componenti esterne. Questa informazione è introdotta nella struttura di un modello generativo e conduce ad una forma generativa di Independent Component Analysis (gICA) che viene utillizzata direttamente per classificare il segnale. Questo modello viene confrontato con un approccio discriminativo più comunemente usato, in cui dal segnale EEG viene estratta informazione rilevante successivamente donata ad un classificatore.

All'inizio, gli utilizzatori di un sistema BCI possono avere molteplici modi realizzare uno stato mentale. Inoltre le condizione psicologiche e fisiologiche possono cambiare da una sessione di registrazione all'altra e da un giorno all'altro. Di conseguenza, il segnale EEG corrispondente può variare sensibilmente. Come primo tentativo di risolvere questo problema, utilizziamo una mistura di modelli gICA in cui il segnale EEG è suddiviso in diversi regimi, ognuno dei quali corrisponde ad un diverso modo di realizzare uno stato mentale.

Potenzialmente, un limite del modello gICA è il fatto che la natura temporale del segnale EEG non è presa in considerazione. Di conseguenza, analizziamo un'estensione di questo modello in cui ogni componente indipendente viene modellata utilizzanto un modello autoregressivo.

Il resto della tesi concerne l'analisi dei segnali EEG e, in particolare, l'estrazione di processi dinamici indipendenti da più elettrodi. Nel campo di ricerca sul BCI, un tale metodo di decomposizione ha varie possibili applicazioni. In particolare, può essere utilizzato per rimuovere artefatti dal segnale, per analizzare le sorgenti nel cervello e in definitiva per aiutare la visualizzazione e l'interpretazione del segnale. Introduciamo una forma particolare di linear Gaussian state-space model che soddisfa varie proprietà, come la possibilità di specificare un numero arbitrario di processi indipendenti e la possibilità di ottenere processi in particolari bande di frequenza. Discutiamo poi un'estensione di questo modello per il caso in cui non conosciamo a priori il numero corretto di processi che hanno generato la serie temporale e la conoscenza circa il loro contenuto di frequenza non è precisa. Quest'estensione è fatta utilizzando un'analisi di Bayes. Il modello che ne deriva può automaticamente determinare il numero e la complessità della dinamica nascosta, con una preferenza per la soluzione più semplice, ed è in grado di trovare processi indipendenti con particolare contenuto di frequenza. Un contributo importante in questo lavoro è lo sviluppo di un nuovo algoritmo per realizzare l'inferenza che è numericamente stabile e più semplice che altri presenti in letteratura.

Parole Chiave

EEG, Brain Computer Interfaces, Classificazione Generativa, Classificazione Discriminativa, Independent Component Analysis, Processi Dinamici Indipendenti, Bayesian Linear Gaussian State-Space Models.

Abstract

This thesis explores latent-variable probabilistic models for the analysis and classification of electroenchephalographic (EEG) signals used in Brain Computer Interface (BCI) systems. The first part of the thesis focuses on the use of probabilistic methods for classification. We begin with comparing performance between 'black-box' generative and discriminative approaches. In order to take potential advantage of the temporal nature of the EEG, we use two temporal models: the standard generative hidden Markov model, and the discriminative input-output hidden Markov model. For this latter model, we introduce a novel 'apposite' training algorithm which is of particular benefit for the type of training sequences that we use. We also asses the advantage of using these temporal probabilistic models compared with their static alternatives.

We then investigate the incorporation of more specific prior information about the physical nature of EEG signals into the model structure. In particular, a common successful assumption in EEG research is that signals are generated by a linear mixing of independent sources in the brain and other external components. Such domain knowledge is conveniently introduced by using a generative model, and leads to a generative form of Independent Components Analysis (gICA). We analyze whether or not this approach is advantageous in terms of performance compared to a more standard discriminative approach, which uses domain knowledge by extracting relevant features which are subsequently fed into classifiers.

The user of a BCI system may have more than one way to perform a particular mental task. Furthermore, the physiological and psychological conditions may change from one recording session and/or day to another. As a consequence, the corresponding EEG signals may change significantly. As a first attempt to deal with this effect, we use a mixture of gICA in which the EEG signal is split into different regimes, each regime corresponding to a potentially different realization of the same mental task.

An arguable limitation of the gICA model is the fact that the temporal nature of the EEG signal is not taken into account. Therefore, we analyze an extension in which each hidden component is modeled with an autoregressive process.

The second part of the thesis focuses on analyzing the EEG signal and, in particular, on extracting independent dynamical processes from multiple channels. In BCI research, such a decomposition technique can be applied, for example, to denoise EEG signals from artifacts and to analyze the source generators in the brain, thereby aiding the visualization and interpretation of the mental state. In order to do this, we introduce a specially constrained form of the linear Gaussian state-space model which satisfies several properties, such as flexibility in the specification of the number of recovered independent processes and the possibility to obtain processes in particular frequency ranges. We then discuss an extension of this model to the case in which we don't know a priori the correct number of hidden processes which have generated the observed time-series and the prior knowledge about their frequency content is not precise. This is achieved using an approximate variational Bayesian analysis. The resulting model can automatically determine the number and appropriate complexity of the underlying dynamics, with a preference for the simplest solution, and estimates processes with preferential spectral properties. An important contribution from our work is a novel 'sequential' algorithm for performing smoothed inference, which is numerically stable and simpler than others previously published.

Keywords

EEG, Brain Computer Interfaces, Generative Classification, Discriminative Classification, Independent Component Analysis, Independent Dynamical Processes, Bayesian Linear Gaussian State-Space Models.

Contents

Acknowledgments

I would like to thank all people who contributed to the realization of this thesis. First of all, David Barber, who supervised me during this Ph.D. I am very grateful to him for the big amount of time that he dedicated to this work, for his patience, for his continual support and motivation and for all topics in machine learning and graphical models that I could learn and discuss with him.

I would also like to thank Samy Bengio, who supervised me during the first two years IDIAP, for his important help and support. I am also grateful to José del R. Millán, who introduced me to the challenging BCI research area.

I am very grateful go to my family, that was close to me during these years.

I would like to thank all people who spent time with me in Martigny. In particular, Christos Dimitrakakis, Bertrand Mesot, Jennifer Petree, Daniel Gatica-Perez, Marios Athineos and Alessandro Vinciarelli.

I finally would like to thank my friends Mauro Ruggeri, Rossana Bertucci, Andrea di Ferdinando, Idina Bolognesi, Giuseppe Umberto Marino and Pep Mouriño.

Chapter 1

Introduction

1.1 Motivation

Non-invasive EEG-based Brain Computer Interface (BCI) systems allow a person to control devices by using the electrical activity of the brain, recorded at electrodes placed over the scalp. A principle motivation for research in this direction is to provide physically-impaired people, who lack accurate muscular control but have intact brain capabilities, with an alternative way of communicating with the outside world. Current possible applications of such systems are: the selection of buttons or letters from a virtual keyboard [Sutter (1992); Birbaumer et al. (2000); Middendorf et al. (2000); Obermaier et al. (2001b); Millán (2003)]; the control of a cursor on a computer screen [Kostov and Polak (2000); Wolpaw et al. (2000)]; the control of a motorized wheelchair [Renkens and Millán (2002)] and the basic control of a hand neuroprosthesis [Pfurtscheller et al. (2000a)].

In BCI research, EEG[1] is preferred to other techniques for analyzing brain function, primarily since it has a relatively fine temporal resolution (on the millisecond scale), enabling rapid estimates of the user's mental state. In addition, the acquisition system is portable, economically affordable and, importantly, non-invasive. However, EEG has the drawback of being relatively weak, and also results from the amassed activity of many neurons, so that it is difficult to perform a precise spatial analysis. EEG is also easily masked by artifacts such as mains-electrical interference and DC level drift. Other common artifacts include user movements, such as eye-movements and blinks, swallowing, etc., inaccuracy of electrode placement and other external artifacts. Furthermore, research in this area is limited by the scarce neurophysiological knowledge about the brain mechanisms generating the outgoing signal.

[1] For the rest of this Section, for EEG we will intend scalp recorded EEG, as opposed to EEG recorded by electrodes implanted in the cortex.

Improvements in BCI research will thus depend on different factors: identification of training protocols and feedback that help the user to achieve and maintain good control of the system; achievement of new insights about brain function; development of better electrodes; design of systems that are easy to use; and, finally, application of more appropriate models for analyzing EEG signals. One important aspect is the development of models for EEG analysis which incorporate prior information about the signal. These models can be used to improve the spatial resolution and to remove noise in the EEG, to select certain EEG characteristics and to aid the visualization and interpretation of the signal. Our belief is that this is an area of potential improvement over most current methods of EEG analysis, and will be therefore a focal point of this thesis.

There exist two main types of EEG-based BCI systems, namely systems which use brain activity generated in response to specific visual or auditory stimuli and systems which use activity spontaneously generated by the user. For example, a common stimulus-driven BCI system uses P300 activity for controlling a virtual keyboard [Donchin et al. (2000)]. The user looks at the letter on the keyboard he/she wishes to communicate. The system randomly highlights parts of the keyboard. When, by chance, that part of the keyboard corresponding to the user's choice is highlighted, a so-called P300 mental response is evoked. This response is relatively robust and easy to recognize in the EEG recordings. A disadvantage with this kind of stimulus-driven BCI systems is the fact that the user cannot operate the system in a free manner. For this reason, systems which use spontaneous brain activity are advantageous [Millán (2003)]. In the spontaneous approach, the user is asked to imagine one of a limited set of mental tasks (i.e. moving either the left or right hand). Based on the EEG recordings, these recognized commands can be used to control a cursor or provide an alternative interface to a virtual keyboard. The advantage of this spontaneous activity approach is that the interface is potentially more immediate and flexible to operate since the system may, in principle, be used to directly recognize the mental state of the user. However, compared to stimulus-driven EEG systems, spontaneous EEG systems present some additional difficulties, such as inconsistencies in the user's mental state, due to change of strategies, fatigue, motivation and other physiological and psychological factors. These issues make the correspondence between electrode activity and mental state more difficult to achieve than with stimulus-driven systems. Despite these additional difficulties, this thesis primarily concentrates on the analysis of spontaneous activity since we believe that, ultimately, this may lead to a more flexible BCI system.

1.2 Goals

This thesis investigates several aspects which are related to the design of principled techniques for analyzing and classifying EEG signals. In order to provide a system which is completely under

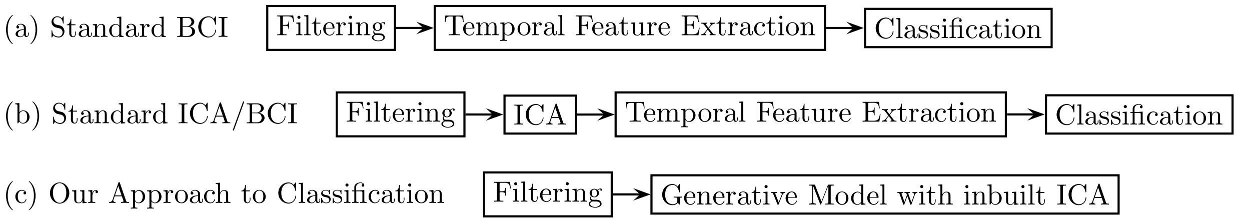

(a) Standard BCI
Filtering
Temporal Feature Extraction
Classification
(b) Standard ICA/BCI
Filtering
ICA
Temporal Feature Extraction
Classification
(c) Our Approach to Classification
Filtering
Generative Model with inbuilt ICA

Figure 1.1: Structure of approaches used in the thesis. Chapter 3 concentrates on using a traditional approach (a) to EEG classification, based on a series of independent steps, without using any form of independence analysis. In Chapters 4 and 5 we compare model (a) and an ICA extended version of (a) (model (b)) with a unified model (c) using a generative method.

user control, we will concentrate our attention on spontaneous EEG mainly recorded using an asynchronous protocol (see Section 2.3.2). Whilst the methods are developed specifically with EEG in mind, they are of potential interest for other forms of signals as well. The development of the models used in the thesis is outlined in Fig. 1.1. One of the main difficulties in EEG analysis is the issue of signal corruption by artifacts and activity not related to the mental task. A straightforward way to alleviate some of these difficulties is to use a filtering step to remove unwanted frequencies from the signal. This is used in most of the models that we consider in the thesis. In the final two Chapters, we will address the issue of filtering more specifically.

The first issue that we want to investigate is the classification of EEG signals using standard 'black-box' methods from the machine learning literature. This relates to model (a) in Fig. 1.1. More specifically, we are interested in a comparison between generative and discriminative approaches when no prior information about the EEG signals is directly incorporated into the structure of the models. To take potential advantage of the temporal nature of EEG, we use two temporal models: the generative Hidden Markov Model (HMM) and the discriminative Input-Output Hidden Markov Model (IOHMM). The application of the IOHMM to classification of sequences in which a class is assigned to each element using the standard training algorithm is inappropriate for the type of EEG data that we consider. Therefore, we investigate a novel 'apposite' objective function and we compare it with another solution proposed in the literature, in which a class is assigned only at the end of the training sequence.

The second goal is to investigate the incorporation of prior beliefs about the EEG signal into the structure of a generative model. In particular, we are interested in using forms of Independent Components Analysis (ICA). On a very high level, a common assumption is that EEG signals can be seen as resulting from activity located at different parts in the brain, or from other independent components, such as artifacts or external noise. This can also be motivated from a physical viewpoint in which the electromagnetic sources within the brain undergo, to a good approximation, linear and instantaneous mixing to form the scalp recorded EEG potentials [Grave de Peralta Menendez et al. (2005)]. We'll look at two approaches to incorporating such an ICA assumption. The most standard approach is depicted in Fig. 1.1b, in which an additional ICA step is used to find independent components from the filtered data. To perform ICA, standard packages such as FastICA [Hyvärinen (1999)] are used. Our particular interest is an alternative method in which independence is build into a single model , see Fig. 1.1c, using a generative approach. This model can be then used as a classifier. The idea is that this unified approach may be potentially advantageous since the independent components are identified along with a model of the data.

In the final part of the thesis, our goal is to build a novel signal *analysis* tool, which can be used by practitioners to visualize independent components which underlie the generation of the observed EEG signals. Such a tool can be used to denoise EEG from artifacts, to spatially filter the signal, to select mental-task related subsignals and to analyze the source generators in the brain, thereby aiding the visualization and interpretation of the mental state.

In general subsignal extraction is an ill-posed problem [Hyvärinen et al. (2001)]. Therefore, the subsignals that are extracted depend on the assumptions of the procedure. In EEG, extracting independent components is hampered by high levels of noise and artifacts corrupting the signal and we need to encode strong beliefs about the structure of the components in order to have confidence in the results.

Most current approaches to extracting EEG components first use filtering on each channel *independently* to select frequencies of interest, followed by a standard ICA method. This two-stage procedure is potentially disadvantageous since the overall assumption of the nature of a component is difficult to determine. In addition, the common approach of performing an initial filtering step may remove important information from the signal useful for identifying independent components. Our interest therefore is to make a *single* model which builds in directly that each component is independent and possesses certain desired spectral properties. In this way, we hope to better understand the assumptions behind each component model and thereby have more confidence in the estimation.

Knowing the number of components in the signal is a key issue in independent component analysis. In most standard packages used for EEG analysis, such as FastICA [Hyvärinen (1999)], the number of components is fixed to the number of channel observations. However, in the case of EEG it is quite reasonable to assume that there may be more or less independent components than channels. We therefore are interested in methods which do not put constraints on the number of components that can be recovered, and that, in addition, can estimate this number. Whilst there exist methods which, in principle, can estimate a number of components which differs from the number of channels, these models either have a complexity which grows exponentially with the number of components [Attias (1999)], or assume a particular distribution for the hidden components which is quite restrictive [Hyvärinen (1998); Girolami (2001); Lewicki and Sejnowski (2000)]. Furthermore, these methods fix the number of components, while we will be interested in determining this number. Finally, these methods do not consider temporal dynamics and are not able to add additional assumptions, specifically spectral constraints. On the other hand temporal ICA methods exist, see for example Pearlmutter and Parra (1997); Penny et al. (2000); Ziehe and Müller (1998), but they do not encode specific spectral properties, nor are they suitable for overcomplete representations.

In order to achieve a flexible form of temporal ICA method, which can automatically estimate the number of components and that is able to encode specific forms of spectral information, we will use a Bayesian linear Gaussian state-space model, developing a novel inference approach which is simpler than other approaches previously proposed and can take advantage of the existing literature on Kalman filtering and smoothing.

In summary, the general goal of this thesis is to introduce methods to incorporate basic prior knowledge into a principled framework for the analysis and classification of EEG signals. This will generally be performed using a probabilistic framework, for which the incorporation of prior knowledge is particularly convenient.

1.3 Organization

The thesis is organized into two main parts. Chapters 3, 4 and 5 concern the *classification* of mental tasks, whilst Chapters 6 and 7 deal with signal *analysis* by extracting independent dynamical processes from an unfiltered multivariate EEG time-series.

Chapter 2 We give a short introduction on different methods for recording brain function and an overview of current approaches for BCI research. We also discuss the state-of-the-art in EEG classification for BCI systems.

Chapter 3 We compare a generative approach versus a discriminative probabilistic approach for the discrimination of three different types of spontaneous EEG activity.

Chapter 4 This Chapter concerns the direct use of a generative ICA model of EEG signals as a classifier. There are several aspects which are considered: first, how this approach relates to other more traditional approaches which commonly view ICA-type methods only as a preprocessing step. Another aspect considered is how the incorporation of prior information into the generative model is beneficial in terms of performance with respect to a discriminative approach. Finally, we investigate if a mixture of the model proposed can solve the issue of EEG variations during different recording sessions and during different days due to inconsistency of user's strategy in performing the mental task and physiological and psychological changes.

Chapter 5 This Chapter extends the generative ICA model to include an autoregressive process, in order to asses the advantage of exploiting the temporal structure of the EEG signals.

Chapter 6 This Chapter outlines an approach for the analysis of EEG signals by forming a decomposition into independent dynamical processes. We do this by introducing a constrained version of a linear Gaussian state-space model. This may be then used for extracting independent processes underlying EEG signal and select processes which contain specific spectral frequencies. We discuss some of the drawbacks of standard maximum-likelihood training, including the difficulty of automatically determining the number and complexity of the underlying processes.

Chapter 7 Here we extend the model introduced in Chapter 6 by performing a Bayesian analysis which enables to specify a given model structure by incorporating prior information about the model parameters. This extension allows to automatically determine the number and appropriate complexity of the underlying dynamics (with a preference for the simplest solution) and to estimate independent dynamical processes with preferential spectral properties.

Chapter 8 In this Chapter we draw conclusions about the work presented in the previous Chapters and we outline possible future directions.

Chapter 2

Present-day BCI Systems

In this Chapter we give an overview and background of different methodologies for measuring brain activity and discuss advantages and disadvantages of EEG as a technique for BCI. We then explain the different types of EEG signals and recording protocols which are currently used in BCI research. Finally we present the state-of-the-art in BCI related EEG classification.

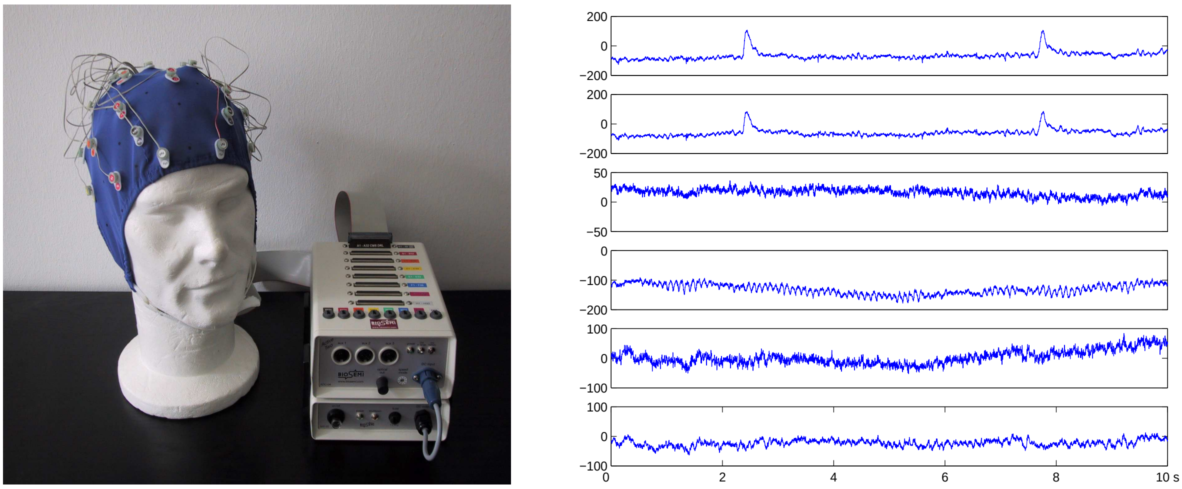
2.1 Measuring Brain Function

There are several methods for measuring brain function. Each technique has different characteristics and its own region of applicability. We shortly describe the main methods and discuss their properties in relation to EEG.

Electroencephalography

Electroencephalographic (EEG) signals are a measure of the electrical activity of the brain recorded from electrodes placed on the cortex or on the scalp. A comprehensive introduction on EEG can be found in Niedermeyer and Silva (1999). While implanted electrodes can pick up the activity of single neurons, scalp electrodes encompass the activity of many neurons. The poor spatial resolution of scalp EEG (limited to 1 centimeter [Nunez (1995)]) is due to the low conductivity of the skull, the cerebrospinal fluid and the meninges, which cause a reduction and dispersion of the activity originated in the cortex. Scalp EEG is also very sensitive to subject movement and external noise. One important strength of EEG is the temporal resolution, which is in the range of milliseconds [Nunez (1995)]. Unlike PET and fMRI, that rely on blood flow which may be decoupled from the brain electrical activity, EEG measures brain activity directly. In summary, EEG has the following characteristics:

- It measures directly brain function.



(a) (b)

Figure 2.1: (a): The portable Biosemi 32-channel system used for recording some of the EEG data analyzed in this thesis. (b): 10 seconds of EEG data recorded with this system while a person is performing continual mental generation of words starting with a certain letter from two (left and right hemisphere) frontal, two central and two parietal electrodes (50 Hz mains contamination has been removed). The first two electrodes present two blinking artifacts, while the fourth electrode presents strong rhythmical activity centered at 10 Hz.

- It has a high temporal resolution, in the range of milliseconds.
- The spatial resolution is in the range of centimeters for scalp electrodes, while implanted electrodes can measure the activity of single neurons.
- Scalp electrodes are non-invasive while implanted electrodes are invasive.
- The required equipment is portable.

In Fig. 2.1a we show a scalp EEG acquisition system using 32 electrodes. This system has been used for recording some EEG data used in this thesis. In Fig. 2.1b we plot 10 seconds of EEG activity recorded from frontal, central and parietal electrodes in the left and right hemispheres (see Fig. 2.2a), while a person is performing continual mental generation of words starting with a certain letter. This multichannel recording is typical of EEG (50 Hz mains contamination has been removed). The EEG traces exhibit two isolated low frequency events in the first two channels, which correspond to eye-blink artifacts. In addition, the fourth channel presents strong rhythmic activity around 10 Hz, which indicates that the underlying area of the right hemisphere is not activated during this cognitive task (see Section 2.3.2).

Functional Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) uses radio waves and magnetic fields to provide an image of internal organs and tissues. A specific type of MRI, called the Blood-Oxygen-Level-Dependent Functional Magnetic Resonance Imaging (BOLD-fMRI) [Huettel et al. (2004)], measures the quick metabolic changes that take place in the active parts of the brain, by measuring regional differences in oxygenated blood. Increased neural activity causes an augmented need of oxygen, which is provided by the neighboring blood vessels. The temporal resolution of this technique is of the order of 0.1 seconds and the spatial resolution of the order of a few millimeters. BOLD-fMRI is very sensitive to head movement. A disadvantage of BOLD-fMRI is the fact that it measures neural activity indirectly, and it is therefore susceptible to influence by non-neural changes in the brain. BOLD-fMRI is non-invasive and does not involve the injection of radioactive materials as other techniques that measure metabolic changes (i.e. PET). In conclusion, BOLD-fMRI has the following main characteristics:

- It measures indirectly brain function.
- It has a moderate temporal resolution, around 0.1 seconds.
- It has high spatial resolution, in the order few millimeters.
- It is a non-invasive technique.
- It requires a large-scale non-portable equipment.

Positron Emission Tomography

Positron Emission Tomography (PET) estimates the local cerebral blood flow, oxygen and glucose consumption and other regional metabolic changes, in order to identify the active regions of the brain. As BOLD-fMRI, it therefore provides an indirect measure of neural activity. The spatial resolution of PET is of the order of few millimeters. However, the temporal resolution varies from minutes to hours [Nunez (1995)]. The main drawback of PET is that it requires the injection of a radioactive substance into the bloodstream. In summary, PET has the following characteristics:

- It measures indirectly brain function.
- It has a low temporal resolution, in the range of minutes to hours.

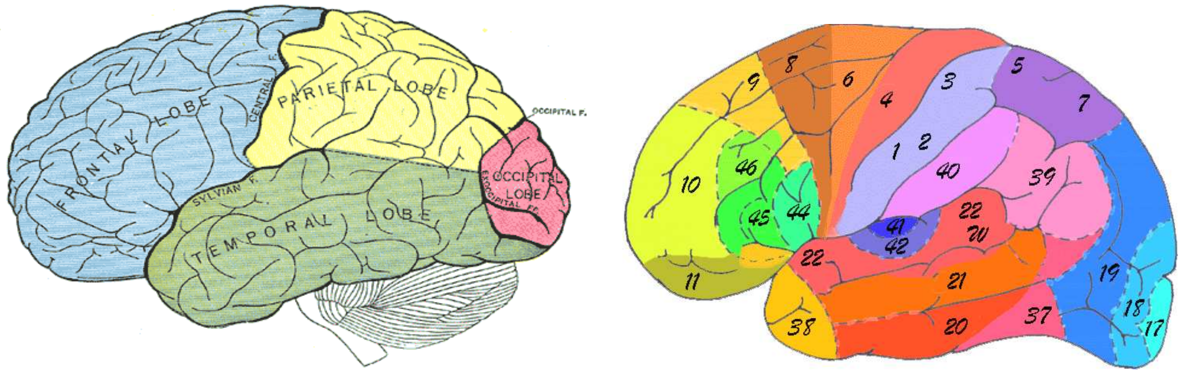
- It has high spatial resolution, in the order of few millimeters.
- It is an invasive technique.
- It requires a large-scale non-portable equipment.

Magnetoencephalography

Magnetoencephalography (MEG) measures the magnetic field components perpendicular to the scalp generated by the brain activity, with gradiometers placed at a certain distance (from 2 to 20 mm) from the scalp [Malmivuo et al. (1997)]. MEG, as scalp EEG, is more sensitive to neocortical sources than other sources farther from the sensors. It has a temporal resolution of the range of milliseconds. The spatial resolution of MEG is subject of controversial discussions. Indeed, it is widely believed that MEG has better spatial resolution than EEG because the skull has low conductivity to electric current but is transparent to magnetic fields. However, it would seem better founded that the spatial resolution of MEG is limited to 1 cm [Nunez (1995)]. The controversial debate about this point is discussed in Crease (1991); Malmivuo et al. (1997). One important advantage of MEG over EEG is the fact that the measured signals are not distorted by the body. However, the signal strengths are extremely small and specialized shielding is required to eliminate the magnetic interference of the external environment. As EEG, MEG is a direct measure of brain function. It is believed that MEG provides complementary information to scalp EEG, even if this is also a controversial point [Malmivuo et al. (1997)]. In conclusion, MEG has the following characteristics:

- It measures directly brain function.
- It has a high temporal resolution, in the order milliseconds.
- It has a low spatial resolution, limited to 1 cm.
- It is a non-invasive technique.
- It requires a large-scale non-portable equipment.

Researchers often combine EEG or MEG with fMRI or PET to obtain both high temporal and spatial resolution. For BCI research, scalp EEG is the most widely used methodology because it is non-invasive, it has a high temporal resolution, and the acquisition system is portable and cheap relative to MEG, PET and fMRI, which are still very expensive technologies.



(a)

(b)

Figure 2.2: (a): The cerebral cortex of the brain is divided into four distinct sections: frontal lobe, parietal lobe, temporal lobe and occipital lobe. The frontal lobe contains areas involved in cognitive functioning, speech and language. The parietal lobe contains areas involved in somatosensory processes. Areas involved in the processing of auditory information and semantics are located in the temporal lobe. The occipital lobe contains areas that process visual stimuli. (b): A more detailed map of the cortex covering the lobes contains 52 distinct areas, as defined by Brodmann [Brodmann (1909)]. Some important areas involved in the mental tasks used in BCI are: area 4, which corresponds to the primary motor area; area 6, which is the premotor or supplemental motor area. These two areas are involved in motor activity and planning of complex, coordinated movements. Areas 8 and 9 are also related to motor function. Other areas are: the Broca's area (44), which is involved in speech production, and the Wernicke's area (22), which is involved in the understanding and comprehension of spoken language. Areas 17, 18 and 19 are involved in visual projection and association. Areas 1, 2, 3 and 40 are related to somatosensory projection and association.

2.2 Frequency Range Terminology

EEG recordings often present rhythmical patterns. One example is the rhythmical activity centered at 10 Hz when motor areas of the cortex (see Fig. 2.2b, areas 4, 6, 8, 9) are not active. Despite the large levels of noise present in EEG recordings, identifying rhythmic activity is relatively straightforward using spectral analysis [Proakis and Manolakis (1996)]. For this reason, many approaches to BCI using EEG search for the presence/absence of rhythmic activity in certain frequencies and locations.

A rough characterization of EEG waves associated to different brain function exists, although the terminology is imprecise and sometimes abused, since EEG waves are often classified as belonging to a certain frequency range on the basis of mere visual inspection rather than by using a precise frequency analysis. Bearing this in mind, we can define six main types of waves, namely: δ, θ, α, μ, β and γ waves.

δ **waves** are the lowest brain waves (below 4 Hz). They are present in deep sleep, infancy and some organic brain disease. They appear occasionally and last no more than 3-4 seconds.

θ **waves** are in the frequency range from 4 Hz to 8 Hz and are related to drowsiness, infancy, deep sleep, emotional stress and brain disease.

α **waves** are rhythmical waves that appear between 8 and 13 Hz. They are present in most adults during a relaxed, alert state. They are best seen over the occipital area but they also appear in the parietal and frontal regions of the scalp (see Fig. 2.2a). Alpha waves attenuate with drowsiness and open eyes [Neidermeyer (1999)].

μ **waves or Rolandic** μ **rhythms** are in the frequency range of the α waves. However, they are not always present in adults. They are seen over the motor cortex (see Fig. 2.2b, areas 4, 6, 8, 9) and attenuate with limb movement [Neidermeyer (1999)] (see Section 2.3.2).

β **waves** appear over 13 Hz and are associated to thinking, concentration and attention. Some β rhythms are reduced with cognitive processing and limb movement (see Section 2.3.2).

γ **waves** appear in the frequency range approximately 26-80 Hz. Gamma rhythms are related to high mental activity, perception, problem solving, fear and consciousness.

2.3 Present-Day EEG-based BCI

The first human scalp recordings were made in 1928[1] by Hans Berger, who discovered that characteristic patterns of EEG activity were associated with different levels of consciousness [Berger (1929)]. From that time on, EEG has been used mainly to evaluate neurological disorders and to analyze brain function. The idea of an EEG-based communication system was first introduced by Vidal in the 1970s. Vidal showed that visual evoked potentials could provide a communication channel to control the movement of a cursor [Vidal (1973, 1977)]. However, the field was relatively dormant until recently, when the discovery of the mechanisms and spatial location of many brain-wave phenomena and their relationships with specific aspects of brain function yielded the possibility to develop systems based on the recognition of specific electro-physiological signals. Furthermore, a variety of studies, which started with the intention to explore therapeutic applications of EEG, demonstrated that people can learn to control certain features of their EEG activity. Finally, the development of computer hardware and software

[1] Spontaneous brain activity in the brain of animals was measured much earlier [Finger (1994)].

made possible the online analysis of multichannel EEG. All these aspects caused an explosion of interest in this research area. A detailed review of present-day BCI approaches can be found in Kübler et al. (2002); Wolpaw et al. (2002); Curran and Stokes (2003); Millán (2003).

Present day EEG-based BCIs can be classified into three main groups, according to the type of EEG signal that they use and the position of the electrodes: those using scalp recorded EEG waveforms generated in response to specific stimuli (*exogenous* EEG); those using scalp recorded spontaneous EEG signals, that is EEG waveforms that occur during normal brain function (*endogenous* EEG); and those using implanted electrodes.

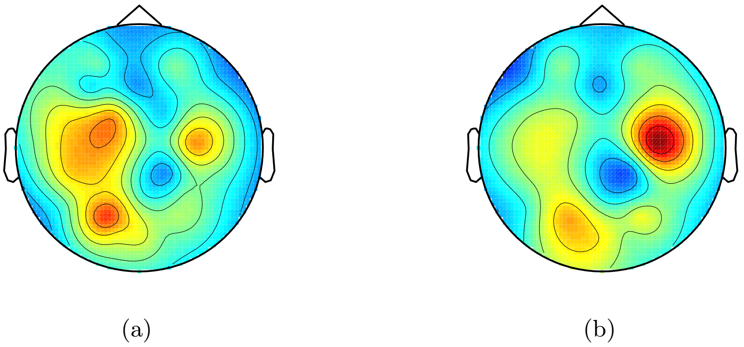
2.3.1 Exogenous EEG-based BCI

Exogenous EEG activity (also called Evoked Potentials (EP) [Rugg and Coles (1995)]) is generated in response to specific stimuli. This activity is relatively easy to detect and in most cases does not requires any user training. However, the main drawback of BCI systems based on exogenous EEG is the fact that they do not allow spontaneous control by the user. There are two main type of EP used in BCI:

Visual Evoked Potentials There are BCI systems which use the amplitude of a visual evoked EEG signal to determine gaze direction. One example is given in Sutter (1992), where the user faces a virtual keyboard in which letters flash one at a time. The user looks directly at the letter that he/she wants to select. The visual evoked potential recorded from the scalp when the selected letter flashes is larger than when other letters flash, so that the system can deduce the desired choice.

Other systems are based on the fact that looking at a stimulus blinking at a certain frequency evokes an increase in EEG activity at the same frequency in the visual cortex [Middendorf et al. (2000); Cheng et al. (2002)]. For example, in the system described in Middendorf et al. (2000), several virtual buttons appear on the screen and flash at different frequencies. The users look at the button that they want to choose and the system recognizes the button by measuring the frequency content in the EEG.

P300 Evoked Potentials Some BCI researchers [Farwell and Donchin (1998); Donchin et al. (2000)] use P300 evoked potentials, that is positive peaks at a latency of about 300 milliseconds generated by infrequent or particularly significant auditory, visual or somatosensory stimuli, when alternated with frequent or routine stimuli.



(a) (b)

Figure 2.3: Topographic distribution of power in the α band while a person is performing repetitive left (a) and right (b) imagined movement of the hand. The topographic plots have been obtained by interpolating the values at the electrodes using the eeglab toolbox [http://www.sccn.ucsd.edu/eeglab]. Red regions indicate the presence of strong rhythmical activity. We can notice the different topography of the α oscillations for the two mental tasks.

2.3.2 Endogenous EEG-based BCI

BCI based on endogenous brain activity requires a training period in which users learn strategies to generate the mental states associated to the control of the system. The duration of the training depends on both the algorithms used to analyze the EEG and the ability of the user to operate the system. These systems are very sensitive to the physiological and psychological condition of the user, i.e. motivation, fatigue, etc. There are two main types of endogenous EEG signals which are considered for BCI applications, namely Bereitschaftspotential and EEG rhythms [Jahanshahi and Hallett (2003)].

Bereitschaftspotential

A commonly used spontaneous EEG signal in BCI is the Bereitschaftspotential (BP) [Birbaumer et al. (2000); Blankertz et al. (2002)]. BP is a slowly decreasing cortical potential which develops 1-1.5 seconds prior to limb movement. The BP has a different spatial distribution depending on the used limb. For example, roughly speaking, BP shows larger amplitude contralateral to the moving finger. Therefore, the difference in the spatial distribution of BP can be used as an indicator of left or right limb movement. The same kind of activity is also present when the movement is only imagined.

EEG Rhythms

Most researchers working on endogenous EEG-based BCI focus on brain oscillations associated with sensory and cognitive processing and motor behavior [Anderson (1997); Pfurtscheller et al. (2000b); Roberts and Penny (2000); Wolpaw et al. (2000); Millán et al. (2002)]. When a region of the brain is not actively involved in a processing task, it tends to synchronize the firing of its neurons, giving rise to several rhythms such as the *Rolandic μ rhythm*, in the α band (7-13 Hz), and the *central β rhythm*, above 13 Hz, both originating over the sensorymotor cortex. Sensory and cognitive processing or movement of the limbs are usually associated to a decrease in μ and β rhythms. A similar blocking, which involves similar brain regions, is present with motor imagery, that is when a subject only imagines to make a movement but this movement does not take place [Pfurtsheller and Neuper (2003)]. While some β rhythms are harmonics of the μ rhythms, some of them have different spatial location and timing, and thus they are considered independent EEG features [Pfurtscheller and da Silva (1999)]. Some cognitive tasks commonly used in BCI are arithmetic operations, music composition, rotation of geometrical objects, language, etc. The spatial distribution of these rhythms is different according to the location of the limb and to the type of cognitive processing. In Fig. 2.3 we show the differences in the scalp distribution of EEG rhythms (α band) while a user is performing imagination of movement of the left (Fig. 2.3a) and right (Fig. 2.3b) hand. Red regions indicate the presence of strong rhythmical activity.

There exist two different protocols used to analyze motor-planning related EEG, namely *synchronous* and *asynchronous* protocols.

Synchronous Protocol Many endogenous BCI systems operate in a synchronous mode. This means that, at an instant of time, the user is asked to make a specific (imagined) movement for a fixed amount of time determined by the system. In general, a short interval between two consecutive movements is given to the user, in order for him/her to go back to baseline brain activity. The EEG data from each movement is then classified.

This synchronous protocol has the limitation that the user is restricted to communicating in time intervals defined by the system, which may result in a slow and non-flexible BCI system.

Asynchronous Protocol In an asynchronous protocol, the user repetitively performs a certain task, without any resting interval, and the system performs classification at fixed intervals without knowledge of when each motor plan has started. In principle, this kind of system is more flexible, but the resulting EEG signal is more complex and difficult to analyze than in the synchronous case.

2.3.3 Intracranial EEG-based BCI

EEG signals recorded at the scalp provide a non-invasive way of monitoring brain activity. However scalp EEG pick up activity of a broad region in the cortex. There exist many BCI systems which use micro-electrodes surgically implanted in the cortex to record action potentials of single neurons. There are two main types of systems which use implanted electrodes: motor and cognitive-based systems. Motor-based systems record activity from motor areas related to limb movement. In some case, the neural firing rates controlled by the user is used to move for example a cursor on a screen [Kennedy et al. (2000)]. In some other case, the recorded activity is used to determine motor parameters or patterns of muscular activations [Schwartz and Moran (2000); Nicolelis (2001); Donoghue (2002); Carmena et al. (2003); Santucci et al. (2005)]. Cognitive-based systems record activity related to higher level cognitive processes that organize behavior [Pesaran and Andersen (2006)].

2.4 State-of-the-Art in EEG Classification

Current scalp EEG-based BCI systems use a variety of different algorithms to determine the user's intention from the EEG signal. Determining the state-of-the art for classification methods is hampered for the following reasons:

> EEG signals are noisy and subject to high variability, and the amount of available labelled training data is often low. A classifier which therefore performs better on one dataset may give different results on another dataset.
>
> In the case of systems based on spontaneous brain activity, variability is present as a consequence of the user's specific physical and psychological conditions. For this reason, training and testing should be performed on *different* sessions and/or day in order to ascertain a more realistic generalization performance of the algorithms. There are a few studies reporting difference in performance under different training and testing conditions, see for example Anderson and Kirby (2003). In Chapter 4 we will also discuss this issue. Most researchers report results in a less realistic scenario in which training and testing is done on data recorded very close in time.
>
> Different EEG signals and protocols may require different classification strategies, which makes comparison of techniques more complex.

Historically, few datasets have been publicly available for BCI research. For this reason, we limit our overview of methods and results to the datasets from the BCI competitions, which initiated in 2001 with the intention of standardizing comparison between competing methods. Currently, there are three competition datasets [BCI Competition I (2001); BCI Competition II (2003); BCI Competition III (2004)]. Most training and testing datasets are recorded very close in time and during the same day. For this reason, reported performances are likely to be optimistic compared to the performances one would expect in a realistic scenario. Furthermore, competition participants are free to select electrodes and features before performing classification, so that it becomes difficult to understand if the difference in the results is due to feature and electrode selection or to the classification method. Finally, depending on the dataset and protocol used, different methods need to be applied.

Despite these caveats, the BCI competition provides the main comparison arena for the algorithms, and we therefore here discuss the approaches taken by the winners of the two most recent competitions.

2.4.1 BCI Competition II

Dataset Ia This data was recorded while a person was moving a cursor up or down on a computer screen using Bereitschaftspotential (BP). Cortical positivity (negativity) led to a downward (upward) movement.

The winner of the competition [Mensh et al. (2004)] used BP and spectral features [Proakis and Manolakis (1996)] from high β power band and linear discriminant analysis (LDA) [Mardia (1979)]. Comparable results were obtained by G. Dornhege and co-workers who used regularized LDA [Friedman (1989)] on the intensity of evoked response [Blankertz et al. (2004)]. Similar results were also obtained by K.-M. Chung and co-workers who used a Support Vector Machine (SVM) classifier[2] [Cristianini and Taylor (2000)] on the raw data [Blankertz et al. (2004)].

Dataset Ib This data was recorded while a person was moving a cursor up and down on a computer screen using BP, as in dataset Ia.

The winner, V. Bostanov, used a stepwise LDA on wavelet [Chui (1992)] transformed data [Blankertz et al. (2004)]. However, the results are barely better than using random guessing, so that their significance is lost.

Dataset IIa The users used μ and β-rhythm amplitude to control the vertical movement of a cursor toward a target located at the edge of the video screen.

The winner used bandpass filtering, Common Spatial Patterns (CSP)[3] [Fukunaga (1990);

[2] It is not reported if a linear or non-linear SVM has been used.

[3] For a two class problem, CSP finds a linear transformation of the data which maximizes the variance for one class while minimizing it for the other class. More specifically, if Σ_1 and Σ_2 are the covariances of class 1 and 2

Ramoser et al. (2000); Dornhege et al. (2003)], and regularized LDA [Blanchard and Blankertz (2004)].

Dataset IIb In this dataset, the user faces a 6×6 matrix of characters, whose rows and columns are jointly highlighted at random. The user selects the character he/she wants to communicate by looking at it. Only infrequently is the desired character highlighted by the system. This infrequent stimulus produces a particular EEG signal (see P300 evoked potential in Section 2.3.1). The goal is to understand which character the user wants to select by analyzing the P300 response.

Five out of seven participants of the competition obtained 100% accuracy in predicting which characters the user wanted to select. They used a Gaussian SVM on bandpass filtered data [Meinicke et al. (2002)]; continuous wavelet transform, scalogram peak detection and stepwise LDA; and regularized LDA on spatio-temporal features [Blankertz et al. (2004)].

Dataset III The data was recorded while a person was controlling a feedback bar in one dimension by imagination of left or right hand movement. The task was to provide classification at each time-step.

The winner used a multivariate Gaussian distribution on bandpass filtered data for each class with Bayes rule [Blankertz et al. (2004)].

Dataset IV In this dataset, the user had to perform two tasks: depressing a keyboard key with a left or right finger.

The winner applied CSP and LDA to extract three types of features derived from BP, μ and β rhythms, and used a linear perceptron for classification [Wang et al. (2004)].

2.4.2 BCI Competition III

Dataset I The data was recorded while a person was performing imagined movements of either the left small finger or the tongue.

The winner used a combination of band-power, CSP or waveform mean and LDA for feature extraction and a linear SVM for classification.

Dataset II The data was recorded using a P300 speller paradigm as in dataset IIb of BCI Competition II.

respectively, CSP finds a matrix W and a diagonal matrix D such that $W\Sigma_1 W^{\mathsf{T}} = D$ and $W\Sigma_2 W^{\mathsf{T}} = I - D$ (the symbol T indicates the transpose operator). Then a CSP model for class 1 is given by selecting the columns of W which correspond to the biggest eigenvalues (elements of D), while a CPS model for class 2 is given by selecting the columns of W which correspond to the smallest eigenvalues.

Preprocessing and Feature Extraction
Classification

Figure 2.4: Standard Approach to BCI. Preprocessing removes artifacts from the data. Feature extraction is commonly made to represent the strength of predefined spectral features in the data. These features are then passed to standard classification systems.

The winner used a linear SVM on bandpass filtered data.

Dataset IIIa The user had to perform imagery left hand, right hand, foot or tongue movements.

The winner used Fisher ratios over channel-frequency-time bins, μ and β passband filters, CSP and classified using an SVM[4].

Dataset IIIb The user had to perform motor imagery (left hand, right hand) with online feedback.

The winner combined BP and α and β features. Classification was performed by fitting a multivariate Gaussian distribution to each task and using Bayes rule.

Dataset IVa In this dataset, the user had to perform three tasks: imagination of left hand, right hand and right foot movement.

The winner used a combination of CSP, autoregressive coefficients and temporal waves of the BP and classified using LDA.

Dataset IVc In this dataset, the user had to perform three tasks: imagination of left hand, right foot and tongue movements. The test data was recorded more than three hours after the training data, with the tongue task replaced by the relax task. The goal was to classify a trial as belonging to the left, right or relax task, even if no training data for the relax task was available.

The winner used CSP, and LDA for classification.

Dataset V This data was recorded while a user was performing imagination of left and right hand movements and generation of words beginning with the same random letter.

The best results were found using a distance-based classifier [Cuadras et al. (1997)] and an SVM with a Gaussian kernel on provided power spectral density features.

[4] The winner does not specify if a linear or non-linear SVM has been used.

2.4.3 Discussion of Competing Methodologies

From the competition results, we can conclude that the best performances were obtained by using various LDA approaches and linear or Gaussian SVM classifiers. However, these are more or less the only methods used by the competitors and it would seem that the difference in the results may be attributed more to the electrode selection and feature extraction than to the classifiers themselves.

From the competition, it is also clear that linear methods are widely used. An interesting debate about the relative benefit of linear and non-linear methods for BCI is presented in [Müller et al. (2003)]. In this paper, it is suggested that linear classifiers may be a good approach for EEG due to their simplicity and given that they are presumably less prone to overfitting caused by noise and outliers. However, from the experimental results, it is not clear which approach is to be preferred. For example, in Garrett et al. (2003), the authors report the results of LDA and two non-linear classifiers, MLP [Bishop (1995)] and SVM, applied to the classification of spontaneous EEG during five mental tasks, showing that non-linear classifiers produce better classification results. However, in Penny and Roberts (1997), the authors compare the use of a committee of Bayesian neural networks with LDA for two mental tasks, reporting no advantage of the non-linear neural network over LDA. The difference in the conclusions reported in Garrett et al. (2003) and Penny and Roberts (1997) may be due to the different number of mental tasks used in the two sets of experiments. Indeed, while Garrett et al. (2003) use five mental tasks, Penny and Roberts (1997) analyze only two mental tasks. The first problem may be more complicated and the use of a non-linear method may be beneficial. This seems to be confirmed in Hauser et al. (2002), where the authors compare the use of a linear SVM with Elman [Elman (1990)] and time-delay neural networks [Waibel et al. (1989)] for three mental tasks, reporting poor performance of the linear SVM. They suggest to use a non-linear static classifier [Millán et al. (2002)].

It is interesting to note that all proposed methods use the approach displayed in Fig. 2.4, in which filtering is performed to remove unwanted frequencies, after which features are extracted and then fed into a separate classifier. In this thesis, specifically in Chapters 4 and 5, we will explore a rather different approach in which information about the EEG and the mental tasks is not used to extract features but rather embedded directly into a model, which may subsequently be used for direct classification of the EEG time-series.

Chapter 3

HMM and IOHMM for EEG Classification

The work presented in this Chapter is an extension of Chiappa and Bengio (2004).

3.1 Introduction

This Chapter discusses the classification of EEG signals into associated mental tasks. This will also be the subject of the following two Chapters, which discuss various alternative classification procedures.

There are two standard approaches to classification, *discriminative* and *generative*, which we outline below, and the goal is to evaluate these approaches using some baseline models in the machine learning literature. Both generative and discriminative approaches have potential advantages and disadvantages, as we shall explain, and in this Chapter we will evaluate how they perform when implemented using limited prior information about the form of EEG signals. Since the EEG signals are inherently temporal, we will consider a classical generative temporal model, the Hidden Markov Model (HMM), and a relatively new discriminative temporal model, the Input-Output Hidden Markov Model (IOHMM). A central contribution of this Chapter is a novel form of training algorithm for the IOHMM, which considerably improves the performance relative to the baseline standard algorithm. Of additional interest is the value of using such temporal models over related static versions. We will therefore evaluate whether or not the HMM improves on the mixture of Gaussians model and whether or not the IOHMM improves on the Multilayer Perceptron.

Generative Approach

In a generative approach, we define a model for generating data v belonging to particular mental task $c \in 1, \ldots, C$ in terms of a distribution $p(v|c)$. Here, v will correspond to a time-series of multi-channel EEG recordings, possibly preprocessed. The class c will be one of three mental tasks (imagined left/right hand movements and imagined word generation). For each class c, we train a separate model $p(v|c)$, with associated parameters Θ_c, by maximizing the likelihood of the observed signals for that class. We then use Bayes rule to assign a novel test signal v^* to a certain class c according to:

$$p(c|v^*) = \frac{p(v^*|c)p(c)}{p(v^*)}.$$

That model c with the highest posterior probability $p(c|v^*)$ is designated the predicted class.

Advantages In general, the potential attraction of a generative approach is that prior information about the structure of the data is often most naturally specified through $p(v|c)$. However, in this Chapter, we will not explicitly incorporate prior information into the structure of the model, but rather use a limited form of preprocessing to extract relevant frequency information. Incorporating prior information directly into the structure of the generative model will be the subject of Chapters 3 and 4.

Disadvantages A potential disadvantage of the generative approach is that it does not directly target the central issue, which is to make a good classifier. That is, the goal of generative training is to model the observation data v as accurately as possible, and not to model the class distribution. If the data v is complex, or high-dimensional, it may be that finding a suitable generative data model is a difficult task. Furthermore, since each generative model is separately trained for each class, there is no competition amongst the models to explain the data. In particular, if each class model is quite poor, there may be little confidence in the reliability of the prediction. In other words, training does not focus explicitly on the differences between mental tasks, but rather on accurately modelling the distribution of the data associated to each mental task.

The generative temporal model used in this Chapter is the Hidden Markov Model (HMM) [Rabiner and Juan (1986)]. Here the joint distribution $p(v_{1:T}|c)$ is defined for a sequence of multivariate observations $v_{1:T} = \{v_1, \cdots, v_T\}$. The HMM is a natural candidate as a generative temporal model due to its widespread use in time-series modeling. Additionally, the HMM is well-understood, robust and computationally tractable.

Discriminative Approach

In a discriminative probabilistic approach we define a single model $p(c|v)$ common to all classes, which is trained to maximize the probability of the class label c. This is in contrast to the generative approach above, which models the data and not the class. Given novel data v^*, we then directly calculate the probabilities $p(c|v^*)$ for each class c, and assign v^* to the class with the highest probability.

Advantages A clear potential advantage of this discriminative approach is that it directly addresses the issue that we are interested in solving, namely making a classifier. We are here therefore modelling the discrimination boundary, as opposed to the data distribution in the generative approach. Whilst the data from each class may be distributed in a complex way, it could be that the discrimination boundary between the classes is relatively easy to model.

Disadvantages A potential drawback of the discriminative approach is that the model is usually trained as 'black-box' classifier, with no prior knowledge of how the signal is formed built into the model structure.

In principle, one could use a generative description $p(v|c)$, building in prior information, and form a joint distribution $p(v, c)$, from which a discriminative model $p(c|v)$ may be obtained using Bayes rule. Subsequently, the parameters Θ_c for this model could be found by maximizing the discriminative class probability. This approach is rarely taken in the machine learning literature since the resulting functional form of $p(c|v)$ is often complex and training is difficult.

For this reason, here we do not encode prior knowledge into the model structure or parameters, but rather specify an explicit model $p(c|v)$ with the requirement of having a tractable functional form for which training is relatively straightforward.

The discriminative probabilistic approach considered in this Chapter is the Input-Output Hidden Markov Model (IOHMM) [Bengio and Frasconi (1996)]. The IOHMM is a natural temporal discriminative model to consider since it is tractable and has shown good performance in dealing with complex time-series [Bengio et al. (2001)].

As we shall see, the IOHMM nominally requires a class label (output variable) for each time-step t. Since in our EEG data each training sequence corresponds to only a single class, model resources are wasted on ensuring that consecutive outputs are in the same class. We therefore introduce a novel training algorithm for the IOHMM that compensates for this difficulty and greatly improves the generalization accuracy of the model.

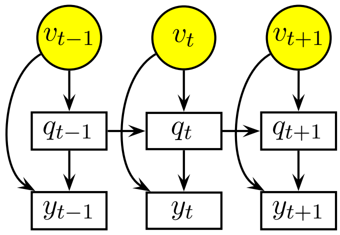


Figure 3.1: Graphical model of the IOHMM. Nodes represent the random variables and arrows indicate direct dependence between variables. In our case, the output variable y_t is discrete and represents the class label, while the input variable v_t is the continuous (feature extracted from the) EEG observation. The yellow nodes indicate that these variables are given, so that no associated distributions need to be defined for $v_{1:T}$.

3.2 Discriminative Training with IOHMMs

An Input-Output Hidden Markov Model (IOHMM) is a probabilistic model in which, at each time-step $t \in 1, \ldots, T$, an output variable y_t is generated by a hidden discrete variable q_t, called the state, and an input variable v_t [Bengio and Frasconi (1996)]. The input variables represent an observed (preprocessed) EEG sequence and the output variables represent the classes.

The joint distribution of the state and output variables, conditioned on the input variables, is given by:

$$p(q_{1:T}, y_{1:T}|v_{1:T}) = p(y_1|v_1, q_1)p(q_1|v_1)\prod_{t=2}^{T} p(y_t|v_t, q_t)p(q_t|v_t, q_{t-1})\,,$$

whose graphical model [Lauritzen (1996)] representation is depicted in Fig. 3.1. Thus an IOHMM is defined by *state-transition* probabilities $p(q_t|v_t, q_{t-1})$, and *emission* probabilities $p(y_t|v_t, q_t)$. An issue in the IOHMM is how to make these transition and emission distributions functionally dependent on the continuous input v_t. In this work we use a nonlinear parameterization which has proven to be powerful in previous applications [Bengio et al. (2001)]. More specifically, we model the input-dependent state-transition distributions using:

$$p(q_t = i|v_t, q_{t-1} = j) = \frac{e^{z^i}}{\sum_k e^{z^k}}\,, \tag{3.1}$$

where $z^k = \sum_{j=0}^{W} w_{kj} f\left(\sum_{i=0}^{U} u_{ji} v_t^i\right)$ and f is a nonlinear function. The emission distributions $p(y_t = c|v_t, q_t = j)$ are modeled in a similar way. This parameterization is called a Multilayer Perceptron (MLP) [Bishop (1995)] in the machine learning literature. The denominator in Eq. (3.1) ensures that the distribution is correctly normalized.

The IOHMM enables us to specify, for each time-step t, a class label y_t. Alternatively, since in our EEG data each training sequence corresponds to only a single class, we may assign a single class label for the whole sequence. As we will see, in this case the label need to be assigned at the end of the sequence and the variables corresponding to unobserved outputs (class labels) for times less than T are marginalized away to form a suitable likelihood. These two standard approaches are outlined below.

3.3 Continual Classification using IOHMMs

For our EEG discrimination task, features from a window of EEG sequence will be extracted and will represent an input v_t of the IOHMM. Therefore, a single input v_t already conveys some class information. In this case, a reasonable approach consists of specifying the class label for each input of the sequence. The log-likelihood objective function is[1]:

$$\mathcal{L}(\Theta) = \log \prod_{m=1}^{M} p(y^m_{1:T} | v^m_{1:T}, \Theta) \,, \tag{3.2}$$

where Θ denotes the model parameters and m indicates the m-th training example.

After learning the parameters Θ, a test sequence is assigned to the class c^* such that:

$$c^* = \arg\max_c p(y_1 = c, \dots, y_T = c | \Theta).$$

For notational convenience, in the rest of the Section 3.3 we will describe the learning using a single sequence, the generalization to several sequences being straightforward.

3.3.1 Training for Continual Classification

A common approach to maximize log-likelihoods in latent variable models is to use the Expectation Maximization (EM) algorithm [McLachlan and Krishnan (1997)]. However, in our case the usual M-step cannot be carried out in closed form, due the constrained form of the transition and emission distributions. We therefore use a variant, the Generalized Expectation Maximization (GEM) algorithm [McLachlan and Krishnan (1997)]:

Generalized EM At iteration i, the following two steps are performed:

E-step Compute $\mathcal{Q}(\Theta, \Theta^{i-1}) = \langle \log p(q_{1:T}, y_{1:T} | v_{1:T}, \Theta) \rangle_{p(q_{1:T}|v_{1:T}, y_{1:T}, \Theta^{i-1})}$,

[1] We assume, for notational simplicity, that all sequences have the same length T. This will be the case in the experiments considered in this Chapter.

M-step Find a value Θ^i such that $\mathcal{Q}(\Theta^i, \Theta^{i-1}) \geq \mathcal{Q}(\Theta^{i-1}, \Theta^{i-1})$.

In the above $\langle \cdot \rangle_{p(\cdot)}$ denotes the expectation operator with respect to the distribution $p(\cdot)$. The inequality in the M-step ensures that the log-likelihood is not decreased at each iteration and that, under fairly general conditions, the sequence of values $\{\Theta^i\}$ converges to a local maximum Θ^*.

The conditional expectation of the complete data log-likelihood $\mathcal{Q}(\Theta, \Theta^{i-1})$ can be expressed as:

$$\begin{aligned}\mathcal{Q}(\Theta, \Theta^{i-1}) = &\sum_{t=1}^{T} \langle \log p(y_t|v_t, q_t, \Theta) \rangle_{p(q_t|v_{1:T}, y_{1:T}, \Theta^{i-1})} \\ &+ \sum_{t=2}^{T} \langle \log p(q_t|v_t, q_{t-1}, \Theta) \rangle_{p(q_{t-1:t}|v_{1:T}, y_{1:T}, \Theta^{i-1})} \\ &+ \langle \log p(q_1|v_1, \Theta) \rangle_{p(q_1|v_{1:T}, y_{1:T}, \Theta^{i-1})} \ .\end{aligned} \tag{3.3}$$

Thus the E-step requires $p(q_t|v_{1:T}, y_{1:T}, \Theta^{i-1})$ and $p(q_{t-1:t}|v_{1:T}, y_{1:T}, \Theta^{i-1})$. Computing these marginals is a form of inference and is achieved using the recursive formulas presented in Section 3.3.2. We perform a generalized M-step using a gradient ascent method[2]:

$$\Theta^i = \Theta^{i-1} + \lambda \frac{\partial \mathcal{Q}(\Theta, \Theta^{i-1})}{\partial \Theta}\Big|_{\Theta=\Theta^{i-1}} .$$

Here λ is the *learning rate* parameter, which will be chosen using a validation set. The derivatives of $\log p(y_t|q_t, v_t, \Theta)$, $\log p(q_t|q_{t-1}, v_t, \Theta)$ and $\log p(q_1|v_1, \Theta)$ with respect to the network weights w_{ij} and u_{ij} are achieved using the chain rule (back-propagation algorithm [Bishop (1995)]).

3.3.2 Inference for Continual Classification

In Bengio and Frasconi (1996), the terms $p(q_t|v_{1:T}, y_{1:T})$ (and $p(q_{t-1:t}|v_{1:T}, y_{1:T})$) are computed using a parallel approach, which consists of a set of forward recursions for computing the term $p(q_t, y_{1:t}|v_{1:t})$ and a set of backward recursions for computing $p(y_{t+1:T}|v_{t+1:T}, q_t)$. The two values are then combined to compute $p(q_t|v_{1:T}, y_{1:T})$. To be consistent with other smoothed inference procedures in this thesis, we present here an alternative backward pass in which $p(q_t|v_{1:T}, y_{1:T})$ is directly recursively computed using $p(q_t|v_{1:t}, y_{1:t})$.

[2] In our implementation, we use only a single gradient update. Multiple gradient updates would correspond to a more complete M-step. However, in our experience, convergence using the single gradient update form is reasonable.

Forward Recursions:

The filtered state posteriors $p(q_t|v_{1:t}, y_{1:t})$ can be computed recursively in the following way:

$$\begin{aligned}
p(q_t|v_{1:t}, y_{1:t}) &\propto p(q_t, y_t|v_{1:t}, y_{1:t-1}) \\
&= p(y_t|v_{1:t}, q_t, y_{1:t-1})p(q_t|v_{1:t}, y_{1:t-1}) \\
&= p(y_t|v_t, q_t) \sum_{q_{t-1}} p(q_{t-1:t}|v_{1:t}, y_{1:t-1}) \\
&= p(y_t|v_t, q_t) \sum_{q_{t-1}} p(q_t|v_{1:t}, q_{t-1}, y_{1:t-1})p(q_{t-1}|v_{1:t}, y_{1:t-1}) \\
&= p(y_t|v_t, q_t) \sum_{q_{t-1}} p(q_t|v_t, q_{t-1})p(q_{t-1}|v_{1:t-1}, y_{1:t-1}) \,,
\end{aligned}$$

where the proportionality constant is determined by normalization.

Backward Recursions:

In the standard backward recursions presented in the IOHMM literature, $p(y_{t+1:T}|v_{t+1:T}, q_t)$ is computed independently of $p(q_t, y_{1:t}|v_{1:t})$ computed in the forward recursions. These two terms are subsequently combined to obtain $p(q_t|v_{1:T}, y_{1:T})$. Here we give an alternative backward recursion in which $p(q_t|v_{1:T}, y_{1:T})$ is directly computed as a function of $p(q_{t+1}|v_{1:T}, y_{1:T})$, using the filtered state posteriors. Specifically, we compute the smoothed state posterior recursively using:

$$\begin{aligned}
p(q_t|v_{1:T}, y_{1:T}) &= \sum_{q_{t+1}} p(q_{t:t+1}|v_{1:T}, y_{1:T}) \\
&= \sum_{q_{t+1}} p(q_t|v_{1:T}, q_{t+1}, y_{1:T})p(q_{t+1}|v_{1:T}, y_{1:T}) \qquad (3.4) \\
&= \sum_{q_{t+1}} p(q_t|v_{1:t+1}, q_{t+1}, y_{1:t})p(q_{t+1}|v_{1:T}, y_{1:T}) \,.
\end{aligned}$$

The term $p(q_t|v_{1:t+1}, q_{t+1}, y_{1:t})$ can be computed as:

$$\begin{aligned}
p(q_t|v_{1:t+1}, q_{t+1}, y_{1:t}) &\propto p(q_{t:t+1}|v_{1:t+1}, y_{1:t}) \\
&= p(q_{t+1}|v_{1:t+1}, q_t, y_{1:t})p(q_t|v_{1:t+1}, y_{1:t}) \\
&= p(q_{t+1}|v_{t+1}, q_t)p(q_t|v_{1:t}, y_{1:t}) \,,
\end{aligned}$$

where the proportionality constant is determined by normalization. The joint distribution $p(q_{t:t+1}|v_{1:T}, y_{1:T})$ is found from Eq. (3.4) before summing over q_{t+1}.

In the next Section we will see that the continual classification objective function (3.2) is problematic and we will introduce a novel alternative procedure.

3.3.3 Apposite Continual Classification

We described a training algorithm for the IOHMM which requires the specification a class y_t for each input v_t. In this case the objective function to maximize is:

$$\log \prod_{m=1}^{M} p(y_1^m = c^m, \ldots, y_T^m = c^m | v_{1:T}^m, \Theta) \,, \tag{3.5}$$

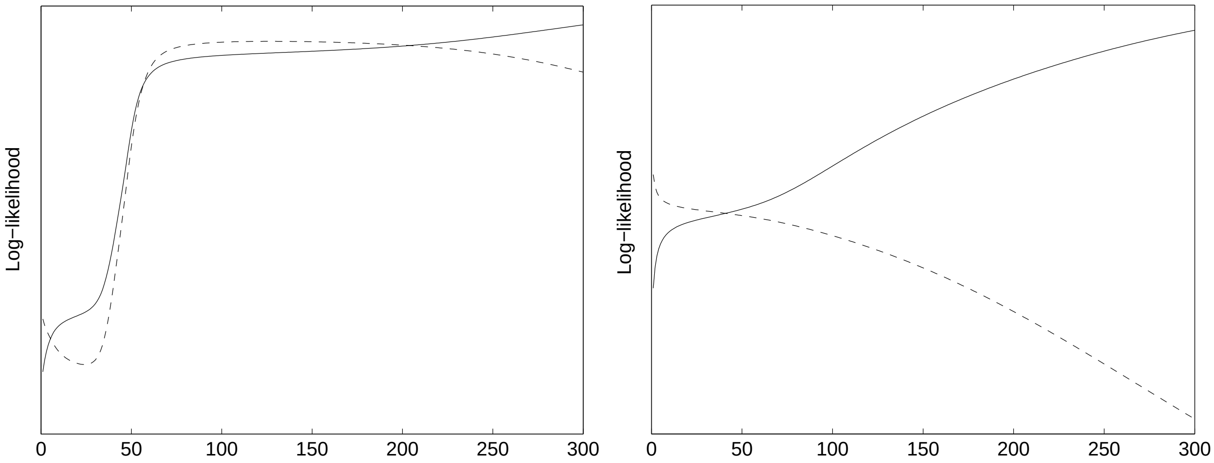
where c^m is the correct class label. During testing we compute $p(y_1 = c, \ldots, y_T = c | v_{1:T}, \Theta)$ for each class c and assign the test sequence $v_{1:T}$ to the class which gives the highest value. Ideally, we would like the distance between the probability of the correct and incorrect class to increase during the training iterations. The log-likelihood of an incorrect assignment is defined by:

$$\log \prod_{m=1}^{M} \sum_{i^m=1, i^m \neq c^m}^{C} p(y_1^m = i^m, \ldots, y_T^m = i^m | v_{1:T}^m, \Theta) \,. \tag{3.6}$$

However, the fact that we specify the same class label for the whole sequence of inputs may force the model resources to be spent in this characteristic, with the consequence that the model focuses on predicting the *same* class at each time-step t, instead of focusing on *which* class is predicted.

Example Problem with Standard Training

We will illustrate the problem with an example. We are interested in discriminating among three mental tasks from the corresponding EEG sequences. We train an IOHMM model on the EEG sequences from different classes using the objective function (3.5). In Fig. 3.2a we plot, with a solid line, the value of the log-likelihood (3.5) at different training iterations. As we can see, the log-likelihood (3.5) increases at each iteration, as expected. Using the same model parameters, at each iteration we compute the probability of the incorrect class (3.6) (Fig. 3.2a, dashed line). As we can see, at the beginning and end of training the model focuses on increasing the distance between (3.5) and (3.6). However, there are transient iterations in which the distance between (3.5) and (3.6) becomes smaller. Since, during training, we present to the model only one type of input sequence whose elements have all the same class label, this characteristic dominates learning and the discriminative power of the IOHMM is partially lost.



(a)

(b)

Figure 3.2: Evolution log-likelihood evaluated for different specifications of the class label. (a): Standard Continual Classification. (b): Apposite Continual Classification. Solid line (-): Log-likelihood values when the correct class labels are specified (Eq. (3.5)). Dashed line (- -): Log-likelihood values when incorrect identical class labels are specified (Eq. (3.6)).

The Apposite Objective

To avoid the problems mentioned with continual classification training, we need to adjust the training to discriminate between joint probabilities of identical outputs. A candidate objective function to achieve this is:

$$\mathcal{D}(\Theta) = \log \prod_{m=1}^{M} \frac{p(y_1^m = c^m, \ldots, y_T^m = c^m | v_{1:T}^m, \Theta)}{\sum_{i^m=1}^{C} p(y_1^m = i^m, \ldots, y_T^m = i^m | v_{1:T}^m, \Theta)}\,, \tag{3.7}$$

where c^m is the correct class label. This objective function encourages the model to discriminate between the generation of identical correct class labels and the generation of identical incorrect class labels. To maximize Eq. (3.7) we cannot use a GEM, since the presence of the denominator means that Jensen's inequality cannot be used to justify convergence to a local maximum of the objective function [Neal and Hinton (1998)]. We therefore use gradient ascent of $\mathcal{D}(\Theta)$. Computing directly the derivatives of $\mathcal{D}(\Theta)$ is complicated due to the coupling of the parameters caused by the hidden variables $q_{1:T}$. However, we can simplify the problem in the following way: we notice that, by denoting with $\mathcal{L}_c(\Theta)$ and $\mathcal{N}(\Theta)$ the numerator and denominator of $\mathcal{D}(\Theta)$ for a single sequence, we can write:

$$\begin{aligned}\frac{\partial \mathcal{D}(\Theta)}{\partial \Theta} &= \frac{\partial \log \mathcal{L}_c(\Theta)}{\partial \Theta} - \frac{\partial \log \mathcal{N}(\Theta)}{\partial \Theta} \\ &= \frac{\partial \log \mathcal{L}_c(\Theta)}{\partial \Theta} - \frac{1}{\mathcal{N}(\Theta)} \sum_{i=1}^{C} \frac{\partial \mathcal{L}_i(\Theta)}{\partial \Theta} \\ &= \frac{\partial \log \mathcal{L}_c(\Theta)}{\partial \Theta} - \sum_{i} \frac{\mathcal{L}_i(\Theta)}{\mathcal{N}(\Theta)} \frac{\partial \log \mathcal{L}_i(\Theta)}{\partial \Theta}\,.\end{aligned}$$

That is, ultimately, we only need to compute the derivatives of the terms $\log \mathcal{L}_i(\Theta)$. This is advantageous since the presence of the logarithm breaks the likelihood terms into separate factors. In order to find their derivatives, we use the following result:

$$\begin{aligned}\frac{\partial}{\partial \Theta} \log p(y_{1:T}|v_{1:T}) &= \frac{1}{p(y_{1:T}|v_{1:T})} \frac{\partial}{\partial \Theta} \sum_{q_{1:T}} p(y_{1:T}, q_{1:T}|v_{1:T}) \\ &= \frac{1}{p(y_{1:T}|v_{1:T})} \sum_{q_{1:T}} p(y_{1:T}, q_{1:T}|v_{1:T}) \frac{\partial \log p(y_{1:T}, q_{1:T}|v_{1:T})}{\partial \Theta} \\ &= \left\langle \frac{\partial \log p(y_{1:T}, q_{1:T}|v_{1:T})}{\partial \Theta} \right\rangle_{p(q_{1:T}|y_{1:T}, v_{1:T})}\,.\end{aligned}$$

In the final expression above, thanks to the logarithm, we can break the derivative into individual terms, as in in the complete data log-likelihood (3.3). In this way we have transformed the difficult problem of finding the derivative of the original objective function into a simpler problem. Inferences required for the averages above can be performed using the results in Section 3.3.2.

Advantage of Apposite Training

We trained an IOHMM model on the same EEG data of the example discussed above, but using the new apposite objective function $\mathcal{D}(\Theta)$. In Fig. 3.2b we plot with a solid line the evolution of the log-likelihood of sequences consisting of identical correct class labels (Eq. (3.5)), while the dashed line indicates the log-likelihood of sequences consisting of identical but incorrect class labels (Eq. (3.6)). It can be see that the distance between the values 3.5 and 3.6 increases with the training iterations, as desired. Hence, we believe that this novel training criterion may significantly improve the classification ability of the IOHMM.

3.4 Endpoint Classification using IOHMMs

An alternative way of training an IOHMM model to classify sequences, and avoid the problem with continual classification, is to assign a single class label for the whole sequence. In this case the class label need to be given at the end of the sequence. Indeed assigning a single output label at a time $t \neq T$ would imply that $p(y_t|v_{1:T}) = p(y_t|v_{1:t})$, that is future information about the input sequence is not taken into account for determining the posterior class probability. In this case, training maximizes the following conditional log-likelihood:

$$\mathcal{L}(\Theta) = \log \prod_{m=1}^{M} p(y_T^m | v_{1:T}^m, \Theta)\,. \tag{3.8}$$

Once trained, the model may be applied to a novel sequence to find the most likely endpoint class.

For notational convenience, in the rest of the Section 3.4 we will describe the learning using a single sequence.

3.4.1 Training for Endpoint Classification

Analogously to Section 3.3.1, in order to maximize Eq. (3.8) we can use a Generalized EM procedure:

Generalized EM At iteration i, the following two steps are performed:

E-step Compute $\mathcal{Q}(\Theta, \Theta^{i-1}) = \langle \log p(q_{1:T}, y_T | v_{1:T}, \Theta)\rangle_{p(q_{1:T}|v_{1:T}, y_T, \Theta^{i-1})}$,

M-step Find a value Θ^i such that $\mathcal{Q}(\Theta^i, \Theta^{i-1}) \geq \mathcal{Q}(\Theta^{i-1}, \Theta^{i-1})$.

The conditional expectation of the complete data log-likelihood $\mathcal{Q}(\Theta, \Theta^{i-1})$ can be expressed as:

$$\begin{aligned}\mathcal{Q}(\Theta, \Theta^{i-1}) = {} & \langle \log p(y_T | q_T, v_T, \Theta)\rangle_{p(q_T|v_{1:T}, y_T, \Theta^{i-1})} \\ & + \sum_{t=2}^{T} \langle \log p(q_t | q_{t-1}, v_t, \Theta)\rangle_{p(q_{t-1:t}|v_{1:T}, y_T, \Theta^{i-1})} \\ & + \langle \log p(q_1 | v_1, \Theta)\rangle_{p(q_1|v_{1:T}, y_T, \Theta^{i-1})}\,.\end{aligned}$$

Thus the E-step requires $p(q_T|v_{1:T}, y_T, \Theta^{i-1})$ and $p(q_{t-1:t}|v_{1:T}, y_T, \Theta^{i-1})$. These can be computed as follows.

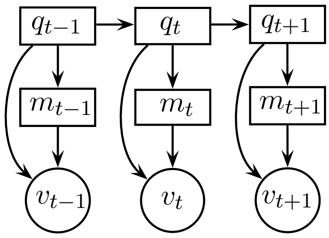


Figure 3.3: Graphical representation of a hidden Markov model with mixture of Gaussian emission distributions.

3.4.2 Inference for Endpoint Classification

The term $p(q_T|v_{1:T}, y_T)$ can be computed in the following way:

$$p(q_T|v_{1:T}, y_T) \propto p(q_T, y_T|v_{1:T}) = p(y_T|v_T, q_T) \sum_{q_{T-1}} p(q_T|v_T, q_{T-1})p(q_{T-1}|v_{1:T-1}) \, .$$

The filtered state posterior $p(q_t|v_{1:t})$ $(t < T)$ needed above can be computed using the following forward recursion:

$$p(q_t|v_{1:t}) = \sum_{q_{t-1}} p(q_t|v_t, q_{t-1})p(q_{t-1}|v_{1:t-1}) \, .$$

The smoothed state posterior $p(q_t|v_{1:T}, y_T)$ can be computed by backward recursion:

$$\begin{aligned} p(q_t|v_{1:T}, y_T) &= \sum_{q_{t+1}} p(q_t|v_{1:T}, q_{t+1}, y_T)p(q_{t+1}|v_{1:T}, y_T) \\ &= \sum_{q_{t+1}} p(q_t|v_{1:t+1}, q_{t+1})p(q_{t+1}|v_{1:T}, y_T) \, , \end{aligned}$$

where

$$p(q_t|v_{1:t+1}, q_{t+1}) \propto p(q_{t:t+1}|v_{1:t+1}) = p(q_{t+1}|v_t, q_t)p(q_t|v_{1:t}) \, .$$

3.5 Generative Training with HMMs

In this Section we present the generative alternative to the previously described discriminative models. Readers familiar with HMMs may wish to skip this Section.

A Hidden Markov Model (HMM) [Rabiner and Juan (1986)] is a probabilistic model in which, at each time-step t, an observed variable v_t is generated by an hidden discrete variable q_t, called the state, which evolves according to a Markovian dynamics. As done in most cases in which the output variables are continuous, we assume that the visible variable is distributed as a mixture of Gaussians, that is:

$$p(v_t|q_t) = \sum_{m_t} p(v_t|q_t, m_t)p(m_t|q_t),$$

where $p(v_t|q_t, m_t)$ is a Gaussian distribution with mean μ_{q_t,m_t} and covariance Σ_{q_t,m_t}. The joint distribution is given by:

$$p(v_{1:T}, m_{1:T}, q_{1:T}) = p(v_1|m_1, q_1)p(m_1|q_1)p(q_1) \prod_{t=2}^{T} p(v_t|q_t, m_t)p(m_t|q_t)p(q_t|q_{t-1}),$$

whose graphical model representation is depicted in Fig. 3.3. A different model with associated parameters Θ_c is trained for each class $c \in 1, \ldots, C$ by maximizing the log-likelihood $\log \prod_{m \in M_c} p(v^m_{1:T}|\Theta_c)$ of the M_c observed training sequences.

During testing, a novel sequence is assigned to the class whose model gives the highest joint density of observations:

$$c^* = \arg\max_c p(v_{1:T}|\Theta_c).$$

In the next Section we describe how the model parameters are learned.

3.5.1 Inference and Learning in the HMM

In the HMM, the conditional expectation of the complete data log-likelihood $\mathcal{Q}(\Theta, \Theta^{i-1})$ for a single sequence can be expressed as:

$$\begin{aligned}\mathcal{Q}(\Theta, \Theta^{i-1}) = &\sum_{t=1}^{T} \langle \log p(v_t|q_t, m_t, \Theta)\rangle_{p(q_t,m_t|v_{1:T},\Theta^{i-1})} \\ &+ \sum_{t=1}^{T} \langle \log p(m_t|q_t, \Theta)\rangle_{p(q_t,m_t|v_{1:T},\Theta^{i-1})} \\ &+ \sum_{t=2}^{T} \langle \log p(q_t|q_{t-1}, \Theta)\rangle_{p(q_{t-1:t}|v_{1:T},\Theta^{i-1})} \\ &+ \langle \log p(q_1|\Theta)\rangle_{p(q_1|v_{1:T},\Theta^{i-1})} .\end{aligned}$$

Thus the E-step ultimately consists of estimating $p(q_t, m_t|v_{1:T}, \Theta^{i-1})$ and $p(q_{t-1:t}|v_{1:T}, \Theta^{i-1})$. This inference is achieved using the recursive formulas given in Appendix A.1. These formulas differ from the standard forward-backward algorithm in the HMM literature[3] [Rabiner and Juan

[3] In the standard forward-backward algorithm, $p(q_t, v_{1:t})$ is computed in the forward pass, while $p(v_{t+1:T}|q_t)$ is computed in the backward pass. The two values are then combined to compute $p(q_t|v_{1:T})$. Then $p(q_t, m_t|v_{1:T})$ is

(1986)] in that, in the forward pass, the filtered posterior $p(q_t, m_t|v_{1:t})$ is computed and then, in the backward pass, the smoothed posterior $p(q_t, m_t|v_{1:T})$ is directly recursively estimated from the filtered posterior. This approach is analogous to the one presented for the IOHMM in Section 3.3.2.

The M-step consists of setting $\frac{\partial \mathcal{Q}(\Theta,\Theta^{i-1})}{\partial \Theta}$ to zero, which can be solved in closed form. The updates are presented in Appendix A.1.

3.5.2 Previous Work using HMMs

Hidden Markov models have been already applied to EEG signals (see, for example Flexer et al. (2000); Zhong and Ghosh (2002)). Specifically to BCI research, HMMs have been used for classifying motor imaginary movements [Obermaier et al. (1999, 2001a); Obermaier (2001); Obermaier et al. (2003)]. The idea was to model changes of μ and β rhythms using a temporal model. In this case the EEG signal was filtered, different features were extracted (band-power, adaptive autoregressive coefficients and Hjort parameters [Hjort (1970)]), and then fed into an HMM model. One HMM model for each mental task was created and used in a generative way as in our case. The EEG data was recorded using a synchronous protocol (see Section 2.3.2), in which the users had to follow a fixed scheme for performing the mental task followed by some seconds of resting. The HMM model showed some improvement over linear discriminant analysis [Mardia (1979)]. In our case, we use an asynchronous recording protocol in which the user concentrates repetitively on a mental action for a given amount of time and switches directly to the next task, without any resting period. In this case, the patterns of EEG activity may be different.

3.6 EEG Data and Experiments

In this Section we will compare the discriminative approach, using the IOHMM model in which a class label is assigned for each observation v_t and trained with the apposite objective function (3.7), and the generative approach, using the HMM described in Section 3.5. Whilst HMMs have been previously applied to EEG classification, as far as we are aware, the application of the IOHMM to EEG classification is novel.

We will also evaluate the classification performance on two static alternatives to the IOHMM and HMM, in order to asses the advantage in using temporal models. A natural way to form static alternatives is to drop temporal dependencies $p(q_t|q_{t-1})$. The IOHMM then becomes a model in which the outputs $p(y_t|v_t)$ from an MLP are multiplied to give $p(y_{1:T}|v_{1:T}) =$

computed as $p(q_t, m_t|v_{1:T}) = p(q_t|v_{1:T})p(m_t|q_t, v_t)$.

$\prod_{t=1}^{T} p(y_t|v_t)$. Analogously, the HMM reduces to a Gaussian Mixture-type model in which $p(v_t)$ are combined to give $p(v_{1:T}) = \prod_{t=1}^{T} p(v_t)$. We will call these models MLP and GMM respectively.

These experiments concern classification of the following three mental tasks:

1. Imagination of self-paced left hand movements,
2. Imagination of self-paced right hand movements,
3. Mental generation of words starting with a letter chosen spontaneously by the subject at the beginning of the task.

EEG potentials were recorded with the Biosemi ActiveTwo system [http://www.biosemi.com] using the following electrodes located at standard positions of the 10-20 International System [Jasper (1958)]: FP1, FP2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz and O2 (see Fig. 3.4). The raw potentials were re-referenced to the common average reference in which the overall mean is removed from each channel. The signals were recorded at a sample rate of 512 Hz. Subsequently, the band 8-30 Hz was selected with a 2nd order Butterworth filter [Proakis and Manolakis (1996)]. This preprocessing filter allow us to focus on cognitive and motor-related EEG rhythms. Out of the 32 electrodes, only the following 19 electrodes were considered for the analysis: F3, FC1, FC5, T7, C3, CP1, CP5, P3, Pz, P4, CP6, CP2, C4, T8, FC6, FC2, F4, Fz and Cz (see Fig. 3.4 (red)).

The EEG data was acquired in an unshielded room from two healthy subjects without any experience with BCI systems during three consecutive days. Each day, the subjects performed 5 recording sessions lasting around 4 minutes followed by an interval of around 5 to 10 minutes. During each recording session, around every 20 seconds an operator verbally instructed the subject to continually perform one of the three mental tasks described above.

In order to extract concise information about EEG rhythms we computed the power spectral density (PSD) [Proakis and Manolakis (1996)] from the EEG signal. This is a common approach used in the BCI literature [Millán (2003)]. The PSD was computed over half a second of data with a temporal shift of 250 milliseconds[4]. As input $v_{1:T}$ to the IOHMM model, and output $v_{1:T}$ to the HMM model, we gave 7 consecutive PSD estimates ($T = 7$). This means that each training sequence corresponds to 2 seconds of EEG data.

HMM and IOHMM models were trained on the EEG signal of the first 2 days of recordings, while the first and last sessions of the last day were used for validation and test respectively. We obtained the following number of training, validation and test sequences:

[4] Additional window lengths and shifts not presented here were considered. Similar experimental conclusions were obtained.

	Subject A	Subject B
IOHMM	19.0%	18.5%
MLP	22.5%	23.3%
HMM	25.0%	26.4%
GMM	24.1%	27.5%

Table 3.1: Error rate of Subject A and Subject B using HMM, IOHMM and their static counterparts: GMM and MLP. Random guessing corresponds to an average error of 66.7%.

- Subject A: 4297, 976, 996
- Subject B: 3890, 912, 976

Temporal Models

IOHMM Setup: The validation set was used to select the number of iterations for the gradient updates, the number of possible values for the hidden states (up to 7) and the number of hidden units (between 5 and 50) for the MLP transition and emission networks. The MLP networks had one hidden layer with hyperbolic tangent nonlinearity.

HMM Setup: The validation set was used to choose the number of EM iterations, the number of fully-connected states (in the range from 2 to 7) and the number of Gaussians (between 1 and 15).

Static Models

MLP Setup: As in the IOHMM case, the validation set was used to select the number of iterations for the gradient updates and the number of hidden units between 5 and 50. The MLP had one hidden layer with a hyperbolic tangent nonlinearity.

GMM Setup: As in the HMM case, the validation set was used to choose the number of EM iterations and the number of Gaussians (between 1 and 15).

3.6.1 Results

From the results presented in Table 3.1 we can observe the superior performance of the discriminative approach over the generative approach. This can be explained by the fact that, when using a generative approach, a separate model is trained for each class on examples of that class only. As a consequence, the training focuses on the characteristics of each class and not on

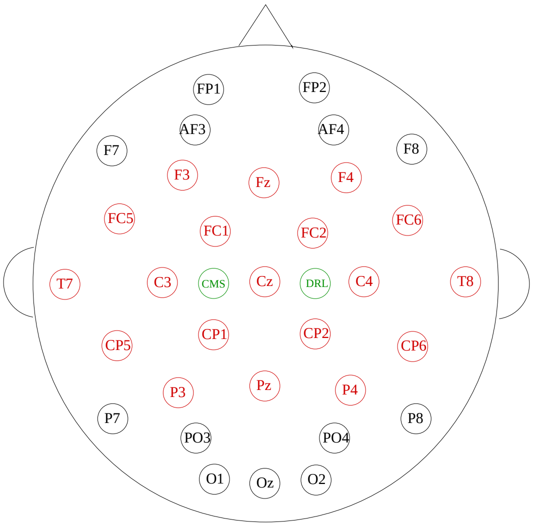


Figure 3.4: Electrode placement in the Biosemi ActiveTwo system [http://www.biosemi.com] used to record the EEG data. In red are displayed the electrodes selected for the experiments. In green are displayed the reference electrodes.

the differences among them. On the contrary, in the discriminative approach, a single model is trained using the data from all the classes.

Another important result of Table 3.1 is the lack of advantage in using the dynamics in the generative approach, since HMMs and their static counterparts GMMs give almost the same performance. On the contrary, in the discriminative approach some improvement when using the dynamics is present, especially for Subject B.

3.7 Apposite Continual versus Endpoint Classification

In Section 3.3 we presented a new training and testing method for the classification of sequences using IOHMMs. The objective function was modified so that training focuses on the improvement of classification performance. This approach was based on the fact that features from raw data were extracted so that each input v_t of the IOHMM conveys strong information about the class. In Section 3.4 we have discussed the alternative in which a class label is given only at the end of the sequence. Whilst this alternative avoid the training problem with continual classification, giving an output only at the end of the sequence may introduce long-term dependency problems.

In order to test which approach has to be preferred we have compared the apposite continual classification algorithm against the alternative endpoint classification algorithm, for the EEG data presented above. The comparison is shown in Table 3.2. We can see that the proposed

	Endpoint IOHMM	Continual Apposite IOHMM
Subject A	34.8%	19.0%
Subject B	36.8%	18.5%

Table 3.2: Error rate the discriminating EEG signals using a standard versus the novel apposite IOHMM training algorithm. The first column gives the performance of the endpoint training algorithm described in Section 3.4 (Eq. (3.8)), while the second column gives the performance of the apposite continual classification algorithm described in Section 3.3 (Eq. (3.7)).

apposite algorithm performs significantly better than the endpoint classification procedure of Section 3.4.

3.8 Conclusion

In this Chapter we have compared the use of discriminative and non-discriminative Markovian models for the classification of three mental tasks. The experimental results suggest that the use of a discriminative approach for classification improves the performance over the non-discriminative approach.

However, the form of generative model used in this Chapter does not encode any strong beliefs about the way the data is generated. In this sense, using a generative model as a 'black box' procedure does not exploit well the potential advantages of the approach. From the experimental results here, it is clear that much stronger and more realistic constraints on the form of the generative model need to be made, and this is a relatively open area. This will be one of the issues addressed in the next and subsequent Chapters. We will see that, by using some prior information about how the EEG signal has been generated, a non-discriminative generative approach can perform as well as or even outperform a discriminative one.

The main technical contribution of this Chapter is a new training algorithm for the IOHMM that encourages model resources to be spent on discriminating between sequences in which the same class labels is specified for all the time-steps. Furthermore, the apparently difficult problem of computing the gradient of the new objective function was transformed into subproblems which require computing the same kind of derivative as in the M-step of the EM algorithm. The new apposite training algorithm significantly improves the performance relative to the standard endpoint approach previously presented in the literature.

Chapter 4

Generative ICA for EEG Classification

The work presented in this Chapter has been published in Chiappa and Barber (2005a, 2006).

4.1 Introduction

This Chapter investigates the incorporation of prior beliefs about the EEG signal into the structure of a generative model which is used for direct classification of EEG time-series. In particular, we will look at a form of Independent Components Analysis (ICA) [Hyvärinen et al. (2001)]. On a very high level, a common assumption is that EEG signals can be seen as resulting from the activity of independent components in the brain, and from external noise. This can also be motivated from a physical viewpoint in which the electromagnetic sources within the brain undergo, to a good approximation, linear and instantaneous mixing to form the scalp recorded EEG potentials [Grave de Peralta Menendez et al. (2005)]. For these reasons ICA seems an appropriate model of EEG signals and has been extensively applied to related tasks. One important application of ICA to EEG (and MEG) is addressed at the identification of artifacts [Jung et al. (1998); Vigário (1997); Vigário et al. (1998a)]. Another classical use of ICA is for the analysis of underlying brain sources. For example, ICA was able to separate somatosensory and auditory brain responses in vibrotactile stimulation [Vigário et al. (1998b)], and to isolate different components of auditory evoked potentials [Vigário et al. (1999)]. In Makeig et al. (2002), ICA was used to test between two different hypotheses of the genesis of EEG evoked by visual stimuli. More specifically related to BCI research, several studies have addressed the issue of whether an ICA decomposition can enhance differences in the mental tasks such as to improve the performance of brain-actuated systems. In Makeig et al. (2000), the authors analyze a visual attention task and show that ICA finds μ-components which show a spectral reactivity to motor events stronger than the one measured from scalp channels. They suggest that ICA can be used for optimizing brain-actuated control. In Delorme and Makeig (2003), ICA is used for analyzing EEG data recorded from subjects which attempt to regulate power at 12 Hz over the left-right central scalp. For classification of EEG signals, ICA is often used on the filtered data as a denoising technique or as a feature extractor for improving the performance of a separate classifier. For example, in Hoya et al. (2003), ICA is used to remove ocular artifacts, while Hung et al. (2005) used ICA to extract task-related independent components. In all these cases, ICA is applied as a preprocessing step in order to extract cleaner or more informative features. The temporal features of the spatially preprocessed data are then used as inputs to a standard classifier.

In contrast to these approaches, Penny et al. (2000) introduce a combination of Hidden Markov Models and ICA as a generative model of the EEG data and give a demonstration of how this model can be applied directly to the detection of when switching occurs between the two mental conditions of baseline activity and imaginary movement.

We are interested in a similar generative approach in which independence is built into the structure of the model. However, we want to begin with a simplified model with no temporal dependence between the hidden components, since we are interested in investigating whether a static generative ICA method for direct classification improves on a more standard approach in which an ICA decomposition is applied as a preprocessing step and a separate method is used for classification. The idea is that this unified approach may be potentially advantageous over the standard approach, since the independent components are identified along with a model of the data. The more complex temporal extension will be considered in the next Chapter.

We will consider two different datasets, which involve classifying EEG signals generated by word or movement tasks, as detailed in Section 4.3. Our approach will be to fit, for each person, a generative ICA model to each separate task, and then use Bayes rule to form a classifier. The training criterion will be to maximize the class conditional likelihood. This approach will be compared with the more standard technique of using a Support Vector Machine (SVM) [Cristianini and Taylor (2000)] trained with power spectral density features. We will compare two temporal feature types, one computed from filtered data and the other computed from filtered data preprocessed by ICA using the FastICA package [Hyvärinen (1999)].

The goal is to investigate the potential advantage of using an ICA transformation for improving BCI performance. In addition we investigate if the use of a more principled model in which the independence is directly incorporated into the model structure is advantageous with respect to the more standard approach in which ICA is used as a preprocessing prior to classification. The comparison will be performed under several training and testing conditions, in order to take into account variations in the EEG signal during different days.

Finally, we investigate if a mixture of the model proposed can help to solve the issue of EEG variations during different recording sessions and during different days due to inconsistency of strategy in performing the mental tasks and physiological and psychological changes.

4.2 Generative Independent Component Analysis (gICA)

Under the linear ICA assumption, signals v_t^j recorded at time $t = 1, \ldots, T$ at scalp electrodes $j = 1, \ldots, V$ are formed from a linear and instantaneous superposition of electromagnetic activity h_t^i in the cortex, generated by independent components $i = 1, \ldots, H$, that is:

$$v_t = Wh_t + \eta_t \,.$$

Here the mixing matrix W mimics the mixing and attenuation of the source signals. The term η_t potentially models additive measurement noise. For reasons of computational tractability[1], we consider here only the limit of zero noise. The empirical observations v_t are made zero-mean by a preprocessing step, which obviates the need for a constant output bias, and allows us to assume that h_t also has zero mean. Hence we can define $p(v_t|h_t) = \delta(v_t - Wh_t)$, where $\delta(\cdot)$ is the Dirac delta function. It is also convenient to consider square W, so that $V = H$. Our aim is to fit a model of the above form to each class of task c. In order to do this, we will describe each class specific model as a joint probability distribution, and use maximum likelihood as the training criterion. Whilst this is a hidden variable model ($h_{1:T_c}$ are hidden), thanks to the δ function, we can easily integrate out the hidden variables to form the likelihood of the visible variable $p(v_{1:T_c})$ directly [MacKay (1999)]. Given the above assumptions, the density of the observed and hidden variables for data from class c is[2]:

$$p(v_{1:T}, h_{1:T}|c) = \prod_{t=1}^{T} p(v_t|h_t, c) \prod_{i=1}^{H} p(h_t^i|c) = \prod_{t=1}^{T} \delta(v_t - W_c h_t) \prod_{i=1}^{H} p(h_t^i|c) \,. \tag{4.1}$$

Here $p(h_t^i|c)$ is the prior distribution of the activity of source i, and is assumed to be stationary. By integrating the joint density (5.1) over the hidden variables h_t we obtain:

$$p(v_{1:T}|c) = \prod_{t=1}^{T} \int_{h_t} \delta(v_t - W_c h_t) \prod_{i=1}^{H} p(h_t^i|c) = |\det W_c|^{-T} \prod_{t=1}^{T} \prod_{i=1}^{H} p(h_t^i|c) \,, \tag{4.2}$$

where $h_t = W_c^{-1} v_t$.

[1] Non zero noise may be dealt with at the expense of approximate inference [Hjen-Srensen et al. (2001)].

[2] To simplify the notation we assume that, for each class c, the observed sequence has the same length T.

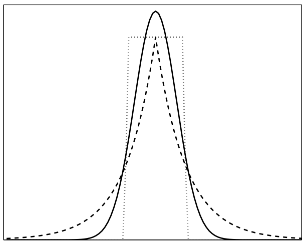


Figure 4.1: Generalized exponential distribution for $\alpha = 2$ (solid line), $\alpha = 1$ (dashed line) and $\alpha = 100$ (dotted line), which correspond to Gaussian, Laplacian and approximately uniform distributions respectively.

There is an important difference between standard applications of ICA and the use of a generative ICA model for classification. In a standard usage of ICA, the sole aim is to estimate the mixing matrix W_c from the data. In that case, it is not necessary to model accurately the source distribution $p(h_t^i|c)$. Indeed, the statistical consistency of estimating W_c can be guaranteed using only two types of fixed prior distributions: one for modelling sub-Gaussian and another for modelling super-Gaussian h_t^i [Cardoso (1998)]. However, the aim of our work is to perform classification, for which an appropriate model for the source distribution of each component h_t^i is fundamental. As in Lee and Lewicki (2000) and Penny et al. (2000), we use the generalized exponential family which encompasses many types of symmetric and unimodal distributions[3]:

$$p(h_t^i|c) = \frac{f(\alpha_{ic})}{\sigma_{ic}} \exp\left(-g(\alpha_{ic}) \left|\frac{h_t^i}{\sigma_{ic}}\right|^{\alpha_{ic}}\right) ,$$

where

$$f(\alpha_{ic}) = \frac{\alpha_{ic}\Gamma(3/\alpha_{ic})^{1/2}}{2\Gamma(1/\alpha_{ic})^{3/2}} , \quad g(\alpha_{ic}) = \left(\frac{\Gamma(3/\alpha_{ic})}{\Gamma(1/\alpha_{ic})}\right)^{\alpha_{ic}/2} ,$$

and $\Gamma(\cdot)$ is the Gamma function. Although unimodality appears quite a restrictive assumption, our experience on the tasks we consider is that it is not inconsistent with the nature of the underlying sources, as revealed by a histogram analysis of $h_t = W_c^{-1} v_t$. The parameter σ_{ic} is the standard deviation[4], while α_{ic} determines the sharpness of the distribution as shown in Fig. 4.1. In the unconstrained case, where a separate model is fitted to data from each class independently, we aim to maximize the class-conditional log-likelihood

$$\mathcal{L}(c) = \log p(v_{1:T}|c) .$$

[3] Importantly, this is able to model both super and sub Gaussian distributions, which are required to isolate the independent components.

[4] Due to the indeterminacy of the variance of h_t^i (h_t^i can be multiplied by a scaling term a as long as the i^{th} column of W_c is multiplied by $1/a$), σ_{ic} could be set to one in the general model described above. However this cannot be done in the constrained version $W_c \equiv W$ considered in the experiments.

In the case where parameters are tied across the different models, for example if the mixing matrix is kept constant over the different models ($W_c \equiv W$), the objective becomes instead $\sum_c \mathcal{L}(c)$. By setting to zero the derivatives of $\mathcal{L}(c)$ with respect to σ_{ic}, we obtain the following closed-form solution:

$$\sigma_{ic} = \left(\frac{g(\alpha_{ic})\alpha_{ic}}{T} \sum_{t=1}^{T} |h_t^i|^{\alpha_{ic}} \right)^{1/\alpha_{ic}} .$$

After substituting this optimal value of σ_{ic} into $\mathcal{L}(c)$, the derivatives with respect to the parameters α_{ic} and W_c^{-1} are used in the scaled conjugate gradient method described in Bishop (1995). These are:

$$\frac{\partial \mathcal{L}(c)}{\partial \alpha_{ic}} = \frac{T}{\alpha_{ic}} + \frac{T}{\alpha_{ic}^2} \frac{\Gamma'(1/\alpha_{ic})}{\Gamma(1/\alpha_{ic})} + \frac{T}{\alpha_{ic}^2} \log \left(\frac{\alpha_{ic} \sum_{t=1}^{T} |h_t^i|^{\alpha_{ic}}}{T} \right) - \frac{T \sum_{t=1}^{T} |h_t^i|^{\alpha_{ic}} \log |h_t^i|}{\alpha_{ic} \sum_{t=1}^{T} |h_t^i|^{\alpha_{ic}}}$$

$$\frac{\partial \mathcal{L}(c)}{\partial W_c^{-1}} = T \left(W_c^{\mathsf{T}} - \sum_{t=1}^{T} b_t v_t^{\mathsf{T}} \right) , \quad \text{with} \quad b_t^i = \frac{\text{sign}(h_t^i)|h_t^i|^{\alpha_{ic}-1}}{\sum_{t=1}^{T} |h_t^i|^{\alpha_{ic}}} ,$$

where the prime symbol $'$ indicates differentiation and the symbol T indicates the transpose operator. After training, a novel test sequence $v_{1:T}^*$ is classified using Bayes rule $p(c|v_{1:T}^*) \propto p(v_{1:T}^*|c)$, assuming $p(c)$ is uniform.

4.3 gICA versus SVM and ICA-SVM

4.3.1 Dataset I

This dataset concerns classification of the following three mental tasks:

1. Imagination of self-paced left hand movements,

2. Imagination of self-paced right hand movements,

3. Mental generation of words starting with a letter chosen spontaneously by the subject at the beginning of the task.

EEG potentials were recorded with the Biosemi ActiveTwo system [http://www.biosemi.com], using the following electrodes located at standard positions of the 10-20 International System [Jasper (1958)]: FP1, FP2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz and O2 (see Fig. 4.2). The raw potentials were re-referenced to the common average reference in which the overall mean is removed from each channel. The signals were recorded at a sample rate of 512 Hz. Subsequently, the band 6-26 Hz was selected with a 2nd order Butterworth filter [Proakis

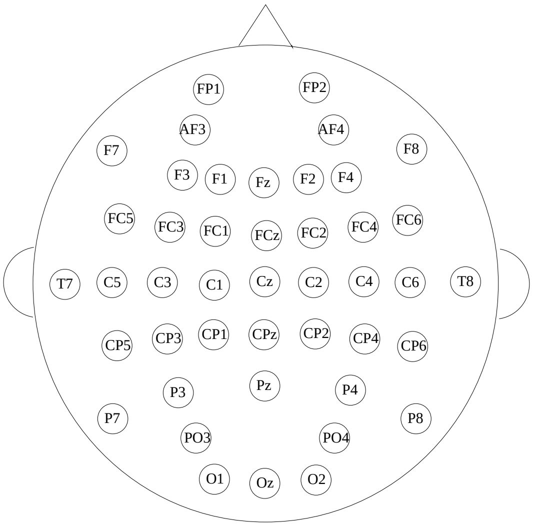


Figure 4.2: Electrode placement. Dataset I has been recorded using electrodes FP1, FP2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz and O2. Dataset II has been recorded using electrodes F3, F1, Fz, F2, F4, FC5, FC3, FC1, FCz FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6, O1 and O2.

and Manolakis (1996)]. This preprocessing filter allow us to focus on μ and β rhythms. Out of the 32 electrodes, only the following 17 electrodes were considered for the analysis: F3, Fz, F4, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P3, PZ, P4 (see Fig. 4.2). This electrode selection was done on the basis of prior knowledge and a preliminary performance analysis. The data was acquired in an unshielded room from two healthy subjects without any previous experience with BCI systems. During an initial day the subjects familiarized themselves with the system, aiming to produce consistent mental states for each task. This data was not used for the training or analysis of the system. In the following two days several sessions were recorded for analysis, each lasting around 4 minutes followed by an interval of around 5 to 10 minutes. During each recording session, around every 20 seconds an operator verbally instructed the subject to continually perform one of the three mental tasks described above.

In a practical scenario, it is envisaged that a user will have an initial intense training period after which, ideally, very little retraining or re-calibration of the system should be required. The performance of BCI systems needs to be robust to potential changes in the manner that the user performs a mental task from session to session, and indeed from day to day. Methods which are highly sensitive to such variations are unsuitable for a practical BCI system. We therefore performed two sets of experiments. In the first case, training, validation and testing were performed on data recorded within the same day, but using separate sessions. The detailed

	Day 1					Day 2								
	Subjects A B C					Subjects A B				Subject C				
Training	1-2-3		4-5			1-2		3-4		1-2-3		4-5		
Validation	4	5	2-3	1-2	1-3	3	4	1	2	4	5	2-3	1-2	1-3
Testing	5	4	1	3	2	4	3	2	1	5	4	1	3	2

Table 4.1: Dataset I covers two days of data: 5 recording sessions on Day 1 for all subjects; for Day 2, Subjects A and B have 4 sessions and Subject C 5 sessions. The table describes how we split these sessions into training, validation and test sessions for the within-the-same-day experiments.

train, validation and test setting is given in Table 4.1. In the second set of experiments, we used the first day to train and validate the models, with test performance being evaluated on the second day alone and vice-versa. In particular, the first three sessions of one day were used for training and the last session(s) for validation. Classification of the three mental tasks was performed using a window of one second of signal. That is, from each session we extracted around 210 samples of 512 time-steps, obtaining the following number of test examples: 1055, 1036 and 1040 for Day 1; 850, 836 and 1040 for Day 2 (Subjects A, B and C respectively).

The non-temporal gICA model described in Section 4.2 was compared with two temporal feature approaches: SVM and ICA-SVM. The purpose of these experiments is to consider whether or not using gICA can provide state-of-the-art performance compared to more standard methods based on using temporal features. Also of interest is whether or not standard ICA preprocessing would improve the performance of temporal feature classifiers.

gICA For gICA, no temporal features need to be extracted and the signal $v_{1:T}$ (downsampled to 64 samples per second) is used, as described in Section 4.2. Since we assume that the scalp signal is generated by a linear mixing of sources in the cortex, provided the data is acquired under the same conditions, it would seem reasonable to further assume that the mixing is the same for all classes ($W_c \equiv W$), and this constrained version was therefore also considered. The number of iterations for training the gICA parameters was determined using a validation set[5].

SVM For the SVM method, we first need to find the temporal features which will subsequently be used as input to the classifier. Several power spectral density representations were

[5] The maximization of the log-likelihood (4.2) is a non-convex problem, thus the choice of the initial parameters may be important. We analyzed two cases in which the W_c matrix was initialized to the identity or to the matrix found by FastICA [Hyvärinen (1999)] using the hyperbolic tangent (randomly initialized), while the exponents of the generalized exponential distribution α were set to 1.5. In both cases we obtained similar performance. We therefore decided to initialize W_c to the identity matrix in all subsequent experiments.

considered. The best performance was obtained using Welch's periodogram method in which each pattern was divided into half-second length windows with an overlap of 1/4 of second, from which the average of the power spectral density (PSD) over all windows was computed. This gave a total of 186 feature values (11 for each electrode) as input for the classifier. Each class was trained against the others, and the kernel width (from 50 to 20000) and the parameter C (from 10 to 200) were found using the validation set.

ICA-SVM The data is first transformed by using the FastICA algorithm [Hyvärinen (1999)] with the hyperbolic tangent nonlinearity and an initial W matrix equal to the identity, then processed as in the SVM approach above.

Results

A comparison of the performance of the spatial gICA against the more traditional methods using temporal features is shown in Table 4.2. The setup of exactly how each training and test sessions were used is given in Table 4.1. Together with the mean, we give the standard deviation of the error on the test sessions, which indicates the variability of performance obtained in different sessions. For gICA, using a different mixing matrix W_c for each mental task generally improves performance. Thus, in the following, we consider only gICA W_c for the comparison with the other standard approaches.

Subject A For this subject, for which the best overall results are found, all three models give substantially the same performance, without loss when training and testing on different days.

Subject B When training and testing on the same day, gICA W_c and ICA-SVM perform similarly, and better than the SVM. However, when training on Day 2 and testing on Day 1, the performance of all models degenerates but more heavily for gICA W_c. ICA-SVM still gives some advantage over SVM. This situation is reversed when training on Day 1 and testing on Day 2.

Subject C For this subject the general performance of the methods is poor. Bearing this in mind, the SVM performs slightly better on average than gICA W_c and ICA-SVM when training and testing on the same day, whereas the two ICA models perform similarly. For training and testing on different days, on average, gICA slightly outperforms the ICA-SVM method, with the best results being given by the plain SVM method. A possible reason for this is that, in this subject, finding reliably the independent components is a challenging

Subject A	gICA W_c	gICA W	SVM	ICA-SVM
Train Day 1, Test Day 1	33.8±6.5%	34.7±5.8%	35.8±5.2%	34.7±5.5%
Train Day 2, Test Day 1	34.2±5.3%	36.1±5.0%	33.3±5.1%	32.8±5.6%
Train Day 2, Test Day 2	24.7±7.5%	26.8±7.1%	24.5±5.9%	25.1±6.3%
Train Day 1, Test Day 2	23.6±4.7%	24.6±5.0%	22.7±4.5%	24.0±2.4%

Subject B	gICA W_c	gICA W	SVM	ICA-SVM
Train Day 1, Test Day 1	31.4±7.1%	34.9±7.4%	38.4±5.2%	32.9±6.1%
Train Day 2, Test Day 1	45.6±5.1%	49.1±3.7%	42.1±4.7%	36.6±7.2%
Train Day 2, Test Day 2	32.5±4.4%	35.1±5.1%	36.7±3.0%	28.9±2.3%
Train Day 1, Test Day 2	31.4±2.3%	35.7±3.3%	39.3±4.3%	40.5±1.6%

Subject C	gICA W_c	gICA W	SVM	ICA-SVM
Train Day 1, Test Day 1	50.5±2.8%	49.4±4.2%	45.5±3.1%	49.0±3.4%
Train Day 2, Test Day 1	52.7±3.6%	55.7±3.3%	48.1±4.7%	52.5±3.8%
Train Day 2, Test Day 2	43.1±2.6%	45.0±4.2%	44.3±4.4%	44.8±3.5%
Train Day 1, Test Day 2	50.2±2.5%	55.3±4.2%	48.7±3.5%	54.9±2.9%

Table 4.2: Mean and standard deviation of the test errors in classifying three mental tasks using gICA with a separate W_c for each class (gICA W_c), gICA with a matrix W common to all classes (gICA W), SVM trained on PSD features (SVM) and SVM trained on PSD features computed from FastICA transformed data (ICA-SVM). Random guessing corresponds to an average error of 66.7%.

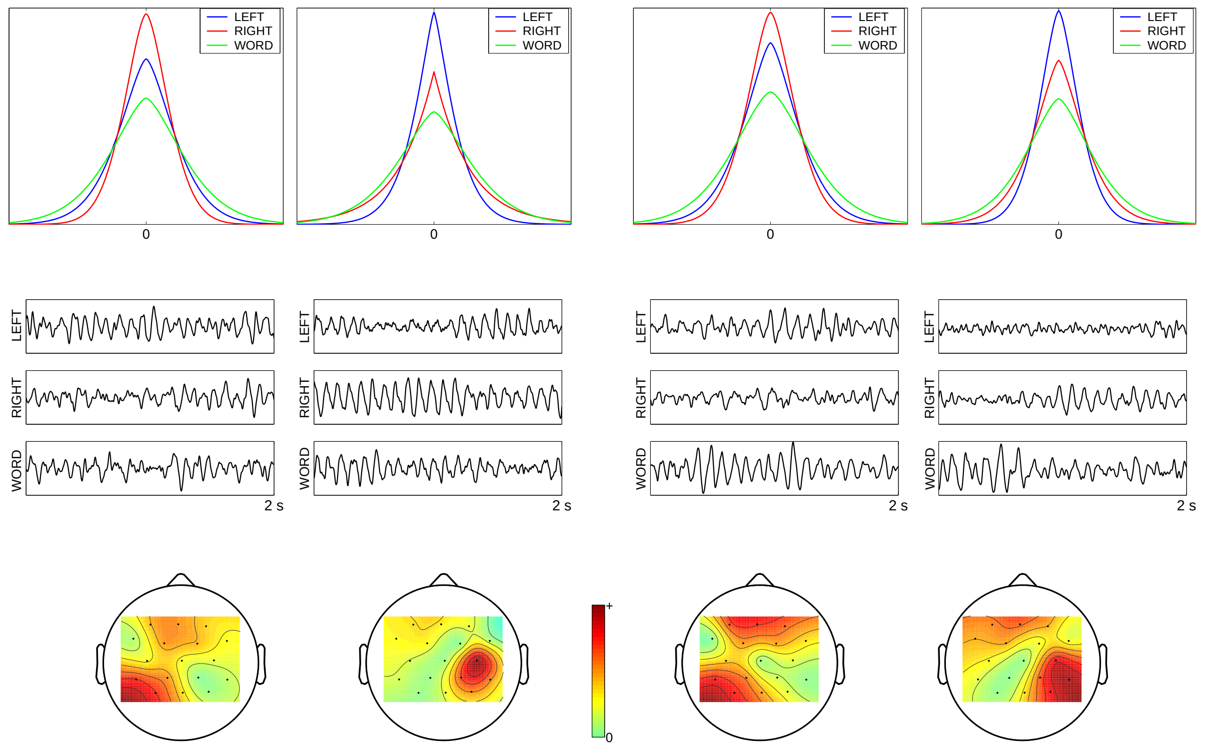
task with convergence difficulties often expressed by FastICA, and the performance of the classifier may be hindered by this numerical instability.

In summary:

1. Training and testing on different days may significantly degrade performance. This indicates that some subjects may be either fundamentally inconsistent in their mental strategies, or the recording situation is not consistent. This more realistic scenario is to be compared with relatively optimistic results from more standard same-day training and testing benchmarks [BCI Competition I (2001); BCI Competition II (2003); BCI Competition III (2004)].

2. ICA preprocessing generally improves classification performance. However, in poorly performing subjects, the convergence of FastICA was problematic, indicating that the ICA components were not reliably estimated, and thereby degrading performance.

3. gICA and ICA-SVM have similar overall performance. This indeed suggests that, for this dataset, state-of-the-art performance can be achieved using gICA, compared with temporal feature based approaches.



Comp. a_1 Comp. a_2 Comp. b_1 Comp. b_2

Figure 4.3: Estimated source distributions and scalp projection of two hidden components for Subject A (Comp. a1, Comp. a2) and Subject B (Comp. b1, Comp. b2). The larger the *width* of the hidden distribution the more that component contributes to the scalp activity. Plotted beneath are two seconds of the same two hidden components, selected at random from the test data, and plotted for each of the three class models. The topographic plots have been obtained by interpolating the values at the electrodes (black dots) using the eeglab toolbox [http://www.sccn.ucsd.edu/eeglab]. Due to the indeterminacy of the hidden component variance, y-axis scale has been removed.

Visualizing the Independent Components

Whilst black-box classification methods such as the SVM give reasonable results, one of the potential advantages of the gICA method is that the parameters and hidden variables of the model are interpretable. Indeed, the absolute value of the 17 elements of the i^{th} column of W indicates the amount of contribution to EEG activity in the 17 electrodes coming from the i^{th} component. Our interest here is to see if the contribution to activity found by the gICA method which is most relevant for discrimination indeed corresponds to known neurological information about different cortical activity under separate mental tasks[6]. To explore this we used the gICA model with a matrix W common to all classes in order to have a correspondence between independent components of different classes. We then selected the column w^i of W whose corresponding hidden component distribution $p(h^i_t|c)$ showed large variation with the class c. Values of w^i which are close to zero indicate a low contribution to the activity from component i, whilst values of w^i away (either positive or negative) indicate stronger contributions. The distributions $p(h^i_t|c)$ and scalp projections $|w^i|$ are shown in Fig. 4.3 for two components. Visually, the projections of components a_1 and b_1 are most similar. For these two components, the word task has the strongest activation (width of the distribution), followed by the left task and the right task. This suggests that for these two subjects a similar spatial contribution to scalp activity from this component occurs when they are asked to perform the tasks. To a lesser extent, visually components a_2 and b_2 are similar in their scalp projection, and again the order of class activation in the two components is the same (word task followed by right and left tasks). Examining both the spatial and temporal nature of the components, a_1 and a_2 seem thus to represent a rhythmic contribution to activity which is more strongly present in the part of the cortex not involved in generating a motor output, that is (roughly speaking), the left hemisphere when the subject imagines to move his left hand and the right hemisphere when the subject imagines to move his right hand. When the subject concentrates on the word task, this rhythmic activity seems to be stronger than for the left and right tasks in both hemispheres.

4.3.2 Dataset II

The second dataset analyzed in this work was provided for the BCI competition 2003 [BCI Competition II (2003); Blankertz et al. (2002, 2004)]. The user had to perform one of two tasks: depressing a keyboard key with a left or right finger. This dataset differs from the previous one in that here the movements are real and not imagined, the assumption being that similar brain activity occurs when the corresponding movement is imagined only.

EEG was recorded from one healthy subject during 3 sessions lasting 6 minutes each. Sessions were recorded during the same day at intervals of some minutes. The key depression occurred in a self-chosen order and timing. For the competition, 416 epochs of 500 ms EEG were provided, each ending 130 ms before an actual key press, at a sampling rate of 1000 and 100 Hz. The epochs were randomly shuffled and split into a training-validation set and a test set consisting of 316 and 100 epochs respectively. EEG was recorded from 28 electrodes: F3, F1, Fz, F2, F4, FC5, FC3, FC1, FCz FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2,

[6] Note that actual cortical activity is generated by all 17 components. Therefore the actual cortical activity for each mental task is not considered here, but rather that contribution which appears to vary most with respect to the different tasks.

CP4, CP6, O1 and O2 (see Fig. 4.2).

In this dataset, in addition to μ and β rhythms, another important EEG feature related to movement planning, called the Bereitschaftspotential (BP), can be considered[7]. BP is a slowly decreasing cortical potential which develops 1-1.5 seconds prior to a movement. The BP shows larger amplitude contralateral to the moving finger. The difference in the spatial distribution of BP is thus an important indicator of left or right finger movement. In order to include such a feature in the ICA or gICA approach, it is likely that a non-symmetric prior (or a non symmetric FastICA approach) would need to be considered. We apply only the symmetric gICA (and FastICA) models to a preprocessed form of this dataset in which we filter to consider only μ-β bands, thereby removing any large scale shape effects such as the BP[8]. For the other methods not solely based on ICA, we retained possible BP features for a point of comparison to see if the use of BP features indeed is critical for reasonable performance on this database. The following methods were considered:

μ-β-gICA The μ-β filtered data is used as input to the generative ICA model described in Section 4.2.

BP-SVM This method focuses on the use of the BP as the features for a classifier. Here we preprocessed raw data in the 'BP band' (350 dimensional feature vector, 25 for each of the 14 electrodes). A Gaussian kernel was used and its width learned (in the range 10-5000), together with the strength of the margin constraint C (in the range 10-200), on the basis of the validation set.

μ-β-SVM This method focuses on the μ-β band, which precludes therefore any use of a BP for classification. The data was first filtered in the μ-β band as described above. Then the power spectral density was computed (168 dimensional feature vector).

BP-μ-β-SVM Here the combination of BP features and μ-β spectral features were used as input to an SVM classifier.

[7] It was not possible to consider this feature in the previous dataset recorded using an synchronous protocol.

[8] We analyzed 100 Hz sampled data. The raw potentials were re-referenced to the common average reference. Then, the following 14 electrodes were selected: C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4 and CP6. For analyzing μ and β rhythms, each epoch was zero-mean and filtered in the band 10-32 Hz with a 2nd order Butterworth (zero-phase forward and reverse) digital filter. For BP, each epoch was low-pass filtered at 7 Hz using the same filtering setting, then the first 25 frames of each epoch were disregarded. This preprocessing was based on a preliminary analysis taking into consideration the best performance obtained in the BCI competition 2003 on this dataset [Wang et al. (2004)].

μ-β-gICA W	μ-β-gICA W_c	BP-SVM	μ-β-SVM
16.0±1.2%	17.0±2.3%	21.6±1.5%	25.4±3.1%

BP-μ-β-SVM	μ-β-ICA-SVM	BP-μ-β-ICA-SVM
18.8±0.8%	22.2±2.3%	16.2±0.8%

Table 4.3: Mean and standard deviation of the the test errors in classifying two finger movement tasks. Random guessing corresponds to an error of 50%.

μ-β-ICA-SVM Here the μ-β filtered data is further preprocessed using FastICA to form features to the SVM classifier.

BP-μ-β-ICA-SVM Here the combination of BP features with μ-β-ICA features forms the input to the SVM classifier.

Results

The comparison between these models is given in Table 4.3, in which we present the mean test error and standard deviation obtained by using 5-fold cross-validation[9]. Given the low number of test samples, it is difficult to present decisive conclusions. However, by comparing μ-β-SVM and μ-β-ICA-SVM, we note that using an ICA decomposition on μ-β filtered data improves performance. For this dataset, gICA-type models obtain superior performance to methods in which ICA is used as preprocessing. Finally, and perhaps most interestingly, the performance of gICA on μ-β is comparable with the results obtained by *combining* μ-β and BP features (BP-μ-β-ICA-SVM). The results from the gICA method are comparable to the best results previously reported for this dataset[10].

[9] For each of the methods, we split the training data into 5 sets and performed cross-validation for hyperparameters by training on 4 sets and validating on the fifth. The resulting model was then evaluated on the separate test set. This procedure was repeated for the other four combinations of choosing 4 training and 1 validation set from the 5 sets. The mean and standard deviation of the 5 resulting models (for each method) are then presented.

[10] The winner of the BCI competition 2003 applied a spatial subspace decomposition filter and Fisher discriminant analysis to extract three types of features derived from BP and μ-β rhythms, and used a linear perceptron for classification. The final accuracy on the test was 16.0% [Wang et al. (2004)].

4.4 Mixture of Generative ICA

Although the performance of gICA is reasonable, if used in any BCI system, it would still achieve far from perfect performance. Whilst the reason for this may simply be inherently noisy data, another possibility is that the subject's reaction when asked to think about a particular mental task drifts significantly from one session and/or day to another. It is also natural to assume that a subject has more than one way to think about a particular mental task. The idea of using a mixture model is to test the hypothesis that the data may be naturally split into regimes, within which a single model may accurately model the data, although this single model is not able to model accurately all the data. This motivates the following model for a single sequence of observations

$$p(v_{1:T}|c) = \sum_{m=1}^{M_c} p(v_{1:T}|m, c)p(m|c)\,,$$

where m describes the mixture component. The number of mixture components M_c will typically be rather small, being less than 5. We will then fit a separate mixture model to data for each class c.

4.4.1 Parameter Learning

To ease the notation a little, from here we drop the class dependency. Analogously to Section 3.3.1, in order to estimate the parameters σ_{im}, α_{im}, W_m and $p(m)$, we can use a generalized EM algorithm, which enables us to perform maximum likelihood in the context of latent or hidden variables, in this case being played by m. In the mixture case we have a set of sequences $v^s_{1:T}$, $s = 1, \dots, S$ each of the same length T. The expected complete data log-likelihood is given by:

$$\mathcal{L} = \left\langle \log \prod_{s=1}^{S} p(v^s_{1:T}|m)p(m) \right\rangle_{p(m|v^s_{1:T})} = \sum_{s=1}^{S} \left\langle \sum_{t=1}^{T} \log |\det W_m^{-1}| p(W_m^{-1} v^s_t) + \log p(m) \right\rangle_{p(m|v^s_{1:T})} \,, \quad (4.3)$$

where S indicates the number of sequences and $\langle\cdot\rangle$ indicates the expectation operator. Here v^s_t is the vector of observations at time t from sequence s. In the E-step, inference is performed in the following way:

$$p(m|v^s_{1:T}) = \frac{p(v^s_{1:T}|m)p(m)}{\sum_{m'=1}^{M} p(v^s_{1:T}|m')p(m')}\,.$$

In the M-step, the prior is updated as:

$$p(m) = \frac{1}{S}\sum_{s=1}^{S} p(m|v^s_{1:T})\,.$$

The maximum-likelihood solution of σ_{im} has the following form:

$$\sigma_{im} = \left(\frac{g(\alpha_{im})\alpha_{im} \sum_{s=1}^{S} p(m|v_{1:T}^{s}) \sum_{t=1}^{T} |h_t^{im}|^{\alpha_{im}}}{\sum_{s=1}^{S} Tp(m|v_{1:T}^{s})} \right)^{1/\alpha_{im}} .$$

The substitution of this solution into $\mathcal{L}$ gives:

$$\begin{aligned} \mathcal{L} = & \sum_{s=1}^{S} T \sum_{m=1}^{M} p(m|v_{1:T}^{s}) \Big(\log|\det W_m^{-1}| + \sum_{i=1}^{H} \log \frac{\alpha_{im}}{2\Gamma(1/\alpha_{im})} - \sum_{i=1}^{H} \frac{1}{\alpha_{im}} \log \alpha_{im} \\ & - \sum_{i=1}^{H} \frac{1}{\alpha_{im}} \log \frac{\sum_{s=1}^{S} p(m|v_{1:T}^{s}) \sum_{t=1}^{T} |h_t^{im}|^{\alpha_{im}}}{\sum_{s=1}^{S} Tp(m|v_{1:T})} - \sum_{i=1}^{H} \frac{1}{\alpha_{im}} \Big) + \sum_{s=1}^{S} \sum_{m=1}^{M} p(m|v_{1:T}^{s}) \log p(m) . \end{aligned}$$

The other parameters are updated using a scaled conjugate gradient methods. The derivatives of $\mathcal{L}$ with respect to α_{im} and W_m^{-1} are given by:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \alpha_{im}} = & \Big(\frac{1}{\alpha_{im}} + \frac{1}{\alpha_{im}^2} \frac{\Gamma'(1/\alpha_{im})}{\Gamma(1/\alpha_{im})} + \frac{1}{\alpha_{im}^2} \log \frac{\alpha_{im} \sum_{s=1}^{S} p(m|v_{1:T}) \sum_{t=1}^{T} |h_t^{im}|^{\alpha_{im}}}{\sum_{s=1}^{S} Tp(m|v_{1:T}^{s})} \\ & - \frac{\sum_{s=1}^{S_c} p(m|v_{1:T}^{s}) \sum_{t=1}^{T} |h_t^{im}|^{\alpha_{im}} \log|h_t^{im}|}{\alpha_{im} \sum_{s=1}^{S} p(m|v_{1:T}^{s}) \sum_{t=1}^{T} |h_t^{icm}|^{\alpha_{im}}} \Big) \sum_{s=1}^{S} Tp(m|v_{1:T}^{s}) \\ \frac{\partial \mathcal{L}}{\partial W_m^{-1}} = & \sum_{s=1}^{S} Tp(m|v_{1:T}^{s}) \left(W_m^{\mathsf{T}} - \sum_{t=1}^{T} b_t (v_t^{s})^{\mathsf{T}} \right) , \end{aligned}$$

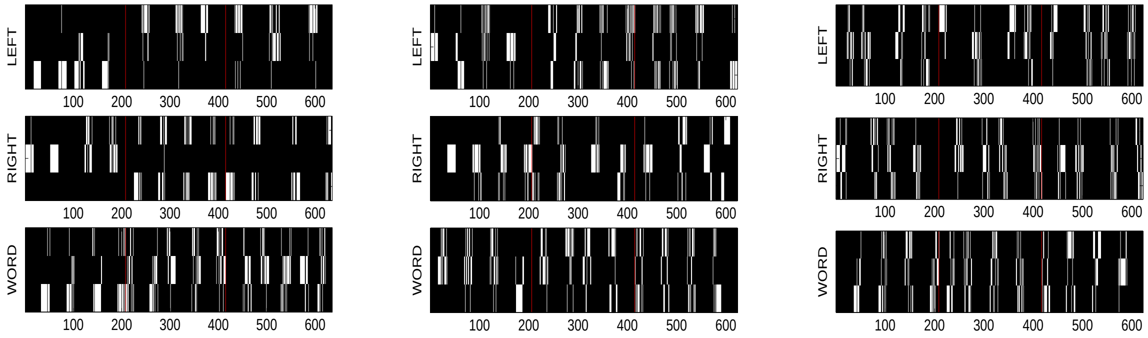
where

$$b_t^i = \frac{\mathrm{sign}(h_t^{im})|h_t^{im}|^{\alpha_{im}}}{\sum_{s=1}^{S} p(m|v_{1:T}) \sum_{t=1}^{T} |h_t^{im}|^{\alpha_{im}}} .$$

4.4.2 gICA versus Mixture of gICA

Dataset I

We first fitted a mixture of three gICA models to the first three sessions of Day 1. The aim here is that this may enable us to visualize how each subject switches between mental strategies, and therefore to form an idea of how reliably each subject is performing. These results are presented in Fig. 4.4, where switching for each subject between the three different mixture components is shown. Interestingly, we see that for Subjects A and B and all three tasks, only a single component tends to be used during the first session, suggesting a high degree of consistency in the way that the mental tasks were realized. For Subject C, a lesser degree of reliability is present. This situation changes so that, in the latter two sessions, a much more rapid switching occurs



Subject A Subject B Subject C

Figure 4.4: We show here results of fitting a separate mixture model with three components to each of the three tasks for the first three sessions of Day 1. Time (in seconds) goes from left to right. At any time, only one of the three classes (corresponding to the verbal instruction to the subject), and only one of the three hidden states for that class (the one with the highest posterior probability), is highlighted in white. The plot shows how the subjects change in their strategy for realizing a particular mental task with time. The vertical lines indicate the boundaries of the training sessions, which correspond to a gap of 5-10 minutes.

(indeed this switching happens much more quickly than the time prescribed for a mental task). This suggests that the consistency with which subjects perform the mental tasks deteriorates with time, highlighting the need to potentially account for such drift in approach.

To see whether or not this results in an improved classification, we trained the mixture of gICA model, as described above, on the dataset. Table 4.4 compares the performance between gICA and the mixtures of gICA models using a separate W_c matrix for each class. The number of mixture components (ranging from 2 to 5) was chosen from the validation set. The W_c was initialized adding a small amount of noise to W_c found using one mixture. Whilst the mixture of ICA model seems to be reasonably well motivated, disappointingly, only a minor improvement with respect to the single mixture case is found on Subjects A and B. It is not clear why the performance improvement is so modest. This may be due to the fact that whilst drift is indeed an issue and better modelled by this approach, the model does not capture the online nature of adaptation that may occur in practice. That is, a stationary mixture model may be inadequate for capturing the dynamic nature of changes in user mental strategies.

Subject A	gICA W_c	MgICA W_c
Train Day 1, Test Day 1	33.8±6.5%	31.1±4.9%
Train Day 2, Test Day 1	34.2±5.3%	33.6±5.0%
Train Day 2, Test Day 2	24.7±7.5%	22.3±6.4%
Train Day 1, Test Day 2	23.6±4.7%	22.4±3.0%

Subject B	gICA W_c	MgICA W_c
Train Day 1, Test Day 1	31.4±7.1%	30.6±3.8%
Train Day 2, Test Day 1	45.6±5.1%	40.0±10.0%
Train Day 2, Test Day 2	32.5±4.4%	29.1±3.0 %
Train Day 1, Test Day 2	31.4±2.3%	29.5±6.0 %

Subject C	gICA W_c	MgICA W_c
Train Day 1, Test Day 1	50.5±2.8%	52.2±4.8%
Train Day 2, Test Day 1	52.7±3.6%	52.2±2.7%
Train Day 2, Test Day 2	43.1±2.6%	44.6±3.2%
Train Day 1, Test Day 2	50.2±2.5%	51.6±1.6%

Table 4.4: Mean and standard deviation of the test errors in classifying three mental tasks using gICA with a separate W_c for each class (gICA W_c) and a mixture of gICA with a separate W_c for each class (MgICA W_c).

Dataset II

The result of using a mixture model with a separate W_c for each class is $19.4 \pm 2.6\%$. Compared with the results presented from the single gICA and other methods in Table 4.3, this result is disappointing, being a little (though not significantly) worse than the single gICA method. Here, the number of mixture components (from 2 to 5) is chosen on the basis of the validation set and this should, in principle, avoid overfitting. However, the validation error for a single component is often a little better than for a number of mixture components greater than 1, suggesting indeed that the model is overfitting slightly.

4.5 Conclusions

In this work we have presented an analysis on the use of a spatial generative Independent Component Analysis (gICA) model for the discrimination of mental tasks for EEG-based BCI systems. We have compared gICA against other standard approaches, where temporal information from a window of data (power spectral density) is extracted and then processed using an SVM classifier. Our results suggest that using gICA alone is powerful enough to produce good performance for the datasets considered. Furthermore, using ICA as a preprocessing step for power spectral density SVM classifiers also tends to improve the performance, giving roughly the same performance as gICA. An important point is that performance generally degrades when one trains a method on one day and tests on another, although for some subjects this is less apparent. This more realistic scenario is a more severe test of BCI methods and, in our view, merits further consideration. For this reason, we investigated whether or not a mixture model, which may cope with potentially severe changes in mental strategy, may improve performance. Indeed, the use of mixture models appears to be well-founded since, based on the training data alone, switching between mixture components tends to increase with time. However the resulting performance improvements for classification were rather modest (or even slightly worse), suggesting that the model is overfitting slightly. Indeed, the model does not deal well with the potentially dynamic nature of change. An online version of training may be a reasonable way to avoid this difficulty, by which some form of continual recalibration based on feedback is provided.

An arguable limitation of the gICA model considered in this Chapter is that the temporal nature of EEG is not taken into account. We will address this issue in the next Chapter, where we model each hidden component with an autoregressive model.

Chapter 5

Generative Temporal ICA for EEG Classification

The work presented in this Chapter has been published in Chiappa and Barber (2005b).

5.1 Introduction

In Chapter 4 we investigated the incorporation of prior beliefs about how the EEG signal is generated into the structure of a generative model. More specifically, we made the assumption that a multichannel EEG signal v_t results from linear mixing of independent sources in the brain and other external components h_t^i, $i = 1, \ldots, H$. The resulting model is a form of generative Independent Component Analysis (gICA) which was used to classify spontaneous EEG. We have seen that this model performs similarly or better than standard 'black-box' classification methods, and similarly to a model in which ICA is used as a preprocessing step before extracting spectral features which are then classified by a separate discriminative model. This is noteworthy, since in the gICA model no temporal features are used and the model is trained on only filtered EEG data. As a consequence, we could randomly shuffle the elements $v_{1:T}$ and obtain the same classification performance. Indeed, each hidden variable h_t^i was considered to be temporally independent and identically distributed, that is $p(h_{1:T}^i) = \prod_{t=1}^{T} p(h_t^i)$. An open question is therefore whether we can improve the performance of gICA by extending this model to take into account temporal information. A motivation for that is the fact that temporal modeling of the hidden components has shown to improve separation in the case of other types of signals, such as speech data [Pearlmutter and Parra (1997)].

In this Chapter we therefore further investigate the use of a generative ICA model for classification addressing the specific issue of whether modeling the temporal dynamics of the hidden

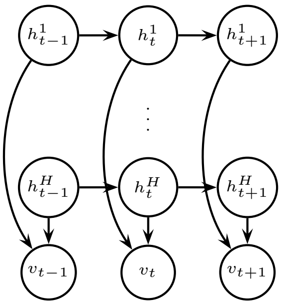


Figure 5.1: Graphical representation of an ICA model with temporal dependence between the hidden variables (order $m = 1$).

variables improves the discriminative performance of the generative ICA model. In particular, we will model each hidden component with an autoregressive process, since this was successfully previously applied and the resulting model is tractable.

As in Chapter 4, our approach will be to fit, for each person, a generative ICA model to each separate task, and then use Bayes rule to form directly a classifier. This model will be compared with its static special case, where no temporal information is taken into account, namely the gICA model of Chapter 4. In addition, we will compare it with two standard techniques in which power spectral density features are extracted from the temporal EEG data and fed into a Multilayer Perceptron (MLP) [Bishop (1995)] and Support Vector Machine (SVM) [Cristianini and Taylor (2000)].

5.2 Generative Temporal ICA (gtICA)

In Section 4.2 we introduced a Generative Independent Component Analysis (gICA) model, in which a vector of observations v_t is assumed to be generated by statistically independent (hidden) random variables h_t via an instantaneous linear transformation:

$$v_t = Wh_t \,,$$

with W assumed to be a square matrix. In this model, each hidden component h^i_t was considered to be temporally independent identically distributed, that is:

$$p(h^i_{1:T}) = \prod_{t=1}^{T} p(h^i_t).$$

In particular each component was modeled using a generalized exponential distribution. We now consider temporal dependencies between different time-steps. A reasonable temporal model for the hidden variables which has shown to improve separation in the case of other types of signals is the autoregressive process. We therefore model the i^{th} hidden component h^i_t with a linear autoregressive model of order m, defined as:

$$h^i_t = \sum_{k=1}^{m} a^i_k h^i_{t-k} + \eta^i_t = \hat{h}^i_t + \eta^i_t \,,$$

where η^i_t is the noise term. The graphical representation of this model is shown in Fig. 5.1. Analogously to Chapter 4, our aim is to fit a model of the above form to each class of task c using maximum likelihood as the training criterion. Given the above assumptions, we can factorize the density of the observed and hidden variables as follows[1]:

$$p(v_{1:T}, h_{1:T}|c) = \prod_{t=1}^{T} p(v_t|h_t, c) \prod_{i=1}^{H} p(h^i_t|h^i_{t-1:t-m}, c) \,. \tag{5.1}$$

Using $p(v_t|h_t) = \delta(v_t - Wh_t)$, where $\delta(\cdot)$ is the Dirac delta function, we can easily integrate (5.1) over the hidden variables $h_{1:T}$ to form the likelihood of the observed sequence $v_{1:T}$:

$$p(v_{1:T}|c) = |\det W_c|^{-T} \prod_{t=1}^{T} \prod_{i=1}^{H} p(h^i_t|h^i_{t-1:t-m}, c) \,, \tag{5.2}$$

where $h_t = W_c^{-1} v_t$. We model $p(h^i_t|h^i_{t-1:t-m}, c)$ with the generalized exponential distribution, that is:

$$p(h^i_t|h^i_{t-1:t-m}, c) = \frac{f(\alpha_{ic})}{\sigma_{ic}} \exp\left(-g(\alpha_{ic}) \left|\frac{h^i_t - \hat{h}^i_t}{\sigma_{ic}}\right|^{\alpha_{ic}}\right) \,,$$

where

$$f(\alpha_{ic}) = \frac{\alpha_{ic}\Gamma(3/\alpha_{ic})^{1/2}}{2\Gamma(1/\alpha_{ic})^{3/2}} \,, \quad g(\alpha_{ic}) = \left(\frac{\Gamma(3/\alpha_{ic})}{\Gamma(1/\alpha_{ic})}\right)^{\alpha_{ic}/2} \,,$$

and $\Gamma(\cdot)$ is the Gamma function. As we have seen in Section 4.2, the generalized exponential can model many types of symmetric and unimodal distributions. The logarithm of the likelihood (5.2) is summed over all training sequences belonging to each class and then maximized by using a scaled conjugate gradient method [Bishop (1995)]. This requires computing the derivatives with respect to all the parameters, that is, the mixing matrix W_c, the autoregressive coefficients a^i_k, and the parameters of the exponential distribution σ_{ic} and α_{ic}. After training, a novel test

[1] This is a slight notation abuse for reasons of simplicity. The model is only defined for $t > m$. This is true for all subsequent dependent formulae.

sequence $v^*_{1:T}$ is classified using Bayes rule $p(c|v^*_{1:T}) \propto p(v^*_{1:T}|c)$, assuming $p(c)$ is uniform.

5.2.1 Learning the Parameters

The normalized log-likelihood of a set of sequences of class c is given by

$$\mathcal{L}(c) = \frac{1}{S_c(T-m)} \sum_{s=1}^{S_c} \log p(v^s_{m+1:T}|h^s_{1:m}, c)\,,$$

where s indicates the s^{th} training pattern of class c. We write $p(v^s_{m+1:T}|h^s_{1:m}, c)$, rather than the notational abuse $p(v^s_{1:T}|c)$ in the previous text, since this takes care of the initial time steps which would otherwise be problematic. In the following, $h^s_t = W_c^{-1} v^s_t$, for $t = 1, \ldots, T$. Dropping the pattern index s, the component index i and the class index c we find that the maximum likelihood solution for σ is:

$$\sigma = \left(\frac{g(\alpha)\alpha}{S(T-m)} \sum_{s=1}^{S} \sum_{t=m+1}^{T} |h_t - \hat{h}_t|^{\alpha} \right)^{1/\alpha} .$$

After substituting this value into $\mathcal{L}$, we obtain:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{1}{\alpha} + \frac{1}{\alpha^2} \frac{\Gamma'(1/\alpha)}{\Gamma(1/\alpha)} + \frac{1}{\alpha^2} \log \left(\frac{\alpha \sum_{s=1}^{S} \sum_{t=m+1}^{T} |h_t - \hat{h}_t|^{\alpha}}{S(T-m)} \right) - \frac{\sum_{s=1}^{S} \sum_{t=p+1}^{T} |h_t - \hat{h}_t|^{\alpha} \log |h_t - \hat{h}_t|}{\alpha \sum_{s=1}^{S} \sum_{t=m+1}^{T} |h_t - \hat{h}_t|^{\alpha}}$$

$$\frac{\partial \mathcal{L}}{\partial W^{-1}} = W^{\mathsf{T}} - \sum_{s=1}^{S} \sum_{t=m+1}^{T} \left(b_t v_t^{\mathsf{T}} + \hat{B}_t \right) ,$$

where b_t is a vector of elements

$$b^i_t = \frac{\mathrm{sign}\left(h^i_t - \hat{h}^i_t\right) \left|h^i_t - \hat{h}^i_t\right|^{\alpha_i - 1}}{\sum_{s=1}^{S} \sum_{t=m+1}^{T} |h^i_t - \hat{h}^i_t|^{\alpha}_i},$$

and $\hat{B}_t$ is a matrix of rows

$$\hat{B}^i_t = \frac{\mathrm{sign}\left(h^i_t - \hat{h}^i_t\right) \left|h^i_t - \hat{h}^i_t\right|^{\alpha_i - 1} \sum_{k=1}^{m} a^i_k v^{\mathsf{T}}_{t-k}}{\sum_{s=1}^{S} \sum_{t=m+1}^{T} |h^i_t - \hat{h}^i_t|^{\alpha}_i} .$$

Finally, the derivative with respect to the autoregressive coefficient a_k is given by:

$$\frac{\partial \mathcal{L}}{\partial a_k} = \frac{\sum_{s=1}^{S} \sum_{t=m+1}^{T} \mathrm{sign}(h_t - \hat{h}_t) |h_t - \hat{h}_t|^{\alpha - 1} h_{t-k}}{\sum_{s=1}^{S} \sum_{t=m+1}^{T} |h_t - \hat{h}_t|^{\alpha}} .$$

5.3 gtICA versus gICA, MLP and SVM

EEG potentials were recorded with the Biosemi ActiveTwo system (http://www.biosemi.com), using 32 electrodes located at standard positions of the 10-20 International System [Jasper (1958)], at a sample rate of 512 Hz. The raw potentials were re-referenced to the Common Average Reference in which the overall mean is removed from each channel. Subsequently, the band 6-16 Hz was selected with a 2nd order Butterworth filter [Proakis and Manolakis (1996)]. Only the following 19 electrodes were considered for the analysis: F3, FC1, FC5, T7, C3, CP1, CP5, P3, Pz, P4, CP6, CP2, C4, T8, FC6, FC2, F4, Fz and Cz.

The data were acquired in an unshielded room from two healthy subjects without any previous experience with BCI systems. During an initial day the subjects learned how to perform the mental tasks. In the following two days, 10 recordings, each lasting around 4 minutes, were acquired for the analysis. During each recording session, every 20 seconds an operator instructed the subject to perform one of three different mental tasks. The tasks were:

1. Imagination of self-paced left hand movements,
2. Imagination of self-paced right hand movements,
3. Mental generation of words starting with a letter chosen spontaneously by the subject at the beginning of the task.

The time-series obtained from each recording session was split into segments of signal lasting one second. This was the time length in which classification was performed. The first three sessions of recording of each day were used for training the models while the other two sessions where used alternatively for validation and testing. We obtained around 420 test examples for each day.

The temporal gICA model was compared with its static equivalent gICA and with two standard approaches for EEG classification, in which for each segment the power spectral density was extracted and then processed using an MLP and a SVM.

gtICA In the temporal gICA model, the data $v_{1:T}$ (downsampled to 64 samples per second) was used, without extracting any temporal feature. The validation set was used to choose the number of iterations of the scaled conjugate gradient and the order m of the autoregressive model (from 1 to 8). Since we assume that the scalp signal is generated by linear mixing of sources in the cortex, provided the data are acquired under the same conditions, it would seem reasonable to further assume that the mixing is the same for all classes ($W_c \equiv W$) and this constrained version was also considered. The static gICA model is obtained as a special case of the temporal gICA model in which the autoregressive order m is set to 0.

	Subject A		Subject B	
	Day 1	Day 2	Day 1	Day 2
gICA W	40.0±0.6%	34.8±22.2%	28.5±6.6%	31.5±2.0%
gtICA W	40.2±3.0%	36.7±22.2%	27.8±4.9%	30.8±2.7%
gICA W_c	37.1±0.6%	36.0±24.6%	25.6±2.4%	30.8±3.0%
gtICA W_c	38.8±2.3%	36.2±23.6%	27.1±5.2%	28.2±0.0%
MLP	37.1±2.1%	38.1±21.4%	30.5±4.0%	34.2±2.1%
SVM	35.1±0.9%	38.1±20.3%	32.4±5.5%	36.6±1.7%

Table 5.1: Mean and standard deviation of the test errors in classifying three mental tasks using Generative Static ICA (gICA), Generative Temporal ICA (gtICA), MLP and SVM. W_c uses a separate matrix for each class, as opposed to a common matrix W. Classification is performed on 1 second length data. Random guessing corresponds to an average error of 66.7%. From the standard deviation, we can observe big difference in performance of Subject A, Day 2 on the two testing sessions.

MLP For the MLP we extracted temporal features which were used as input to the classifier. More specifically, we estimated the power spectral density using the following Welch's periodogram method: each pattern of one second length was divided into a quarter of second long windows with an overlap of 1/8 of second. Then the overall average was computed. A softmax, one hidden layer MLP was trained using cross-entropy, with the validation set used to choose the number of iterations, the number of hyperbolic tangent hidden units (ranging from 1 to 100) and the learning rate of the gradient ascent method.

SVM In the SVM, the same features as in the MLP were given as input to the classifier. Each class was trained against the others. A Gaussian SVM was considered, with kernel width (ranging from 1 to 20000) and the parameter C (ranging from 10 to 200) found using the validation set.

Results

A comparison of the performance of the tgICA versus its static equivalent gICA, the MLP and SVM is shown in Table 5.1. Together with the mean, we give the standard deviation of the error on the two test sessions, which indicates the variability of performance obtained in different sessions.

Disappointingly, by modeling the independent components with an autoregressive process we don't obtain an improvement in discrimination with respect to the static case. Indeed the performance of the generative temporal ICA model and its static equivalent is similar. It may be that a simple autoregressive model is not suitable for the EEG data, due to non-stationarity

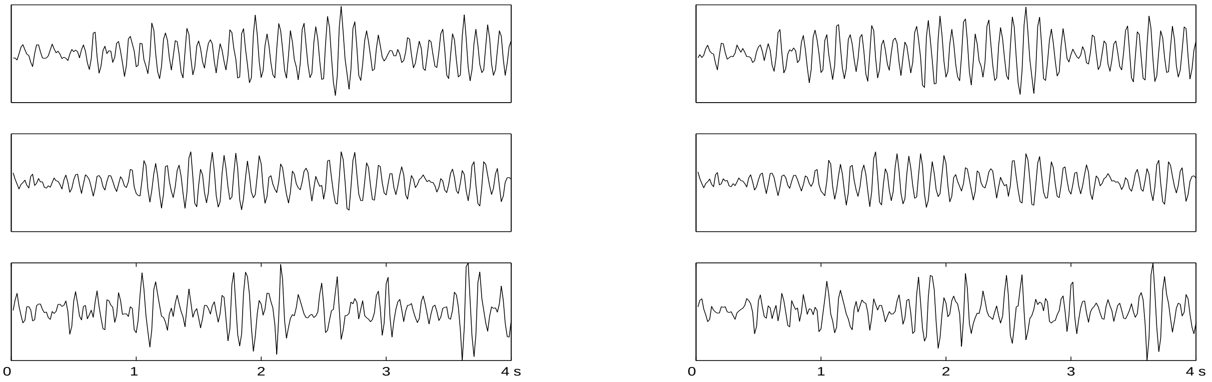


Figure 5.2: 4 seconds of three selected hidden components for Subject A, Day 2 using generative static ICA (left) and generative temporal ICA (right). Due to the indeterminacy of variance of the hidden components, y-axes scale has been removed.

or changes in the hidden dynamics.

On the other hand, the static generative ICA approach, in which a different matrix W_c for each class is computed, performs as well as or better than the temporal feature approach using MLPs and SVMs.

Visualizing Independent Components

We are interested in knowing the difference in the components estimated by using the generative temporal ICA and a generative static ICA methods.

For Subject A, we used the second day's data to select the three hidden components whose distribution varied most across the three classes, using the ICA model with a matrix W common to all classes. In the generative static ICA model, the three components were selected by looking at the distribution $p(h_t^i)$, while in the temporal ICA model they were selected by looking at the conditional distribution $p(h_t^i|h_{t-1:t-m}^i)$ for the order m that gave the best performance in the test set. The time courses (4 seconds of the word task) of the selected hidden components are shown in Fig. 5.2. As we can see, the time courses between the static components (left) and temporal components (right) are very similar. In general we found a high correspondence among almost all the 19 components of the static and temporal ICA model. The components for which a correspondence was not found don't show differences in the autoregressive coefficients and in the conditional distribution, and are thus not relevant for discrimination. Finally note that the hidden components found by the generative temporal ICA don't look smoother as we would expect when modeling the dynamics of the hidden sources.

5.4 Conclusions

In this Chapter we applied a generative temporal Independent Component Analysis model to the discrimination of three mental tasks. In particular, the temporal dynamics of each hidden component was modeled by an autoregressive process. We have compared this model with its static equivalent introduced in Chapter 4, in order to address the issue of whether the use of temporal information can improve the discriminative power of the generative ICA model. Taking into account temporal information was shown to be advantageous for separating other types of signals not well separable using a static ICA method. However, this approach does not seem to bring additional discriminant information when ICA is used as a generative model for direct classification. By analyzing the components extracted by the temporal and static ICA model, we have seen that similar discriminative components are extracted. The reason may be that a simple linear dynamical model is not suitable for our EEG data, due to strong non-stationarity in the hidden dynamics. In this case, it may be more appropriate to use a switching model which can handle changes of regime in the EEG dynamics [Bar-Shalom and Li (1998)].

Chapter 6

EEG Decomposition using Factorial LGSSM

6.1 Introduction

The previous Chapters of this thesis focused on probabilistic methods for classifying EEG data. For the remainder of the thesis, we will concentrate on analyzing the EEG signal and, in particular, on extracting independent dynamical processes from multiple channels. Decomposing a multivariate time-series into a set of independent subsignals is a central goal in signal processing and is of particular interest in the analysis of biomedical signals. Here, accepting the common assumption in EEG-related research and in agreement with the previous Chapters, we focus on a method which assumes that EEG is generated by a *linear instantaneous* mixing of independent components, which include both biological and noise components. In BCI research, such a decomposition method, has several potential applications:

- It can be used to denoise EEG signals from artefacts and to select the mental-task related subsignals. These subsignals are spatially filterered into independent processes which can be more informative for the discrimination of different types of EEG data.
- It can be used to analyze the source generators in the brain, aiding the visualization and interpretation of the mental states.

The main properties that we want to include in our model, and which are missing in most decomposition methods, are:

- Flexibility in choosing the number of subsignals that can be recovered.

- The possibility to obtain dynamical systems in particular frequency ranges.
- The use of the temporal structure of the EEG which, in many cases, can be of help in obtaining a good decomposition. This means that we will need to take into account the dynamics of the components s_t^i. The component will be modelled as independent in the following sense:
$$p(s_{1:T}^i, s_{1:T}^j) = p(s_{1:T}^i)p(s_{1:T}^j), \qquad \text{for } i \neq j.$$

A model which satisfies the desired properties and which, in addition, has the advantage of being computationally tractable and easy to parameterize may be obtained from a specially constrained form of a Linear Gaussian State-Space Model.

6.1.1 Linear Gaussian State-Space Models (LGSSM)

A linear (discrete-time) Gaussian state-space model [Durbin and Koopman (2001)] assumes that, at time t, an observed signal $v_t \in \mathcal{R}^V$ (assumed zero mean) is generated by linear mixing of a hidden dynamical system $h_t \in \mathcal{R}^H$ corrupted by Gaussian white noise, that is:

$$v_t = Bh_t + \eta_t^v, \qquad \eta_t^v \sim \mathcal{N}(0, \Sigma_V)\,,$$

where $\mathcal{N}(0, \Sigma_V)$ denotes a Gaussian distribution with zero mean and covariance Σ_V. The dynamics of the underlying system is linear but corrupted by Gaussian noise:

$$\begin{aligned} &h_1 \sim \mathcal{N}(\mu, \Sigma) \\ &h_t = Ah_{t-1} + \eta_t^h, \qquad \eta_t^h \sim \mathcal{N}(0, \Sigma_H), \qquad t > 1\,. \end{aligned}$$

The purpose is to infer properties of the hidden process $h_{1:T}$ from the knowledge of the observations $v_{1:T}$. The linearity and Gaussian-noise assumptions make the LGSSM tractable while providing enough generality to represent many real-world systems. For this reason LGSSMs are widely used in many different disciplines [Grewal and Andrews (2001)].

Previous applications of the LGSSM in BCI-related Research

The most common use of a LGSSM in BCI-related research is to estimate the autoregressive coefficients of an EEG time-series, considered as the hidden variables of a LGSSM. Tarvainen et al. (2004) computed the spectrogram from the estimated time-varying coefficients for tracking α rhythms, while in Schlögl et al. (1999) abrupt increases in the prediction-error covariance of such a model were used as detectors of artefacts. In Georgiadis et al. (2005), the LGSSM was used as an alternative to other filtering techniques for denoising P300 evoked potentials. The

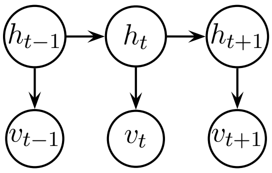


Figure 6.1: Graphical representation of a state-space model. The variables h_t are continuous. Most commonly the visible output variables v_t are continuous.

purpose was to explicitly model the fact that trial-to-trial P300 variability is partly due to different artifacts, level of user's attention, etc.; but also partly due to changes in the dynamics of the underlying system. In all these works the model parameters $\Theta = \{A, B, \Sigma_H, \Sigma_V, \mu, \Sigma\}$ were assumed to be known. Galka et al. (2004) proposed to use a LGSSM for solving the inverse problem of estimating the sources in the brain from EEG recording, incorporating both temporal and spatial smoothness constraints in the solution. In this case, the output matrix B was the standard 'lead field matrix' used in inverse modeling, while A was properly structured so that only neighboring sources could interact. All neighbors were assumed to evolve with the same dynamics.

The next Section of this Chapter reviews the general theory of the LGSSM, for which inference results in the classical Kalman filtering and smoothing algorithms. This is done by using a probabilistic definition of the LGSSM, which gives a simple way of finding the smoothing recursions and the moments required in the learning of the system parameters. The development of the theory from this perspective constitutes the basis for the Bayesian extension of the LGSSM presented in Chapter 7. Section 6.3 gives the update formulas for learning the system parameters using EM maximum likelihood. In Section 6.4 we introduce a constrained LGSSM for finding independent hidden processes of an EEG time-series. We show how, on artificial data, this temporal model is able to recover independent processes, as opposed to other static techniques. We then apply the model to raw EEG data for extracting independent mental processes. Finally we discuss the issues of identifying the correct number of underlying hidden sources and biasing the parameters towards a desired dynamics.

6.2 Inference in the LGSSM

An equivalent probabilistic definition of the LGSSM is the following:

$$p(h_{1:T}, v_{1:T}) = p(v_1|h_1)p(h_1)\prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1}),$$

where $p(h_t|h_{t-1}) = \mathcal{N}(Ah_{t-1}, \Sigma_H)$ and $p(v_t|h_t) = \mathcal{N}(Bh_t, \Sigma_V)$. The graphical representation of this model is given in Fig. 6.1. Here we made the assumption that η^h_t and η^v_t are mutually uncorrelated jointly Gaussian white noise sequences, that is $\left\langle \eta^h_t (\eta^v_{t'})^\mathsf{T} \right\rangle_{p(\eta^h_t, \eta^v_{t'})} = 0$ for all t and t'. Furthermore, h_1 is not correlated with η^h_t and η^v_t. In Section 6.2.1 and Section 6.2.2 we derive the forward and backward recursions for the filtered and smoothed state estimates. They correspond to the standard predictor-corrector [Mendel (1995)] and Rauch-Tung-Striebel [Rauch et al. (1965)] recursions respectively, but they are found using the probabilistic definition of the LGSSM. The advantage in using this non-standard approach is the simplicity in the way the smoothed state estimates and the cross-moments required for EM are computed. In particular, computing the cross-moments with this approach is computationally less expensive than using the standard approach [Shumway and Stoffer (2000)].

6.2.1 Forward Recursions for the Filtered State Estimates

In this section, we are interested in computing the mean $\hat{h}^t_t$ and covariance P^t_t of $p(h_t|v_{1:t})$. This can be computed recursively using:

$$p(v_t, h_t|v_{1:t-1}) = p(v_t|h_t) \underbrace{\int_{h_{t-1}} p(h_t|h_{t-1}) p(h_{t-1}|v_{1:t-1})}_{p(h_t|v_{1:t-1})}.$$

This relation expresses $p(h_t, v_t|v_{1:t-1})$ (and consequently $p(h_t|v_{1:t})$) as a function of $p(h_{t-1}|v_{1:t-1})$. From $p(h_{t-1}|v_{1:t-1})$ the predictor $p(h_t|v_{1:t-1})$ is computed and then a correction is applied with the term $p(v_t|h_t)$ to incorporate the new measurement v_t. The mean and covariance of $p(h_t|v_{1:t-1})$ as a function of the mean and covariance of $p(h_{t-1}|v_{1:t-1})$ can be found by using the linear system equations:

$$\begin{aligned}
\hat{h}^{t-1}_t &= \left\langle Ah_{t-1} + \eta^h_t \right\rangle_{p(h_{t-1}|v_{1:t-1})} = A\hat{h}^{t-1}_{t-1} \\
P^{t-1}_t &= \left\langle (A\tilde{h}_{t-1} + \eta^h_t)(A\tilde{h}_{t-1} + \eta^h_t)^\mathsf{T} \right\rangle_{p(h_{t-1}|v_{1:t-1})} = AP^{t-1}_{t-1}A^\mathsf{T} + \Sigma_H \,,
\end{aligned}$$

where $\tilde{h}^{t-1}_{t-1} = h_{t-1} - \hat{h}^{t-1}_{t-1}$. We can compute the joint density $p(v_t, h_t|v_{1:t-1})$ by using the linear system equations, as before:

$$\begin{aligned}
\langle v_t \rangle_{p(v_t|v_{1:t-1})} &= B\hat{h}^{t-1}_t \\
\left\langle \tilde{v}^{t-1}_t (\tilde{h}^{t-1}_t)^\mathsf{T} \right\rangle_{p(v_t, h_t|v_{1:t-1})} &= BP^{t-1}_t
\end{aligned}$$

$$\left\langle \tilde{v}_t^{t-1}(\tilde{v}_t^{t-1})^{\mathsf{T}} \right\rangle_{p(v_t|v_{1:t-1})} = BP_t^{t-1}B^{\mathsf{T}} + \Sigma_V \,.$$

The joint covariance of $p(v_t, h_t|v_{1:t-1})$ is:

$$\begin{pmatrix} BP_t^{t-1}B^{\mathsf{T}} + \Sigma_V & BP_t^{t-1} \\ P_t^{t-1}B^{\mathsf{T}} & P_t^{t-1} \end{pmatrix} .$$

Using the formulas for conditioning in Gaussian distributions (see Appendix A.4.2) we find that $p(h_t|v_{1:t-1}, v_t)$ has mean and covariance:

$$\begin{aligned} \hat{h}_t^t &= \hat{h}_t^{t-1} + P_t^{t-1}B^{\mathsf{T}}(BP_t^{t-1}B^{\mathsf{T}} + \Sigma_V)^{-1}(v_t - B\hat{h}_t^{t-1}) = \hat{h}_t^{t-1} + K(v_t - B\hat{h}_t^{t-1}) \\ P_t^t &= P_t^{t-1} - P_t^{t-1}B^{\mathsf{T}}(BP_t^{t-1}B^{\mathsf{T}} + \Sigma_V)^{-1}BP_t^{t-1} = (I - KB)P_t^{t-1} \,, \end{aligned}$$

where $K = P_t^{t-1}B^{\mathsf{T}}(BP_t^{t-1}B^{\mathsf{T}} + \Sigma_V)^{-1}$. In the experiments, we will use another equivalent expression for P_t^t, called the Joseph's stabilized form:

$$P_t^t = (I - KB)P_t^{t-1}(I - KB)^{\mathsf{T}} + K\Sigma_V K^{\mathsf{T}}.$$

The final forward recursive updates are:

$$\begin{aligned} \hat{h}_t^{t-1} &= A\hat{h}_{t-1}^{t-1} \\ P_t^{t-1} &= AP_{t-1}^{t-1}A^{\mathsf{T}} + \Sigma_H \\ \hat{h}_t^t &= \hat{h}_t^{t-1} + K(v_t - B\hat{h}_t^{t-1}) \\ P_t^t &= (I - KB)P_t^{t-1}(I - KB)^{\mathsf{T}} + K\Sigma_V K^{\mathsf{T}} \end{aligned}$$

where $\hat{h}_1^0 = \mu$ and $P_1^0 = \Sigma$.

6.2.2 Backward Recursions for the Smoothed State Estimates

To find a recursive formula for the smoothed state estimates we use the fact that

$$p(h_t|v_{1:T}) = \int_{h_{t+1}} p(h_t|h_{t+1}, v_{1:t})p(h_{t+1}|v_{1:T}) \,.$$

The term $p(h_t|h_{t+1}, v_{1:t})$ can be obtained by conditioning the joint distribution $p(h_t, h_{t+1}|v_{1:t})$ with respect to h_{t+1}. The joint covariance of $p(h_t, h_{t+1}|v_{1:t})$ is given by:

$$\begin{pmatrix} P_t^t & P_t^t A^{\mathsf{T}} \\ AP_t^t & AP_t^t A^{\mathsf{T}} + \Sigma_H \end{pmatrix} .$$

From the formulas of Gaussian conditioning, we find that $p(h_t|h_{t+1}, v_{1:t})$ has mean and covariance:

$$\begin{aligned}
&\hat{h}_t^t + P_t^t A^\mathsf{T}(AP_t^tA^\mathsf{T} + \Sigma_H)^{-1}(h_{t+1} - A\hat{h}_t^t)\\
&P_t^t - P_t^tA^\mathsf{T}(AP_t^tA^\mathsf{T} + \Sigma_H)^{-1}AP_t^t\,.
\end{aligned}$$

This is equivalent to the following linear system:

$$h_t = \overleftarrow{A}_t h_{t+1} + \overleftarrow{m}_t + \overleftarrow{\eta}_t\,,$$

where $\overleftarrow{A}_t = P_t^tA^\mathsf{T}(AP_t^tA^\mathsf{T}+\Sigma_H)^{-1}$, $\overleftarrow{m}_t = \hat{h}_t^t - \overleftarrow{A}_t(A\hat{h}_t^t)$ and $p(\overleftarrow{\eta}_t|v_{1:t}) = \mathcal{N}(0, P_t^t - P_t^tA^\mathsf{T}(AP_t^tA^\mathsf{T} + \Sigma_H)^{-1}AP_t^t$. By definition $p(\overleftarrow{\eta}_t, h_{t+1}|v_{1:T}) = p(\overleftarrow{\eta}_t|v_{1:t})p(h_{t+1}|v_{1:T})$. This 'time reversed' dynamics is particularly useful for easily deriving the recursions. Indeed, by using the defined linear system, we easily find that:

$$\begin{aligned}
h_t^T &= \hat{h}_t^t + \overleftarrow{A}_t(\hat{h}_{t+1}^T - A\hat{h}_t^t)\\
P_t^T &= \overleftarrow{A}_tP_{t+1}^T\overleftarrow{A}_t^\mathsf{T} + P_t^t - P_t^tA^\mathsf{T}(AP_t^tA^\mathsf{T} + \Sigma_H)^{-1}AP_t^t = P_t^t + \overleftarrow{A}_t(P_{t+1}^T - P_{t+1}^t)\overleftarrow{A}_t^\mathsf{T}\,.
\end{aligned}$$

We also notice that using the 'time reversed' system we can easily compute the cross-moment:

$$\left\langle h_{t-1}h_t^\mathsf{T}\right\rangle_{p(h_{t-1:t}|v_{1:T})} = \overleftarrow{A}_{t-1}P_t^T + \hat{h}_{t-1}^T(\hat{h}_t^T)^\mathsf{T}$$

that will be used in Section 6.3. This approach is simpler and computationally less expensive than the one presented in Roweis and Ghahramani (1999); Shumway and Stoffer (2000). As in the forward case, we can use the following more stable formulation of the smoothed covariance $P_t^T = (I - \overleftarrow{A}_tA)P_t^t(I - \overleftarrow{A}_tA)^\mathsf{T} + \overleftarrow{A}_t(P_{t+1}^T + \Sigma_H)\overleftarrow{A}_t^\mathsf{T}$. The final backward recursive updates are:

$$\begin{aligned}
\overleftarrow{A}_t &= P_t^tA^\mathsf{T}(P_{t+1}^t)^{-1}\\
\hat{h}_t^T &= \hat{h}_t^t + \overleftarrow{A}_t(\hat{h}_{t+1}^T - A\hat{h}_t^t)\\
P_t^T &= (I - \overleftarrow{A}_tA)P_t^t(I - \overleftarrow{A}_tA)^\mathsf{T} + \overleftarrow{A}_t(P_{t+1}^T + \Sigma_H)\overleftarrow{A}_t^\mathsf{T}
\end{aligned}$$

6.3 Learning the Parameters of a LGSSM

The parameters of a LGSSM can be learned by maximum likelihood using the Expectation Maximization (EM) algorithm [Shumway and Stoffer (1982)]. At each iteration i, EM maximizes the expectation of the complete data log-likelihood for the M training sequences $v_{1:T_m}^m$ (we omit the dependency on m):

$$\mathcal{Q}(\Theta, \Theta^{i-1}) = \left\langle \log \prod_{m=1}^{M} p(v_{1:T}, h_{1:T}|\Theta) \right\rangle_{p(h_{1:T}|v_{1:T}, \Theta^{i-1})} .$$

The update rules, derived in Appendix A.5, are:

$$\begin{aligned}
\Sigma_H &= \frac{\sum_{m=1}^{M}\sum_{t=2}^{T}\left(\left\langle h_t h_t^{\mathsf{T}}\right\rangle - A\left\langle h_{t-1}h_t^{\mathsf{T}}\right\rangle - \left\langle h_t h_{t-1}^{\mathsf{T}}\right\rangle A^{\mathsf{T}} + A\left\langle h_{t-1}h_{t-1}^{\mathsf{T}}\right\rangle A^{\mathsf{T}}\right)}{M(T-1)}\\
\Sigma_V &= \frac{\sum_{m=1}^{M}\sum_{t=1}^{T}\left(v_t v_t^{\mathsf{T}} - B\left\langle h_t\right\rangle v_t^{\mathsf{T}} - v_t\left\langle h_t^{\mathsf{T}}\right\rangle B^{\mathsf{T}} + B\left\langle h_t h_t^{\mathsf{T}}\right\rangle B^{\mathsf{T}}\right)}{MT}\\
\Sigma &= \frac{\sum_{m=1}^{M}\left(\left\langle h_1 h_1^{\mathsf{T}}\right\rangle - \left\langle h_1\right\rangle \mu^{\mathsf{T}} - \mu\left\langle h_1^{\mathsf{T}}\right\rangle + \mu\mu^{\mathsf{T}}\right)}{M}\\
\mu &= \frac{\sum_{m=1}^{M}\left\langle h_1\right\rangle}{M}\\
A &= \sum_{m=1}^{M}\sum_{t=2}^{T}\left\langle h_t h_{t-1}^{\mathsf{T}}\right\rangle\left(\sum_{m=1}^{M}\sum_{t=2}^{T}\left\langle h_{t-1}h_{t-1}^{\mathsf{T}}\right\rangle\right)^{-1}\\
B &= \sum_{m=1}^{M}\sum_{t=1}^{T} v_t\left\langle h_t^{\mathsf{T}}\right\rangle\left(\sum_{m=1}^{M}\sum_{t=1}^{T}\left\langle h_t h_t^{\mathsf{T}}\right\rangle\right)^{-1},
\end{aligned}$$

where $\langle h_t \rangle = \hat{h}_t^T$, $\left\langle h_t h_t^{\mathsf{T}}\right\rangle = P_t^T + \hat{h}_t^T(\hat{h}_t^T)^{\mathsf{T}}$ and $\left\langle h_{t-1}h_t^{\mathsf{T}}\right\rangle = \overleftarrow{A}_{t-1}P_t^T + \hat{h}_{t-1}^T(\hat{h}_t^T)^{\mathsf{T}}$.

We have concluded the general theory of the LGSSM. We now present a specially constrained LGSSM that will enable us to extract independent processes from an EEG time-series.

6.4 Identifying Independent Processes with a Factorial LGSSM

Our idea is to use a LGSSM to decompose a multivariate EEG time-series v_t^n, $t = 1, \ldots, T$, $n = 1, \ldots, V$ into a set of of C simpler components generated by independent dynamical systems. More precisely, we seek to find a set of scalar components s_t^i such that:

$$p(s_{1:T}^i, s_{1:T}^j) = p(s_{1:T}^i)p(s_{1:T}^j), \qquad \text{for } i \neq j.$$

The components generate the observed time-series through a noisy linear mixing $v_t = Ws_t + \eta_t^v$. This is a form of Independent Components Analysis (ICA) [Hyvärinen et al. (2001)] although differs from the more standard assumption of independence at each time-step $p(s_{1:T}^i, s_{1:T}^j) = \prod_{t=1}^{T} p(s_t^i)p(s_t^j)$. In order to make independent dynamical subsystems, we force the transition

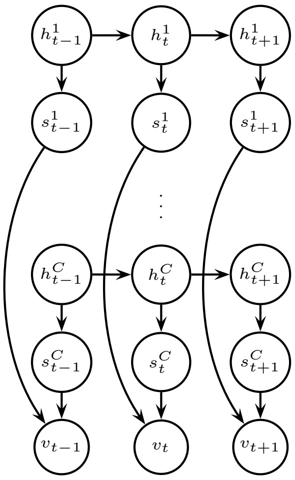


Figure 6.2: The variable h^c_t represents the vector dynamics of component c, which are projected by summation to form the dynamics of the scalar s^c_t. These components are linearly mixed to form the visible observation vector v_t.

matrix A, and the state noise covariances Σ_H and Σ of a LGSSM, to be block-diagonal. In other words, we constrain the evolution of the hidden states h_t to be of the form:

$$\begin{pmatrix} h^1_t \\ \vdots \\ h^C_t \end{pmatrix} = \begin{pmatrix} A^1 & & 0 \\ & \ddots & \\ 0 & & A^C \end{pmatrix} \begin{pmatrix} h^1_{t-1} \\ \vdots \\ h^C_{t-1} \end{pmatrix} + \eta^h_t \,, \qquad \eta^h_t \sim \mathcal{N}(0, \Sigma_H)\,, \tag{6.1}$$

where

$$\Sigma_H = \begin{pmatrix} \Sigma^1_H & & 0 \\ & \ddots & \\ 0 & & \Sigma^C_H \end{pmatrix}\,,$$

and h^c_t is a $H_c \times 1$ dimensional vector representing the state of dynamical system c. This means that the original vector of hidden variables h_t is made of a set of C subvectors h^c_t, each evolving according to its dynamics. A one dimensional component s^c_t for each independent dynamical system is formed from $s^c_t = \mathbf{1}^{\mathsf{T}}_c h^c_t$, where $\mathbf{1}_c$ is a $H_c \times 1$ unit vector. We can represent this in the following matrix form:

$$\begin{pmatrix} s^1_t \\ \vdots \\ s^C_t \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{1}^{\mathsf{T}}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{1}^{\mathsf{T}}_C \end{pmatrix}}_{P} \begin{pmatrix} h^1_t \\ \vdots \\ h^C_t \end{pmatrix} \,. \tag{6.2}$$

The resulting emission matrix is constrained to be of the form

$$B = WP\,, \tag{6.3}$$

where W is the $V \times C$ mixing matrix and P is the $C \times H$ projection given above with $H = \sum_c H_c$. Such a constrained form for B is needed to provide interpretable scalar components. The graphical structure of this model is presented in Fig. 6.2. Unlike a general LGSSM, in which the parameters, and consequently the hidden states, cannot be uniquely determined, in this constrained model each component s_t^i can be determined up to a scale factor. This is discussed in Section 6.4.1.

6.4.1 Identifiability of the System Parameters

Unconstrained Case

In general, the parameters $\Theta = \{A, B, \Sigma_H, \Sigma_V, \mu, \Sigma\}$ of an unconstrained LGSSM cannot be uniquely identified. Indeed, for any invertible matrix D, we can define a new model with parameters $\tilde{\Theta} = \{\tilde{A}, \tilde{B}, \tilde{\Sigma}_H, \tilde{\Sigma}_V, \tilde{\mu}, \tilde{\Sigma}\}$ as:

$$\begin{aligned}
\tilde{A} &= D^{-1}AD\\
\tilde{B} &= BD\\
\tilde{\Sigma}_H &= D^{-1}\Sigma_H D^{-\mathsf{T}}\\
\tilde{\Sigma}_V &= \Sigma_V\\
\tilde{\mu} &= D^{-1}\mu\\
\tilde{\Sigma} &= D^{-1}\Sigma D^{-\mathsf{T}}\,.
\end{aligned}$$

The original hidden variables become $\tilde{h}_t = D^{-1}h_t$. This model is equivalent to the original one, in the sense that it gives the same value of the likelihood $p(v_{1:T}|\tilde{\Theta}) = p(v_{1:T}|\Theta)$. This can be easily seen by observing that $p(v_{1:T}|\tilde{\Theta})$ can be factorized into:

$$p(v_{1:T}) = p(v_1|\tilde{\Theta})\prod_{t=2}^{T} p(v_t|v_{1:t-1}, \tilde{\Theta}).$$

Each term $p(v_t|v_{1:t-1}, \tilde{\Theta})$ has mean and covariance given by:

$$\begin{aligned}
&\tilde{B}\tilde{\hat{h}}_t^{t-1} = B\hat{h}_t^{t-1}\\
&\tilde{B}\tilde{P}_t^{t-1}\tilde{B}^{\mathsf{T}} + \Sigma_V = BDD^{-1}P_t^{t-1}D^{-\mathsf{T}}D^{\mathsf{T}}B^{\mathsf{T}} + \Sigma_V = BP_t^{t-1}B^{\mathsf{T}} + \Sigma_V\,.
\end{aligned}$$

This means that maximum likelihood will not give a unique solution for the parameters and, as a consequence, we cannot estimate the original hidden variables h_t, since they are indistinguishable from $\tilde{h}_t$.

Factorial LGSSM

In our case we put constraints on the model parameters for finding independent components s_t. These constraints make each component s_t^c identifiable up to a scale factor. Indeed a new model with parameters $\tilde{\Theta} = \{\tilde{A} = A, \tilde{W} = WD, \tilde{P} = D^{-1}P, \tilde{\Sigma}_H = \Sigma_H, \tilde{\Sigma}_V = \Sigma_V, \tilde{\mu} = \mu, \tilde{\Sigma} = \Sigma\}$, which defines new components $\tilde{s}_t = D^{-1}s_t$, is equivalent to the original one only when D is diagonal, otherwise $\tilde{P}$ will not be block-diagonal. In other words, the only alternative solution has the original component s_t^c rescaled by a factor d_{cc}.

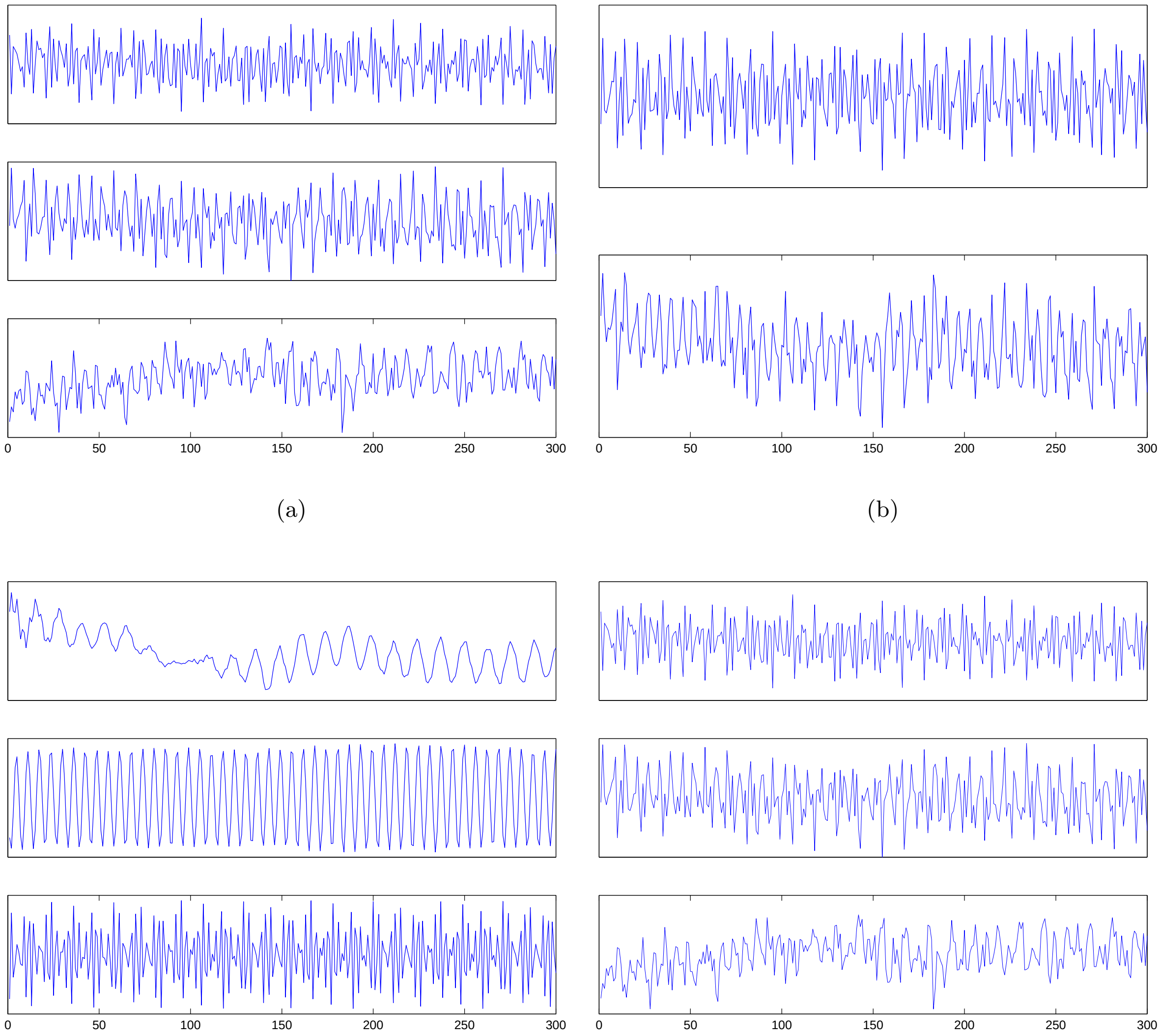
Finally, we observe that constraining all nonzero elements of the projection matrix P to be equal to one is not restrictive. Indeed, if our time-series has been generated by a model with block-diagonal A, Σ_H, Σ and output matrix $B = WP$, where P is a general block-diagonal projection $P = diag((p^1)^\mathsf{T}, \ldots, (p^C)^\mathsf{T})$ with p^c a $H_c \times 1$ dimensional vector, we can define a new model with parameters $\tilde{\Theta} = \{\tilde{A} = D^{-1}AD, \tilde{W} = W, \tilde{P} = PD, \tilde{\Sigma}_H = D^{-1}\Sigma_H D^{-\mathsf{T}}, \tilde{\Sigma}_V = \Sigma_V, \tilde{\mu} = D^{-1}\mu, \tilde{\Sigma} = D^{-1}\Sigma D^{-\mathsf{T}}\}$ with $D = diag(diag(p^1)^{-1}, \ldots, diag(p^C)^{-1})$. The transition matrix $\tilde{A}$ and noise covariances $\tilde{\Sigma}_H$ and $\tilde{\Sigma}$ will still be block-diagonal. This model gives the same components $\tilde{s} = s$.

6.4.2 Artificial Experiment

The FLGSSM described above has no restrictions on the size of the hidden space. Thus, in principle, it can recover a number of components greater that the number of observations. This problem is called *overcomplete* separation. Most blind source separation methods restrict W to be square, that is the number of components and observations is the same. Overcomplete separation is a very difficult task, and the hope is that, in some case, the restriction imposed by the dynamics will aid in finding the correct solution. We linearly mixed three components into two dimensional observations, with addition of Gaussian noise with covariance

$$\Sigma_V = \begin{pmatrix} 0.0164 & 0.005 \\ 0.0054 & 0.0333 \end{pmatrix}.$$

The original components and the noisy observations are displayed in Fig. 6.3a and Fig. 6.3b respectively. We compared the FLGSSM described above with another model that can perform source separation in the overcomplete case with the presence of Gaussian output noise, namely Independent Factor Analysis (IFA) [Attias (1999)]. As the FLGSSM, IFA assumes that the



(a)

(b)

(c)

(d)

Figure 6.3: (a): Original components s_t. (b): Observations resulting from mixing the original components, $v_t = W s_t + \eta_t^v$. (c): Recovered components using the FLGSSM. (d): Recovered components found using IFA.

observations are generated by a noisy linear mixing of hidden components:

$$v_t = W s_t + \eta_t^v, \qquad \eta_t^v \sim \mathcal{N}(0, \Sigma_V)\,.$$

However, each s_t is made of C statistically independent factors s_t^c which do not evolve according to a linear dynamical model, but they are assumed to be temporally independent identically distributed, that is $p(s_{1:T}^c) = \prod_{t=1}^T p(s_t^c)$. In particular, each factor s_t^c is distributed as a mixture of M_c Gaussians:

$$p(s_t^c) = \sum_{m^c=1}^{M_c} p(s_t^c|m^c)p(m^c)\,,$$

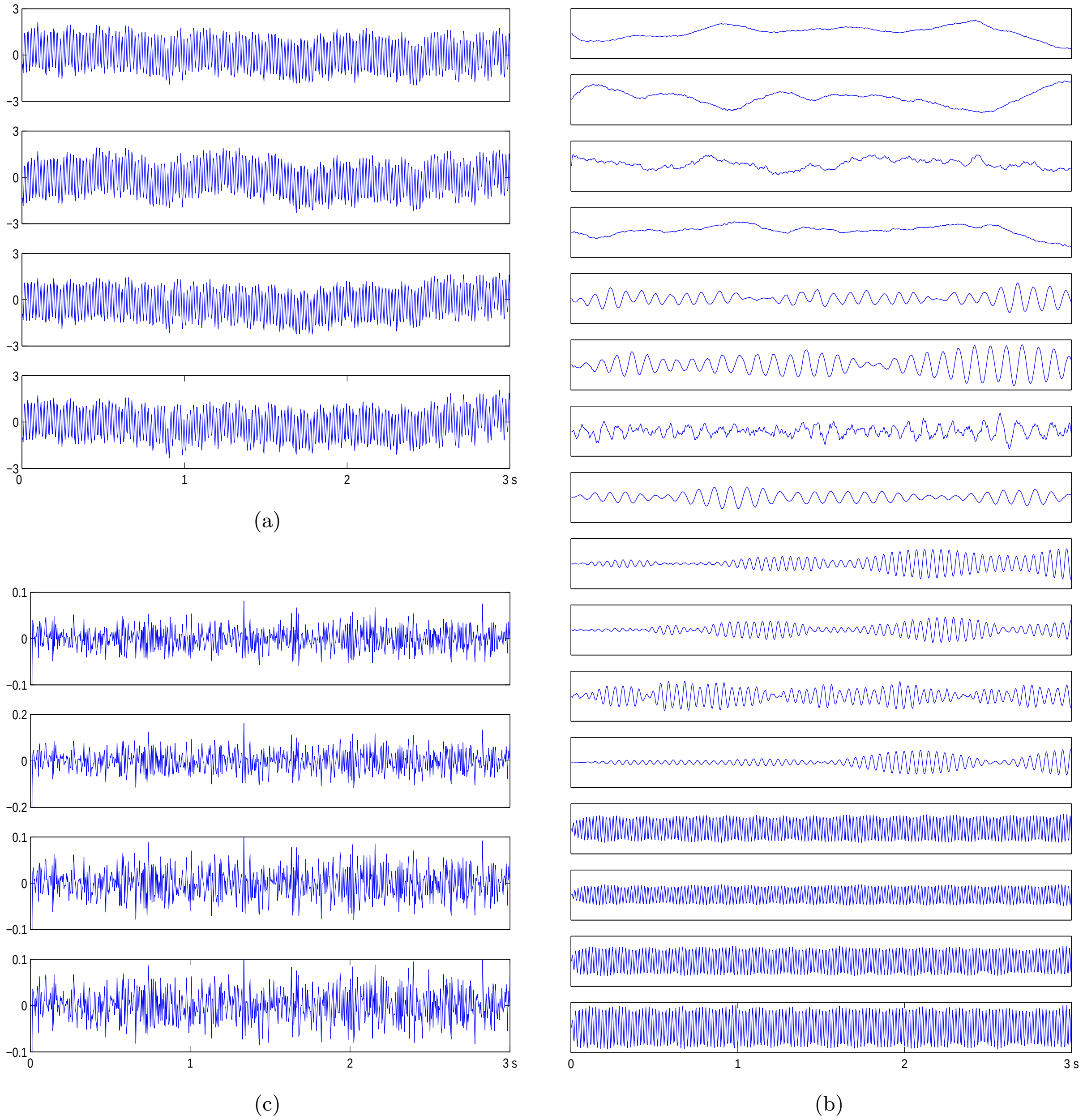
where $p(s_t^c|m^c)$ is Gaussian. The use of a mixture of Gaussians solves the problem of invariance under rotation[1] and makes the hidden components uniquely identifiable. The parameters are learned with the EM algorithm. Thus this model is similar to our FLGSSM, the difference being in the way the hidden components are modeled. In our case, we use linear dynamics, while IFA use a mixture of Gaussians.

For this example, we used four Gaussians for each hidden factor. In the FLGSSM, the size of each independent process H_c was set to three. Fig. 6.3c and Fig. 6.3d show the components estimated by the FLGSSM and IFA respectively. We can see that the FLGSSM gives good estimates, while IFA does not give satisfactory estimates of the components. Thus, this example shows how the use of temporal information can aid separation in difficult cases, like the overcomplete one. Of course, the FLGSSM may fail when the hidden dynamics it is too complicated to be modeled by a linear Gaussian model.

6.4.3 Application to EEG Data

We apply here the FLGSSM to a sequence of raw unfiltered EEG data recorded from four channels located in the right hemisphere while a person is performing imagined movement of the right hand. We are interested in extracting motor related EEG rhythms, mainly centered at 10 and 20 Hz. The EEG data is shown in Fig. 6.4a. As we can see, the interesting information is completely masked by the presence of 50 Hz mains contamination and by low frequency drift terms (DC level). To incorporate prior information about the noise and frequencies of interest, we defined A^c to be a block-diagonal matrix, with each block being a rotation at a desired

[1] The use of a Gaussian distribution, with mean μ and diagonal covariance Σ, for s_t would give a model $v_t = W s_t + \eta_t^v$ which is invariant under a rotation matrix R such that $RR^{\mathsf{T}} = I$. Indeed a new model $v_t = \tilde{W}\tilde{s}_t + \eta_t^v$, where $\tilde{W} = W U_\Sigma^{\mathsf{T}} R$, $\tilde{s}_t = R^{\mathsf{T}} U_\Sigma^{-\mathsf{T}} s_t$ and U_Σ is the Cholesky decomposition of Σ, has the same likelihood, given that $p(v_t)$ has mean $W U_\Sigma^{\mathsf{T}} R R^{\mathsf{T}} U_\Sigma^{-\mathsf{T}} \mu = W\mu$ and covariance $W U_\Sigma^{\mathsf{T}} R R^{\mathsf{T}} U_\Sigma^{-\mathsf{T}} \Sigma U_\Sigma^{-1} R R^{\mathsf{T}} U_\Sigma W^{\mathsf{T}} + \Sigma_V = W \Sigma W^{\mathsf{T}} + \Sigma_V$, and the $\tilde{s}$ are still independent.



(a)

(c)

(b)

Figure 6.4: (a): Three seconds of raw EEG signals recorded form the right hemisphere while a person is performing imagined movement of the right hand. (b): Components extracted by the FLGSSM. (c): Reconstruction error sequences $v_t - Bh_t$.

frequency ω, that is:

$$\gamma \begin{pmatrix} \cos(2\pi\omega/N) & \sin(2\pi\omega/N) \\ -\sin(2\pi\omega/N) & \cos(2\pi\omega/N) \end{pmatrix},$$

where N is the number of samples per second. The constant $\gamma < 1$ damps the oscillation. Projecting this to one dimension describes a damped oscillation with frequency ω. The noise η_t^h affects both the amplitude and phase of the oscillator. By stacking such oscillators together into a single component, we can bias components to have particular frequencies. In this case $s_t^c = \mathbf{1}_c^\mathsf{T} h_t^c$, where $\mathbf{1}_c$ is the vector $(1, 0, \cdots, 1, 0)^2$. For the EEG data, we used 16 block-diagonal matrices with frequencies [0.5], [0.5], [0.5], [0.5], [10,11], [10,11], [10,11], [10,11], [20,21], [20,21], [20,21], [20,21], [50], [50], [50], [50] Hz. The extracted components are plotted in Fig. 6.4b. The model successfully extracted the components at the specified frequencies giving reconstruction error sequences $v_t - Bh_t$ which do not contain activity in the 10-20 Hz range (see Fig. 6.4c). In this example, as it is commonly the case with EEG signals, we did not have a priori knowledge about the number of hidden components. We therefore specified a large number, hoping that irrelevant components would appear as noise. However, it is probable that the phase of the 50 Hz activity does not change for all electrodes, thus the actual number of 50 Hz noise components may be smaller than the four specified above. More importantly, it is likely that the number of hidden processes which generate the 10 and 20 Hz activity measured at the scalp is smaller that four, given that the electrodes are located in the same area of the scalp. Thus it is important to have a model that can estimate automatically the correct number of components. In addition, if prior knowledge about the relevant frequencies is not accurate, we would like a model which may eventually find a solution different from a given prior matrix A^c. This motivates the Bayesian approach introduced in the next Chapter. There we constrain the model to look for the simplest possible explanation of the visible variables. Furthermore, we will give the possibility to specify matrices A^c as priors for the learned dynamics. We will see that, for this EEG data (Fig. 6.4a), this model will prune out many of the 10 and 20 Hz components, while two A^c matrices, biased to be close to rotation at 50 Hz, will move away from the given prior to model other frequencies in the EEG data.

6.5 Conclusions

The aim of this Chapter was to decompose a multichannel EEG recording into subsignals generated by independent dynamical processes. We proposed a specially constrained Linear Gaussian State-Space Model (LGSSM) for which an arbitrary number of components can be extracted. The model exploits the temporal evolution of the components which is helpful for the separation. On artificial data, we have demonstrated that, by using the dynamics of the hidden variables,

[2] Notice that the use of a fixed projection and a fixed block rotation matrix does not result in loss of generality. Indeed, if the case in which originally the projection vector is $p^c = (p_1^c, 0, \cdots, p_{H_c/2}^c, 0)$ we can redefine a new model in which $\tilde{p}^c = p^c D$ with $D = diag(p_1^c, p_1^c, \cdots, p_{H_c/2}^c, p_{H_c/2}^c)^{-1}$. This rescaling will not modify the transition matrix $\tilde{A}^c = D^{-1} A^c D = A^c$.

this model can solve the difficult problem of overcomplete separation of noisy mixtures when another standard static model for blind source separation fails. When applying the model to a sequence of raw EEG data, we could successfully extract relevant mental task information at particular frequencies. In this example, as in most cases in which we aim at extracting independent components from EEG and other sequences, we did not know the correct number of hidden components. For this reason, we specified a number sufficiently high to ensure that the desired information is correctly extracted and does not appear as output noise. Furthermore, we fixed the transition matrix to specific rotations for extracting components at particular frequencies even if this prior information could be inaccurate. This Chapter therefore raises two important issues:

- When we don't know a priori the correct number of hidden processes which have generated the observed time-series, it would be desirable to have a model that can automatically prefer the smallest number of them.
- In many cases we are interested in specific dynamical systems. For example, we may want to extract components in certain frequency ranges, even if this prior information is not precise. Thus, rather than fixing the transition matrices A^c, we would like to learn them but with a bias toward a certain dynamics.

These problems will be addressed in the next Chapter, where we will introduce a Bayesian extension of the FLGSSM.

Chapter 7

Bayesian Factorial LGSSM

The work presented in this Chapter has been published in Chiappa and Barber (2007).

7.1 Introduction

In Chapter 6 we discussed a method for finding independent dynamical systems underlying multiple channels of observation. In particular, we extracted one dimensional subsignals to aid the interpretability of the decomposition. The proposed method, called Factorial Linear Gaussian State-Space Model (FLGSSM), is a specially constrained linear Gaussian state-space model with many desiderable properties such as flexibility in choosing the number of extracted independent processes, the use of temporal information and the possibility to specify a dynamics. However, this model has some limitations. More specifically, the number of independent processes has to be set a priori, whereas in EEG analysis we rarely know the correct number. Furthermore, it would preferable to specify a preferential prior dynamics while keeping some flexibility in the model to move away from it. In order to overcome these limitations, in this Chapter we propose a Bayesian analysis of the FLGSSM. The advantage of the Bayesian approach is that it enables us to specify a preference for the model structure, through a proper choice of the prior $p(\Theta)$. In particular, in our model we will specify a prior on the mixing matrix W such that the number of independent processes that contribute to the observations is as small as possible, and a prior for the transition matrix A to contain a specific frequency structure. This will enable us to automatically determine the number and appropriate complexity of the underlying dynamics, with a preference for the simplest solution, and to estimate independent processes with preferential spectral properties.

For completeness, we will first discuss the Bayesian treatment for a general LGSSM. We will then derive the Bayesian Factorial LGSSM used for finding independent dynamical processes.

On artificially generated data, we will demonstrate the ability of the model to recover the correct number of independent hidden processes. Then we will present an application to unfiltered EEG signals to discover low complexity components with preferential spectral properties, demonstrating improved interpretability of the extracted components over related methods.

7.2 Bayesian LGSSM

We remind the reader that a LGSSM is a model of the form:

$$\begin{aligned}
&h_1 \sim \mathcal{N}(\mu, \Sigma) \\
&h_t = Ah_{t-1} + \eta_t^h, \qquad && \eta_t^h \sim \mathcal{N}(\mathbf{0}_H, \Sigma_H), \qquad t > 1 \\
&v_t = Bh_t + \eta_t^v, && \eta_t^v \sim \mathcal{N}(\mathbf{0}_V, \Sigma_V)\,,
\end{aligned}$$

where $\mathbf{0}_X$ denotes an X-dimensional zero vector. In the standard maximum likelihood approach, as used in Chapter 6, the parameters $\Theta = \{A, B, \Sigma_H, \Sigma_V, \mu, \Sigma\}$ of the LGSSM are estimated by maximizing the data likelihood $p(v_{1:T}|\Theta)$. Maximum likelihood suffers from the problem of not taking into account model complexity and cannot be reliably used to determine the best model structure. In contrast, the Bayesian approach considers Θ as a random vector with a prior distribution $p(\Theta)$. Hence we have a distribution over parameters, rather than a single optimal solution. One advantage of the Bayesian approach is that it enables us to specify what kinds of parameters Θ we would a priori prefer. The parameters Θ in general depend on a set of hyperparameters $\hat{\Theta}$. Thus the likelihood can be written as:

$$p(v_{1:T}|\hat{\Theta}) = \int_{\Theta} p(v_{1:T}|\hat{\Theta}, \Theta)p(\Theta|\hat{\Theta})\,. \tag{7.1}$$

In a full Bayesian treatment we would define additional prior distributions over the hyperparameters $\hat{\Theta}$. Here we take instead the type II Maximum likelihood ('evidence') framework, in which the optimal set of hyperparameters is found by maximizing $p(v_{1:T}|\hat{\Theta})$ with respect to $\hat{\Theta}$ [MacKay (1995); Valpola and Karhunen (2002); Beal (2003)].

7.2.1 Priors Specification

For the parameter priors, we define Gaussians on the columns of A and B:

$$p(A|\alpha, \Sigma_H) \propto \prod_{j=1}^{H} e^{-\frac{\alpha_j}{2}(A_j - \hat{A}_j)^{\mathsf{T}}\Sigma_H^{-1}(A_j - \hat{A}_j)}, \qquad p(B|\beta, \Sigma_V) \propto \prod_{j=1}^{H} e^{-\frac{\beta_j}{2}(B_j - \hat{B}_j)^{\mathsf{T}}\Sigma_V^{-1}(B_j - \hat{B}_j)}\,,$$

which has the effect of biasing the transition and emission matrices to desired forms $\hat{A}$ and $\hat{B}$. This specific dependence on Σ_H and Σ_V is chosen in order to obtain simple forms of the required statistics, as we shall see. The conjugate priors for the inverse covariances Σ_H^{-1} and Σ_V^{-1} are Wishart distributions [Beal (2003)][1]. In the simpler case of assuming diagonal inverse covariances these become Gamma distributions [Beal (2003); Cemgil and Godsill (2005)]. The hyperparameters are $\hat{\Theta} = \{\alpha, \beta\}$[2].

7.2.2 Variational Bayes

If we were able to compute $p(v_{1:T}|\hat{\Theta})$ and $p(\Theta, h_{1:T}|v_{1:T}, \hat{\Theta})$ we could use, for example, an Expectation Maximization (EM) algorithm for finding the hyperparameters. However, despite the above Gaussian priors, the integral in Eq. (7.1) is intractable. This is a common problem in Bayesian theory and several methods can be applied for approximating Eq. (7.1). One possibility is to use Markov chain Monte Carlo methods that approximate the integral by sampling. The main problem with these methods is that they are slow given the high number of samples required to obtain a good approximation. Here we take the variational approach, as discussed by Beal (2003). The idea is to approximate the distribution over the hidden states and the parameters with a simpler distribution. Using Jensen's inequality, the log-likelihood can be lower bounded as:

$$\mathcal{L} = \log p(v_{1:T}|\hat{\Theta}) \geq -\left\langle \log q(\Theta, h_{1:T}) \right\rangle_{q(\Theta,h_{1:T})} + \left\langle \log p(v_{1:T}, h_{1:T}, \Theta|\hat{\Theta}) \right\rangle_{q(\Theta,h_{1:T})} . \tag{7.2}$$

For certain simplifying choices of the variational distribution q, we hope to achieve a tractable lower bound on the likelihood, which we may then optimize with respect to q and $\hat{\Theta}$. The key approximation in Variational Bayes is:

$$q(\Theta, h_{1:T}) \equiv q(\Theta)q(h_{1:T}).$$

This assumption allows other simplifications to follow, without further loss of generality. In particular:

$$\begin{aligned}\mathcal{L} \geq & -\left\langle \log q(\Theta) \right\rangle_{q(\Theta)} - \left\langle \log q(h_{1:T}) \right\rangle_{q(h_{1:T})} \\ & + \left\langle \log p(v_{1:T}, h_{1:T}|\Theta) p(A, \Sigma_H|\hat{\Theta}) p(B, \Sigma_V|\hat{\Theta}) \right\rangle_{q(h_{1:T})q(\Theta)} .\end{aligned}$$

Since A, Σ_H and B, Σ_V separate in Eq. (7.2), optimally $q(\Theta) = q(A, \Sigma_H)q(B, \Sigma_V)$. Hence:

[1] For expositional simplicity, we do not put priors on μ and Σ.

[2] For simplicity, we keep the parameters of the Wishart priors fixed.

$$
\begin{aligned}
\mathcal{L} \geq & - \left\langle D(q(A|\Sigma_H), p(A|\Sigma_H, \hat{\Theta})) \right\rangle_{q(\Sigma_H)} - D(q(\Sigma_H), p(\Sigma_H|\hat{\Theta})) \\
& - \left\langle D(q(B|\Sigma_V), p(B|\Sigma_V, \hat{\Theta})) \right\rangle_{q(\Sigma_V)} - D(q(\Sigma_V), p(\Sigma_V|\hat{\Theta})) \\
& + H_q(h_{1:T}) + \langle \log p(v_{1:T}, h_{1:T}|\Theta) \rangle_{q(h_{1:T})q(A,\Sigma_H)q(B,\Sigma_V)} \\
\equiv & \mathcal{F}(q(\Theta, h_{1:T}), \hat{\Theta}) \,,
\end{aligned}
$$

where $D(q(x), p(x))$ is the Kullback-Leibler (KL) divergence $\langle \log q(x)/p(x) \rangle_{q(x)}$.

Variational EM In summary, at each iteration i, we perform the following steps:

Variational E-step $q^i(\Theta, h_{1:T}) = \arg\max_{q(\Theta,h_{1:T})} \mathcal{F}(q(\Theta, h_{1:T}), \hat{\Theta}^{i-1})$,

Variational M-step $\hat{\Theta}^i = \arg\max_{\hat{\Theta}} \mathcal{F}(q^i(\Theta, h_{1:T}), \hat{\Theta})$.

Thanks to the factorial form $q(\Theta, h_{1:T}) = q(A|\Sigma_H)q(\Sigma_H)q(W|\Sigma_V)q(\Sigma_V)q(h_{1:T})$, the E-step above may be performed using a co-ordinate wise procedure in which each optimal factor is determined by fixing the other factors. The procedure is described below. The initial parameters $\hat{\Theta}$ are set randomly.

Determining $q(B|\Sigma_V)$

The contribution to the objective function $\mathcal{F}$ from $q(B|\Sigma_V)$ is given by:

$$
\left\langle - \log q(B|\Sigma_V) - \frac{1}{2} \sum_{t=1}^{T} \left\langle (v_t - Bh_t)^T \, \Sigma_V^{-1} \, (v_t - Bh_t) \right\rangle_{q(h_t)} + \log p(B|\Sigma_V) \right\rangle_{q(B|\Sigma_V)q(\Sigma_V)} .
$$

For given Σ_V, the above can be interpreted as the negative KL divergence between $q(B|\Sigma_V)$ and a Gaussian distribution in B. Hence, optimally, $q(B|\Sigma_V)$ is a Gaussian, for which we simply need to find the mean and covariance. The covariance $[\Sigma_B]_{ij,kl} \equiv \langle (B_{ij} - \langle B_{ij} \rangle) \, (B_{kl} - \langle B_{kl} \rangle) \rangle$ (averages wrt $q(B|\Sigma_V)$) is given by:

$$
[\Sigma_B]_{ij,kl} = [H_B^{-1}]_{jl} \, [\Sigma_V]_{ik} \,,
$$

where

$$
[H_B]_{jl} \equiv \sum_{t=1}^{T} \left\langle h_t^j h_t^l \right\rangle_{q(h_t)} + \beta_j \delta_{jl}.
$$

The mean is given by $\langle B \rangle = N_B H_B^{-1}$, where $[N_B]_{ij} \equiv \sum_{t=1}^{T} \left\langle h_t^j \right\rangle v_t^i + \beta_j \hat{B}_{ij}$.

Determining $q(\Sigma_V)$

By specifying a Wishart prior for the inverse covariance, conjugate update formulae are possible. In practice, it is more common to specify a diagonal inverse covariance $\Sigma_V^{-1} = diag(\rho)$, where each diagonal element follows a Gamma prior [Beal (2003); Cemgil and Godsill (2005)]:

$$p(\rho|b_1, b_2) = Ga(b_1, b_2) = \prod_{i=1}^{V} \frac{b_2^{b_1}}{\Gamma(b_1)} \rho_i^{b_1-1} e^{-b_2\rho_i} \, .$$

In this case $q(\rho)$ factorizes and the optimal updates are:

$$q(\rho_i) = Ga\left(b_1 + \frac{T}{2}, b_2 + \frac{1}{2}\left(\sum_t (v_t^i)^2 - [G_B]_{i,i} + \sum_j \beta_j \hat{B}_{ij}^2\right)\right) \, ,$$

where $G_B \equiv N_B H_B^{-1} N_B^{\mathsf{T}}$.

Determining $q(A|\Sigma_H)$

The contribution of $q(A|\Sigma_H)$ to the objective function $\mathcal{F}$ is given by:

$$\left\langle -\log q(A|\Sigma_H) - \frac{1}{2}\sum_{t=2}^{T} \left\langle (h_t - Ah_{t-1})^{\mathsf{T}} \Sigma_H^{-1} (h_t - Ah_{t-1}) \right\rangle_{q(h_{t-1:t})} + \log p(A|\Sigma_H) \right\rangle_{q(A|\Sigma_H)q(\Sigma_H)} \, .$$

As for $q(B|\Sigma_V)$, optimally $q(A|\Sigma_H)$ is a Gaussian with covariance $[\Sigma_A]_{ij,kl}$ given by:

$$[\Sigma_A]_{ij,kl} = [H_A^{-1}]_{jl} \, [\Sigma_H]_{ik} \, ,$$

where

$$[H_A]_{jl} \equiv \sum_{t=1}^{T-1} \left\langle h_t^j h_t^l \right\rangle_{q(h_t)} + \alpha_j \delta_{jl}.$$

The mean is given by $\langle A \rangle = N_A H_A^{-1}$, where $[N_A]_{ij} \equiv \sum_{t=2}^{T} \left\langle h_{t-1}^j h_t^i \right\rangle + \alpha_j \hat{A}_{ij}$.

Determining $q(\Sigma_H)$

Analogously to Σ_V, for $\Sigma_H^{-1} = diag(\tau)$ with prior $Ga(a_1, a_2)$ the updates are:

$$q(\tau_i) = Ga\left(a_1 + \frac{T-1}{2}, a_2 + \frac{1}{2}\left(\sum_{t=2}^{T} \left\langle (h_t^i)^2 \right\rangle - [G_A]_{i,i} + \sum_j \alpha_j \hat{A}_{ij}^2\right)\right) \, ,$$

where $G_A \equiv N_A H_A^{-1} N_A^{\mathsf{T}}$.

Unified Inference on $q(h_{1:T})$

By differentiating $\mathcal{F}$ with respect to $q(h_{1:T})$ under normalization constraints, we obtain that optimally $q(h_{1:T})$ is Gaussian since its log is quadratic in $h_{1:T}$, being namely[3]:

$$\begin{aligned} &-\frac{1}{2}\sum_{t=1}^{T}\left\langle (v_t - Bh_t)^{\mathsf{T}}\Sigma_V^{-1}(v_t - Bh_t)\right\rangle_{q(B,\Sigma_V)} \\ &-\frac{1}{2}\sum_{t=2}^{T}\left\langle (h_t - Ah_{t-1})^{\mathsf{T}}\,\Sigma_H^{-1}\,(h_t - Ah_{t-1})\right\rangle_{q(A,\Sigma_H)} \, . \end{aligned} \tag{7.3}$$

Optimally, $q(A|\Sigma_H)$ and $q(B|\Sigma_V)$ are Gaussians, so we can easily carry out the averages. The further averages over $q(\Sigma_H)$ and $q(\Sigma_V)$ are also easy due to conjugacy. Whilst this defines the distribution $q(h_{1:T})$, quantities such as $q(h_t)$ need to be inferred from this distribution. Clearly, in the non-Bayesian case, the averages over the parameters are not present, and the above simply represents an LGSSM whose visible variables have been clamped into their evidential states. In that case, inference can be performed using any standard method. Our aim, therefore, is to represent the *averaged* Eq. (7.3) directly as an LGSSM $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$, for some suitable parameter settings.

Mean + Fluctuation Decomposition

A useful decomposition is to write:

$$\left\langle (v_t - Bh_t)^{\mathsf{T}}\Sigma_V^{-1}(v_t - Bh_t)\right\rangle_{q(B,\Sigma_V)} = \underbrace{(v_t - \langle B\rangle\, h_t)^{\mathsf{T}}\left\langle \Sigma_V^{-1}\right\rangle (v_t - \langle B\rangle\, h_t)}_{mean} + \underbrace{h_t^{\mathsf{T}} S_B h_t}_{fluctuation} \;,$$

and similarly:

$$\left\langle (h_t - Ah_{t-1})^{\mathsf{T}}\Sigma_H^{-1}(h_t - Ah_{t-1})\right\rangle_{q(A,\Sigma_H)} = \underbrace{(h_t - \langle A\rangle\, h_{t-1})^{\mathsf{T}}\left\langle \Sigma_H^{-1}\right\rangle (h_t - \langle A\rangle\, h_{t-1})}_{mean} + \underbrace{h_{t-1}^{\mathsf{T}} S_A h_{t-1}}_{fluctuation} \,,$$

where the parameter covariances are $S_B = VH_B^{-1}$ and $S_A = HH_A^{-1}$. The mean terms simply represent a clamped LGSSM with averaged parameters. However, the extra contributions from the fluctuations mean that Eq. (7.3) cannot be written as a clamped LGSSM with averaged parameters. In order to deal with these extra terms, our idea is to treat the fluctuations as

[3] For simplicity of exposition, we ignore the contribution from h_1 here.

arising from an augmented visible variable, for which Eq. (7.3) can then be considered as a clamped LGSSM.

Inference Using an Augmented LGSSM

To represent Eq. (7.3) as a LGSSM $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$, we augment v_t and B as:

$$\tilde{v}_t = vert(v_t, \mathbf{0}_H, \mathbf{0}_H), \qquad \tilde{B} = vert(\langle B \rangle, U_A, U_B),$$

where U_A is the Cholesky decomposition of S_A, so that $U_A^\mathsf{T} U_A = S_A$. Similarly, U_B is the Cholesky decomposition of S_B. The equivalent LGSSM $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$ is then completed by specifying[4]

$$\tilde{A} \equiv \langle A \rangle, \quad \tilde{\Sigma}_H \equiv \left\langle \Sigma_H^{-1} \right\rangle^{-1}, \quad \tilde{\Sigma}_V \equiv diag(\left\langle \Sigma_V^{-1} \right\rangle^{-1}, I_H, I_H), \quad \tilde{\mu} \equiv \mu, \quad \tilde{\Sigma} \equiv \Sigma.$$

The validity of this parameter assignment can be checked by showing that, up to negligible constants, the exponent of this augmented LGSSM has the same form as Eq. (7.3). Now that this has been written as an LGSSM $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$, standard inference routines in the literature may be applied to compute $q(h_t) = \tilde{q}(h_t|\tilde{v}_{1:T})$ [Bar-Shalom and Li (1998); Park and Kailath (1996); Grewal and Andrews (2001)][5].

In Algorithm 1 we give the FORWARD and BACKWARD procedures to compute $\tilde{q}(h_t|\tilde{v}_{1:T})$. We present two variants of the FORWARD pass. Either we may call procedure FORWARD with parameters $\tilde{A}, \tilde{B}, \tilde{\Sigma}_H, \tilde{\Sigma}_V, \tilde{\mu}, \tilde{\Sigma}$ and the augmented visible variables $\tilde{v}_t$ in which we use steps 1a, 2a, 5a and 6a. This is exactly the predictor-corrector form of a Kalman filter (see Section 6.2). Otherwise, in order to reduce the computational cost, we may call procedure FORWARD with the parameters $\langle A \rangle, \langle B \rangle, \left\langle \Sigma_H^{-1} \right\rangle^{-1}, \left\langle \Sigma_V^{-1} \right\rangle^{-1}, \mu, \Sigma$ and the original visible variable v_t in which we use steps 1b (where $U_{AB}^\mathsf{T} U_{AB} \equiv S_A + S_B$), 2b, 5b and 6b. The two algorithms are mathematically equivalent. Computing $q(h_t) = \tilde{q}(h_t|\tilde{v}_{1:T})$ is then completed by calling the common BACKWARD pass, which corresponds to the Rauch-Tung-Striebel pass (see Section 6.2).

The important point here is that the reader may supply any standard Kalman filtering and smoothing routine, and simply call it with the appropriate parameters. In some parameter regimes, or in very long time-series, numerical stability may be a serious concern, for which several stabilized algorithms have been developed over the years, for example the square-root

[4] Strictly, we need a time-dependent emission $\tilde{B}_t = \tilde{B}$, for $t = 1, \ldots, T-1$. For time T, $\tilde{B}_T$ has the Cholesky factor U_A replaced by $\mathbf{0}_{H,H}$.

[5] Note that, since the augmented LGSSM $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$ is designed to match the *fully* clamped distribution $q(h_{1:T})$, filtering $\tilde{q}(h_{1:T}|\tilde{v}_{1:T})$ does not correspond to filtering $q(h_{1:T})$.

Algorithm 1 LGSSM: Forward and backward recursive updates. The smoothed posterior $p(h_t|v_{1:T})$ is returned in the mean $\hat{h}_t^T$ and covariance P_t^T.

```
procedure Forward
    1a: $P \leftarrow \Sigma$
    1b: $P \leftarrow (\Sigma^{-1} + S_A + S_B)^{-1} = (I - \Sigma U_{AB}\left(I + U_{AB}^\mathsf{T}\Sigma U_{AB}\right)^{-1} U_{AB}^\mathsf{T}) \equiv D\Sigma$
    2a: $\hat{h}_1^0 \leftarrow \mu$
    2b: $\hat{h}_1^0 \leftarrow D\mu$
    3: $K \leftarrow PB^\mathsf{T}(BPB^\mathsf{T} + \Sigma_V)^{-1}$, $P_1^1 \leftarrow (I - KB)P$, $\hat{h}_1^1 \leftarrow \hat{h}_1^0 + K(v_t - B\hat{h}_1^0)$
    for $t \leftarrow 2, T$ do
        4: $P_t^{t-1} \leftarrow AP_{t-1}^{t-1}A^T + \Sigma_H$
        5a: $P \leftarrow P_t^{t-1}$
        5b: $P \leftarrow D_t P_t^{t-1}$, where $D_t \equiv (I - P_t^{t-1}U_{AB}\left(I + U_{AB}^\mathsf{T} P_t^{t-1} U_{AB}\right)^{-1} U_{AB}^\mathsf{T})$
        6a: $\hat{h}_t^{t-1} \leftarrow A\hat{h}_{t-1}^{t-1}$
        6b: $\hat{h}_t^{t-1} \leftarrow D_t A\hat{h}_{t-1}^{t-1}$
        7: $K \leftarrow PB^\mathsf{T}(BPB^\mathsf{T} + \Sigma_V)^{-1}$, $P_t^t \leftarrow (I - KB)P$, $\hat{h}_t^t \leftarrow \hat{h}_t^{t-1} + K(v_t - B\hat{h}_t^{t-1})$
    end for
end procedure
procedure Backward
    for $t \leftarrow T - 1, 1$ do
        $\overleftarrow{A}_t \leftarrow P_t^t A^\mathsf{T}(P_{t+1}^t)^{-1}$
        $P_t^T \leftarrow P_t^t + \overleftarrow{A}_t(P_{t+1}^T - P_{t+1}^t)\overleftarrow{A}_t^\mathsf{T}$
        $\hat{h}_t^T \leftarrow \hat{h}_t^t + \overleftarrow{A}_t(\hat{h}_{t+1}^T - A\hat{h}_t^t)$
    end for
end procedure
```

forms [Morf and Kailath (1975); Park and Kailath (1996); Grewal and Andrews (2001)]. By converting the problem to a standard form, we have therefore unified and simplified inference, so that future applications may be more readily developed.

Relation to Previous Approaches

An alternative approach to the one above, and taken in Beal (2003); Cemgil and Godsill (2005), is to recognize that the posterior is:

$$\log q(h_{1:T}) = \sum_{t=2}^{T} \phi_t(h_{t-1}, h_t) + const.$$

for suitably defined quadratic forms $\phi_t(h_{t-1}, h_t)$. Here the potentials $\phi_t(h_{t-1}, h_t)$ encode the averaging over the parameters A, B, Σ_H, Σ_V. The approach taken in Beal (2003) is to recognize this as a pairwise Markov chain, for which the Belief Propagation recursions may be applied. The

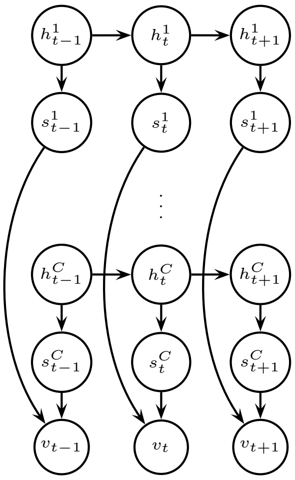


Figure 7.1: The variable h^c_t represents the vector dynamics of component c, which are projected by summation to form the dynamics of the scalar s^c_t. These components are linearly mixed to form the visible observation vector v_t.

backward pass from Belief Propagation makes use of the observations $v_{1:T}$, so that any approximate online treatment would be difficult. The approach in Cemgil and Godsill (2005) is based on a Kullback-Leibler minimization of the posterior with a chain structure, which is algorithmically equivalent to Belief Propagation. Whilst mathematically valid procedures, the resulting algorithms do not correspond to any of the standard forms in the Kalman filtering/smoothing literature, whose properties have been well studied [Verhaegen and Dooren (1986)].

Finding the Optimal $\hat{\Theta}$

Differentiating $\mathcal{F}$ with respect to $\hat{\Theta}$ we find that, optimally:

$$\alpha_j = \frac{H}{\left\langle \left(A_j - \hat{A}_j\right)^{\mathsf{T}} \Sigma_H^{-1} \left(A_j - \hat{A}_j\right) \right\rangle_{q(A_j,\Sigma_H)}}, \ \beta_j = \frac{V}{\left\langle \left(B_j - \hat{B}_j\right)^{\mathsf{T}} \Sigma_V^{-1} \left(B_j - \hat{B}_j\right) \right\rangle_{q(B_j,\Sigma_V)}}.$$

The other hyperparameters can be found similarly to the EM maximum likelihood derivation of a LGSSM (see Appendix A.5), and they are given by: $\mu = \langle h_1 \rangle_{q(h_1)}$, $\Sigma = \left\langle (h_1 - \mu)(h_1 - \mu)^{\mathsf{T}} \right\rangle_{q(h_1)}$.

We have completed the Bayesian treatment of a general LGSSM. We are going to describe the Bayesian Factorial LGSSM used in the experiments.

7.3 Bayesian FLGSSM

We remind the reader that, in a Factorial LGSSM, A and Σ_H are block-diagonal matrices (see Eq. (6.1)), and independent dynamical processes are generated by $s_t = Ph_t$ (see Eq. (6.2)), with P a block-diagonal matrix. That is, the output matrix is parameterized as $B = WP$. This model is shown in Fig. 7.1.

Since we do not have any particular preference for the structure of the noise, we do not define a prior for Σ_H and Σ_V, which are instead considered as hyperparameters. On the other hand, ideally, the number of components effectively contributing to the observed signal should be small. We can essentially turn off a component component by making the associated column of W very small. This suggests the following Gaussian prior:

$$p(W|\beta) = \prod_{j=1}^{C} \left(\frac{\beta_j}{2\pi}\right)^{V/2} e^{-\frac{\beta_j}{2}\sum_{i=1}^{V} W_{ij}^2} .$$

Similarly, we can bias each dynamical system to be close to a desired transition $\hat{A}$ (possibly zero) by using:

$$p(A^c|\alpha_c) = \left(\frac{\alpha_c}{2\pi}\right)^{H_c^2/2} e^{-\frac{\alpha_c}{2}\sum_{i,j=1}^{H_c}\left(A_{ij}^c - \hat{A}_{ij}^c\right)^2}$$

for each component c, so that $p(A|\alpha) = \prod_c p(A^c|\alpha_c)$. Finding the optimal $q(W)$, $q(A)$ and $q(h_{1:T})$ is discussed below.

Determining $q(W)$

The contribution to the (modified) objective function $\mathcal{F}$ from $q(W)$ is given by:

$$\left\langle -\log q(W) - \frac{1}{2}\sum_{t=1}^{T}\left\langle (v_t - WPh_t)^T \Sigma_V^{-1} (v_t - WPh_t)\right\rangle_{q(h_t)} + \log p(W|\beta)\right\rangle_{q(W)} .$$

This can be interpreted as the negative KL divergence between $q(W)$ and a Gaussian distribution in W. Hence, optimally, $q(W)$ is a Gaussian.

The covariance $[\Sigma_W]_{ij,kl} \equiv \langle (W_{ij} - \langle W_{ij}\rangle)(W_{kl} - \langle W_{kl}\rangle)\rangle$ (averages wrt $q(W)$) is given by the inverse of the quadratic contribution:

$$\left[\Sigma_W^{-1}\right]_{ij,kl} = \left[\Sigma_V^{-1}\right]_{ik} \sum_t \left\langle \tilde{h}_t^j \tilde{h}_t^l \right\rangle_{q(h_t)} + \beta_j \delta_{ik}\delta_{jl} ,$$

where $\tilde{h}_t = Ph_t$ and δ_{ij} is the Kronecker delta function. The mean is given by:

$$\langle W_{ij} \rangle_{p(W_{ij})} = \sum_{k,l,n,t} [\Sigma_W]_{ijkl} \left[\Sigma_V^{-1}\right]_{kn} \left\langle \tilde{h}_t^l \right\rangle_{q(h_t)} v_t^n \,.$$

Determining $q(A)$

The contribution of $q(A)$ to the objective function is given by:

$$\left\langle -\log q(A) - \frac{1}{2} \sum_{t=2}^{T} \left\langle (h_t - Ah_{t-1})^{\mathsf{T}} \, \Sigma_H^{-1} \, (h_t - Ah_{t-1}) \right\rangle_{q(h_{t-1:t})} + \log p(A|\alpha) \right\rangle_{q(A)} .$$

Since the dynamics are independent, optimally we have a factorized distribution $q(A) = \prod_c q(A^c)$, where $q(A^c)$ is Gaussian with covariance $[\Sigma_{A^c}]_{ij,kl} \equiv \left\langle (A^c_{ij} - \left\langle A^c_{ij} \right\rangle)(A^c_{kl} - \langle A^c_{kl} \rangle) \right\rangle$ (averages wrt $q(A^c)$). Momentarily dropping the dependence on the component c, the covariance for each component is:

$$\left[\Sigma_A^{-1}\right]_{ijkl} = \left[\Sigma_H^{-1}\right]_{ik} \sum_{t=2}^{T} \left\langle h^j_{t-1} h^l_{t-1} \right\rangle_{q(h_{t-1})} + \alpha \delta_{ik} \delta_{jl} \,,$$

and the mean is:

$$\langle A_{ij} \rangle_{q(A_{ij})} = \sum_{k,l} [\Sigma_A]_{ij,kl} \left(\alpha \hat{A}_{kl} + \sum_n \left[\Sigma_H^{-1}\right]_{kn} \sum_{t=2}^{T} \left\langle h^l_{t-1} h^n_t \right\rangle_{q(h_{t-1:t})} \right) ,$$

where in the above all parameters and the variable h should be interpreted as pertaining to dynamic component c only.

Inference on $q(h_{1:T})$

A small modification of the mean + fluctuation decomposition for B occurs, namely:

$$\left\langle (v_t - Bh_t)^{\mathsf{T}} \Sigma_V^{-1} (v_t - Bh_t) \right\rangle_{q(W)} = (v_t - \langle B \rangle h_t)^{\mathsf{T}} \Sigma_V^{-1} (v_t - \langle B \rangle h_t) + h_t^{\mathsf{T}} P^{\mathsf{T}} S_W P h_t \,,$$

where $\langle B \rangle \equiv \langle W \rangle P$ and $S_W = V H_W^{-1}$. The quantities $\langle W \rangle$ and H_W are obtained as above with the replacement $h_t \leftarrow Ph_t$. To represent the above as a LGSSM, we augment v_t and B as

$$\tilde{v}_t = vert(v_t, \mathbf{0}_H, \mathbf{0}_C), \qquad \tilde{B} = vert(\langle B \rangle, U_A, U_W P),$$

where U_W is the Cholesky decomposition of S_W. The equivalent LGSSM is then completed by specifying $\tilde{A} \equiv \langle A \rangle$, $\tilde{\Sigma}_H \equiv \Sigma_H$, $\tilde{\Sigma}_V \equiv diag(\Sigma_V, I_H, I_C)$, $\tilde{\mu} \equiv \mu$, $\tilde{\Sigma} \equiv \Sigma$, and inference for $q(h_{1:T})$ performed using Algorithm 1. This demonstrates the elegance and unity of the approach in Section 7.2.2, since no new algorithm needs to be developed to perform inference, even in this special constrained parameter case.

Finding the Optimal $\hat{\Theta}$

Differentiating $\mathcal{F}$ with respect to α_c and β_j we find that, optimally:

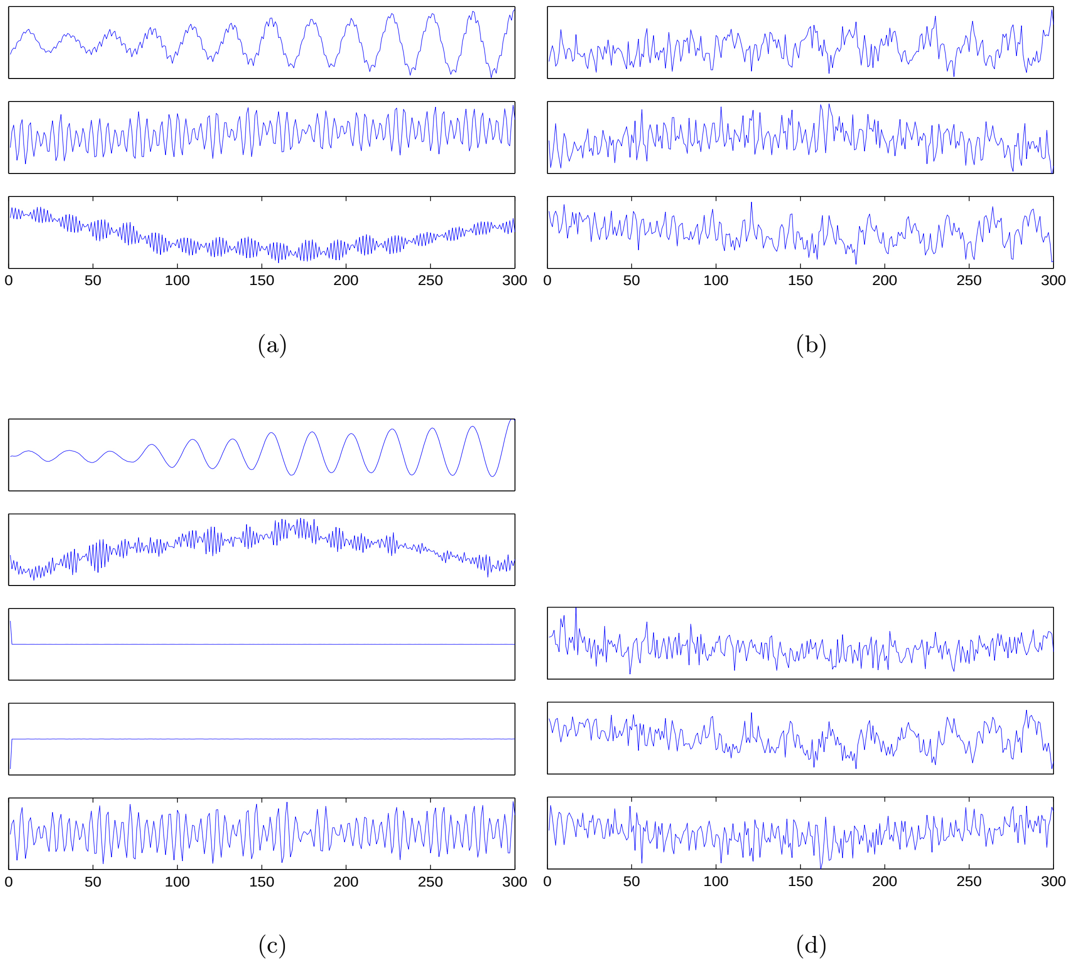
$$\alpha_c = \frac{H_c^2}{\sum_{i,j} \left\langle [A^c - \hat{A}^c]_{ij}^2 \right\rangle_{q(A^c)}}, \qquad \beta_j = \frac{V}{\sum_i \left\langle W_{ij}^2 \right\rangle_{q(W)}} .$$

The other hyperparameters are given by:

$$\begin{aligned}
\Sigma_H^c &= \frac{1}{T-1} \sum_{t=2}^{T} \left\langle \left(h_t^c - A^c h_{t-1}^c\right) \left(h_t^c - A^c h_{t-1}^c\right)^{\mathsf{T}} \right\rangle_{q(A^c) q(h_{t-1:t}^c)} \\
\Sigma_V &= \frac{1}{T} \sum_{t=1}^{T} \left\langle (v_t - W P h_t) (v_t - W P h_t)^{\mathsf{T}} \right\rangle_{q(W) q(h_t)} \\
\Sigma &= \left\langle (h_1 - \mu)(h_1 - \mu)^{\mathsf{T}} \right\rangle_{q(h_1)} \\
\mu &= \langle h_1 \rangle_{q(h_1)} .
\end{aligned}$$

7.3.1 Demonstration

In a proof of concept experiment, we used a LGSSM to generate 3 components with random 5×5 transition matrices A^c, $h_1 \sim \mathcal{N}(\mathbf{0}_H, I_H)$ and $\Sigma_H = I_H$. The components were mixed into 3 observations $v_t = W s_t + \eta_t^v$, for W chosen with elements from a zero mean unit variance Gaussian distribution, and $\Sigma_V = I_V$. We then trained a different LGSSM with 5 components and dimension $H_c = 7$. To bias the model to find the simplest components, we used $\hat{A}^c \equiv \mathbf{0}_{H_c,H_c}$ for all components. In Fig. 7.2a and Fig. 7.2b we see the original components and the noisy observations respectively. The observation noise is so high that a good estimation of the components is possible only by taking the dynamics into account. In Fig. 7.2c we see the estimated components from our method after 400 iterations. Two of the 5 components have been removed and the remaining three are a reasonable estimation of the original components. The FastICA [Hyvärinen et al. (2001)] result is given in Fig. 7.2d. In fairness, FastICA cannot deal with noise and also seeks temporally independent components, whereas in this example the components are slightly correlated. Nevertheless, this example demonstrates that, whilst a



(a)

(b)

(c)

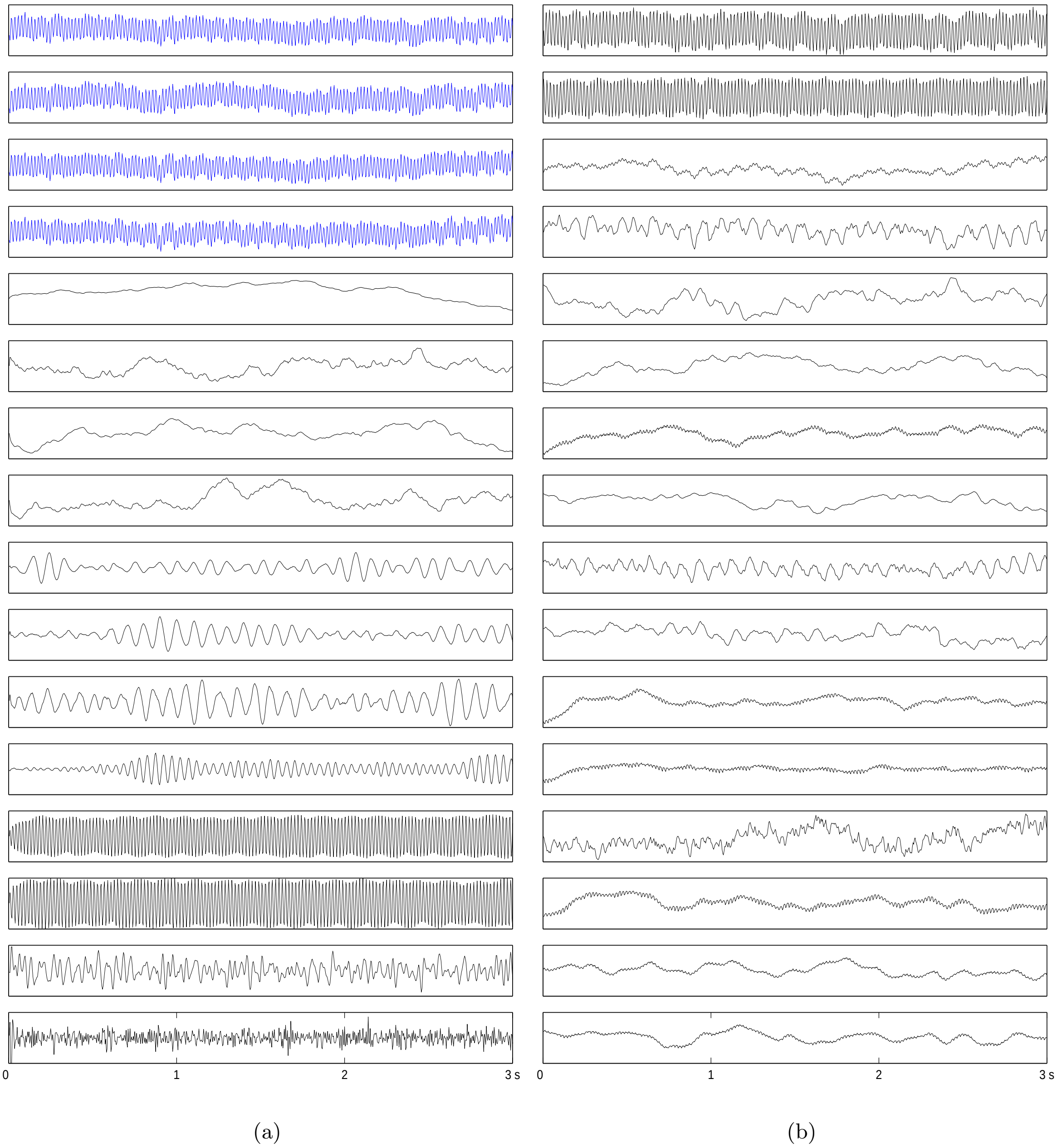
(d)

Figure 7.2: (a): Original (correlated) components s_t. (b): Observations resulting from mixing the original components, $v_t = Ws_t + \eta_t^v$, $\eta_t^v \sim \mathcal{N}(0, I)$. (c): Recovered components using our method. (d): Independent components found using FastICA.

standard method such as FastICA indeed produces independent components, this may not be a satisfactory result, since there is no search for simplicity of the underlying dynamical system, nor indeed may independence at each time point be a desirable criterion.

7.3.2 Application to EEG Analysis

In Fig. 7.3a (blue), we show three seconds of EEG data recorded from 4 channels (located in the right hemisphere) while a subject is performing imagined movement of his right hand. This is the same data used in Section 6.4.3 (Fig. 6.4a). Each channel shows low frequency drift terms,



(a)

(b)

Figure 7.3: (a): The top four (blue) signals are the original unfiltered EEG channel data. The remaining 12 subfigures are the components s_t estimated by our method. (b): The 16 factors estimated by NDFA after convergence (800 iterations).

together with the presence of 50 Hz mains contamination, which mask the information related to the mental task, mainly centered at 10 and 20 Hz. Standard ICA methods such as FastICA do not find satisfactory components based on raw 'noisy' data, and preprocessing with bandpass filters is usually required. However, even with prefiltering, the number of components is usually restricted in ICA to be equal to the number of channels. In EEG this is potentially too restrictive since there may be many independent oscillators of interest underlying the observations and we would like some way to automatically determine the effective number of such oscillators. In agreement with the approach used in Section 6.4.3, we used 16 components. To preferentially find components at particular frequencies, we specified a block diagonal matrix $\hat{A}^c$ with each block being a rotation at the desired frequency. The frequencies for the 16 components were [0.5], [0.5], [0.5], [0.5], [10,11], [10,11], [10,11], [10,11], [20,21], [20,21], [20,21], [20,21], [50], [50], [50], [50] Hz respectively. After training, the Bayesian approach removed 4 unnecessary components from the mixing matrix W, that is one [10,11] Hz and three [20,21] Hz components. The temporal evolution of the 12 retained components is presented in Fig. 7.3a (black). We can see that effectively the first 4 components contain dominant low frequency drift, the following 3 contain [10,11] Hz, while the 8th contains [20,21] Hz centered activity. Out of the 4 components initialized to 50 Hz, only 2 retained 50 Hz activity, while the A^c of last two components have changed in order to model other frequencies present in the signals. In the non-Bayesian FLGSSM this activity was considered as output noise (see Fig. 6.4c). In order to asses the advantage of using prior frequencies for extracting task-related information and the potential limitations of using a linear model, we have compared our method with another temporal method, namely Nonlinear Dynamical Factor Analysis (NDFA) [Valpola and Karhunen (2002); Särelä et al. (2001)]. NDFA is a model of the form:

$$\begin{aligned}
&s_1 \sim \mathcal{N}(\mu, \Sigma) \\
&s_t = g(s_{t-1}) + \eta_t^h, \qquad && \eta_t^h \sim \mathcal{N}(0, \Sigma_H), \qquad t > 1 \\
&v_t = f(s_t) + \eta_t^v, && \eta_t^v \sim \mathcal{N}(0, \Sigma_V)\,,
\end{aligned}$$

where f and g are non linear mappings modeled by MLP networks [Bishop (1995)]. The functional form of f and g is:

$$\begin{aligned}
f(s_t) &= B\tanh[As_t + a] + b \\
g(s_{t-1}) &= s_{t-1} + D\tanh[Cs_{t-1} + c] + d\,,
\end{aligned}$$

where A, B, C and D are the weight matrices of the hidden and output layers of the MLP networks and a, b, c and d are the corresponding bias vectors. Whilst being an attractive and powerful method, standard NDFA places no constraint that the observations are formed from mixing independent *scalar* dynamic components, which makes interpretation of the resulting factors difficult. Furthermore, NDFA does not directly constrain the factors to contain particular frequencies so that in Särelä et al. (2001), in order to extract rhythmic activity in MEG signals, a bias is incorporated by initializing the model with band-filtered principal components of the data. In addition, NDFA uses nonlinear state dynamics and mixing, which hampers inference and makes the incorporation of known constraints more complex. We extracted 16 factors using a NDFA model in which both MLPs had one hidden layer of 30 neurons. The other parameters were set to the default values. In Fig. 7.3b we show the temporal evolution of the resulting factors. The first 10 factors from the top give the strongest contribution to the observations. In agreement with our method, there are 2 main 50 Hz components (first two factors), even if a small 50 Hz activity is present also in other factors, namely 7, 11, 12 and 14. The slow drift has not been isolated and is present in almost all factors. The rhythmic information related to hand movement, namely [10,20] Hz activity, is spread over factors 3, 4, 9, 10 and 13, which however contain also other frequencies. From this example we can conclude that, while the two methods give similar results, the prior specification of independent dynamical processes at particular frequencies has helped our model to better isolate the activity of interest into a smaller number of components, and, among these components, to separate the contribution of oscillators at different frequencies, that is 10 Hz and 20 Hz oscillators.

7.4 Conclusion

We presented a method to identify independent dynamical processes in noisy temporal data, based on a Bayesian procedure which automatically biases the solution to finding a small number of components with preferential spectral properties. This procedure is related to other temporal models previously proposed in the literature, but has the particular property that the components are themselves projections from higher dimensional independent linear dynamical systems. Here we concentrated on the projection to a single dimension since this aids interpretability of the signals, being of particular importance for applications in biomedical signal analysis. A particular advantage of the linear dynamics approach is the tractability of inference. With an application to raw EEG data, we have shown that this method is able to automatically extract non-redundant independent rhythmical activity related to a mental task from multiple channels of observation. The ease of incorporating desired spectral preferences in the extracted components is shown to be of some benefit compared to the related NDFA method.

Previous implementations of the variational Bayesian Linear Gaussian State-Space Model (LGSSM) have been achieved using Belief Propagation, which differs from inference in the Kalman filtering/smoothing literature, for which highly efficient and stabilized procedures exist. A central contribution of this Chapter is to show how inference *can* be written using the standard Kalman filtering/smoothing recursions by augmenting the original model. Additionally, a minor modification to the standard Kalman filtering routine may be applied for computational efficiency. This simple way to implement the approximate Bayesian procedure is of considerable interest given the widespread applications of LGSSMs.

One disadvantage of the current model is that some signals (or artifacts in EEG) may be so complex that they are difficult to model with a stationary state-space model. One possibility would be extend the Bayesian analysis to a switching factorial LGSSM. A Bayesian treatment of this model could be relatively straightforward. Furthermore, even if such a model is intractable, a novel stable approximation has recently been introduced [Barber (2004)].

Chapter 8

Conclusions and Future Directions

This thesis investigated several aspects related to the design of techniques for analyzing and classifying EEG signals. The general goal was to test different approaches for classification and to introduce and analyze methods which incorporate basic prior knowledge about EEG signals into a principled framework. This was performed using a probabilistic framework. In general, working with EEG signals is a very challenging and difficult task, and general statements have to be taken with some care. Bearing this in mind, here we draw some tentative conclusions and suggest some possible research directions.

The first issue that we investigated was the classification of EEG rhythms generated by three mental tasks using standard 'black-box' methods from the machine learning literature. In particular, we performed a comparison between a generative and a discriminative approach for the case in which no prior information about the EEG signal is incorporated into the model structure. We used an approach which is common in the analysis of EEG rhythms and consists of extracting spectral features in a defined frequency band from the raw temporal data which are then fed into a separate classifier. To take potential advantage of the temporal nature of EEG, we used two temporal models: the generative Hidden Markov Model (HMM) and discriminative Input-Output HMM (IOHMM). From a technical point of view, we contributed to the development of the IOHMM model by introducing a new 'apposite' training algorithm, for the case in which a class for each input of the training sequence is specified. This was necessary, since in our EEG data each training sequence corresponds only to a single class and, in this case, using maximum likelihood training is inappropriate, since it would waste model resources on maintaining consecutive same-class outputs. Our apposite objective function encourages model resources rather to be spent on discriminating between sequences in which the same class label is replicated through all the time-steps. Furthermore, the apparently difficult problem of computing the gradient of the new objective function was transformed into simpler subproblems.

The new apposite training algorithm significantly improves the performance relative to the standard approach previously presented in the literature in which class label is given only at the end of the sequence.

From the classification performance, the discriminative approach taken is preferential to the generative one. Whilst the reason for the disadvantage of the generative approach arguably lies in the lack of competition between the models based on each class, the advantages of the generative framework, namely the ease of incorporation of prior knowledge about the problem domain, were not well exploited.

The next work addressed the issue of whether the incorporation of prior knowledge about the problem domain into a generative model can be beneficial for classification with respect to using a 'black-block' generative model. The prior beliefs about the EEG signal that we used is the widely accepted assumption that EEG signals result from linear mixing of independent activity in the brain and from other components, such as artifacts or external noise. The resulting model is a form of generative Independent Component Analysis (gICA) which can be used for direct classification of EEG. We have applied this model to the classification of two different EEG datasets. The first dataset was recorded while users were performing three mental tasks, while the second dataset contained two real motor tasks. For users which perform sufficiently well, gICA performs better than a discriminative model in which independence information about EEG is not used; while for users which perform already relatively well gICA gives no advantage; and for users which perform badly the performance of gICA may even be slightly worse. We have also seen that a standard approach, in which ICA is performed before extracting spectral features which are fed into a discriminative classifier, and gICA, which uses the filtered EEG times-series, perform similarly. This may appear surprising, since a common belief is that classifying directly the EEG time-series is not very robust and extracting power spectral density information combined with a discriminative approach is more powerful. In general, we believe that, given the nature of the problem, the accuracy of spectral information does not play a big role in terms of performance and that, given the obtained results, principle methods which classify directly the EEG time-series deserve more attention.

A potentially important advantage of using a generative approach is the fact that a different mixing matrix for each mental task can be created. In this respect, the generative ICA model is more powerful than a model in which the same matrix is computed for all classes and then features are extracted and fed into a classifier. This intuition seems to be supported also by the experimental results in which we compared the performance of gICA, in which a unique mixing matrix is computed for all classes, with the case in which a different matrix for each mental task is estimated. Another aspect which was highlighted by the experiments is the fact that the incorporation of independence assumptions seems to be more beneficial for within-day experiments, but the advantage may be lost when training and testing are performed in different days. We should take into consideration that the persons that partecipated to the experiments never used a BCI system before, thus we expect a change in the EEG signals between sessions and days which is higher than the one that would appear after some user training. We have tried to address this problem of EEG variability using a mixture of gICA models. In this case, several gICA models were created, each with a certain probability of having generated the observations. In principle, this model is more powerful than a simple gICA and can be used when EEG signal can be grouped into different regimes, as may be the case when the user employs different mental strategies or the recording conditions change. However, in the experiments the resulting mixture model does not improve much on the basic method.

In the second part of this thesis we developed a method for identifying independent dynamical sources. In particular, such a tool can be used to denoise EEG from artifacts, to spatially filter the signal, to select mental-task related subsignals and to analyze the source generators in the brain, thereby aiding the visualization and interpretation of the mental states. Unlike many other independent component extraction methods in which the number of components and channels has to be the same, in our model the number of components that can be extracted is independent of the number of channels. Especially when working with relatively few electrodes, the constraint of current extraction methods on the component number can be a strong limitation, since the EEG signal contains potentially many independent activities. As a consequence, current approaches normally require prior filtering on each channel to select only frequencies of interest and remove other activity, in order to reduce the total number of components. Our Factorial Linear Gaussian State-Space Model (FLGSSM) does not have this limitation and therefore can work on the raw unfiltered EEG data. This can be beneficial, since filtering may unintentionally remove important information from the signal useful for identifying independent components. A strength of our approach is a Bayesian procedure which automatically biases the solution to finding a small number of components with preferential spectral properties. With an application to raw unfiltered EEG data, we have shown that this method is able to automatically extract non redundant independent rhythmical activity related to a mental task from multiple channels of observations and isolate task-related information better than other related temporal models.

From a technical point of view, an important difference between our variational Bayesian treatment of the LGSSM and others previously presented in the literature is the simplicity and stability of the recursive formulas for computing the statistics of the hidden variables. Indeed the problem of estimating required statistics was transformed into the problem of estimating the hidden posteriors of a LGSSM with modified output dimension. This could thus be performed using standard forward and backward recursions.

A potential limitation of the Bayesian FLGSSM is the fact that, while it can model very complicated distributions, components which have a too complex or changing dynamics (such as in the case of some EEG artifacts) will be considered as noise and modeled by a Gaussian distribution. A possible solution would be to extend the Bayesian analysis to a switching FLGSSM [Bar-Shalom and Li (1998)]. In a switching factorial linear Gaussian state-space model, at each time step, a hidden variable determines which of a set of FLGSSMs is more suitable for generating the observations. This model is used for complex time-series which are characterized by different dynamical regimes. Even if this model is intractable, a stable form of approximation has recently been introduced [Barber (2004)]..

Another interesting direction would be to use the proposed Bayesian approach to classify the EEG signals. However, this requires improvements on the current algorithm to speedup convergence and make the model more practical. At the current stage, this model is more appropriate for the analysis of the signal since this does not require the use of as much EEG data as needed for training a classifier.

The type of prior information about EEG generation that we have incorporated in our generative approach was the simple assumption that the EEG signal is the result of a linear instantaneous mixing of independent components. There are more specific assumptions which are currently used in EEG analysis. In particular there exists a research area which deals with the problem of estimating the generators of the brain electromagnetic activity measured at the scalp. This problem is called EEG inverse problem [Grave de Peralta Menendez et al. (2005)]. In this case, the matrix which defines the mapping from the internal generators into the scalp is given by the physical laws which describe the propagation of the electromagnetic fields from the brain to the scalp. That is, unlike the method discussed in this thesis, the mixing matrix is given on the basis of prior physical knowledge. However, the number of generators in the brain is assumed to be much higher than the number of electrodes. The goal is therefore to estimate the best solution out of the infinite possibilities by adding additional constraints to the problem. Common approaches use least-squared type methods and incorporate spatial, temporal, neurophysiological or biophysical constraints [Galka et al. (2004); Grave de Peralta Menendez et al. (2005)]. The type of solution depends strongly on the kind of constraints which are incorporated. Recently a temporal model was used and was shown to improve over static reconstruction [Galka et al. (2004)]. If we use the given mixing matrix in our factorial LGSSM, we obtain an estimator of the hidden generators which use temporal information. A Bayesian approach similar to the one used in the thesis could improve the quality of the extracted solutions obtained by current approaches by incorporating prior knowledge about temporal dynamics.

Appendix A

A.1 Inference and Learning the Gaussian HMM

In the Gaussian HMM learning is performed using the Expectation Maximization (EM) algorithm [McLachlan and Krishnan (1997)]. At iteration i, the complete data log-likelihood is given by:

$$\begin{aligned}\mathcal{Q}(\Theta, \Theta^{i-1}) &= \langle \log p(q_{1:T}, m_{1:T}, v_{1:T}, \Theta)\rangle_{p(q_{1:T}, m_{1:T}, |v_{1:T}, \Theta^{i-1})}\\ &= \sum_{t=1}^{T} \langle \log p(v_t|q_t, m_t, \Theta)\rangle_{p(q_t, m_t|v_{1:T}, \Theta^{i-1})}\\ &+ \sum_{t=1}^{T} \langle \log p(m_t|q_t, \Theta)\rangle_{p(q_t, m_t|v_{1:T}, \Theta^{i-1})}\\ &+ \sum_{t=2}^{T} \langle \log p(q_t|q_{t-1}, \Theta)\rangle_{p(q_{t-1:t}|v_{1:T}, \Theta^{i-1})}\\ &+ \langle \log p(q_1|\Theta)\rangle_{p(q_1|v_{1:T}, \Theta^{i-1})} \ .\end{aligned}$$

The terms $p(q_t, m_t|v_{1:T}, \Theta^{i-1})$ and $p(q_{t-1:t}|v_{1:T}, \Theta^{i-1})$ are found using the following recursions (where we omit the dependence on Θ^{i-1}):

Forward Recursions:

$$\begin{aligned}p(q_t, m_t|v_{1:t}) &\propto p(q_t, m_t, v_t|v_{1:t-1})\\ &= p(v_t|q_t, m_t)p(m_t|q_t)\sum_{q_{t-1}} p(q_{t-1:t}|v_{1:t-1})\\ &= p(v_t|q_t, m_t)p(m_t|q_t)\sum_{q_{t-1}} p(q_t|q_{t-1}, v_{1:t-1})p(q_{t-1}|v_{1:t-1})\\ &= p(v_t|q_t, m_t)p(m_t|q_t)\sum_{q_{t-1}} p(q_t|q_{t-1})\sum_{m_{t-1}} p(q_{t-1}, m_{t-1}|v_{1:t-1})\,,\end{aligned}$$

where the proportionality constant is found by normalization.

Backward Recursions:

$$\begin{aligned}
p(q_t, m_t|v_{1:T}) &= \sum_{q_{t+1}} p(q_{t:t+1}, m_t|v_{1:T})\\
&= \sum_{q_{t+1}} p(q_t, m_t|q_{t+1}, v_{1:T}) \sum_{m_{t+1}} p(q_{t+1}, m_{t+1}|v_{1:T})\\
&= \sum_{q_{t+1}} p(q_t, m_t|q_{t+1}, v_{1:t}) \sum_{m_{t+1}} p(q_{t+1}, m_{t+1}|v_{1:T})\\
&= \sum_{q_{t+1}} \frac{p(q_{t:t+1}, m_t|v_{1:t})}{\sum_{q_t} p(q_{t:t+1}, m_t|v_{1:t})} \sum_{m_{t+1}} p(q_{t+1}, m_{t+1}|v_{1:T})\\
&= \sum_{q_{t+1}} \frac{p(q_{t+1}|q_t, v_{1:t}) p(q_t, m_t|v_{1:t})}{\sum_{q_t} p(q_{t+1}|q_t, v_{1:t}) p(q_t, m_t|v_{1:t})} \sum_{m_{t+1}} p(q_{t+1}, m_{t+1}|v_{1:T})\\
&= \sum_{q_{t+1}} \frac{p(q_{t+1}|q_t) p(q_t, m_t|v_{1:t})}{\sum_{q_t} p(q_{t+1}|q_t) p(q_t, m_t|v_{1:t})} \sum_{m_{t+1}} p(q_{t+1}, m_{t+1}|v_{1:T})\,.
\end{aligned}$$

In order to find an update for the parameters, we have to find the derivative of $\mathcal{Q}(\Theta, \Theta^{i-1})$ with respect to the mean and covariance of $p(v_t|q_t, m_t)$; with respect to $p(q_t|q_{t-1})$ under the constraint $\sum_{q_t} p(q_t|q_{t-1}) = 1$; with respect to $p(q_1)$ under the constraint $\sum_{q_1} p(q_1) = 1$; and with respect to $p(m_t|q_t)$ under the constraint $\sum_{m_t} p(m_t|s_t) = 1$. The final updates are given by:

$$\begin{aligned}
\mu_{q_t,m_t} &= \frac{\sum_{t=1}^T v_t p(q_t, m_t|v_{1:T}, \Theta^{i-1})}{\sum_{t=1}^T p(q_t, m_t|v_{1:T}, \Theta^{i-1})}\\
\Sigma_{q_t,m_t} &= \frac{\sum_{t=1}^T (v_t - \mu_{q_t,m_t})(v_t - \mu_{q_t,m_t})^\mathsf{T} p(q_t, m_t|v_{1:T}, \Theta^{i-1})}{\sum_{t=1}^T p(q_t, m_t|v_{1:T})}\\
p(q_t|q_{t-1}) &= \frac{\sum_{t=2}^T p(q_{t-1:t}|v_{1:T}, \Theta^{i-1})}{\sum_{t=2}^T \sum_{q_t} p(q_{t-1:t}|v_{1:T}, \Theta^{i-1})}\\
p(q_1) &= \frac{\sum_{t=1}^T p(q_1|v_{1:T}, \Theta^{i-1})}{T}\\
p(m_t|q_t) &= \frac{\sum_{t=1}^T p(q_t, m_t|v_{1:T}, \Theta^{i-1})}{\sum_{t=1}^T \sum_{m_t} p(q_t, m_t|v_{1:T}, \Theta^{i-1})}\,.
\end{aligned}$$

A.2 Matrix Inversion Lemma

If the matrices A, B, C, D satisfy

$$B^{-1} = A^{-1} + C^{\mathsf{T}} D^{-1} C \,, \tag{A.1}$$

where all inverses are assumed to exist, then

$$B = A - AC^{\mathsf{T}}(CAC^{\mathsf{T}} + D)^{-1}CA \,. \tag{A.2}$$

Indeed by pre-multiplying A.1 by B and post-multipling by A and C^{T} we obtain:

$$A = B + BC^{\mathsf{T}} D^{-1} CA \tag{A.3}$$

$$AC^{\mathsf{T}} = BC^{\mathsf{T}} + BC^{\mathsf{T}} D^{-1} CAC^{\mathsf{T}} = BC^{\mathsf{T}} D^{-1}(D + CAC^{\mathsf{T}}) \,. \tag{A.4}$$

Post-multiplying A.4 by $(D + CAC^{\mathsf{T}})^{-1}CA$ and subtracting the resulted quantities from A we obtain:

$$A - AC^{\mathsf{T}}(D + CAC^{\mathsf{T}})^{-1}CA = A - BC^{\mathsf{T}} D^{-1} CA = B \,,$$

where for the last equality we have used A.3. If A and B are $n \times n$ matrices, C is $m \times n$ and D is $m \times m$, then the computation of B using A.2 requires the inversion of one $m \times m$ matrix, while the computation from A.1 requires the inversion of one $m \times m$ and two $n \times n$ matrices.

A.3 Gaussian Random Variables: Moment and Canonical Representation

The **moment** representation of a Gaussian distribution is of form:

$$p(x) = \frac{1}{\sqrt{\det(2\pi\Sigma_x)}} e^{-\frac{1}{2}(x-\mu_x)^{\mathsf{T}}\Sigma_x(x-\mu_x)} ,$$

where μ_x and Σ_x are the mean and covariance. The **canonical** representation of a Gaussian distribution is of form:

$$p(x) \propto e^{-\frac{1}{2}(x^{\mathsf{T}} M_x x - 2x^{\mathsf{T}} m_x)} .$$

The relation between the two representation is given by: $\mu_x = M_x^{-1} m_x$ and $\Sigma_x = M_x^{-1}$.

A.4 Jointly Gaussian Random Variables

Let x and y be jointly Gaussian random variables, $x \sim \mathcal{N}(\mu_x, \Sigma_x)$, $y \sim \mathcal{N}(\mu_y, \Sigma_y)$. Let Σ_{xy} denote the cross-covariance, that is $\Sigma_{xy} = \left\langle (x-\mu_x)(y-\mu_y)^{\mathsf{T}} \right\rangle_{p(x,y)}$.
Consider the random vector z formed by concatenating the two variables x and y, $z = [x, y]$. Then $\mu_z = [\mu_x, \mu_y]$ and

$$\Sigma_z = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{pmatrix} .$$

The components of Σ_z^{-1} can be expressed as:

$$\Sigma_z^{-1} = \begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix} ,$$

with

$$\begin{aligned} A &= (\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})^{-1} = \Sigma_x^{-1} + \Sigma_x^{-1}\Sigma_{xy}C\Sigma_{yx}\Sigma_x^{-1} && \text{(A.5)} \\ B &= -A\Sigma_{xy}\Sigma_y^{-1} = -\Sigma_x\Sigma_{xy}C \\ C &= (\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})^{-1} = \Sigma_y^{-1} + \Sigma_y^{-1}\Sigma_{xy}A\Sigma_{yx}\Sigma_y^{-1} , && \text{(A.6)} \end{aligned}$$

where in Eq. (A.5) and Eq. (A.6) we have made use of the matrix inversion lemma. This can be easily shown by forming $\Sigma_z\Sigma_z^{-1} = I$ and equating elements on both sides.

A.4.1 The Marginal Density Function

Consider two jointly Gaussian random variables x and y with joint distribution expressed in the canonical form as:

$$p(x,y) \propto e^{-\frac{1}{2}\left[\begin{pmatrix} x \\ y \end{pmatrix}^{\mathsf{T}} \underbrace{\begin{pmatrix} N_x & N_{xy} \\ N_{yx} & N_y \end{pmatrix}}_{M_z} \begin{pmatrix} x \\ y \end{pmatrix} -2 \begin{pmatrix} x \\ y \end{pmatrix}^{\mathsf{T}} \underbrace{\begin{pmatrix} n_x \\ n_y \end{pmatrix}}_{m_z} \right]} .$$

This means that the moment form of the mean is $\mu_z = M_z^{-1}m_z$ and covariance is $\Sigma_x = M_z^{-1}$. By using Eq. (A.5), we find that the marginal distribution $p(x)$ has mean and covariance:

$$\begin{aligned} \mu_x &= (N_x - N_{xy}N_y^{-1}N_{yx})^{-1}(n_x - N_{xy}N_y^{-1}n_y) && \text{(A.7)} \\ \Sigma_x &= (N_x - N_{xy}N_y^{-1}N_{yx})^{-1} , \end{aligned}$$

and the canonical form is:

$$\begin{aligned} m_x &= n_x - N_{xy} N_y^{-1} n_y \\ M_x &= N_x - N_{xy} N_y^{-1} N_{yx} \,. \end{aligned}$$

A.4.2 The Conditional Density Function

Let x and y be jointly Gaussian vectors with means μ_x and μ_y and covariance Σ_x and Σ_y respectively. Then $p(x|y)$ is also Gaussian with mean and covariance:

$$\begin{aligned} \mu_{x|y} &= \langle x \rangle_{p(x|y)} = \mu_x + \Sigma_{xy} \Sigma_y^{-1} (y - \mu_y) \\ \Sigma_{x|y} &= \left\langle (x - \mu_{x|y})(x - \mu_{x|y})^\mathsf{T} \right\rangle_{p(x|y)} = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \end{aligned}$$

This can be shown by using the following formula:

$$p(x|y) = \frac{p(x,y)}{p(y)},$$

from which we obtain

$$p(x|y) = \frac{1}{\sqrt{\frac{\det(2\pi\Sigma_z)}{\det(2\pi\Sigma_y)}}} e^{-\frac{1}{2}(z-\mu_z)^\mathsf{T} \begin{pmatrix} A & B \\ B^T & C - \Sigma_y^{-1} \end{pmatrix} (z-\mu_z)} \,.$$

The quadratic exponent can be written as:

$$\begin{aligned} &(x-\mu_x)^\mathsf{T} A (x-\mu_x) + 2(x-\mu_x)^\mathsf{T} B (y-\mu_y) + (y-\mu_y)^\mathsf{T} (C - \Sigma_y^{-1})(y-\mu_y) \\ &= (x-\mu_x)^\mathsf{T} A (x-\mu_x) - 2(x-\mu_x) A \Sigma_{xy} \Sigma_y^{-1} (y-\mu_y) + (y-\mu_y)^\mathsf{T} \Sigma_y^{-1} \Sigma_{yx} A \Sigma_{xy} \Sigma_y^{-1} (y-\mu_y) \\ &= (x - \mu_x - \Sigma_{xy}\Sigma_y^{-1}(y-\mu_y))^\mathsf{T} A (x - \mu_x - \Sigma_{xy}\Sigma_y^{-1}(y-\mu_y)) \,, \end{aligned}$$

which looks like a quadratic expression in $(x - \mu_x)$.
We have to show that $\det \Sigma_z / \det \Sigma_y = \det(\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})$. In order to do that, we consider the following factorization:

$$\Sigma_z = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{pmatrix} = \begin{pmatrix} L & \Sigma_{xy} \\ 0 & \Sigma_y \end{pmatrix} \begin{pmatrix} I & 0 \\ U & I \end{pmatrix} = \begin{pmatrix} L + \Sigma_{xy} U & \Sigma_{xy} \\ \Sigma_y U & \Sigma_y \end{pmatrix} .$$

which implies that $U = \Sigma_y^{-1}\Sigma_{yx}$ and $L = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} = \Sigma_{x|y}$. From the factorization we see that $\det \Sigma_z = \det L \det \Sigma_y = \det \Sigma_{x|y} \det \Sigma_y$.

A.5 Learning Linear Gaussian State-Space Model Parameters

The EM algorithm for learning the parameters of a linear Gaussian state-space model was first introduced in Shumway and Stoffer (1982). The expectation of the complete-data log-likelihood for M sequences $v^m_{1:T_m}$ has the following form (we omit the dependency on m):

$$
\begin{aligned}
\mathcal{Q} &= \left\langle \log \prod_{m=1}^{M} p(v_{1:T}, h_{1:T}) \right\rangle_{p(h_{1:T}|v_{1:T})} \\
&= \sum_{m=1}^{M}\sum_{t=1}^{T} \langle \log p(v_t|h_t)\rangle_{p(h_t|v_{1:T})} + \sum_{m=1}^{M}\sum_{t=2}^{T} \langle \log p(h_t|h_{t-1})\rangle_{p(h_{t-1:t}|v_{1:T})} + \langle \log p(h_1)\rangle_{p(h_1|v_{1:T})} \\
&= MT \log \frac{1}{\sqrt{|2\pi\Sigma_V|}} - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T} \left\langle (v_t - Bh_t)^{\mathsf{T}} \Sigma_V^{-1} (v_t - Bh_t) \right\rangle_{p(h_t|v_{1:T})} \\
&+ M(T-1) \log \frac{1}{\sqrt{|2\pi\Sigma_H|}} - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=2}^{T} \left\langle (h_t - Ah_{t-1})^{\mathsf{T}} \Sigma_H^{-1} (h_t - Ah_{t-1}) \right\rangle_{p(h_{t-1:t}|v_{1:T})} \\
&+ M \log \frac{1}{\sqrt{|2\pi\Sigma|}} - \frac{1}{2}\sum_{m=1}^{M} \left\langle (h_1 - \mu)^{\mathsf{T}} \Sigma^{-1} (h_1 - \mu) \right\rangle_{p(h_1|v_{1:T})} .
\end{aligned}
$$

The updates of the parameters are estimated by setting to zero the following derivatives:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \Sigma_H^{-1}} &= \frac{1}{2} M(T-1)\Sigma_H - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=2}^{T} \left\langle (h_t - Ah_{t-1})(h_t - Ah_{t-1})^{\mathsf{T}} \right\rangle_{p(h_{t-1:t}|v_{1:T})} \\
\frac{\partial \mathcal{Q}}{\partial \Sigma_V^{-1}} &= \frac{1}{2} MT\Sigma_V - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T} \left\langle (v_t - Bh_t)(v_t - Bh_t)^{\mathsf{T}} \right\rangle_{p(h_{1:T}|v_{1:T})} \\
\frac{\partial \mathcal{Q}}{\partial \Sigma^{-1}} &= \frac{1}{2} M\Sigma - \frac{1}{2}\sum_{m=1}^{M} \left\langle (h_1 - \mu)(h_1 - \mu)^{\mathsf{T}} \right\rangle_{p(h_1|v_{1:T})} \\
\frac{\partial \mathcal{Q}}{\partial \mu} &= \Sigma^{-1} \left(\sum_{m=1}^{M} \langle h_1 \rangle_{p(h_1|v_{1:T})} - \mu \right) \\
\frac{\partial \mathcal{Q}}{\partial A} &= \Sigma_H^{-1} \sum_{m=1}^{M}\sum_{t=2}^{T} \left\langle (h_t - Ah_{t-1}) h_{t-1}^{\mathsf{T}} \right\rangle_{p(h_{t-1:t}|v_{1:T})} \\
\frac{\partial \mathcal{Q}}{\partial B} &= \Sigma_V^{-1} \sum_{m=1}^{M}\sum_{t=1}^{T} \left\langle (v_t - Bh_t) h_t^{\mathsf{T}} \right\rangle_{p(h_t|v_{1:T})} .
\end{aligned}
$$

Bibliography

C. W. Anderson. Effects of variations in neural network topology and output averaging on the discrimination of mental tasks from spontaneous electroencephalogram. *Journal of Intelligent Systems*, 7:165–190, 1997.

C. W. Anderson and M. Kirby. EEG subspace representations and feature selection for brain-computer interfaces. In *1st IEEE Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, 2003.

H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1999.

Y. Bar-Shalom and X.-R. Li. *Estimation and Tracking: Principles, Techniques and Software.* Artech House, 1998.

D. Barber. A stable switching Kalman smoother. IDIAP-RR 89, IDIAP, 2004.

BCI Competition I. liinc.bme.columbia.edu/competition.htm, 2001.

BCI Competition II. ida.first.fhg.de/projects/bci/competition, 2003.

BCI Competition III. ida.first.fraunhofer.de/projects/bci/competition_iii, 2004.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

Y. Bengio and P. Frasconi. Input-output HMMs for sequence processing. *IEEE Transactions on Neural Networks*, 7:1231–1249, 1996.

Y. Bengio, V.-P. Lauzon, and R. Ducharme. Experiments on the applications of IOHMMs to model financial returns series. *IEEE Transactions on Neural Networks*, 12:113–123, 2001.

H. Berger. Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87:527–570, 1929. (Translation from P. Gloor: Hans Berger on the electroencephalogramm of man. *Electroencephalography and Clinical Neurophysiology*, Supp. 28, pages 37-73).

N. Birbaumer, A. Kübler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor. The thought translation device (TTD) for completely paralyzed patients. *IEEE Transactions on Rehabilitation Engineering*, 8:190–193, 2000.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

G. Blanchard and B. Blankertz. BCI competition 2003 - Data set IIa: spatial patterns of self-controlled brain rhythm modulations. *IEEE Transactions on Biomedical Engineering*, 51: 1062–1066, 2004.

B. Blankertz, G. Curio, and K.-R. Müller. Classifying single trial EEG: Towards brain computer interfacing. In *Advances in Neural Information Processing Systems*, pages 157–164, 2002.

B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schölgl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer. The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *IEEE Transactions on Biomedical Engineering*, 51:1044–1051, 2004.

K. Brodmann. *Vergleichende Lokalisationslehre der Großhirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues.* Barth, Leipzig, 1909.

J.-F. Cardoso. On the stability of source separation algorithms. In *Workshop on Neural Networks for Signal Processing*, pages 13–22, 1998.

J. M. Carmena, M. A. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. F. Dimitrov, P. G. Patil, C. S. Henriquez, and M. A. L. Nicolelis. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biology*, 1:193–208, 2003.

A. T. Cemgil and S. J. Godsill. Probabilistic phase vocoder and its application to interpolation of missing values in audio signals. In *13th European Signal Processing Conference*, 2005.

M. Cheng, X. Gao, S. Gao, and D. Xu. Design and implementation of a brain-computer interface with high transfer rates. In *Workshop on Neural Networks for Signal Processing*, pages 1181–1186, 2002.

S. Chiappa and D. Barber. Generative independent component analysis for EEG classification. In *13th European Symposium on Artificial Neural Networks*, pages 297–302, 2005a.

S. Chiappa and D. Barber. Generative temporal ICA for classification in asynchronous BCI systems. In *2nd International IEEE EMBS Conference On Neural Engineering*, pages 514–517, 2005b.

S. Chiappa and D. Barber. EEG classification using generative independent component analysis. *Neurocomputing*, 69:769–777, 2006.

S. Chiappa and D. Barber. Bayesian factorial linear Gaussian state-space models for biosignal decomposition. *Signal Processing Letters*, 2007. To appear.

S. Chiappa and S. Bengio. HMM and IOHMM modeling of EEG rhythms for asynchronous BCI systems. In *12th European Symposium on Artificial Neural Networks*, pages 199–204, 2004.

C. K. Chui. *An Introduction to Wavelets.* Academic Press, 1992.

R.P. Crease. Images of conflict: MEG vs. EEG. *Science*, 253:374–375, 1991.

N. Cristianini and J. S. Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

C. Cuadras, J. Fortiana, and F. Oliva. The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification*, 14:117–136, 1997.

E. A. Curran and M. J. Stokes. Learning to control brain activity: A review of the production and control of EEG components for driving brain-computer interface (BCI) systems. *Brain and Cognition*, 51:326–336, 2003.

A. Delorme and S. Makeig. EEG changes accompanying learned regulation of 12-Hz EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engeneering*, 11:133–137, 2003.

E. Donchin, K. M. Spencer, and R. Wijesinghe. The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering*, 8: 174–179, 2000.

J.P. Donoghue. Connecting cortex to machines: recent advances in brain interfaces. *Nature Neuroscience*, 5:1085–1088, 2002.

G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Increase information transfer rates in BCI by CSP extension to multi-class. In *Advances in Neural Information Processing Systems (NIPS)*, pages 733–740, 2003.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods.* Oxfor University Press, 2001.

J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

L. A. Farwell and E. Donchin. Talking off the top your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70:510–523, 1998.

S. Finger. *Origin of Neuroscience: A History of Explorations Into Brain Function.* Oxford University Press, 1994.

A. Flexer, P. Sykacek, I. Rezek, and G. Dorffner. Using hidden Markov models to build an automatic, continuous and probabilistic sleep stager. In *IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, pages 3627–3631, 2000.

J. H. Friedman. Regularized discriminant analysis. *Journal of American Statistical Association*, 84:165–175, 1989.

K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

A. Galka, O. Yamashita, T. Ozaki, R. Biscay, and P. Valdés-Sosa. A solution to the dynamical inverse problem of EEG generation using spatiotemporal Kalman filtering. *NeuroImage*, 23: 435–453, 2004.

D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *Neural Systems and Rehabilitation Engineering*, 11:141–144, 2003.

S. Georgiadis, P. O. Ranta-aho, M. P. Tarvainen, and P. A. Karjalainen. Single-trial dynamical estimation of event-related potentials: A Kalman filter-based approach. *IEEE Transactions on Biomedical Engineering*, 52:1397–1406, 2005.

M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13:2517–2532, 2001.

R. Grave de Peralta Menendez, S. Gonzalez Andino, M. M. Murray, G. Thut, and T. Landis. *Non-invasive Estimation of Local Field Potentials: Methods and Applications.* Oxford University Press, 2005.

M. S. Grewal and A. P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB.* John Wiley and Sons, Inc., 2001.

A. Hauser, P.-E. Sottas, and J. del R. Millán. Temporal processing of brain activity for the recognition of EEG patterns. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 1125–1130, 2002.

B. Hjort. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29:206–310, 1970.

T. Hoya, G. Hori, H. Bakardjian, T. Nishimura, T. Suzuki, Y. Miyawaki, A. Funase, and J. Cao. Classification of single trial EEG signals by a combined principal + independent component analysis and probabilistic neural network approach. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 197–202, 2003.

http://www.biosemi.com.

http://www.sccn.ucsd.edu/eeglab.

S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging.* Sinauer Associates, 2004.

C.-I Hung, P.-L. Lee, Y.-T. Wu, H.-Y. Chen, L.-F Chen, T.-C. Yeh, and J.-C. Hsieh. Recognition of motor imagery electroencephalography using independent component analysis and machine classifiers. *Annals of Biomedical Engineering*, 33:1053–1070, 2005.

A. Hyvärinen. Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22:49–67, 1998.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley and Sons, 2001.

P. Hjen-Srensen, L. K. Hansen, and O. Winther. Mean field implementation of Bayesian ICA. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, pages 439–444, 2001.

M. Jahanshahi and M. Hallett. *The Bereitschaftspotential, movement-related cortical potentials.* Kluwer Academic/Plenum, 2003.

H. H. Jasper. Ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–373, 1958.

T. P. Jung, C. Humphries, T. W. Lee, S. Makeig, M. J. McKeown, V. Iragui, and T. Sejnowski. Extendend ICA removes artifacts from electroencephalografic recordings. *Advances in Neural Information Processing Systems*, pages 894–900, 1998.

P. R. Kennedy, R. A. Bakay, M. M. Moore, K. Adams, and J. Goldwaithe. Direct control of a computer from the human central nervous system. *IEEE Transactions on Rehabilitation Engineering*, 8:198–202, 2000.

A. Kostov and M. Polak. Parallel man-machine training in development of EEG-based cursor control. *IEEE Transactions on Rehabilitation Engineering*, 8:203–205, 2000.

A. Kübler, B. Kotchoubey, J. Kaiser, J. R. Wolpaw, and N. Birbaumer. Brain-computer communication: unlocking the locked in. *Psychological Bulletin*, 3:358–375, 2002.

S. Lauritzen. *Graphical Models.* Oxford Science Publications, 1996.

T.-W. Lee and M. S. Lewicki. The generalized Gaussian mixture model using ICA. In *International Workshop on Independent Component Analysis*, pages 239–244, 2000.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

D. J. C. MacKay. Ensemble learning and evidence maximization. Unpublished manuscipt: www.variational-bayes.org, 1995.

D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Unpublished manuscipt: www.inference.phy.cam.ac.uk/mackay/BayesICA.html, 1999.

S. Makeig, S. Enghoff, T.-P. Jung, and T. J. Senjowski. A natural basis for efficient brain-actuated control. *IEEE Transactions on Rehabilitation Engineering*, 8:208–211, 2000.

S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295:690–694, 2002.

J. Malmivuo, V. Suihko, and H. Eskola. Sensitivity distributions of EEG and MEG measurements. *IEEE Transactions on Biomedical Engineering*, 44:196–208, 1997.

K. V. Mardia. *Multivariate Analysis.* Academic Press, 1979.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley and Sons, 1997.

P. Meinicke, M. Kaper, F. Hoppe, and H. Ritter. Improving transfer rates in brain computer interfacing: A case study. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1107–1114, 2002.

J. M. Mendel. *Lessons in estimation theory for signal processing, communications, and control.* Prentice Hall PTR, 1995.

B. Mensh, J. Werfel, and H. S. Seung. BCI competition 2003–Data set Ia: Combining gamma-band power with slow cortical potentials to improve single-trial classification of electroencephalographic signals. *IEEE Transactions on Biomedical Engineering*, 51:1052–1055, 2004.

M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones. Brain-computer interfaces based on steady-state visual evoked response. *IEEE Transactions on Rehabilitation Engineering*, 8: 211–213, 2000.

J. del R. Millán. Adaptive brain interfaces. *Communications of the ACM*, 46:74–80, 2003.

J. del R. Millán, J. Mouriño, M. Franzé, F. Cincotti, M. Varsta, J. Heikkonen, and F. Babiloni. A local neural classifier for the recognition of EEG patterns associated to mental tasks. *IEEE Transactions on Neural Networks*, 13:678–686, 2002.

M. Morf and T. Kailath. Square-root algorithms for least-squares estimation. *IEEE Transactions on Automatic Control*, 20:487–497, 1975.

K.-R. Müller, C. Anderson, and G. Birch. Linear and non-linear methods in brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitative Engineering*, 11:162–165, 2003.

R. M. Neal and G. E. Hinton. *Learning in Graphical Models.* A view of the EM algorithm that justifies incremental, sparse, and other variants, pages 355–368. Kluwer Academic, 1998.

E. Neidermeyer. *Electroencephalography: basic principles, clinical applications and related fields.* The normal EEG of the waking adult, pages 149–173. Lippincott Williams and Wilkins, 1999.

M.A.L. Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408:361–365, 2001.

E. Niedermeyer and F. Lopes Da Silva. *Electroencephalography: basic principles, clinical applications and related fields.* Lippincott Williams and Wilkins, 1999.

P. L. Nunez. *Neocortical Dynamics and Human EEG Rhythms.* Oxford University Press, 1995.

B. Obermaier. *Design and implementation of an EEG based virtual keyboard using hidden Markov models.* Ph.D. thesis, University of Technology, Graz, Austria, 2001.

B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller. Hidden Markov models for online classification of single trial EEG data. *Pattern Recognition Letters*, 22:1299–1309, 2001a.

B. Obermaier, C. Guger, and G. Pfurtscheller. HMM used for the offine classification of EEG data. *Biomedizinsche Technik*, 44:158–162, 1999.

B. Obermaier, G. R. Müller, and G. Pfurtscheller. "Virtual Keyboard" controlled by spontaneous EEG activity. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 636–641, 2001b.

B. Obermaier, G. R. Müller, and G. Pfurtscheller. "Virtual Keyboard" controlled by spontaneous EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11:422–426, 2003.

P. Park and T. Kailath. New square-root smoothing algorithms. *IEEE Transactions on Automatic Control*, 41:727–732, 1996.

B. A. Pearlmutter and L. C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In *Advances in Neural Information Processing Systems*, pages 613–619, 1997.

W. D. Penny and S. J. Roberts. Bayesian neural networks for detection of imagined finger movements from single-trial EEG. Technical report, Department of Electrical Engineering, Imperial College, London, 1997.

W. D. Penny, S. J. Roberts, and R. M. Everson. Hidden Markov independent components for biosignal analysis. In *International Conference on Advances in Medical Signal and Information Processing*, pages 244–250, 2000.

B. Pesaran and S. Musallam R.A. Andersen. Cognitive neural prosthetics. *Current Biology*, 16: 77–80, 2006.

G. Pfurtscheller and F.H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, pages 1842–1857, 1999.

G. Pfurtscheller, C. Guger, G. Müller, G. Krausz, and C. Neuper. Brain oscillations control hand orthosis in a tetraplegic. *Neuroscience Letters*, 292:211–214, 2000a.

G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer. Current trends in Graz brain-computer interface (BCI) research. *IEEE Transactions on Rehabilitation Engineering*, 8:216–219, 2000b.

G. Pfurtsheller and C. Neuper. *The Bereitschaftspotential movement-related cortical potentials.* Movement and ERD/ERS, pages 191–206. Kluwer Academic/Plenum Publishers, 2003.

J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications*. Prentice Hall, 1996.

L. R. Rabiner and B. H. Juan. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 1986.

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

H. E Rauch, G. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronautics Journal (AIAAJ)*, 3:1445–1450, 1965.

F. Renkens and J. del R. Millán. Brain-actuated control of a mobile platform. *7th International Conference on Simulation of Adaptive Behavior, Workshop Motor Control in Humans and Robots*, 2002.

S. Roberts and W. Penny. Real-time brain-computer interfacing: A preliminary study using Bayesian learning. *Medical and Biological Engineering and Computing*, 38:56–61, 2000.

S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.

M. D. Rugg and M. G. H. Coles. *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition*. Oxford University Press, 1995.

D. M. Santucci, J. D. Kralik, M. A. Lebedev, and M. A. L. Nicolelis. Frontal and parietal cortical ensembles predict single-trial muscle activity during reaching movements in primates. *European Journal of Neuroscience*, 22:1529–1540, 2005.

J. Särelä, H. Valpola, R. Vigário, and E. Oja. Dynamical Factor Analysis of Rhythmic Magnetoencephalographic Activity. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, pages 457–462, 2001.

A. Schlögl, P. Anderer, S. J. Roberts, M. Pregenzer, and G. Pfurtscheller. Artifact detection in sleep EEG by the use of Kalman filtering. In *European Medical and Biological Engineering Conference*, pages 1648–1649, 1999.

A. B. Schwartz and D. W. Moran. Arm trajectory and representation of movement processing in motor cortical activity. *European Journal of Neuroscience*, 12:1851–1856, 2000.

R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3:253–264, 1982.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications.* Springer, 2000.

E. E. Sutter. The brain response interface: Communication through visually induced electrical brain responses. *Journal of Microcomputer Applications*, 15:31–45, 1992.

M. P. Tarvainen, J. K. Hiltunen, P.O. Ranta-aho, and P.A. Karjalainen. Estimation of nonstationary EEG with Kalman smoother approach: an application to event-related synchronization (ERS). *IEEE Transactions on Biomedical Engineering*, 51:516–524, 2004.

H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14:2647–2692, 2002.

M. Verhaegen and P. Van Dooren. Numerical aspects of different Kalman filter implementations. *IEEE Transactions of Automatic Control*, 31:907–917, 1986.

J. J. Vidal. Toward direct brain-computer communication. *Annual Review of Biophysics and Bioengineering*, 2:157–180, 1973.

J. J. Vidal. Real-time detection of brain events in EEG. *Proceedings IEEE*, 65:633–641, 1977.

R. Vigário. Extraction of ocular artefacts from EEG using independent components analysis. *Electroencephalography and Clinical Neurophysiology*, 103:395–404, 1997.

R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems*, pages 229–235, 1998a.

R. Vigário, J. Särelä, V. Jousmäki, and E. Oja. Independent component analysis in decomposition of auditory and somatosensory evoked fields. In *Workshop on Independent Component Analysis and Signal Separation*, pages 167–172, 1999.

R. Vigário, J. Särelä, and E. Oja. Independent component analysis in wave decomposition of auditory evoked fields. In *International Conference on Artificial Neural Networks*, pages 287–292, 1998b.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37: 328–339, 1989.

Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao, and F. Yang. BCI competition 2003-Data set IV: An algorithm based on CSSD and FDA for classifying single-trial EEG. *IEEE Transactions on Biomedical Engineering*, 51:1081–1086, 2004.

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

J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan. Brain-computer interface research at the Wadsworth center. *IEEE Transactions on Rehabilitation Engineering*, 8:222–225, 2000.

S. Zhong and J. Ghosh. HMMs and coupled HMMs for multi-channel EEG classification. In *IEEE International Joint Conference on Neural Networks*, pages 1154–1159, 2002.

A. Ziehe and K. Müller. TDSEP–an efficient algorithm for blind separation using time structure. In *International Conference on Artificial Neural Networks*, pages 675–680, 1998.