Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

Basis Decomposition Discriminant ICA

Alejandro Tabas Díaz

Advisors: Laura Igual Muñoz
Lluís Fuentemilla

September, 2013

Basis Decomposition Discriminant ICA

Author: Alejandro Tabas Díaz
Supervisors: Laura Igual Muñoz
Lluís Fuentemilla Garriga

September, 2013

Abstract

In this Master's Thesis, we introduce the methodology Basis-Decomposition Discriminant ICA (BD-DICA), capable of finding the most discriminant Independent Components to characterise a high-dimensional dataset. The algorithm provides for this characterisation for several components with the same structure as the inputs. An adaptation of the algorithm for Feature Extraction is derived in the conclusions of this report.

BD-DICA is constructed as a combination of the Basis-Decomposition ICA (BD-ICA), an architecture for ICA used in fMRI data analysis, and the Basis-Decomposition Fisher's Linear Discriminant (BDFLD), a modified version of the classical FLD introduced in this work. BD-DICA is originally designed to deal with fMRI Data analysis, in which often we have data of about $10^5 - -10^6$ dimensions and a much smaller number of instances. BD-DICA finds interesting projections in the data whose output show a high discriminant power while maximising independence among the obtained projectors. Additional strategies based in a high restriction over the search subspace reduce highly the chances of overfitting.

Experiments with synthetic data show that the method is robust to noise and that it is capable of successfully finding the discriminant generators of the data. Experiments performed with real fMRI data show that the method offers good results with Resting-State fMRI data. Unfortunately, no conclusive results were obtained for Task-Based fMRI data.

A Gradient-Ascend approach to BD-DICA is exposed in detail along the report, including all needed derivatives. In addition, the implementation we used for the experimentation is publicly available running under MATLAB in www.github/qtabs/bddica. Compatibility with Octave is possible with a few adaptations regarding external libraries used by the algorithm.

Contents

Acknowledgements

As many projects of above a certain size, this Master's Thesis have been written thanks to the good will, care and help of many, many people. This is my first confrontation with that large but beautiful monster which is research, and the diverse dead-ends (that I now know are intrinsic to any research project) one finds after days of working in a single idea can be really discouraging when faced by the first time. In those moments, friendship becomes more important than any expertise, as the spirit of eagerness and optimism that only other humans can provide becomes of extreme importance in the darker moments of the development of the project. Because of that, I want to sincerely thank to all those people who have supported me, in one way or another, during this long journey, that now calls for a (small) pause. Now, as I know I am prone to sentimentalism, I beg for the reader to ignore the treaclish parts if it become annoying. Unfortunately, *Man ist was man ist.*

Pér, Elena, Cám, Amiga, Dolçet, Amigo, Karmaleon, Matiensus, Diegoti[1] and in general all of you, guys from the Hypatia years: thank you for your unlimited contributions to my life. Collaboratively, you have shown me the value of rationality; but more importantly, the value of irrationality. You have taught me how to (really) read a book, how to write a report and how to understand the sunny side of all the stuff one finds along the way. Thanks for staying there for me.

Bicho, thanks for supporting me in the darker moments of this project, and thanks for remember to me constantly that I was able to get this (Texan accent now) dammed think done. You are responsible for half of the self-esteem I have, the most important thing one needs to stand up after defeat. Thanks for your (almost) unconditional support and help.

Mother, father, your support has been so complementary and necessary that it practically brought me the idea of the basis decomposition by itself. I cannot imagine a stranger combination than the two of you, and I sincerely do not understand how people with overlapping parents, without such struggling forces to drive one to an extremal lifestyle, can have a balanced way of living. Thanks for being always there for me and for your so different but unconditional support to this thesis and to this live. I love you, guys.

Jakob, Parodrilo, Edu[2]. You have shown me the path of research. You have tempt my nose toward the dark theoretical side of Physics and then you have show me in perfect timing that

[1] Someone told me that it could be a good idea to include the real names of those people. There they go: Israel Saeta, Elena Marina, Laura Camacho, Violeta Menéndez, Ana Dolcet, Daniel Medina, Carme Codina, Paloma Matía, Diego Tejero... your names are really Spanish, guys!

[2] Again: Jacobo Ruíz, Pablo Rodriguez, Eduardo Martín

Acknowledgements

Fundamental Physics is not the only way of doing Theoretical Physics. You have been, in somehow, my mentors during the Hypatia years and I still remember with affection all the conversations we had about the meaning of Physics and the amazing mysteries behind nature.

Mar, Sonja, Marta. Thanks for accept me in your only-girls club as one of your own from the very beginning. Thanks for showing me the half of what I know about fMRI and ICA, thanks for support my ideas beyond imagination and for your constant support in those early Tuesday meetings. You have been my first research group. Part of this Thesis belongs to you.

Laura, Lluís. I do not know how to thank you for the way in which you have accepted me in your arms from the very beginning. Thank you for your energy, for your optimism, for your laughs and for the way in which you have directed my work during those months.

Laura, thanks for your continuous trust on me and my capacities. Thanks for caring about this work as you did and thank you for all the energy you have put in this project. I know I am not completely staying in the field, but I have really enjoyed the journey.

Lluís, thank you for rescuing me, showing that I am prepared to research in Neuroscience just when I was just about to drop the intention into a gravity well. Thank you for becoming interested in my interests, for trusting so hard in my technical background and for the freedom you have let me to lead the course of this work.

To both of you, thanks for letting me the space and independence I needed during one half of this project and your continuous and meticulous guidance and help during the other half. You have been just the perfect Master's Thesis advisers.

This work was economically supported by the project *CARACTERIZACIÓN DE LA OBESIDAD A TRAVÉS DEL ANÁLISIS AUTOMÁTICO DE IMÁGENES DE RESONANCIA MAGNÉTICA FUNCIONAL*. My scholarship was part of the *Convocatoria de ayudas para la realización de proyectos de investigación en ciencias sociales y humanidades concedido por el Vicerectorado de Política Docente y Científica de la Universidad de Barcelona*. Thank you, UB. You are all right.

Anti-Acknowledgements

There is a vast amount of people who have been there for me during those days. Unfortunately, whereas researches and their close relatives work hard to make from this world a more efficient, clean and comfortable place, it seem to be an absurdly vast amount of people (most of the times, unconsciously) working as hard as we are to prevent that from happening. I do not really now if the preamble of a Master's Thesis is the place for writing this kind of critic; but, if such a work, with this combination of educational and academical dependence, is not the place... I cannot imaging where is it.

During the past years the budget for public research and education has been decreasing constantly in Spain. Not only during the recent past as a consequence of the widely adopted policy of austerity in Europe. Even before that, in a exhibition of loosely understood Keynesian strategies the government of our country has systematically trimmed off the already uncertain future of public research institutions in the behalf of the investment in poorly justified constructions.

I believe that, when a particular sector dies in an economy, as it happened it the construction sector of our country (if not dead, let us call it long sleep), it is customary to dedicate resources for the re-education of the working force of that sector to re-invest their forces in other sectors. We have adopted a completely contrary approach. We have created unsustainable employments for that people on the expenses of the education and research budgets. This policies condemn the construction workers to an unavoidable unemployment and condemn the future generations to face the same problem as their parents.

Investment in research and education defines the shape of the economical structure of a country. A rich research opens the doors to new technologies that are eventually responsible for the creation of new innovating enterprises. A rich education allows to such enterprises to be established in our country, as it is possible to hire local educated employees, which is much cheaper than bringing people from the outside. Research creates knowledge than improves the efficiency of all sectors, making them more productive. Education allows to a country to learn how to implement those improvements in the behalf of the whole society. A country not investing in knowledge, is condemn to the necessity of importing it.

Instead, we are approaching to an economy closest to the one of the underdeveloped countries: instead of encouraging the hiring of new employees by sustaining the creation of new enterprises and improving the education of the unemployables, we are decreasing the social security of the population to decrease the risk of hiring; instead of investing in formation of the workers and in a good reliable quality education for the new generations, we are investing in smoke projects

like Madrid 2020 that might or might not create (temporal) employments in the near future. We are becoming the servants of the developed countries, offering them a nice warm beach in the summer and a big casino city for the rest of the year.

Spain seems to have very clear, however, that this two resources are accessory and that they can be trimmed off without consequences. As a response, young researchers are moving out of the country. Not only to increase the reach of their education. Some of them are having troubles to come back. The CSIC is in one of their worst positions in the democratic history of this country and the usual scholarships for researchers are becoming harder and harder to achieve. It is important to be aware that a large amount of public research cannot be conducted by private investors. Some technologies can take as much as 50 years to have a real application in industry, and of course very few enterprises are willing to sustain research in such early periods. However, this research has to be done if we want to enjoy new medical and energetic technologies in the future.

Even more disastrous are the recent policies are centred in the systematic destruction of public education. College is becoming a luxury at which only the intellectual and economical elites of the country can access. It is usually said that this country is overqualified. I believe that the way of solving those kind of problems are, in the first place, by modernising the economical structure of the country, letting place to the kind of occupations our *overqualified* people can do; and second, to offer a reliable alternative for university education by improving the quality of the professional oriented education. We are following the opposite path, by investing in flying infrastructures that never got inaugurated and imposing fares over the professional oriented public education.

Even worst, are the projects to remove the internal democracy of the university, which ultimately pretend to transform the university from a place of knowledge into a place of professional formation. The arrogance of the leaders of this country is not only transforming the kind of society we will be in the future, but it is also getting sure for the decisions they take to cover up to the minimal detail regarding the institutions.

For the reasons exposed before, to all of you, arrogant, uninformed, intuition-driven leaders of our current society: my most sincere anti-acknowledgement (if such term even exist). You have contributed negatively to the develop of this work and the realisation of these kinds of Thesis will become harder and harder in the future because you are not rationality measuring the real cost of your actions. I do not like the Spain you are modelling. And, probably, neither do you. You need either to assume the responsibilities of your positions or to look for another job. You are playing with the future of a whole Spanish generation. And it is a very fragile and valuable toy.

Preface

Dear reader,

this report is the projection of several months of work dedicated to the development of an Analysis tool we call Basis-Decomposition Discriminant ICA. As you will discover along the pages in this text, the development of such tool is not as direct as one might expect. Instead, as usually happens in mathematical research, we found during the journey that some additional non-existent theoretical frameworks were necessary to complete our development.

In my opinion, these additional derivations are the most enjoyable part of this Thesis and they are representative of the spirit in which this work has been developed. However, a *not that enthusiastic about mathematical formalism* reader might find those parts rather boring, if not irritant, and perhaps even unnecessary for the main development. If this is the case, I ask for some patient. I believe that all formal definitions and derivations in this Thesis have a theoretical justification that will be evident, eventually, along the report.

In the other hand, as it is usually the case in Machine Learning, we support our method in two widely used, well established, methodologies with which the reader might be already familiar. However, the amount of available tools is so extensive in the Machine Learning Community that one cannot just assume for everybody to know every algorithm. With this idea in mind, we have decided to briefly reformulate those techniques in the language and formalism developed in this work. This has a double purpose: the unfamiliar reader has the opportunity to catch up without needing to read any additional literature, whereas the reader which is already familiar with the method can appreciate the shades of our formalism and get familiar with our notation.

This set of additional derivations and reformulations takes a large part of this Master's Thesis. In exchange, ultimate derivation of our algorithm will be (or at least this is our intention) as smooth and reasonable as if it where just the only logical solution to our problem.

All of this make of this report a fairly long document for a Master's Thesis (but please take into account that we have been generous with the spacing). I can promise you, however, that we have tried as hard as we could to keep it short and if something is written is because we thought that it would be essential to successfully capture all the shades and aspects of our work. We have skipped unnecessary already formulated proofs that are always refereed in to the source texts, in which the concepts are introduced with a mastery we cannot match and we have tried to summarise as much as possible all reviewed previous literature.

Having said that, if you are in a hurry, probably you can skip some sections if you feel you are

familiarised with their topics. More specifically, if you are familiar with fMRI data and you are exclusively interested in the algorithm, you can skip without regrets the whole Appendix A, in which we perform an alternative ICA-based analysis over a private database introducing the main concepts of fMRI data and the Section 2.1, in which we review some of the state of the art ICA-based techniques for fMRI data analysis.

If you already know the ICA formulation you will probably get bored reading 3.1, but try to not skip anything else during the chapter even if it sounds familiar (like the section on preprocessing) as we use a different formalism we will need later for the rest of the exposition.

If you have already worked with the Fisher's Linear Discriminant and you are familiar with its mathematical formulation, you can safely skip the discussion of Section 4.1. We use standard notation in this exposition as well as in the rest of the Chapter, but you can always go back to the classical formulation if you feel like you are missing something during the rest of the exposition.

Finally, we should recognise (with a certain amount of pain and sorrow) that the Appendix B is probably the most expendable part of this document. It presents the full derivation of the mathematical expressions needed to implement the algorithm (which, in the other hand, is already implemented under the MATLAB/Octave environment). We decided to write this appendix for two reasons. First, we though that the interested reader should have all the necessary tools to make an implementation of its own. Second, as this is after all an academical work, we though that it would be a shame not to write such an important an laborious derivation.

And well, that's it. There is no necessity to bore you any longer with such metadetails. I sincerely hope you enjoy the reading of this document as much as we have enjoyed developing this algorithm and its theoretical framework.

Chapter 1

Introduction

The Machine Learning and Data Mining communities have been widely devoted to the task of developing tools for knowledge discovery and data characterisation. Indeed, in a very large set of applications, we use classical techniques developed for regression, classification or feature selection, tools clearly devoted to automatic learning, to specifically and exclusively find structures behind the data [1].

This usage is particularly common in Neuroimaging [2], a branch of Computer Vision devoted to the analysis of brain imaging data. In this field of application, the final purpose of the analysis is, often, not to construct a classifier but to understand better the processes occurring in the brain and their relations with the behaviour or health condition of the subjects of the experiment.

In this work, we will develop a new technique directly aimed towards data characterisation based in two widely used feature selection algorithms: Independent Component Analysis (ICA) [3] and Linear Discriminant Analysis (LDA) [4]. Our algorithm combines the strategies of both methodologies to find a representation of the data capturing the differences among the samples belonging to different groups.

More specifically, we are trying to find a (small) set of independent vectors such that the projection of our dataset in the basis defined by those vectors maximises the separation among the groups defined in by some data labels. Therefore, we want to perform some kind of Basis Decomposition in Discriminant and Independent basis vectors. We have called Basis-Decomposition Discriminant ICA (BD-DICA) to such process.

A similar approach has been already implemented in [5], where the authors combine the Fisher's Linear Discriminant and the ICA objective function to construct a feature selection algorithm maximising both, the independence of the obtained features and its discriminant power at the same time.

The twist is that we are going to ask to those vectors to be a linear combination of the vector representation of the samples. This approach is not very common in the literature, but it has been used successfully in some architectures of ICA designed to deal with high dimensional

data [6]. We will provide a theoretical framework for this architecture to formally support this approach to our problem.

The practical advantages of such constraint are numerous. It allows to reduce the search space to the one spanned by the vectors of the observations (or a reduced subset of them), therefore decreasing the computational complexity of the algorithm whilst preserving the original representation of the data, which is crucial in some data characterisation tasks.

These properties put altogether are particularly convenient for analysing fMRI data. fMIR is a functional version of Magnetic Resonance Imaging that provides a 3-dimensional real-time image of the activation in the brain [7]. The representation of such a data has an extremely large dimensionality ($\sim 10^5 - 10^6$) which cannot be trivially reduced when we want to find brain areas of activation.

In this report, we will expose some of the workarounds made for dealing with this kind of data in the past and we will try to justify the Basis Decomposition approach in such context.

More specifically, the remaining of this document is structured as follows. First we will introduce the motivation for the problem we want to assess, how this is solved in the previous literature and why we do think that an additional method could be helpful for data characterisation. After this motivation, we introduce a generalisation of this problem in formal terms and the theoretical framework in which we will develop the rest of the tools. All this is done in Chapter 2.

Chapter 3 is devoted to Independent Component Analysis. We will first introduce the classical formulation of ICA in Section 3.1, which should be taken as a brief review of the method. During the rest of the section we complete the details of the formulation using our framework, revealing the key points behind the Basis Decomposition architecture.

In Chapter 4 we do a similar work but this time with the Fisher's Linear Discriminant. In Section 4.1 we introduce the classical formulation of the algorithm in the same way we introduced ICA before, as a review of the basis of the discriminant. In the rest of sections we develop a generalisation of FLD that we will need later to construct the missing part of the BD-DICA algorithm.

This last algorithm is fully developed in Chapter 5. This chapter is divided in three sections. In Section 5.1 we study an already presented version of D-ICA in the Feature Extraction architecture. Then, in Section 5.2 we develop a Basis-Decomposition compatible Fisher's Linear Discriminant based in the generalisation described before. Finally, we present the Basis-Decomposition Discriminant ICA in Section 5.3.

Chapter 6 is devoted to the experimentations we made to validate our method. We use a synthetic and two real fMRI datasets for that purpose.

The conclusions of the thesis are exposed in Chapter 7, which is the closure of this work. Along with the conclusions we present an extended collection of future work tasks, including extensions and improvements to make this algorithm more useful.

This work has two extra appendices. Appendix A describes an analysis made with fMRI data using an alternative approach with classical Machine Learning techniques. We use this appendix

as an excuse to introduce in more detail the nature of fMRI data and to present the alternative way of doing temporal inference using ICA. This work can be fully read without this appendix (and vice versa), but we believe that it can reveal some justifications for some of the decisions we made in the design of the BD-DICA algorithm.

Appendix B shows the minor details of the BD-DICA: the derivatives and dynamics needed for the Gradient Ascend algorithm. While essential to implement the algorithm, those details are not really important for the theoretical discussion.

The algorithm introduced in this work have been implemented for the MATLAB environment. The libraries are clean and well documented, but probably not fully optimised. They will be published soon together with some testing scripts. For the time being, the code is not publicly available, but it is offered attached to this document. As far as we know, the code works perfectly under Octave by itself, but unfortunately we are using some external libraries to compute the PCA for the data that are not fully compatible with Octave. The same problem arises when importing fMRI volumes directly from a NIFTI file, which is made with the support of SMP [8] in the scripts provided with the implementation to import real fMRI data.

Chapter 2

Problem Specification

During this chapter, we will specify clearly the problem we are facing in two different manners. First, we will explain the original situation we wanted to solve: fMRI connectivity analysis. This should be taken as a state of the art of the current existing approaches to this problem. We will review a selection of currently used approaches in this line, so at the end we can discuss the advantages and drawbacks of our algorithm. Note that, in addition to this review, we offer a real case of study we made using classical techniques in the Appendix A.

This fMRI problem is presented, however, as a motivation towards the second, more formal and general, problem. In this last specification some assumptions about the nature of the problem will be made and we will we introduce a theoretical framework in which to develop its solution. We will use this formal representation for the rest of the exposition of the algorithm. Note that, even when actually representing the same problem as the fMRI one with some additional constraints, this second description is much more general than the first one as it is not specifically restricted to fMRI Data. Therefore any other problem fitting in such definition should be suitable, at least in theory, for our algorithm.

2.1 Fundamentals and Background

2.1.1 fMRI and Independent Component Analysis

Functional Magnetic Resonance Imaging (fMRI) is a functional Neuroimaging technique showing a brain-activation related signal as a 3-dimensional brain map (called volume) activating over time [7]. While the spatial resolution of the technique is very poor in physiological terms, with 3-dimensional pixels (called voxels) of about $1mm^3$, the computational representation of those images is way to large to board it without some previous feature selection methodologies[1].

Given the nature of the physiological mechanisms underlying the signal, Independent Component Analysis (ICA) has proven to be a perfect match for analysing fMRI data [9]. This is

[1]To read a more exhaustive description of this kind of data, please consult Section A.1, where we expose a small summary of the technology and assumptions behind fMRI recordings

specially true in Resting-State fMRI, where the participant is asked to think on anything without falling sleep [10] during the experiment. In this kind of recordings we observe a mixing of independent background brain processes revealing brain connectivities.

However, ICA is also a common choice dealing with another kind of fMRI experiments referred as Task-Based, in which the patient is specifically asked to perform some task during the recording [9]. Task-Based fMRI have a wide set of applications. For example, we can use it to detect which areas of the brain are related with a determinate activity [11], or see how this activation differs between healthy or brain-damaged individuals [12].

A large number of pathologies have been found to be directly related, at least at the scales managed by fMRI, with the joint activation of localised clusters of voxels (brain areas) or sets of unconnected clusters activating together through time (networks of activation) [7]. A number of those networks have been isolated for some researchers in Resting-State [13]. This number is not very high (depending on the studies, it can vary from 10 to 30) and, in theory, all the Resting-State brain activity related with brain connectivity is represented in one or several of those structures. Those structures can also be activated in Task-Based fMRI, in addition to the specific brain areas activated by the performed task.

In any of the cases, the relevant activations in the brain are represented by joint activations of a subset of the voxels along time. These subsets vary in size, but they are always much smaller than the whole brain, which means that the vector representation of such patterns is sparse and therefore has a clear super-Gaussian shape (most of the voxels showing barely no activation, a few of them showing activations beyond the zero) [14].

ICA can take advantage of this non-Gaussian shape to find the statistically independent sources of the brain activity to reduce dimensionality [15]. At the same time, it filters the non-relevant signal patterns (associated with noise from the experimentation) of the brain in a single step. Moreover, the extracted Independent Components (ICs) can be used to characterise the data.

Note that, as we have assumed that the non-Gaussianity is shown by the patterns of activation in the 3-dimensional space, we need to apply ICA in such a way that the obtained Independent Components are directly the 3-Dimensional volumes (or a vector representation of them). This particular arrange of ICA is often called Spatial-ICA [15] in the fMRI literature, and it is intimately related with the Basis Decomposition framework we are exposing in this work.

Spatial-ICA draw the Independent Components as linear combinations of of the 3-Dimensional volumes composing the functional data [15]. In other words, it constructs the ICs as if they were elements in the vector space spanned by the snapshots of the fMRI recordings. As a consequence, the extracted ICs are directly comparable with the recordings, and they are usually superimpose to the image of the brain to detect which parts of the organ participate in the component. The method offers, therefore, a very straightforward way of interpreting the results.

2.1.2 Group ICA

As we have seen, ICA allows us to better characterise fMRI data. This tool can be directly applied to a wide spectrum of experiments involving a single specimen. However, often we need

to perform ICA over a set of fMRI recordings corresponding to different patients to obtain Independent Components being common to all of them [9].

The are currently two well-known ways of aggregating fMRI data from different people to perform ICA. The most elegant solution is called Tensor-ICA [6]. The one we will use in this work is called Group-ICA.

Tensor-ICA [6] refers to tensors as generalisation of matrices in the sense that, if a matrix can be seen as a representation of an aggregation of vectors, we can construct tensors as aggregations of matrices, aggregations of aggregations of matrices, and so on. In particular, Tensor-ICA takes each fMRI recording as a matrix, and them it puts all of them together aggregating the matrices corresponding to different recordings. Each of the components of this tensor A can be written as $A_{i,j,k}$, where the first index i moves along the space, the index j moves along the time and the last index k moves along the patients. Then our representation is a tensor with (voxels $\times$ time-points $\times$ patients) dimensions.

A Tensor-ICA representation of the data assumes that all the voxels and temporal snapshots of the patients are perfectly aligned [6]. Some hard preprocessing is usually made over the spatial part of the fMRI signal to guarantee that the first condition holds. The second one can be forced (and usually it is) in Task-Based experiments by presenting the stimuli in all patients at the same temporal volume. This is, however, not possible in Resting-State fMRI [10], where there is no clear benchmark to start the experimentation (i.e. the background processes of the brain are constantly active and we cannot restart them voluntarily).

Group-ICA aggregates fMRI recordings even when this temporal restriction does not hold [14]. The solution of Group-ICA is to treat all volumes of all patients in the same way by temporal concatenating them. The representation is, in this case, a matrix of ((voxels $\cdot$ time-points) $\times$ patients) dimensions.

Note that while the Tensor-ICA representation saves the temporal structure of the data, Group-ICA completely ignores it. That is why we do not need to make any restriction over the temporal dimension in this last techniques (the recordings can even have a different length!).

2.1.3 Inference in fMRI using Group-ICA

To characterise a kind of brain activity is usually necessary to compare two different sources of data, one presenting the given properties and other presenting a normal activity. It is therefore convenient to define some way of performing this kind of inference over the results of Group-ICA. Note that Group-ICA is not a Feature Extraction methodology and its outputs cannot be fed into a classifier. These outputs are Independent Spatial representations characterising a certain pattern of activation in the brain present in all (or some of) the subjects submitted to the analysis. Therefore, we need to extract a subject-specific (or, at least, class-specific) information from those spatial patterns before performing any kind of inference.

Temporal inference

The most common way of performing inference over ICs is to decompose each of the subject-specific fMRI recordings into a temporal dependent weighted sum of the Independent Components. The architecture of such decomposition is illustrated in Figure 2.1. Each of the weights represent, for each time point and subject, the relative importance of a given IC. The temporal concatenation of those weights for a given subject and IC are usually called *time courses*. These time courses are computed training a general linear model (GLM) for each of the subjects [9].

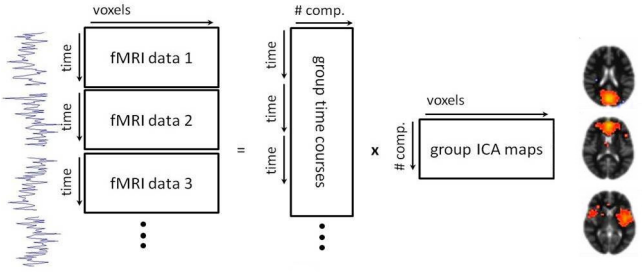

Figure 2.1: Illustration of the GICA architecture considering the mixing coefficients as temporal courses of the given Independent Components.

Once we have obtained the time courses, we can use them to perform inference over the groups and figure out if any of the Independent Components plays a important role in the case of study using classical Feature Selection and Classification techniques [16].

Dealing with the time courses is not trivial either. A typical experiment has about 200 time points, which means that each of the subjects are characterised by a 200 $\times$ number-of-ICs dimensional vector. In addition, as we saw, the time points are not necessarily aligned among different subjects.

One can find several ways of dealing with this problem in the literature, but most of the times the procedure involves translating the whole time course into a single number (e.g. the mean of the difference between the time-courses and some background measure, as the time-courses during the resting periods in a Task-Based experiment). This is the approach we have taken in the case of study described in Appendix A. Other approaches involve Fourier decomposition of the time courses. This approach solves the problem of the alignment and allows us to reduce the dimensionality of the resulting representation as much as we want, by just reducing the number of bins in the frequency histogram [8].

A more sophisticated and more successful representation for the time courses is the amplitude of low-frequency fluctuation (ALFF) measure used in Resting State [17]. The ALFF for a time course is computed as the integral (in frequency) of the square root of the Fourier transform of the time course. Therefore, we are representing each of the ICs as a number containing the *amount of signal* as measured in the frequency space [18].

The name *low-frequency* is given to this technique because usually the time courses are band-pass filtered to preserve only Resting-State relevant signals (which are known lo lie between 0.01Hz and 0.8Hz). In such a way, we remove all frequencies non directly related with brain processing,

which can be caused by other physiological rhythms from systems like the cardiovascular or respiratory (working at 0.1Hz-1.2Hz) or from artefacts from the experimentation itself. In practise, however, those kinds of filtering are performed during the preprocessing of the data and there is no need to take them into account during the rest of the analysis.

Spatial Inference

Sometimes, however, it is more illustrative to represent the differences between the classes in the data using spatial representations. As the Independent Components obtained using ICA are actually common to all subjects, we need to find a way of projecting those ICs into new spatial maps, specific for each subject or group.

A solution is proposed in [19], where the authors use the common ICs to seed a subject-specific search. This search is tuned to maximise group information, therefore maximising at the same time both, independence of the obtained components and presence of group indicators.

Another solution is presented in [20] in which the temporal courses obtained as in 2.1.3 are used to find the subject-specific spatial maps. In order to do that, the method proposes to train yet another GLM to find the best set spatial maps for drawing the whole fMRI record of a subject using the previously found time courses. The whole process is called Dual Regression for evident reasons. As each of the spatial maps found by this technique corresponds to a given time course, which at the same time corresponds to an IC, the subject specific spatial maps can be compared easily without making out any arbitrary correspondence among them.

The inference in this spatial representation is usually done by finding statistically significant differences between the spatial maps corresponding to each of the groups [20]. The output of such process are t-maps showing differences between each of the ICs for each of the groups.

A very different approach is shown in [21], where the authors propose to compute directly the subject-specific ICs for all the individuals using ICA. These spatial maps are then used as basis elements in which to express the signals of the individuals. The coefficients of such basis (actually measured as angles in the Grassmann manifold spanned by the basis) are then fed to classical Machine Learning techniques in the same way in which the time courses are treated in Section 2.1.3. The main difference is that, as the basis is constituted by subject-specific ICs, the most discriminant element is more representative of a group that the usual common Independent Components.

We will not insist more in this spatial representation as we will instead focus in the temporal inference approach for the rest of this work.

2.2 The Discriminant-ICA approach

The approaches shown before for dealing with temporal inference are, however, not fully satisfactory. They require a lot of work to finally isolate a single discriminant Independent Component. It would be useful to have a procedure to directly extract that discriminant IC without the necessity to extract all of them before.

Our work is settled over this approach to the problem. We want to construct a procedure to directly use the information about the labels of the subjects to extract only important Independent Components. In addition, the inclusion of class information at this early stage of the analysis could provide a better representation of the differences among classes instead of focusing just in the statistical properties of the components, leading to a slightly different representation of the Independent Components that might be more useful to characterise the data.

2.2.1 Previous approaches to this problem

We will cite two previous approaches to this problem we found in the literature. The only one directly applicable to fMRI is the one presented in [22], where the authors propose to use Coefficient-Constrained ICA (CC-ICA)[2] to introduce a cost term based in the t-statistics of the groups into the ICA objective function.

This methodology gets very close to what we want, but it has an important drawback: the correcting term can only be used as a correction whose effects are reflected directly in the shape of the obtained ICs. That means that the final results are oriented towards group characterisation, but we still obtain the whole set of Independent Components. Instead, we are looking for a procedure directly ignoring non-important components.

The second, presented in [5], is more similar to our approach. In this work, the authors propose a dual optimisation using an objective function resulting of a weighted sum of the ICA and the LDA objective functions. Unfortunately, this approach is incompatible with Spatial-ICA for reasons that will become obvious later, when we had fully developed our framework, in the detailed formulation of the algorithm in Section 5.1.

2.3 Definition of the problem

In this section we will introduce in a clear and formal way the problem we want to solve. As said before, this problem should not be understood as the start point of our analysis. On the contrary, the formal definition of the problem is as a mathematical formulation of the aggregation of the objective of the method plus all the hypothesis and we have decided to include during the journey. Therefore, before formally introducing the problem, we will take a brief review of the already exposed assumptions and we will introduce and justify some other ones.

2.3.1 Assumptions

Independence and separability assumptions

Until now, we have outlined the problem we want to solve as a supervised version of Independent Component Analysis suitable for the analysis of high-dimensional data such as fMRI. In this first

[2] CC-ICA [23] is a framework to add additional terms into the classical ICA objective function.

outline some assumptions have been already made. We are indeed assuming that the relevant physiological sources of brain activity are non-Gaussian and statistically independent among each other.

We also need to assume that these sources are relatively common to all the subjects involved in the experiment and that some explicit differences exist in the temporal expression of those components between subjects labelled in different classes. Moreover, we need to ask to those interclass differences to be larger than the inter-subject differences. In fact, if that is not the case, we can argue that such physiological source does not show any statistical significant difference among classes.

These two assumptions open the way to a discriminant version of ICA in which the discriminant power of a particular Independent Component can be evaluated using the Fisher's Linear Discriminant, which measures precisely the quotient between the interclass and the intraclass variability for a given representation of the data. As we are interested in the differences measured in the temporal domain, we will ask to the discriminant to measure this quotient based on the time courses of the samples. However, as said before, those time courses can only be obtained using some kind of regression (e.g. a GLM) over the whole set of ICs. This leads us to the stronger following assumption.

The orthogonality assumption

As already said, it is of capital importance to be able to measure those temporal coefficients not only without considering all the ICs, but also in an on-line manner, in order to be able to conduct the search towards components maximising the discriminant. Moreover, as in many applications there is just one physiological relevant source, we need to obtain a way of doing so considering just one component at each time during the search.

To do that, we propose to add an extra assumption to the problem statement: assume that the ICs are orthogonal among each other. In this way, the direct projection (i.e. scalar product) of an fMRI volume over the ICs is directly the temporal coefficient for that volume and that component.

This is, of course, a very strong assumption, but it can be properly justified in the field of fMRI. Indeed, let us consider the two kinds of solution we can obtain in our experiments: we can obtain a Resting-State-like network or a single brain region. In this last case, as brain regions are non-overlapping and the absence of activation can be modelled with a zero in the deactivated voxels, it is clear that the space of solutions is indeed formed by orthogonal vector representations.

The case of the Resting-State networks is not that clear. These networks usually contain several regions and therefore can share some areas of the brain. Fortunately, in the last few years some Neuroscientists have tried to compile a database of such networks to better understand its use. Contrarily to the intuition, the number of such networks is actually quite small: in this work we are using the Biswal's [13] database, which contains 20 networks.

To check if our assumption is plausible, we have tested the 20 Biswal's networks for orthogonalisation. For that, we have imported the networks in a vector representation, we have normalised each of the networks and we have computed the scalar product among them. The results are summarised in a histogram in Figure 2.2, containing only crossed products.

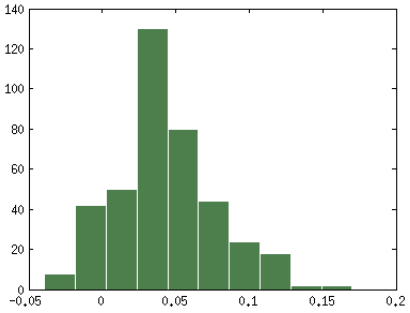

Figure 2.2: Scalar product of the normalised vector representation of the physiological networks indexed by Biswal et al [13]. Note that the histogram has been computed omitting the products of a network with itself

As we can observe, even when the networks are not completely orthogonal, there are very few products exhibiting a greater result than 0.1, and even those do not exceed the limit of 0.2. We think that this result is good enough to allow us to assume that the direct projections of the fMRI data over the considered component is a good approximation to the time course associated to that network.

Having said that, we are not going to impose to the ICs to be orthogonal among each other, since this is an artificial and unnecessary constraint. Moreover, this imposition would affect to the more important statistical independence assumption (as we would be able to deduce some properties of a component by looking at another one). We are just going to assume that the sources of the data are orthogonal when measuring the particular discriminant power of a component, which could be formally justified by the following affirmation: all the sources are independently sampled from a very sparse distribution. However, this formal details are not that important. Just keep in mind that we will use orthogonality without explicitly demanding it.

2.3.2 Formal definition of the problem

Now that we have provided an heuristic idea of the assumptions and problematic behind our method we are in position to formally introduce our problem.

Let $\mathcal{S}$ be a basis spanning the vector space $\mathcal{V} \subseteq \mathbb{R}^D$, where D is the dimension of the N basis vectors $\mathbf{s} \in \mathcal{S}$. Furthermore, let us impose that the representation of those vectors in $\mathbb{R}^D$ follow a non-Gaussian distribution[3].

[3]We will systematically abuse the terminology in that way during this work. What we really want to say is that, for each basis element $\mathbf{s}_i \in \mathcal{S}$, its components ${s_i}^j$, are asked to be sampled by a non-Gaussian distribution.

Along this work, we will use indistinctly the terms *(original) basis element*, *source* and sometimes, erroneously, *Independent Component* to talk about this vectors $\mathbf{s} \in \mathcal{S}$.

Let $\mathcal{X}$ be a set of M vectors $\mathbf{x} \in \mathcal{V}$, each of them representing an observation in a Knowledge Discovery problem. We will use the terms *instances*, *samples* and *observations* to refer to this vectors in this work.

Let $\mathcal{C}$ be a discrete set of 2 or more labels, representing the value of a target characteristic of our observations $\mathcal{X}$. Therefore, for each observation $\mathbf{x}_i$ there will be a class label $c_i \in \mathcal{C}$. We will refer to $\mathcal{D}$ as the set of all the labels c_i defined over a set of observations $\mathcal{X}$ (i.e. the size of $\mathcal{X}$ is the same of $\mathcal{D}$, whilst the set of $\mathcal{C}$ is the same as the number of classes in the problem).

Now, let us define $\mathcal{S}^d \subseteq \mathcal{S}$ as the set of *discriminant basis elements*. The idea is that the coefficients of the samples $\mathbf{x} \in \mathcal{X}$ for the basis elements in the set $\mathcal{S}^d$ are sampled differently for the observations belonging to the different classes (for example, all the coefficients are sampled from a Gaussian distribution, but the means or the variances of the distributions are different for different classes). However, the coefficients corresponding to the basis elements $\mathbf{s} \notin \mathcal{S}^d$ are sampled following the same distributions for all the instances, independently on the instance class labels.

From a point of view of Knowledge Discovery, only the elements in $\mathcal{S}^d$ are important to characterise the differences between subjects belonging to different classes.

Difinition 1 *Given a set of observations $\mathcal{X}$ constructed using an unknown basis $\mathcal{S}$ with non-Gaussian basis elements and a class label for each of those observations $\mathbf{c_i} \in \mathcal{C}$, the* Discriminant Independent Basis Decomposition Problem *is defined as the problem of finding the set of discriminant basis elements $\mathcal{S}^d$.*

2.4 Formalism and notation

Linear algebra plays an important role in this work. Just for clarifying things, let us introduce some concepts before continuing our exposition to fix notation.

We have already said that our observations will be represented by vectors $\mathbf{x}_i \in \mathbb{R}^D$. Now, the jth component of the ith observation will be denoted by ${x_i}^j \in \mathbb{R}$. The components of the transposed vector will be denoted by ${x^i}_j$. Please note that while the vertical position of the indices has been flipped, their horizontal position remains the same. This is of crucial importance in several derivations in this document, as we will use often component notation to describe scalar products.

2.4.1 Projectors

Now, we can define two kinds of linear operators in this configuration. The most intuitive one is the 1-form defined over the dual space of $\mathcal{V}$. We will use the Greek letter ξ to refer to this kinds of forms ($\xi \in \mathcal{V}^*$), which are defined in the following way:

$$\begin{aligned} \xi: \quad \mathcal{V} &\longrightarrow \mathbb{R} \\ \mathbf{x} &\longrightarrow \langle \boldsymbol{\xi}, \mathbf{x} \rangle \end{aligned} \tag{2.1}$$

When considering various of those projectors, we will label them in this way: $\xi^1, \xi^2, ...$ However, the vector representation of such projectors will be written in this other way $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ...$[4].

The components of ξ^i will be denoted as $\xi_i{}^j$ in such a way that the scalar product in the left hand side of Equation 2.1 can be written as (note the transposition):

$$\langle \boldsymbol{\xi}_j, \mathbf{x}_i \rangle = \sum_k \xi^j{}_k \; x_i{}^k \equiv \xi^j{}_k \; x_i{}^k \tag{2.2}$$

The last step in this last equation describes the notation convention of Einstein [24], which establishes that whenever the same index is used twice in a product of two variables, if it appears once up and once down, a sum is implicit in the expression:

$$x_i \; y^i \equiv \sum_i x_i \; y^i \tag{2.3}$$

We will use this notation intensively in this work, since the equations are much more readable and easier to understand this way if several operations are happening at the same time.

2.4.2 w-Applications

The second operator we can define in this context is the operator defining linear combinations over the vectors $\mathbf{x}_i \in \mathcal{V}$. Note that these operators do not belong to any particularly defined vector space and that they are not restricted to have a particular dimension. Also note that the result of any application of this operator over a subspace of $\mathcal{V}$ belongs to the same subspace.

We will call $\mathcal{W}^N$ to the vector space of all these possible linear combinations involving N vectors in $\mathcal{V}$. We will denote those operators as $w \in \mathcal{W}^N$, or simply w-applications. They are defined in the following way when considering the whole set of observations $\mathcal{X}$:

$$\begin{aligned} w: \quad \mathcal{V}^M &\longrightarrow \mathcal{V} \\ \mathcal{X} &\longrightarrow w^i \; \mathbf{x}_i \end{aligned} \tag{2.4}$$

[4]This is not a gratuitous choice of notation, it emphasises the fact that whereas $\xi \in \mathcal{V}^*$, $\boldsymbol{\xi} \in \mathcal{V}$. The reader migh argue that we could construct this last vector representation directly in the dual space. The reason why we did not take that approach is because when developing the BD-ICA algorithm we will use this kind of projectors to represent the candidate Independent Component $\mathbf{s} \in \mathcal{V}$.

Note that, if we need to define several w-applications, we will use w_i or the vector representation $\mathbf{w}_i$ so that the jth component of the application is witten ${w_i}^j$. Therefore:

$$\mathbf{y}_i = {w_i}^j \mathbf{x}_j \tag{2.5}$$

where, of course, $\mathbf{y}_i \in \mathcal{V}$.

This kind of applications will be of capital importance during the development of the Basis Decomposition algorithms, since the characterisation of a dataset is usually represented as a feature vector of the same kind. For instance, in the case of fMRI data, the characterisation of a group of observations is usually a 3-dimensional volume indicating which areas of the brain are important in the context of the experimentation in which the data was recorded.

The dimension of the vector representation $\mathbf{w} \in \mathbb{R}^N$ of those forms will depend on the amount of samples we want to use for defining the proper transformation, but in general we will chose it to be the number of available samples $N = M$ (or a PCA-reduced representation of those samples).

For completeness, we define $\mathcal{W} \equiv \bigcup_N^{inf} \mathcal{W}^N$.

Chapter 3

Independent Component Analysis

Independent Component Analysis [25] is a Blind Source Separation (BSS) [15] methodology capable of separating non-Gaussianity distributed independent sources from a linear mixture. ICA has been widely applied in the Machine Learning Community as a Feature Extraction [26] technique as a substitute of PCA when one cannot assume that the best choice of features show a Gaussian distribution, and therefore a set of decorrelated features are not, in general, statistically independent.

During this chapter, we will review the classical formulation of ICA using the theoretical framework developed in the previous section, and we will emphasise the differences between the architecture of ICA used in Feature Extraction and Basis-Decomposition ICA (BD-ICA), the most common architecture in fMRI data analysis.

3.1 General formulation of ICA

3.1.1 Formulation of the problem

Let $s_{i\{i=1..N\}}$ be a set of N independent random variables following a non-Gaussian distribution. Let $x_{i\{i=1..M\}}$ be a set of M random variables formed as linear combinations of the N s_i such that the observation[1]

$$x_i{}^j = A_i{}^k s_k{}^j \tag{3.1}$$

We will call *mixing matrix* to the matrix $\mathbf{A}$ formed by the coefficients of Equation 3.1 and *mixtures* to the obtained variables x_i.

Now, let suppose that we have access to D ordered samples of the mixtures. ICA tries to recover the D ordered original samples of the sources s_i without knowing the mixing matrix $\mathbf{A}$.

[1] During this section, we will use the terms *observation* and *sample* to refer to each of the samples taken from the random variables defined before

As an ordering is requested, we can express the set of observations for a given random variable x_i in the form of a vector $\mathbf{x}_i$, and analogously $s_i \longrightarrow \mathbf{s}_i$. In this way, we can formulate the ICA problem in a much more familiar form: to recover the original generative vectors $\mathbf{s}_i$.

The approach of ICA to solve that problem is to exploit the facts that the original vectors have non-Gaussian distributed components and that they are independent among each other.

We actually cannot recover the exact sources $\mathbf{s}_i \in \mathcal{S}$. Indeed, their norm and ordering can be altered without affecting the mixtures by a suitable change in the mixing matrix $\mathbf{A}$. As both, the mixing matrix and the basis vectors, are unknown while facing the problem, it is theoretically impossible to deduce either the norm or the ordering of the source vectors.

Therefore we need to restrict ourselves to find what we call a *faithful representation* of the basis vectors as presented in Definition 2. Note that this representation reflects all the information we can extract about the basis vectors with the information contained in the set of mixture vectors $\mathbf{x}_i$.

Difinition 2 *We will say that a set of n basis vectors $\mathbf{y}_i \in \mathcal{Y}$ is a faithful representation of a set of $N \geq n$ (or a complete faithful representation if $n = N$) basis vectors $\mathbf{s}_j \in \mathcal{S}$ iff*

$$\forall\, i \;\; \exists!\, j \quad \textit{such that} \quad \langle \mathbf{y}_i, \mathbf{s}_j \rangle \neq 0 \tag{3.2}$$

Note that in Definition 2 we are allowing only one vector $\mathbf{y_i}$ for each of the original basis vectors $\mathcal{S}$. Also notice that anti-parallel solutions are perfectly allowed (i.e. a negative scalar product is as good as a positive one).

We are now in position to formulate the ICA problem:

Difinition 3 *Given a set of N independent non-Gaussian distributed basis vectors $\mathcal{S}$ and a set of linear mixings $\mathcal{X}$ of such basis vectors, the final purpose of ICA is to find a set $\mathcal{Y}$ being a complete faithful representation of $\mathcal{S}$.*

In general, of course, the exact size of the set $\mathcal{S}$ remains unknown to the experimenter. Several strategies have been proposed to estimate this number, but they are out of the interest and scope of this work.

3.1.2 Independence and non-Gaussianity

The key element in ICA is the maximisation of the statistical independence of the extracted components $\mathbf{y}_i$. This is partially the reason why ICA, being originally a BSS methodology, has become so popular in the Machine Learning community as a dimensionality reduction technique. If the outputs of PCA are decorrelated, the outputs of ICA are statistically independent, a much stronger condition than decorrelation. Note that, whereas this is just an interesting property

for Feature Extraction, it is of vital importance in the sense of Definition 3, since otherwise we cannot guarantee that the obtained components are actually a faithful representation of $\mathcal{S}$.

The next step in our exposition should be, therefore, to find a way of measuring such independence. For that, we will need to use the assumption that the mixings are produced as a linear combination of non-Gaussian sources.

Theorem 1 *Let each of the observations $\mathbf{x}_i \in \mathcal{X}$ be a linear combination of a set of non-Gaussian random variables $\mathbf{s}_i \in \mathcal{S}$ and let $\mathbf{y}_i \in \mathcal{Y}$ be a set of linear combinations of the mixings $\mathcal{X}$. Then, the two following affirmations are equivalent:*

1. *The statistical independence among the member in $\mathcal{Y}$ is maximal.*
2. *The shape of the members of $\mathcal{Y}$ is maximally non-Gaussian.*

An intuition of the proof for this theorem can be draw considering the implications of the Central Limit Theorem (see [27], Chapter 9). In short, a main consequence of that problem is that a mixture of any two mutually independent non-Gaussian distributions is always more Gaussian than any of the original distributions. Therefore, maximising non-Gaussianity of a linear combination of mixtures guarantees that the obtained solution is a fair representation of one of the original sources $\mathbf{s}_i$. A more formal proof of this can be found in [3].

Therefore, to evaluate the independence of a candidate component $\mathbf{y}$ we can just measure its non-Gaussianity. This is a quite interesting result because, among other things, it allows us to focus in a single potential solution at each time (i.e. to maximise global mutual independence we just need to focus in guaranteeing for each of the components to be maximally non-Gaussian).

3.1.3 Measuring non-Gaussianity

We have already argue that a valid strategy to find the correct set of $\mathbf{y}_i$ is to try to maximise its non-Gaussianity. Before going any further, let us define what a non-Gaussianity measures looks like.

Difinition 4 *Let $\mathbf{y} \in \mathbb{R}^D$ be a set of D samples measured from a distribution y. A non-Gaussianity measure is a function*

$$\begin{aligned} J: \quad & \mathbb{R}^D \longrightarrow \mathbb{R} \\ & \mathbf{y} \longrightarrow J(\mathbf{y}) \end{aligned} \tag{3.3}$$

that reaches its minima when y is a Gaussian distribution and it gets larger and larger when the distribution is further and further from a Gaussian shape.

Likelihood approach to measure non-Gaussianity

Respecting its Gaussianity, a non-Gaussian distribution can be either super- or sub-Gaussian. These names represent the fact that the peak in the mode of a super-Gaussian distribution is higher than the one for a Gaussian distribution, provided that both have the same variance.

One way to measure the non-Gaussianity of a set of observations is to maximise its likelihood with a super- or sub-Gaussian distribution. This approach requires us to make further assumptions over the data (i.e. we need to stablish if we are looking for either super or sub-Gaussian distributed sources), and we are not going to use it. However, it can be very helpful in a lot of applications (e.g. fMRI Resting-State networks are known to be super-Gaussian [15]).

Negentropy as a measure of non-Gaussianity

A more sophisticated approach to non-Gaussianity is provided by Information Theory as an application of the entropy. Entropy is a well defined quantity representing the amount of information needed to codify the observations of a given distribution. Information Theory establishes that this information in larger for samples measured form distributions closer to a Gaussian, reaching its maxima for the Gaussian distribution itself (provided that the comparison have been done between distributions with the same mean and variance). A proof for this affirmation in relation with ICA can be found in [3].

It can be proven (we refer the reader again to [3]) that likelihood and Negentropy maximisation are equivalent when the correct assumption about the shape of the distribution have been settled in the likelihood target function. Furthermore, other approaches to this measure like directly maximising statistical independence or minimising mutual information can be directly proven to be equivalent to the objective of maximising Negentropy of the components. Therefore, Negentropy offers a general and reliable approach to ICA.

Difinition 5 *Consider an arbitrary distribution y and a Gaussian distribution ν with same variance and mean that y. Then, we define Negentropy as the following quantity:*

$$J(y) \equiv H(y) - H(\nu) \tag{3.4}$$

where $H(\cdot)$ is the entropy of a given distribution.

It is easy to see that $J(y)$ is a negative quantity that increases with the non-Gaussianity of the distribution y. It is, therefore, a perfect measure of Non-Gaussianity as expressed in Definition 4.

Negentropy is, however, difficult to compute. The entropy of a set of observations must to be computed from its generative probability distribution function (pdf) whose estimation from the data can be computationally prohibitive. Therefore, approximations to Negentropy should be used instead.

Approximations to Negentropy

The most common of those approximations is Kurtosis, the fourth standardised momentum of a distribution. However, Kurtosis depends on a power of 4 with respect to the deviance of the observations, and it is therefore quite unstable facing outliers.

A general approximation for Negentropy was proposed by Hyvarinen [26] by assuming that the distribution of y was near to a Gaussian distribution. This approximation can be simplified to the following expression[2]:

$$J(y) \simeq k(E[G(y)] - E[G(\nu)])^2 \tag{3.5}$$

where $G(y)$ is an arbitrary function of y exhibiting a growth lower than a quadratic function, k is a multiplicative constant that depends on the specific choice of $G(y)$ and $E[x]$ stands for the expectation value of x, which when we only have access to a discrete set of samples $\mathbf{x}$ can be written this way:

$$E[\mathbf{x}] \equiv \frac{1}{N} \sum_{i}^{N} x^i \tag{3.6}$$

A wise choice of $G(y)$ makes this quantity very robust to outliers. A very common choice in this context is:

$$G(y) = -e^{-\frac{y^2}{2}} \tag{3.7}$$

The approximation draw in Equation 3.5 does not, in general, hold. However, it can be proven that it is still a monotonic increasing function of the non-Gaussianity of the distribution of y and, therefore, is a prefect valid choice as a non-Gaussianity measure in the terms of Definition 4.

3.2 Two architectures for ICA

In the Knowledge Discovery context we can understand ICA in two different architectures. The first architecture is largely related with this work and interprets the source vectors $\mathbf{s}_i$ as the basis of the vector space where the observations lie. The second architecture is more related with Dimensionality Reduction and interprets the mixings as the original features of the data and the sources as the target extracted set of features after the Dimensionality Reduction process.

[2]The notation here is a little bit messy, but this cannot be expressed in a more precise way. In the following expression, y is used to name a set of samples of the random variable y. We could make this using a vector as before, but the $G(y)$ function has to be a single variable function. This problem is produced by the unfortunate but common notation for the expectation value, which has to be defined over a set of variables when used to characterise observations

3.2.1 Basis Decomposition oriented ICA

Let $\mathcal{S}$ be a basis spanning the vector space $\mathcal{V}$ whose representations follow a non-Gaussian distribution and $\mathcal{X}$ be a set of M observations represented by vectors $\mathbf{x}_i \in \mathcal{V}$ of dimension D.

Difinition 6 *Basis Decomposition oriented ICA (BD-ICA) is an architecture for ICA that finds a given number n of mutually orthogonal linear applications $w_i \in \mathcal{W}^M$ (as defined in Equation 2.4) such that the resulting linear combinations $\mathbf{y}_i = {w_i}^j \mathbf{x}_j$ present the possible less Gaussianity measure of all the elements in $\mathcal{V}$.*

In other words, Definition 6 finds the *more non-Gaussian* elements in the vector space spanned by the set of observations $\mathcal{V}$.

In the fMRI context, this architecture assumes that there exist some statistically independent spatial patterns in the 3-Dimensional volumes that are activated in a sympathetic way along time. The linear coefficients of this patterns (i.e. the numbers constituting the mixing matrix **A**) can be understood as the time courses we talked about in Section 2.1.3 when performing ICA over a single individual.

Since this architecture assumes spatial independence, the method is usually called *Spatial-ICA* in the Neuroimage literature.

3.2.2 Feature Extraction oriented ICA

The other architecture for ICA is summarised in the following definition:

Difinition 7 *Feature Extraction oriented ICA (FE-ICA) is an architecture for ICA that finds a given number n of mutually orthogonal linear applications $\boldsymbol{\xi}^i \in \mathbb{R}^D$ such (as defined in Equation 2.1) that[3] the resulting projections $y^i_j = \langle \boldsymbol{\xi}^i, \mathbf{x_j} \rangle$ constitute a new representation of $\mathcal{X}$ in a subspace $\mathbb{R}^n \subseteq \mathbb{R}^D$, in such a way that the estimated distribution behind the observations in a given variable of the new feature vectors are as less Gaussian as possible.*

This last definition maximises the non-Gaussianity of the distribution of the features after the projection. The assumption here is not that there exist some generative patterns in the feature space, but that the final features we want to represent our data are maximally non-Gaussian and therefore independent among them. This can be very useful in Feature Selection as it could allow the experimenter to obtain a non-redundant representation of its data before training a classifier.

Nevertheless, this approach can be also useful in Knowledge Discovery. In the fMRI context, we are simply making a very different assumption than in the case of (BD-ICA). The hypothesis here is that there are some patterns behind the data hidden in the temporal domain, being

[3]Note that, in this case, $\boldsymbol{\xi} \in \mathbb{R}^D$, not $\boldsymbol{\xi} \in \mathcal{V}$.

mixed in a linear way in the spatial domain. Therefore, we are assuming that now the temporal patterns are independent. This hypothesis has demonstrated to not be very useful in fMRI image, but other techniques showing a better temporal resolution like Electroencephalography (EEG) imaging show a very good compatibility with this architecture [15]. For these reason, in the Neuroimage context this approach is usually called *Temporal-ICA*[4].

This is also the approach used in the Cocktail Party problem, the best well known example of the BSS family. In this scenario, a set of microphones record a mixture of several conversations in a party. The particular location of the microphones is what determines here the coefficients of the mixing matrix. Analogously, the source vectors are the original speeches of each of the people in the party, whilst the mixings are in this case the recordings of the microphones. Once more, assuming independence along the time dimension, we can recover a faithful representation of the source conversations.

3.3 Preprocessing before ICA

Besides classical noise-reduction and other filtering techniques, some data-transformation like preprocessing is usually made before applying ICA to a given dataset. More specifically, a first run of PCA can make the process much easier and computationally cheaper.

3.3.1 The case of BD-ICA

Consider now the BD-ICA architecture presented in Definition 6. Note first that the search space of ICA is in that case the vector space $\mathcal{V}$. This is an immediate consequence of characterising the Independent Component as w-applications (see Equation 2.4).

Now, according to the assumptions draw in the Problem Definition (Section 2.3.2), $\mathcal{V}$ is spanned by the original N source vectors $\mathbf{s}_i \in \mathcal{S}$. Therefore, we can characterise this vector space with a set of N linearly independent vectors lying in that vector space. That means that we can represent it by just a subset of N linearly independent (or a set of N non-singular linear combinations of) observations.

PCA preprocessing

PCA naturally finds an orthogonal set of decorrelated linear combinations of the observations holding the maximal amount of variance. This last consideration is interesting because the noise can introduce subtle differences in otherwise parallel observations. We want to preserve as much variability as possible of the (actually larger) vector space spanned by the observations.

Therefore, we can use the first N Principal Components (i.e. the N ones with higher eigenvalues[5].) derived from our observation set $\mathcal{X}$ to span the search space of the algorithm, therefore

[4] A very interesting discussion about these two architectures in the Neuroimage context is exposed in [15]

[5] Actually, a possible strategy to estimate the number of Independent Components is to use the eigenvalues of PCA to estimate how many vectors do we need to span the vector space $\mathcal{V}$

reducing the computational cost of the search.

From now on, we will consider the set of reduced whitened samples $\mathbf{z}_i \in \mathcal{Z}$ as the result of applying PCA to the original set of samples $\mathcal{X}$. Let $\mathbf{U}$ be the projection matrix built with the N first eigenvectors computed by PCA. If we apply PCA over the observations $\mathcal{Z}$ is defined in the following way:

$$z_i{}^j = U_i{}^k \, x_k{}^j \,, \quad 1 \leq i \leq N \tag{3.8}$$

Note that, however, any other kind of Dimensionality Reduction can be used to form the $\mathcal{Z}$ subset. The only constraint is that the vectors have to be linearly independent. However, as we will see in a moment, the PCA Feature Extraction technique has some other desirable effects over the search process.

The reduced space $\hat{\mathcal{V}}$

In addition, the output of PCA is formed by a white (i.e. decorrelated normally-distributed) set of samples. As decorrelation in Gaussian distributions means independence, we can use the result of Probability Distribution Theory (see [27], Chapter 7) that establishes that any unitary linear combination of independent normally distributed samples is also normally distributed.

This implies that, if we use a subset of the output of PCA to span the vector space $\mathcal{V}$ and we restrict ourselves to unitary w-applications we are actually reducing the search space to a subset $\hat{\mathcal{V}} \subset \mathcal{V}$ in which all samples are normally distributed[6]. This restriction allows us to fix at the same time the variance of the Independent Components $\mathbf{y}_i \in \mathcal{Y}$, a free parameter in ICA.

We will call $\hat{\mathcal{W}}$ to the set of those unitary w-applications defined over the reduced space of observations. Note that this unitary restriction simply means that the vector representation of the application is unitary and its dimension is N:

$$\hat{\mathcal{W}} = \{w \in \mathcal{W}^N \quad \text{such that} \quad \|\mathbf{w}\|^2 = 1\} \tag{3.9}$$

To maintain a clear document, we will not use any different notation for these applications. If not specified, assume for now on that any used w-application belongs to $\hat{\mathcal{W}}$.

In addition, this reduction of the search space largely simplifies the measure of non-Gaussianity, that no longer needs to be mean/variance independent. In the case of Negentropy, that means that the right side of the difference in Equation 3.5, $E[G(\nu)]$, will remain constant for all considered $w \in \hat{\mathcal{W}}$ Great news if we use a Gradient Ascend based technique (and we will).

[6]Note that, however, $\hat{\mathcal{V}}$ is no longer a vector space.

3.3.2 The case of FE-ICA

All this results hold in the case of FE-ICA (Definition 7) with a few small changes. To begin with, the PCA has to be performed over the features instead of the observations. Then, instead of redefining the observations $\mathbf{x}_i$ we need to redefine the features $\mathbf{x}^i$ (imagine $\mathbf{x}^i$ as columns of an hypothetical matrix formed by gathering as row vectors the mixings $\mathbf{x}_i$). Let $\mathbf{U}$ be the eigenvector matrix from PCA. Then:

$$z_i{}^j = U_k{}^j \; x_i{}^k \; , \quad 1 \leq i \leq M \tag{3.10}$$

where, as before, M is the number of samples.

Also, instead of using an unitary version of the w-applications, we will need to use an unitary version of the 1-forms ξ to fix the variance of the Independent Components and to keep constant the variance and mean of the distributions considered by the non-Gaussianity measure.

The rest of differences are trivial and they are not really necessary to this exposition. The pre-processing for this architecture is largely developed in [26] using a slightly different framework.

3.4 Algorithms for ICA

There are currently two well-developed architecture independent approaches to solve the ICA problem as introduced in Definition 3.

3.4.1 Very briefly: the infomax approach

The most elegant approach is infomax [28]. It uses a Feed-Forward two-layered Neural Network to find the Independent Components. In this approach the weights in each of the neurons in the hidden layer represents an Independent Component. The optimisation algorithm then tries to maximise the information preserved by the network during the transformation produced from the input to the output layer. It can be proven that the preserved information is maximised when the mutual information between the components represented in each of the neurons is minimal if we set the activation function of the neurons to be similar to the cumulative probability function (cdf) of the expected distribution of the ICs.

This approach is equivalent to likelihood maximisation. Actually, it only works if we choose the cdf in the activation function correctly (it doesn't need to be precise, but it should correctly indicate whether if the expected IC is super or sub-Gaussian).

This is the ICA approach used in some of the Discriminant-ICA approaches described in Section 2.2, like [22]. It is also the standard approach implemented in most of the fMRI-oriented ICA software, like FSL-MELODIC [29] or the SPM based GIFT [8].

3.4.2 Projection Pursuit approach

We are not using infomax in this work for two reasons. First of all, we want to extract a small subset of the Independent Components (if not a single one), so we need an approach that let us to extract the components one by one, without worrying about the rest of them.

In the second place, infomax requires to specify the expected kind of non-Gaussianity of the components, making it less general than Negentropy based approaches.

The second approach to ICA satisfy these two requirements. The basic idea of this approach is to find at each time the most promising Independent Component (e.g. the most Non-Gaussian one) until we have obtained a satisfactory set of vectors $\mathcal{Y}$.

In the rest of this section we will develop the Gradient Ascend Negentropy-based version of this algorithm. This is, however, not the state of the art of this version of ICA. Indeed, Hyvarinen et al. have developed a Fixed-Point based algorithm for ICA that outperforms greatly the Gradient Ascend one. This algorithm is called Fast-ICA, presented for the first time in [30][7].

This way of extracting the components is called Projection Pursuit in the BSS literature [15].

3.4.3 Gradient Ascend Methodology for ICA

As discussed, we will now develop a procedure for extracting the ICs with Gradient Ascend using Negentropy as objective function. It can be proven [3] that the Negentropy measure guarantees no spurious maxima (i.e. all local maxima correspond to an Independent Component).

As said before, one Independent Component is extracted at each time and an orthogonalisation restriction between the candidate linear applications and the ones that have been already extracted is used during hill-climbing to guarantee that the subsequent runs do not extract the same ICs more than once.

More specifically, in the architecture of Definition 6 (BD-ICA), we try to find a w-application in each of the runs such that the considered vector $\mathbf{y} = w^i\ \mathbf{x}_i$ corresponds to a maxima in the Negentropy function in the search space $\hat{\mathcal{V}}$. The use of this search space (in opposition to the wider search space $\mathcal{V}$) implies that we are using the reduced observation set of N elements $\mathcal{Z}$ and unitary w-applications $w \in \hat{\mathcal{W}}$, where N is the number of ICs we want to extract.

To find the first Independent Component we first initialise randomly the considered w-application. Then we run iteratively

$$\mathbf{w} \leftarrow \mathbf{w} + \boldsymbol{\nabla} J(\mathbf{y}(\mathbf{w})) \tag{3.11}$$

until some condition has been fulfilled (e.g. the change in $\mathbf{w}$ gets lower than some threshold). In this last equation we have used that $\mathbf{y}(\mathbf{w}) = w^i\ \mathbf{x}_i$.

[7] We talk a little bit more about this in Section 7.2.3

To guarantee that $\mathbf{w}$ remains unitary during the search, we can force it in each iteration by simply running:

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|} \tag{3.12}$$

Other solution is to use a Langrangian approach and add a term into the gradient of the objective function with a multiplier of Lagrange and the term $(\|\mathbf{w}\| - 1)^2$. However, we will not use this approach in this work to keep our algorithm compatible with Second Order approaches to Gradient-Ascend.

Once we obtain the first application w_1 we can search for another one by just imposing w_2 to be orthogonal to w_1. More generally, we will impose to any new application w_i to be orthogonal to all the already found ones w_j for all $j < i$.

This can be done by using Gram-Schmidt orthogonalisation at each step just before the normalisation step:

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \sum_{j<i} \frac{\langle \mathbf{w}_i, \mathbf{w}_j \rangle}{\|w_i\| \|w_j\|} \mathbf{w}_j \tag{3.13}$$

Putting all together we obtain a procedure for extracting the tth w-application corresponding to the tth Independent Component. This procedure is introduced in Algorithm 3.1

Algorithm 3.1: Basis Decomposition ICA

input : $\mathcal{Z}$, set of previously found w-applications $\{w_i\}_{\{i<t\}}$
output: The tth w-application w_t

```
Randomly initialise a unitary vector representation for w_t;
while not convergence do
    Apply Equation 3.11 to w_t using Z;
    Apply Equation 3.13 to w_t using {w_i}_{i<t};
    Apply Equation 3.12 to w_t;
end
```

The explicit expression for the gradient of the Negentropy is derived in Section B.3.

This process is identical for the architecture of Definition 7 (FE-ICA) but considering projectors $\xi \in \mathbb{R}^D$ instead of linear applications $w \in \mathcal{W}$.

3.5 Group-ICA revisited

Now that the ICA base have been formulated, we can review the concepts introduced in Section 2.1.2.

Let first consider a single-subject fMRI recording as presented as an example in Section 3.2.1. The recording of this subject can be expressed as a set of L ordered observations $\mathbf{x} \in \mathcal{X}$. Moreover, in relation with the sources $\mathcal{S}$, we can establish that the coefficients of the mixing matrix $\mathbf{A}$ defined as in Equation 3.1 constitute the time courses for each of the $\mathbf{s} \in \mathcal{S}$. Indeed, if we consider the ith source vector, the set of coefficients ${A_j}^i$ with $j = 1..L$ corresponds to the time courses of that source vector.

Note, however, that this ordering is not important during the search of the ICs, and it only plays an important role when interpreting the results. As we have already argued, we do not even need the whole set of observations to reproduce the ICs.

Now, let us consider an analysis involving the recordings of k subjects, in which $\mathcal{X} = \bigcup_{i=1}^{k} \mathcal{X}_i$, where the set $\mathcal{X}_i$ represents the set of observations belonging to a given subject. In this case, the mixing matrix represents a concatenation of the time courses of the subjects, and therefore we are assuming that the Independent Components are the same for all subjects. That does not mean that all the ICs have to be equally expressed in all of them. On the contrary, if the ith IC is not activated at all for the subject m, we would just observe that the mixing coefficients ${A_j}^i = 0$ for all j such that $\mathbf{x_j} \in \mathcal{X}_m$. This partially explains why we can make some spatial inference based only in the temporal courses of the Independent Components.

Chapter 4

The Fisher's Linear Discriminant

The Fisher's Linear Discriminant (FLD) was introduced in [31] to evaluate the interest of a certain linear function of the features describing the data to characterise different taxonomic classes. The discriminant has been widely used in the Machine Learning Community as the key piece of Linear Discriminant Analysis (LDA), a supervised Feature Projection technique capable of finding the projection of the data showing the best linear separability [4].

In this chapter, we will briefly review the classical discriminant to set the background to develop a generalisation of FLD for arbitrary data transformations. Later, we will show how a Gradient Ascend algorithm would work for such generalised FLD to perform a generalised version of LDA.

4.1 Classical formulation of the FLD

4.1.1 General considerations

In this report, we will understand the term *linear discriminant* as a measure of the linear separability of a given representation of a dataset as reported in [4]. Note that here the *linear* constraint is applied to the frontiers separating the dataset in its classes, not to the particular representation of the dataset.

The Fisher's Linear Discriminant is in this sense a linear discriminant. A kernelised version of FLD [32] extends the method to consider also non-linear frontiers. However, we will limit ourselves to the linear version of method during the rest of this exposition.

In addition, the FLD as presented in [31] only considers representations of the data built using linear 1-dimensional projections and problems with two classes. The generalisation of the FLD for more than two classes have been widely used in the literature [33], but we will consider only the binary case in this work.

Let us now formally introduce the FLD. Let $\mathcal{X}$ be a dataset with observations $\mathbf{x} \in \mathcal{V}$ and labels $c_i \in \mathcal{D}$ defined over the space of (two) classes $\mathcal{C}$. Let $\mathcal{X}_{(c)} \subset \mathcal{X}$ be the set of observations belonging to the class c.

Difinition 8 *The Fisher Linear Discriminant is a Linear Discriminant considering the representation built by projecting the dataset over a projector ξ:*

$$\begin{aligned} \Phi: \quad \mathcal{V}^M, \mathcal{C}^M, \mathcal{V}^* &\longrightarrow \mathbb{R} \\ \mathcal{X}, \mathcal{D}, \boldsymbol{\xi} &\longrightarrow \Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi}) \end{aligned} \tag{4.1}$$

The function $\Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi})$, specified in Equation 4.1, is directly proportional to the distance between the centroids of the projections of each subset $\mathcal{X}_{(c)}$ and an inversely proportional function of the distance among observations belonging to the same subset $\mathcal{X}_{(c)}$.

This definition settles an elegant solution to the problem of how to measure linear separability. Note that, as the distance is by definition a positive quantity, the FLD is a definite positive function.

4.1.2 Analytical formulation

The analytical representation of the FLD as presented in Definition 8 is easy to compute and have been developed widely in the literature (see, for example, [4]). However, we will reproduce here that derivation as its comprehension is crucial to the later generalisation.

Now, let us call $\hat{x} = \langle \boldsymbol{\xi}, \mathbf{x} \rangle$ to the projected samples. Then we can define the mean of each class as

$$\mu_c \equiv \frac{1}{M_c} \sum_{\mathbf{x} \in \mathcal{X}_{(c)}} \hat{x} \tag{4.2}$$

and the scatter of each class

$$\sigma_c \equiv \sum_{\mathbf{x} \in \mathcal{X}_{(c)}} (\mu_c - \hat{x})^2 \tag{4.3}$$

where M_c represents the number of samples in $\mathcal{X}_{(c)}$. Then:

$$\Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi}) \equiv \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \tag{4.4}$$

4.1.3 Matrix Formulation

Usually a matrix representation of the FLD is used in the applications. Let us introduce the within scatter matrix $\mathbf{S_W}$:

$$\mathbf{S_W}(\mathcal{X}, \mathcal{D}) \equiv \frac{1}{M}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{4.5}$$

and between scatter matrix $\mathbf{S_B}$:

$$\mathbf{S_B}(\mathcal{X}, \mathcal{D}) \equiv \frac{1}{M} \sum_{c \in \mathcal{C}} \sum_{x \in \mathcal{X}_c} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^T \tag{4.6}$$

where $\mathbf{m}_c$ represent the mean of the subset $\mathcal{X}_{(c)}$ and $\mathbf{m}$ the mean of $\mathcal{X}$. As many times before, M equals the number of observations in $\mathcal{X}$.

Note that Equation 4.5 captures the mean distance between the samples belonging to the same class whereas Equation 4.6 reflects the mean distance between the centroids of the two classes.

Now, if we project the within scatter matrix $\mathbf{S_W}$:

$$\begin{aligned} \boldsymbol{\xi}\ \mathbf{S_W}\ \boldsymbol{\xi}_t &= \boldsymbol{\xi}\ (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\ \boldsymbol{\xi}_t \\ &= (\boldsymbol{\xi}\ \mathbf{m}_1 - \boldsymbol{\xi}\ \mathbf{m}_2)^2 \\ &= (\mu_1 - \mu_2)^2 \end{aligned} \tag{4.7}$$

The same process for the between scatter matrix $\mathbf{S_B}$ yields:

$$\begin{aligned} \boldsymbol{\xi}\ \mathbf{S_B}\ \boldsymbol{\xi}_t &= \sum_{c \in \mathcal{C}} \sum_{x \in \mathcal{X}_c} \boldsymbol{\xi}(\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^T \boldsymbol{\xi}_t \\ &= \sum_{c \in \mathcal{C}} \sum_{x \in \mathcal{X}_c} (\boldsymbol{\xi}\ \mathbf{x} - \boldsymbol{\xi}\ \mathbf{m}_c)^2 \\ &= \sum_{c \in \mathcal{C}} \sigma_c^2 = \\ &= \sigma_1^2 + \sigma_2^2 \end{aligned} \tag{4.8}$$

Therefore, we can rewrite Equation 4.4 as

$$\Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi}) = \frac{\boldsymbol{\xi}\ \mathbf{S_B}(\mathcal{X}, \mathcal{D})\ \boldsymbol{\xi}_t}{\boldsymbol{\xi}\ \mathbf{S_W}(\mathcal{X}, \mathcal{D})\ \boldsymbol{\xi}_t} \tag{4.9}$$

4.1.4 Dimensionality of the output of LDA

Linear Discriminant Analysis (LDA) aka Fisher's Discriminant Analysis is a technique that, maximising the FLD with respect to the projection vector $\boldsymbol{\xi}$, finds the best possible 1-Dimensional projection of the data to discriminate between the classes.

The fact that the representation is 1-Dimensional makes of LDA a perfect candidate for a linear classifier: we just need to fix a threshold in the resulting feature to use the projection to classify new observations. It can be, however, a drawback when dealing with Feature Projection.

This limitation is inherent to the way in which LDA finds the maxima in the FLD (see [4]). In the multi-class generalisation of the discriminant the limitation also appears in a more general way: the technique allows to extract a total of number-of-classes $-$ 1 projections of the data. Some strategies have been developed to extend this number to an arbitrary number of projections. We will adopt the strategy proposed in [34], where the authors use the same methodology we exposed for the extraction of the Independent Components applied to the LDA problem (i.e. the step shown in Equation 3.13).

Even with those solutions, it is interesting to outline the restriction from a Knowledge Discovery perspective. When we consider two classes, LDA outputs exactly one linear combination of features to distinguish them. This is of course enough if the problem is linearly separable. Consider now the case of three classes. In that case, even if the problem is linearly separable, we need at least two dimensions to successfully separate the data. Actually, all we need is a plane (we do not really care about the particular direction of the two extracted projections, as long as they draw the correct plane).

This considerations are literal in the LDA problem. It can be proven (see, again, [4]) that, for a given dataset with C classes, LDA provides a $C-1$ hyperplane to project the data in the form of $C-1$ projectors. The particular vectors are, however, not uniquely determined, as any set of projectors defining the same manifold draw a maxima in the FLD function.

We will come back to these considerations later in the context of Knowledge Discovery, to show that the best strategy to deal with multi-class data is to subdivide the problem into 2-class problems.

4.2 FLD for arbitrary transformations

In this section we will develop a generalisation of the FLD to consider arbitrary parametrised data transformations. Actually, non-parametrisable transformations are also allowed in this context, but this restriction will be necessary later when deriving a Gradient Ascend methodology for the generalised LDA.

The generalisation is performed in two ways: first, the new version of the FLD will allow non-linear transformations. This can be interesting in order to add some prior knowledge to the data. We will use this generalisation later to include a quadratic term in the representation of the data.

Second, the considered transformations will be extended to considered output spaces with more than one dimension. This last generalisation breaks the restriction exposed in 4.1.4 in a much wider sense than the trick of the orthogonalisation step made during Gradient Ascend. Indeed, as we will consider directly transformations to arbitrary dimensional spaces, we will measure the linear separability directly in that space.

4.2.1 Analytical formulation

Consider a general transformation with parameters $\boldsymbol{\alpha} = (\alpha^1, \alpha^2, ...)$:

$$\begin{aligned} \mathcal{T}_{\boldsymbol{\alpha}} : \quad \mathcal{V} &\longrightarrow \mathbb{R}^R \\ \mathbf{x} &\longrightarrow \mathcal{T}_{\boldsymbol{\alpha}}(\mathbf{x}) \end{aligned} \tag{4.10}$$

where R is the number of dimensions in the transformed space.

If we redefine Equations 4.2 and 4.3 in the following way:

$$\boldsymbol{\mu}_c \equiv \frac{1}{M_c} \sum_{\mathbf{x} \in \mathcal{X}_{(c)}} \mathcal{T}_{\boldsymbol{\alpha}}(\mathbf{x}) \tag{4.11}$$

$$\sigma_c \equiv \sum_{\mathbf{x} \in \mathcal{X}_{(c)}} \|\boldsymbol{\mu}_c - \mathcal{T}_{\boldsymbol{\alpha}}(\mathbf{x})\|^2 \tag{4.12}$$

We can rewrite Equation 4.4 as

$$\Phi_{\text{gen}}(\mathcal{X}, \mathcal{D}, \mathcal{T}) \equiv \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\sigma_1^2 + \sigma_2^2} \tag{4.13}$$

Note that Equation 4.4 is equivalent to this last one if we set the transformation $\mathcal{T}_{\boldsymbol{\alpha}}$ to be a 1-Dimensional projection of the data

$$\mathcal{T}_{\boldsymbol{\xi}}(\cdot) \equiv \langle \boldsymbol{\xi}, \cdot \rangle$$

where we have used projector $\boldsymbol{\xi} \in \mathcal{V}^*$ as the parameter of the transformation.

4.2.2 Matrix representation

We can also develop a matrix representation for this generalisation based in the within and between scatter matrices from Section 4.1.3. However, as the transformation is no longer linear we cannot simply try to transform them in the way we did before.

Instead, consider the following decomposition:

$$\begin{aligned} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 &= \sum_{i=1}^{R} (\mu_1{}^i - \mu_2{}^i)^2 \\ &= (\mu_1{}^i - \mu_2{}^i)(\mu^1{}_i - \mu^2{}_i) \\ &= \text{Tr } \left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\right) \\ &= \text{Tr } \mathbf{S_w}(\mathcal{T}_\alpha(\mathcal{X}), \mathcal{D}) \end{aligned} \tag{4.14}$$

A similar reduction can be done for the between matrix:

$$\begin{aligned} \sigma_1^2 + \sigma_2^2 &= \sum_{c \in \mathcal{C}} \sigma_c^2 \\ &= \sum_{c \in \mathcal{C}} \sum_{x \in \mathcal{X}_c} \text{Tr } \left((\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\mu}_c)(\mathcal{T}_\alpha(\mathbf{x}) - \boldsymbol{\mu}_c)^t\right) \\ &= \text{Tr } \mathbf{S_B}(\mathcal{T}_\alpha(\mathcal{X}), \mathcal{D}) \end{aligned} \tag{4.15}$$

Now, putting altogether, we finally obtain:

$$\Phi_{\text{gen}}(\mathcal{X}, \mathcal{D}, \mathcal{T}_{\boldsymbol{\alpha}}) = \frac{\text{Tr } \mathbf{S_B}(\mathcal{T}_{\boldsymbol{\alpha}}(\mathcal{X}), \mathcal{D})}{\text{Tr } \mathbf{S_W}(\mathcal{T}_{\boldsymbol{\alpha}}(\mathcal{X}), \mathcal{D})} \tag{4.16}$$

One interesting conclusion come out explicitly from this form: the discriminant itself is invariant under rotations in the space formed by the transformation $\mathcal{T}$. This is a natural result if we take into account that the discriminant considers linear separation frontiers. The discriminant is invariant under any kind of affine transformation because it reports linear separability no matter the scale or orientation of the data.

This can be seen as an evidence of the considerations exposed in Section 4.1.4: any two transformations defining the same projection plane are equivalent.

4.3 A Gradient Ascend approach for a generalised LDA

LDA can be solved using analytic matrix methods to minimise the expression in Equation 4.9 (see, again, [4]). This methodology is built over the classical FLD, which allows only linear

projections to a $C-1$ dimensional space. In addition, as it is an analytic method, it does not easily allow us to add other objectives to the optimisation process, which is fundamental for combining the FLD with some independence measure.

A more versatile approach to maximise the FLD is to use Gradient Ascend. It will allow us to use general parametrisable transformations $\mathcal{T}_{\boldsymbol{\alpha}}$ and the addition of other objectives can be trivially achieved by just adding an additional term to the objective function.

To construct the objective function of this version of LDA we only need to use the FLD as described in Equation 4.16. To apply LDA, we first need to define what kind of transformation $\mathcal{T}_{\boldsymbol{\alpha}}$ do we want and on which parameters does it depend. Those are the priors of our model. Then LDA will find the best set of parameters maximising the linear separability of the data in the representation defined by the transformation. The only non-fixed variable in $\Phi(\mathcal{X}, \mathcal{D}, \mathcal{T}_{\boldsymbol{\alpha}})$ is, therefore, $\boldsymbol{\alpha}$.

To construct a Gradient Ascend algorithm to maximise the discriminant consider, for the tst set of transformation parameters $\boldsymbol{\alpha}_t$ we want to find, the Algorithm 4.1 with the following Gradient Ascend step:

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \boldsymbol{\nabla}_{\boldsymbol{\alpha}} \Phi_{\text{gen}}(\mathcal{X}, \mathcal{D}, \mathcal{T}_{\boldsymbol{\alpha}}) \tag{4.17}$$

The derivatives required in this last equation are derived in Section B.1. The expression is, of course, dependent on the derivatives of the specific transformation and parametrisation used in the discriminant.

Algorithm 4.1: LDA for general transformations

input : $\mathcal{X}$, $\mathcal{D}$, $\mathcal{T}_{\alpha}$, set of previously found parameters $\{\boldsymbol{\alpha}_i\}_{\{i<t\}}$
output: The tth vector of parameters $\boldsymbol{\alpha}_t$

1 Randomly initialise a parametrisation vector $\boldsymbol{\alpha}_t$;
2 **while** *not convergence* **do**
3 | Apply 4.17 to $\boldsymbol{\alpha}_t$ using $\mathcal{X}$, $\mathcal{D}$, $\mathcal{T}_{\boldsymbol{\alpha}_t}$;
4 | Apply 3.13 to $\boldsymbol{\alpha}_t$ using $\{\boldsymbol{\alpha}_i\}_{\{i<t\}}$;
5 **end**

Chapter 5

Discriminant Independent Component Analysis

We now have all the tools we need to present a solution for our problem, as defined in Section 2.3. As probably the reader has already noticed, we are going to propose a solution grounded on the dual optimisation of the Basis-Decomposition ICA (see Section 3.2.1) and some version of the Fisher's Linear Discriminant that we have not developed yet.

In this chapter, we will first review Discriminant ICA [5], in which the authors introduce a combination of Feature Extraction ICA (see Section 3.2.2) and the classical FLD to achieve a Feature Extraction oriented Discriminant ICA. Then we will present the Basis Decomposition FLD, a BD oriented version of the classical FLD, compatible with the architecture of BD-ICA. Then, we will finally combine the already introduced BD-ICA and the new discriminant to construct BD-DICA, the solution for the problem introduced in Section 2.3, at the beginning of this report.

5.1 Feature Extraction Discriminant ICA

Discriminant-ICA was introduced in [5] presenting a methodology to combine both, FLD and ICA objective functions to construct a Feature Extraction method in which the projected variables were maximally independent (i.e. non-redundant) among each other while presenting good properties for linear discrimination. This is the purpose of a dual optimisation process involving the Feature Extraction architecture of ICA (Definition 7) and the classical FLD (Definition 8).

This section is a brief review of such work presented in our formalism and notation. Any missed demonstration is therefore refereed to the original paper in [5], in which the subject is considered in a much greater detail.

5.1.1 D-ICA objective function

As usual, we will start to build our Gradient Ascend algorithm by defining its objective function

$$\mathcal{J}^{\mathrm{DICA}}(\boldsymbol{\xi}_i) \equiv J(\mathbf{y}_i(\mathcal{X}, \boldsymbol{\xi}_i)) + \kappa\, \Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi}_i) \tag{5.1}$$

where, as in FE-ICA, ${y_i}^j = {\xi^j}_k\, {x_i}^k$. As before, we have used $J(\mathbf{y}_i)$ for denoting the approximation of Negentropy described in Equation 3.5 and $\Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi}_i)$ is the FLD as described in 4.1. The factor κ is a modulation constant to adjust the importance of the FLD in the algorithm.

5.1.2 D-ICA Algorithm

As the normalisation and orthogonalisation steps needed to reduce unimportant degrees of freedom and let the algorithm find several different components are identical in both techniques, the Gradient Ascend algorithm for this problem is straightforward.

For the tth extracted component:

$$\boldsymbol{\xi}_t \leftarrow \boldsymbol{\xi}_t + \boldsymbol{\nabla}\mathcal{J}^{\mathrm{DICA}}(\boldsymbol{\xi}_t) \tag{5.2}$$

$$\boldsymbol{\xi}_t \leftarrow \boldsymbol{\xi}_t - \sum_{j<t} \frac{\langle \boldsymbol{\xi}_t, \boldsymbol{\xi}_j \rangle}{\|\boldsymbol{\xi}_t\|}\, \boldsymbol{\xi}_j \tag{5.3}$$

$$\boldsymbol{\xi}_t \leftarrow \frac{\boldsymbol{\xi}_t}{\|\boldsymbol{\xi}_t\|} \tag{5.4}$$

where, in the expression of Equation 5.3, we have use the fact that the previously found components $\boldsymbol{\xi}_i$ with $i < t$ had been normalised before.

Note that, as the combination described in Equation 5.1 is purely linear, the derivatives are easily computed using the derivatives of ICA and FLD:

$$\boldsymbol{\nabla}\mathcal{J}^{\mathrm{DICA}}(\boldsymbol{\xi}_i) = \boldsymbol{\nabla} J(\mathbf{y}_i) + \kappa\, \boldsymbol{\nabla}\Phi(\mathcal{X}, \mathcal{D}, \boldsymbol{\xi}_i) \tag{5.5}$$

Usually a set of reduced observations $\mathcal{Z}$ is used instead of the original set $\mathcal{X}$ in order to reduce the computational cost of the algorithm. This reduction is usually performed using PCA over the features of the observations as described in Equation 3.10.

5.1.3 Applications and results

The authors test the D-ICA algorithm with various datasets from the UCI database to explore the effect of putting together such different techniques. What they find is that the effect of the

κ just tune what kind of output they will obtain. The performance results when applying a classifier to the projected data got better when the FLD term has more importance than the ICA term. In return, they obtain a more independent set of features when decreasing κ.

Note that the main application of this algorithm is, as we said, to perform Feature Extraction. Under the assumption that statistically independent features present a good characterisation of the data, the ICA term can push the FLD to avoid extremely overfitted solutions.

5.2 The Basis Decomposition FLD

After studying D-ICA, it is clear that the remaining piece to complete the BD-ICA is a version of FLD compatible with the optimisation problem of BD-ICA, as presented in Section 3.4.3. This section is devoted to the development of such discriminant, which we will call Basis-Decomposition Fisher's Linear Discriminant (BDFDL), in analogy with the BD-ICA architecture.

5.2.1 The Basis Decomposition Transformation

The starting point to develop the BDFDL is to consider the generalised FLD (see Equation 4.13) and find a transformation as defined in Equation 4.10 suiting our interests.

Requisites

Such transformation should satisfy two conditions. First, it should be compatible with the architecture of the Gradient Ascend algorithm of BD-ICA. In order to achieve that, we need to arrange the free parameters of the transformation (i.e. the vector $\boldsymbol{\alpha}$) to coincide with the free parameters of the ICA model (i.e. the coefficients of the w-application).

Second, the transformation should carry the prior information we have about the separability of the problem (i.e. the transformed data should be separable under the assumptions of the problem).

The first condition sets $\mathcal{T}_{\boldsymbol{\alpha}} \equiv \mathcal{T}_{\mathbf{w}}$. The second condition requires some further considerations.

As we described in more depth in Section 2.3.1, the discriminative information of the data should rely in the time courses of the sources. In other words, considering the set of sources $\mathcal{S}$ and the set of observations $\mathcal{X}$, the information in a sample $\mathbf{x}_i \in \mathcal{X}$

$$x_i{}^j = A_i{}^k s_k{}^j$$

interesting for the discriminant is contained in the set of coefficients

$$A_i{}^k, \quad \forall k \text{ such that } \mathbf{s}_k \in \mathcal{S} \tag{5.6}$$

Therefore, our transformation should provide the FLD for a representation of those coefficients.

The shape of the transformation

Unfortunately, the only way of getting those coefficients is to first extract all the components using ICA and then perform a GLM to find the best fit for the coefficients. However, a workaround can be done if we use the assumption of the sources being orthogonal. This condition can be written as:

$$\forall \mathbf{s}_i, \mathbf{s}_j \in \mathcal{S} \quad \langle \mathbf{s}_i, \mathbf{s}_j \rangle = s^i{}_k \; s_j{}^k = K(i) \; \delta^i_j \tag{5.7}$$

where $K(i) = \|\mathbf{s}_i\|^2$ and δ^i_j is the Kronecker delta.

Now, consider the following projection of a sample $\mathbf{x}_i$ over one of the sources $\mathbf{s}_j$:

$$\begin{aligned} \langle \mathbf{s}_j, \mathbf{x}_i \rangle &= x_i{}^k \; s^j{}_k \\ &= A_i{}^l \; s_l{}^k \; s^j{}_k \\ &= A_i{}^l \; K(l) \; \delta^j_l \\ &= A_i{}^j \; K(j) \end{aligned} \tag{5.8}$$

Note that, for a given source vector $\mathbf{s}_j$, the multiplicative coefficient $K(j)$ is the same for all the samples, whilst the coefficients $A_i{}^j$ are characteristic of the sample. Then we can use those projections, now for a general projector ξ, to evaluate how good the candidate is, if we force $\xi \in \mathcal{V}^*$.

To understand this better, let us expand a projector in the basis formed by the sources (as $\boldsymbol{\xi} \in \mathcal{V}$ it should be decomposable in terms of the basis elements of $\mathcal{V}$):

$$\boldsymbol{\xi} = a^j \; \mathbf{s}_j, \quad \mathbf{s}_j \in \mathcal{S} \tag{5.9}$$

The desired result is to find a ξ in which the linear coefficient corresponding with one of the discriminant sources is much higher than the rest. As we do not really know the sources $\mathbf{s}_j$ we cannot know the linear coefficients a_j for a given projector ξ, but as the correct choice offers a better separation of the data in terms of the coefficients of the mixing matrix, we can expect for a good projector ξ to have a high FLD if the coefficients of its base are arranged in the desired way, and vice versa. In simpler terms, the FLD will be higher and higher as the projector ξ gets more and more similar to the discriminant sources, reaching a maxima when $\boldsymbol{\xi} = \mathbf{s} \; \forall \; \mathbf{s} \in \mathcal{S}^d$.

We do not really want $\boldsymbol{\xi} = \mathbf{s}$ but $\boldsymbol{\xi} \propto \mathbf{s}$. Actually, allowing the projector representations to have different norms could lead the optimisation process to a solution with a large norm of $\boldsymbol{\xi}$ rather than a solution with a low angle between $\boldsymbol{\xi}$ and $\mathbf{s}$. Therefore, it is convenient to force $\|\boldsymbol{\xi}\| = 1$.

Sometimes, however, the comparison of the weights in the mixing matrix is not enough. Considering the case in which the discriminant source has a large variance in comparison with the others, but all of them have the same mean. For the discriminant to be able to successfully analyse also those cases, it is convenient to add a quadratic term to the transformation. We will resume this discussion after we have achieved a proper parametrisation for the transformation.

The parametrisation

The next natural step is to parametrise the transformation described before to be compatible with BD-ICA.

As for now, the transformation depends just on the projector $\boldsymbol{\xi}$ just like the classical FLD. However, we do not need for this projector to be defined in $\mathbb{R}^D$, since the sources are confined in $\mathcal{V}$. Actually, having a search space larger than we need is counter productive, since it increases the chances of overfitting.

A simple way of making the BDFDL compatible with BD-ICA while constricting the search space for the ξ to the vector space $\mathcal{V}$ is to parametrise the projector using its basis expansion over the set of observations:

$$\boldsymbol{\xi} = \frac{1}{\|b^j \ \mathbf{x}_j\|} b^j \ \mathbf{x}_j, \quad \mathbf{x}_j \in \mathcal{X}$$

but the coefficients b_j are clearly related with the concept of w-application defined in Section 2.4.2:

$$\boldsymbol{\xi} = \frac{1}{\|w^j \ \mathbf{x}_j\|} w^j \ \mathbf{x}_j, \quad \mathbf{x}_j \in \mathcal{X} \tag{5.10}$$

where $\mathbf{w} \in \mathcal{W}$. Or, in the more usual case in which we are using a reduced set of observations:

$$\boldsymbol{\xi} = \frac{1}{\|w^j \ \mathbf{z}_j\|} w^j \ \mathbf{z}_j, \quad \mathbf{z}_j \in \mathcal{Z} \tag{5.11}$$

and now $\mathbf{w} \in \hat{\mathcal{W}}$. We will use this later case from now on to parametrise the transformation.

Analytical expression

Now that we have justified the decisions taken to conform the transformation, we can define it in formal terms. But first, let us introduce a further constraint over the data samples.

The projection of the observations $\mathbf{x} \in \mathcal{X}$ usually have very different mixing coefficients ${A_i}^j$ that can be scaled in a different manner for each of the instances (consider for example a microphone much more distant than the others from the cocktail party). Therefore, the projected mixing

coefficients cannot be directly compared by the discriminant without a previous normalisation of the data instances.

One way of solving this issue is to project directly a normalised version of the samples $\mathbf{x}\ /\|\mathbf{x}\|$ over $\boldsymbol{\xi}$. Another (more efficient) way of doing it is to normalise the data instances as a previous step in the analysis.

We attempted to use the first approach in the formulation of this work to keep the algorithm as much general as possible, but the theoretical exposition became too dirty and incomprehensible, so we have decided to just assume the second one. Remember however that this assumption is not a requirement of the method but a simplification made to let the derivation clearer. The interested reader can trivially substitute any $\mathbf{x}$ by an $\mathbf{x}/\|\mathbf{x}\|$ if it is needed to keep differently normed instances in $\mathcal{X}$ during the implementation. This changes are shown explicitly in Section B.2.1.

Without more preambles, let us define the Basis-Decomposition Transformation.

Difinition 9 *Let $\boldsymbol{\xi}$ be a candidate element of the basis spanning the vector space $\mathcal{V}$, and a normalised observation $\mathbf{x} \in \mathcal{X}$ lying in that vector space. If the instance is not normalised, the transformation normalises it as part of the process. We define the Basis Decomposition Transformation as a mapping $\mathcal{T}$ in the following way:*

$$\begin{aligned} \mathcal{T}_{\mathbf{w}}^{BD}: \quad \mathcal{V} &\longrightarrow \mathbb{R}^2 \\ \mathbf{x} &\longrightarrow \mathcal{T}_{\mathbf{w}}(\mathbf{x}) \end{aligned} \tag{5.12}$$

where

$$\mathcal{T}_{\mathbf{w}}^{BD}(\mathbf{x}) \equiv \begin{pmatrix} \langle \boldsymbol{\xi}, \mathbf{x} \rangle \\ \langle \boldsymbol{\xi}, \mathbf{x} \rangle^2 \end{pmatrix} \tag{5.13}$$

where $\boldsymbol{\xi} = \frac{1}{\|w^j\ \mathbf{z}_j\|} w^j\ \mathbf{z}_i$, with $\mathbf{z}_i \in \mathcal{V}$ and $w^j \in \hat{\mathcal{W}}$

We have justified already the quadratic term because of the possibility of being the variance of the time courses the discriminant parameter of the class structure.

Another way (more Basis-Decomposition like) of understanding this term is to consider the decomposition of an observation[1] $\mathbf{x}$ in its parallel and perpendicular projections to ξ. Obviously, the first component in Equation 5.13 corresponds to $\mathbf{x}_{\|}$. The second component is a monotonic function of $\mathbf{x}_{\|}$. To see that, consider the norm of $\mathbf{x}_{\|}$:

$$\|\mathbf{x}_{\|}\|^2 = \|\mathbf{x}\|^2 - \|\mathbf{x}_{\perp}\|^2 \tag{5.14}$$

[1] We are not asking now for $\mathbf{x}$ to be normalised to make this derivation as general as possible.

The aggregation of both quadratic and linear terms captures therefore the essence of the effect of projecting a given observation into the subspace spanned by the projector parameter $\boldsymbol{\xi}$ and its perpendicular component. If well it is true that the real perpendicular component is not exactly a quadratic term, this representation will become very advisable when we consider the derivatives of the discriminant for the gradient ascend algorithm. And, after all, we are not using this representation in a classifier. It is just a way of letting the FDL to consider terms beyond the linear projection.

Just one more justification for the quadratic kernel comes from the ALFF representation (see [18]) we have introduced in Section 2.1.3. This representation captures the amplitude measured in the frequency space of the time courses. Naturally, it is completely out of the limits of our generalised version of the FLD to perform Fourier Transforms over the time courses as the instances in the algorithm are each of the 3-Dimensional volumes (i.e. we are limiting ourselves to operations we can make with one volume). However, we can take advantage of the Plancherel's theorem, proven in [35]. This theorem establishes that, for a sequence x formed by N elements x_n and its Discrete Fourier Transform $\hat{x} = \mathcal{F}(x)$:

$$\sum_{n=1}^{N} \|x_n\|^2 = \sum_{k=1}^{N} \frac{1}{N} \|\hat{x}_k\|^2 \tag{5.15}$$

The right side of Equation 5.15 is easy identified as the mean of the Fourier transform of the sequence. Therefore, the expected value of the square of the coefficients in the frequency space (the ones considered by ALFF) is proportional to the expected values of the squares of the time courses. ALFF does not consider these coefficients, however, but its square root. Nevertheless, as in the case of the previous justification, we have enough with a representation made with a monotonic increasing function of the quantity we want to measure.

With all that considerations, it could be argue that rather than a generalised function of the FLD we could use directly a quadratic kernel in the FLD. However, please consider that a quadratic discriminant is not the same that a discriminant finding linear frontiers in the space spanned by a linear and a quadratic term. This last representation is much versatile, as it allows to the Discriminant to detect frontiers made in both, quadratic and linear terms at the same time.

5.2.2 The Basis Decomposition FLD Formulation

Now we have all the pieces to formulate the Basis Decomposition version of the FLD:

Difinition 10 *The Basis Decomposition Fisher's Linear Discriminant is the result of imposing the BD-Transformation (Definition 9) over the generalised FLD described in Equation 4.13:*

$$\begin{aligned} \Phi_{BD}: \quad \mathcal{V}^M, \mathcal{C}^M, \hat{\mathcal{W}} &\longrightarrow \mathbb{R} \\ \mathcal{X}, \mathcal{D}, \mathbf{w} &\longrightarrow \Phi_{BD}(\mathcal{X}, \mathcal{Z}, \mathcal{D}, \mathbf{w}) \end{aligned} \tag{5.16}$$

such that

$$\Phi_{BD}(\mathcal{X}, \mathcal{Z}, \mathcal{D}, \mathbf{w}) = \Phi_{gen}(\mathcal{X}, \mathcal{D}, \mathcal{T}_{\mathbf{w}}^{BD}(\mathcal{Z})) \tag{5.17}$$

where $\mathcal{T}_{\mathbf{w}}^{BD}(\mathcal{Z})$ denotes explicitly the dependency of the transformation with the reduced set of observations.

This discriminant is compatible with the desired architecture of ICA, since the only parameters we can vary in the problem are captured by the linear application ξ. The discriminant has two additional practical advantages. In the first place, as the search space of the projector is limited to those vectors in $\mathcal{V}$, the risk of overfitting is considerably lower that in the case of the original FLD, where the search space is $\mathbb{R}^D$. In addition, as we choose $\mathbf{w} \in \hat{\mathcal{W}}$, we only have N variables in the Gradient Ascend process, so a large dimensionality of the observations does not represent a problem in the efficiency of the algorithm.

Summarising, this discriminant is suitable for problems with a large dimensionality, avoiding the most common issues related with those kinds of datasets: the curse of dimensionality and the computational cost of the analysis.

The algorithm for the generalised-FLD-based Linear Discriminant Analysis has been already presented in Section 4.3. The adaptation is straightforward, we just need to use $\mathbf{w} \in \hat{\mathcal{W}}$ instead of $\boldsymbol{\alpha}$ and add the normalisation step of Equation 3.12 to the Gradient Ascend algorithm to force $\|\mathbf{w}\| = 1$. The specific derivatives for the BD-Transformation will be derived in Section B.2

5.2.3 Some empirical evidence

Before going ahead with this discriminant, we performed some empirical tests over an early implementation to check if it was actually a good candidate for performing inference over our data. These tests were fairly simple, but at the same time they proved the necessity of both, the linear and quadratic terms in the BDFLD.

The experimentation was performed using synthetic data built using some basis vectors, a randomly generated set of time courses and some noise. The construction showed different parameters in the generation of the time courses for the discriminative basis vectors and similar parameters for the rest of networks. Then, all the basis vectors were shown to the BDFLD along with the generated data, to see if the discriminant was able to assign a greater score to the discriminative vectors. We repeated those tests for datasets generated using different amounts of noise.

The results were satisfactory for all tested levels of noise for the construction exposed in this section. However, similar constructions showing only the quadratic or only the linear terms were not as satisfactory. Actually, the discriminant did not score correctly to approximately a half of the networks in both cases.

The details of such experiments are not really important, as we will test the discriminant using more formal methods in Section 6.1. But for now this is presented as one of the main causes contributing to preserve both terms in the discriminant.

5.3 Basis Decomposition Discriminant ICA

The Basis Decomposition Discriminant ICA (BD-DICA) represents our solution for the problem addressed in this work. The construction of the algorithm is quite similar to the already shown Gradient Ascend algorithms. Actually, the procedure is the same as the one we used for BD-ICA (see Section 3.4.3), but considering a different objective function.

5.3.1 The BD-DICA Algorithm

The BD-DICA objective function is just the weighted sum of the objective functions of BD-DICA and BDFDL algorithms[2]:

$$\mathcal{J}(\mathbf{w}) \equiv (1-\kappa)\, J\left(\mathbf{y}_i(\mathcal{Z},\mathbf{w})\right) + \kappa\, \Phi(\mathcal{X},\mathcal{Z},\mathcal{D},\mathbf{w}) \tag{5.18}$$

Note that we have made a slight change in the meaning of κ with respect to [5] in order to better modulate the relative importance of both components of the objective function. In this case, $\kappa = 1$ means to ignore Negentropy and focus only on the discriminant part of the algorithm, whereas $\kappa = 0$ is equivalent to an ICA. This parameter can play an important role in the dynamics of the hill climbing if modulated during the Gradient Ascend.

As announced, the rest of the algorithm is analogous to the one presented in Section 3.4.3 for BD-ICA. Algorithm 5.1 shows the procedure to extract the tth w-application corresponding to the ith Component $\mathbf{y}_t$.

Algorithm 5.1: Basis-Decomposition Discriminant ICA

input : $\mathcal{Z}$, $\mathcal{X}$, $\mathcal{D}$, the set of previously found projections $\{\xi^i\}_{\{i<t\}}$
output: The tth projector $\mathbf{w}_t$

1 Randomly initialise a unitary vector representation for w_t;
2 **while** *not convergence* **do**
3 $\quad \mathbf{w}_t \leftarrow \mathbf{w}_t + \boldsymbol{\nabla}\mathcal{J}(\mathbf{w}_t)$;
4 $\quad \mathbf{w}_t \leftarrow \mathbf{w}_t - \sum_{j<i} \frac{\langle \mathbf{w}_t, \mathbf{w}_j \rangle}{\|\mathbf{w}_t\|}\, \mathbf{w}_j$;
5 $\quad \mathbf{w}_t \leftarrow \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$;
6 **end**

[2]From now on we will use $\mathcal{J}$ to denote the objective function of BD-DICA and Φ for the BDFDL discriminant to keep the notation clear

As before, thanks to the linearity of the derivative, we can decompose the gradient in the Equation in Line 2 of Algorithm 5.1 in the following way:

$$\boldsymbol{\nabla}\mathcal{J}(\mathbf{w}_i) = (1-\kappa)\ \boldsymbol{\nabla}J(\mathbf{w}_i) + \kappa\ \boldsymbol{\nabla}\Phi(\mathbf{w}_i) \tag{5.19}$$

This process is repeated several times until we have obtained a desired number of components. One advantage of this approach is that we can visualise on-line the extracted components and monitor the value of the discriminant for each of them, so we can use this information to construct a stop criterion during the run of the algorithm. However, for most of applications, probably one or two components are enough to characterise the differences between two classes.

5.3.2 Multi-Class extension

An analytical multi-class extension of the algorithm has not been either developed or tested. The missing parts to construct such algorithm will be supply shortly in this section, but first let us consider if such an extension is really useful in the context of knowledge discovery, or if it is better to adopt an one-VS-all strategy to deal with multi-class problems.

One-VS-All strategy

Consider a problem with C classes. We have already introduced the limitation of the classical FLD in the dimensionality of the output, that is constricted to have $C-1$ features. We have also mentioned the implications of a projection of more than one dimension: the projection obtained from the FLD is actually a hyperplane with dimension $C-1$, not a set of $C-1$ well-defined features. In other words, all projectors lying in the hyperplane will be acceptable maxima of the FLD.

This considerations suggest that the use of a multi-class FLD is not the best option to characterise a dataset with more than two classes, if we want a characterisation represented by a vector of the same dimension of the instances.

This is perhaps clearer with the fMRI example, in which we want to characterise ground differences among several groups in the fMRI signal. These differences can be physiologically interpreted if the output of the algorithm is an fMRI volume indicating relevant parts of the brain (in the two-classes problem, usually the part of the brain showing different patterns of activation in the two groups). If we consider an fMRI problem with three classes, we should expect to obtain two equally good solutions from the maximisation of the FLD. These solutions are, however, not interpretable as an fMRI volume but as the vector space spanned by them, as any rotation of such vectors would provide the same score in the FLD.

Therefore, we think that it is better to face the multi-class problem following an all-versus-one strategy, therefore obtaining single-Dimensional projectors ξ^i characterising the differences of one class with respect to the others. This result is much likely to be interpretable and probably it has more sense from a Knowledge Discovery point of view, in which it is more interesting

to obtain detailed information about the classes than having a good representation to separate them.

The extension

We cannot imagine a situation in which a hyperplane over the space of the instances can be of any help to characterise a multi-class dataset, but just in case, we will now shape the natural multi-class extension of our algorithm.

As the only supervised part in the objective function of Equation 5.18 is the FLD, we only need to supply a generalisation for that part of the algorithm. This extension can be easily implemented by just substituting the distance between the centroids of the classes by the weighted sum of the distances between these centroids and the centre of the whole dataset $\boldsymbol{\mu}$:

$$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \longrightarrow \sum_{c \in \mathcal{C}} \frac{M_c}{M} \|\boldsymbol{\mu}_c - \boldsymbol{\mu}\|^2 \tag{5.20}$$

Note that the relative importance of a class c is measured as the portion of samples belonging to the given class $\frac{M_c}{M}$.

This has a direct effect in the definition of the within-scatter matrix (see Equation 4.5):

$$\mathbf{S_W}(\mathcal{X}, \mathcal{D}) \equiv \sum_{c \in \mathcal{C}} \frac{M_c}{M} (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T \tag{5.21}$$

The within scatter measure is trivially substituted by the sum of the scatter of each of the classes.

Chapter 6

Experimentation

In this chapter we will test our algorithm in three different datasets. First, we will use synthetic data satisfying our premises regarding the structure behind the data. This will suffice to prove empirically our theoretical derivation.

Later, we will use two real fMRI datasets, to check if the whole process works in this context. The two dataset correspond to the two kinds of fMRI data analysis problems described in this work: Task-Based and Resting-State.

We will centre our experimentation in Knowledge Discovery and Data Characterisation, and therefore we will not test the performance of the algorithm used as a Feature Extraction methodology (actually, we have not formalised yet the application of the algorithm in Feature Extraction, which we will introduce briefly in the discussion in Section 7).

6.1 Synthetic Data

The validity of the results of the algorithm applied to real data is very difficult to assess, since we cannot be sure about what are the real discriminant patterns in the data, if any. The output of this algorithm is, in addition, difficult to compare with other methodologies since the representation of the characterisation is slightly different than other classical techniques (see Section 2.1.3). Therefore, it is convenient to explore the results of the algorithm knowing beforehand what is the precise solution to the problem.

In this section we will solve a series of synthetic problems showing a different amount of noise and generated accordingly with the current state of the art of Resting-State fMRI. The procedure of the construction of the dataset is first specified, and then we will expose the experimentation and the results of the experiment.

6.1.1 The Generator

The data generator built to perform the synthetic experiments of the algorithm is based in the hypothesis that ALFF is a reliable measure of the differences between groups of subjects. This hypothesis has been successfully tested in several works in Resting-State fMIR analysis (see [18]). This measure has been previously detailed in Section 2.1.3.

To construct our dataset we will use the 20 Resting-State networks isolated by Biswal [13] as the basis constructors of the data (i.e the set of initial basis vectors $\mathcal{S}$). Then, we will select randomly n of those basis vectors to construct the set of discriminant vectors $\mathcal{S}^d \subset \mathcal{S}$. This number is a free parameter of the generator, but following the physiological guidance of the generator we will restrict ourselves to $n = 1, 2, 3$.

This set of networks $\mathcal{S}^d$ will activate differently (i.e. will have a different ALFF measure) for two sets of patients showing different mental conditions. In a clinical case, the abnormal activation of those networks is usually connected with the given disorder (ADHD, schizophrenia...). To avoid such dark topics, at least for a while, we will assume that our networks characterise the Jedi abilities of the affected patients.

The number of subjects is another free parameter of the generator. As it is usual to have about 40 subjects in clinical studies, we used that same amount of subjects in our experimentation. In addition, the distribution of the two groups is usually compensated since, while the experimenter has a limited access to affected patients (there are not a lot of Jedies these days), the access to controls is much easier to achieve and to obtain a balanced dataset is always possible. Therefore, our generator splits equally both groups.

Another important parameter is the length (i.e. the number of volumes) of the experiment. Resting-State experiments are usually of 8 minutes which correspond to $L = 240$ volumes in standard machines (with a temporal resolution of 2 seconds). To spare some memory, however, we have set this number to $L = 100$ in our experimentation.

Now we have most of the pieces to explain the algorithm of our generator, outlined in Algorithm 6.1. We only need to specify how do we draw the frequency distributions. This is made by placing a number of Gaussian with random variances and means along the frequency space. The height of the Gaussian is, however, not completely random. This quantity is chosen randomly with a multiplicative factor a which is different for the normal and the abnormal activations (more specifically, we selected $a_{\text{abnormal}} = 2a_{\text{normal}}$).

The frequency distribution and the corresponding time courses of one of those runs are shown in Figure 6.1 for $a = 1, 2, 4$.

The Gaussian noise added at the end of the process is also a free parameter of the generator. The quantity of noise is measured as the quotient between the norm of the generated signal before adding the noise and the norm of the added Gaussian noise. Note that $\alpha = 1$ means that the Gaussian noise is half of the final signal, but additional irrelevant terms are contained in the remaining half: the apportion of the non-relevant networks, for example.

Algorithm 6.1: Generate synthetic Resting-State data

input : Set of source networks $\mathcal{S}$, number of discriminant sources n, dimension of the data D, number of subjects M, length of the recording L

output: The problem $\{\mathcal{X}, \mathcal{D}\}$ and the solution $\mathcal{S}^d$

Randomly select a subset of n discriminant networks to built $\mathcal{S}^d$;
Split the subjects in two groups: Jedies and controls;
for *subject* $j \in 1..M$ **do**
    for *source vector* $\mathbf{s}_j \in \mathcal{S}$ **do**
        if $s_i \in \mathcal{S}^d$ ***and*** *subject* k *is a Jedi* **then**
            $\mu = \mu_{\text{abnormal}}$
        end
        else
            $\mu = \mu_{\text{normal}}$
        end
        Draw the frequency distribution using the parameter μ of L points $\{d_k^i(j)\}_{k=1..L}$;
        Compute the time course as the Discrete Fourier Transform of the distribution d: $\{b_t^i(j)\}_{t=1..L}$;
    end
    Generate the tth volume of the subject j $\mathbf{x}_t \in \mathcal{X}_j$ as the pondered sum of the source networks $\mathbf{x}_t = {b_t}^i(j)\ \mathbf{s}_i$;
end
Aggregate the volumes by subjects to construct $\mathcal{X}$;
Aggregate the labels of each subject to construct $\mathcal{D}$;
Add some Gaussian noise to all the instances in $\mathcal{X}$ using the same parameters;

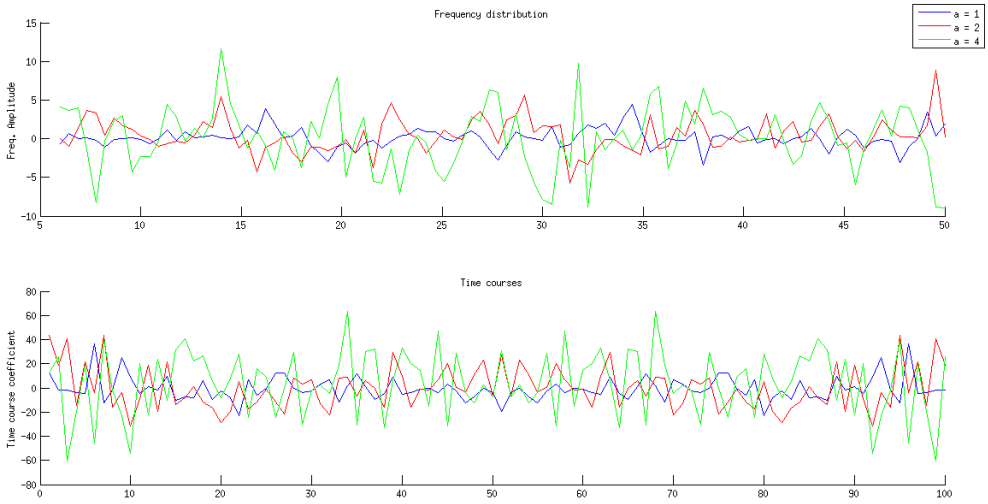

Figure 6.1: Frequency distribution (and its corresponding time course) generated with different Gaussian height baselines a for a run with 100 volumes.

6.1.2 Performed experiments

For the experiments we generated a battery of synthetic datasets with different n (number of discriminant networks) and α (extra amount of Gaussian noise added to the dataset). Specifically, we construct eight datasets with $\alpha = 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1$ for each value of $n = 1, 2, 3$. In total, 24 different datasets sharing all the rest of parameters. We additionally tested the behaviour of the algorithm under optimal conditions ($\alpha = 0, n = 1$).

We should again insist on the fact that the quantity of noise presented is not as small at it seems. To see that, consider the mean difference between the time courses of an abnormally activated network and the time course of a normally activated network δ against the mean value of the amount of noise for a given time course coefficient η. Table 6.1 summarises the amount of noise measured in that way as an average between datasets with the same α. This calculations show that a coefficient of $\alpha = 1$ actually means that the amount of noise per volume is about ten times larger than the mean difference of activation of a discriminant source between a Jedi and a control subject.

The rest of the experimentation was relatively straightforward. We tested the algorithm using different Gradient Ascend parameters and different values for κ. We also asked the algorithm to extract two more networks than the expected ones (i.e. we asked for $n = N + 2$). We used $\kappa = 0.4$ during all the synthetic experimentation.

6.1.3 Results

The results over the synthetic datasets were highly satisfactory. To measure this performance we computed the similarity of the obtained components with the source networks, expecting to

α	0.0005	0.001	0.005	0.01	0.05	0.1	0.5	1
mean abnormal	177	183	198	162	167	172	182	187
mean normal	109	127	101	115	117	112	94	101
η	0.6	1.2	5.0	11	54	98	593	1011
η/δ	0.009	0.02	0.05	0.23	1.1	1.6	6.8	11.7

Table 6.1: This table shows a comparison between the norm of the noise α with the coefficient η/δ. We can see that $\alpha = 1$ is actually equivalent to apply ten times more noise to the dataset than the mean difference between the activation of the discriminant networks in the two groups of subjects. The measures were taken empirically from the data used in the rest of the experimentation.

find a high similarity for the discriminant networks and a low similarity for the rest of them. This similarity was measured as the normalised scalar product of the vector representation of the two networks:

$$\text{sim}(y, s) = \frac{1}{\|\mathbf{y}\| \, \|\mathbf{s}\|} |\mathbf{y}^t \mathbf{s}| \tag{6.1}$$

These similarities are computed for each output of the algorithm and each of the generative networks. Then we label each obtained component with the reference of the generative network and the computed similarity and we compare the labels of the components with the original discriminant networks to see if the output is correct. The mean similarity of the correct outputs is plotted in Figure 6.2 for different levels of noise and amount of discriminant sources.

We can see in Figure 6.2 that the degradation is very small for $\alpha \leq 0.1$, obtaining similarities of sym > 0.95. This degradation increases when $\alpha \sim 1$, probably because the level of noise is destroying the original signal. This seems to indicate that the discriminant part of the objective function does not disrupt the Independent Components too much, probably because in this almost-ideal case the whole IC is entirely responsible for the group differences.

Of course, the similarity is only important when the discriminant source networks are correctly extracted. The found components together with its similarities are represented in the plots of Figure 6.3, in which we also indicate which components actually correspond to discriminant sources.

Note that the algorithm is capable of correctly extracting all the discriminant networks self for the case of $N = 3$, in which a network is missed for $\alpha = 1$. This seems to indicate that the algorithm does not work well with too many discriminant networks and very high levels of noise.

Another interesting effect that can be observed in the results exposed in Figure 6.3 is that the similarity with non-discriminant networks is in general lower than the similarities for the real discriminant ones. This is probably a result of the fact that, when finding non-discriminant networks, both addends in the objective function are working on different problems: the BDFLD wants to get closer to a discriminant position, which now is forbidden by the orthogonalisation step as all discriminant networks have been already found. In the other hand, the Negentropy addend just needs one Independent Component to reach a maxima.

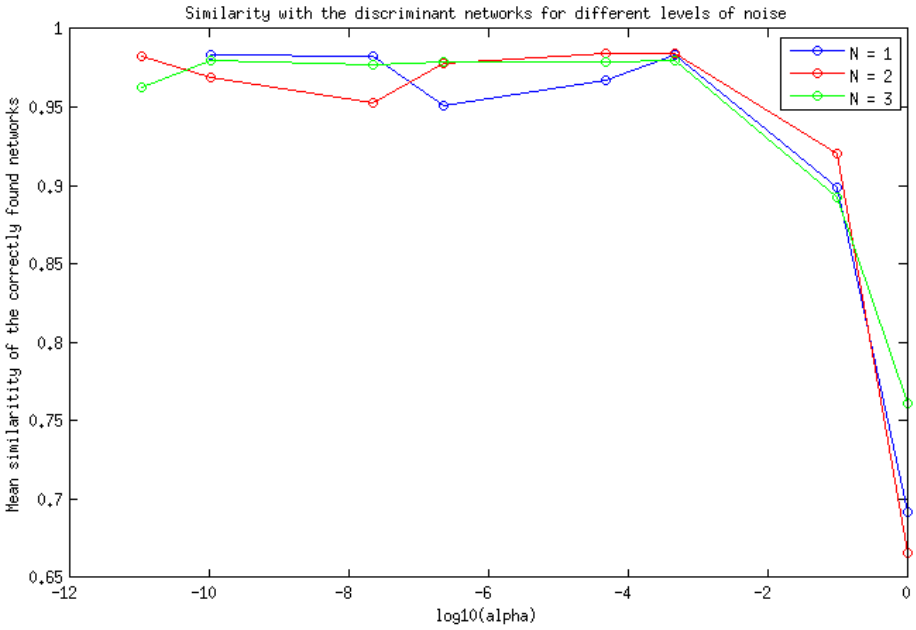

Figure 6.2: Mean similarity of the correctly extracted components for different levels of noise (α) and number of discriminant networks N. Note that the x axis is shown in log scale.

It could also be of interest to check the final value of the BD-DICA objective function of the different solutions. Moreover, we also inspected the final score of the BDFLD and the Negentropy of the solutions. These results are plotted in Figures 6.4, 6.5 and 6.6. The images show that the objective function (Figure 6.4) is not necessarily greater for all the correct solutions, as one might expect. This effect occurs only in cases with $n > 1$ and it seems to be a consequence of the following phenomenon: some linear combinations of networks were obtained as outputs in some of these sets, but they were not considered as solutions because they had a lower similarity with the original networks than previously found entire solutions.

The value of the Negentropy (Figure 6.5) is also more or less independent on the correctness of the solution or the order in which they were extracted. The BDFLD score (Figure 6.6), however, holds much desirable behaviour than the other measures. Indeed, this score is strictly higher for correct solutions than for the other extracted components, including the spurious linear combinations of real sources. This is a very convenient result, because it means that we can use the score of the BDFLD to, in somehow, have a guidance of the value of each solution. This is perhaps the most expected result of all of them, as the role of the determinant is precisely to ensure that the networks are indeed discriminant.

6.1.4 Discussion

The results of the experimentation show what are the limits of our algorithm for this kind of data: a extremely large noise and a large number of components could make the algorithm to

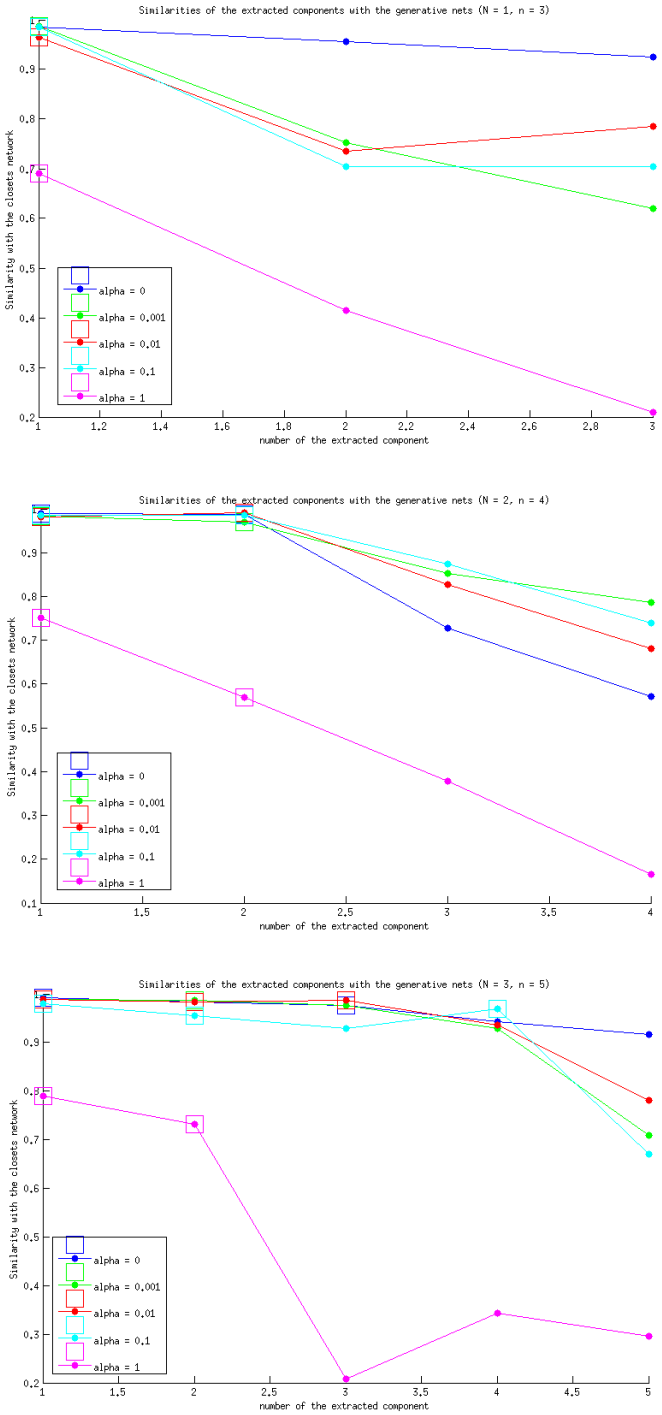

Figure 6.3: Similarities with the nearest network for the extracted components for different levels of noise. Each of the plots correspond to a different number of discriminant sources $N = 1, 2, 3$. Note that in all experiments we extracted $n = N + 2$ networks. The x-axis represents the number of component. The points surrounded by a big square are correct solutions.

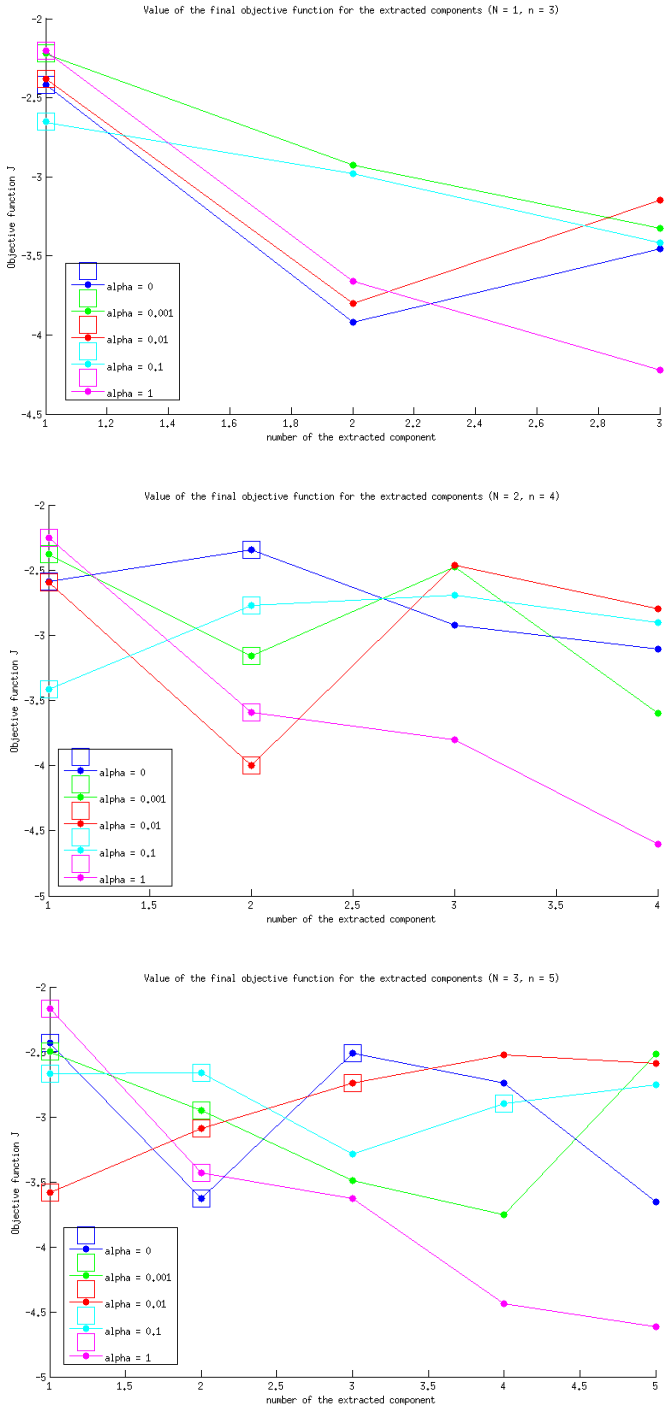

Figure 6.4: Score of the objective function of the BD-DICA algorithm for the extracted components for different levels of noise. Each of the plots correspond to a different number of discriminant sources $N = 1, 2, 3$. Note that in all experiments we extracted $n = N + 2$ networks. The x-axis represents the number of component. The points surrounded by a big square are correct solutions.

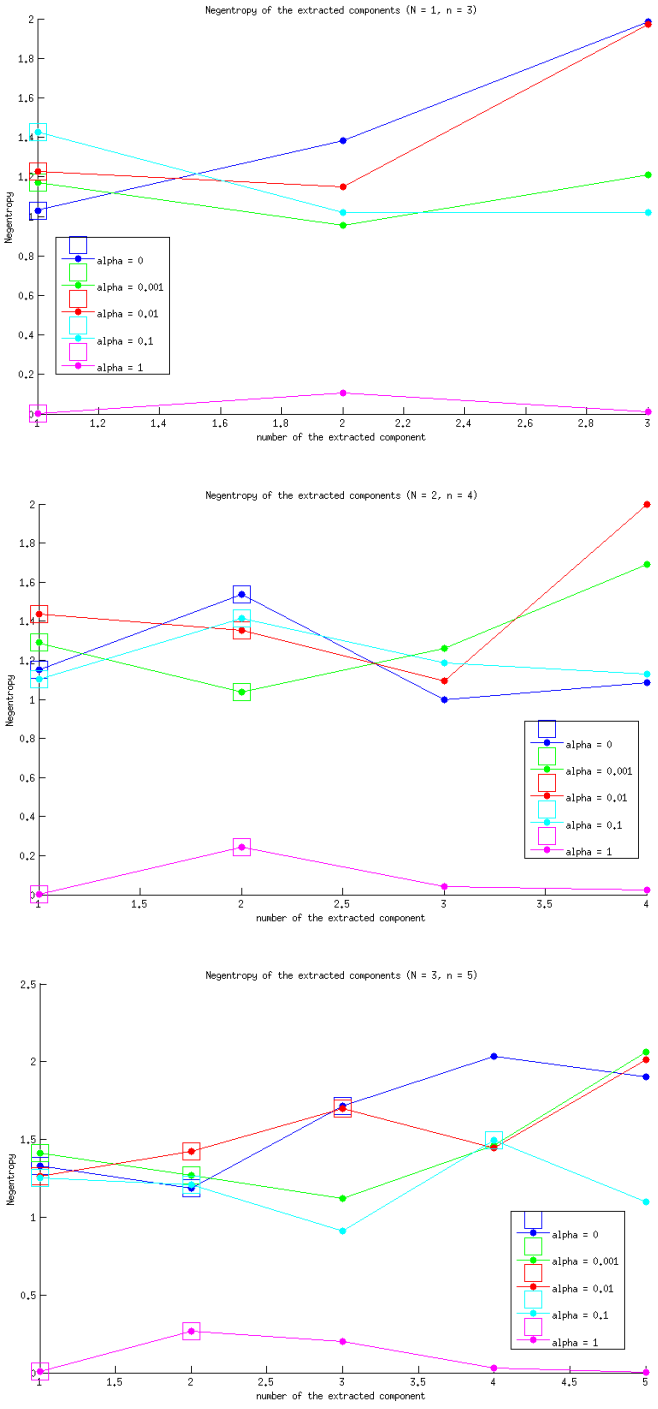

Figure 6.5: Negentropy as measured by the approximation of Equation 3.5 for the extracted components for different levels of noise. Each of the plots correspond to a different number of discriminant sources $N = 1, 2, 3$. Note that in all experiments we extracted $n = N + 2$ networks. The x-axis represents the number of component. The points surrounded by a big square are correct solutions.

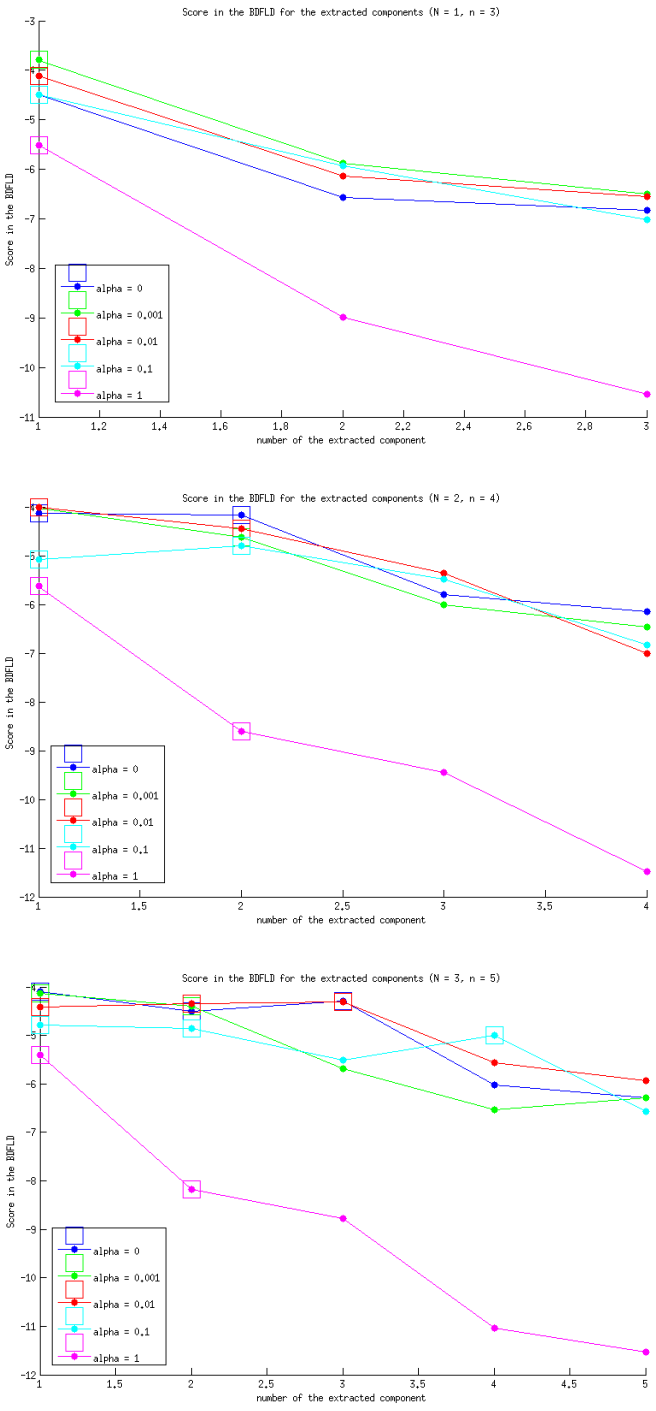

Figure 6.6: Score of the BDFLD discriminant for the extracted components for different levels of noise. Each of the plots correspond to a different number of discriminant sources $N = 1, 2, 3$. Note that in all experiments we extracted $n = N + 2$ networks. The x-axis represents the number of component. The points surrounded by a big square are correct solutions.

miss some solutions. This is not a very restricting problem when dealing with fMRI, as we do not expect for the data to have more than one or two real sources in its structure.

Beside that, we think that BD-DICA offers a very precise representation of the discriminant Independent Components behind the data, even when this sources are not entirely orthogonal, showing an interesting robustness facing different levels of noise.

Of course these results have been obtained using synthetic data satisfying some of the assumptions up to a certain degree, but that does not mean that the problem is trivial. Consider the Figure 6.1 plotting the time courses for different networks of our problem. In our data, we use the baselines $a_1 = 1, a_2 = 2$, which correspond to the red and blue lines in the example. The BDFLD would consider a degraded version (the representation would be exactly the same if the networks were really orthogonal and we had no noise) of those time courses to guess how discriminant are the sources.

The computation time was also relatively satisfactory. Each iteration took about 22 seconds, and depending on the parameters a maxima can be reached in about $10 - 15$ iterations. All of this with a poorly optimised code running over MATLAB. We will talk more about this topic during the general discussion.

A general conclusion we obtained from the experimentation results is that BD-DICA can be understood as a modified BDFLD more than a modified BD-ICA. We prefer to see it this way because it is clear that the leading role in the interpretation of the solutions is contained in the BDFLD. The BD-ICA term in the objective function plays a much discrete role devoted to the avoidance of overfitting, restricting the maxima to those solutions maximising independence meaning.

6.2 Task-Based dataset

The natural continuation of this experimentation is to test the algorithm with a real dataset. As we have oriented most of the exposition towards fMRI analysis, it seems fair to test the algorithm with fMRI data. This is the aim of the following two sections, in which we will present to the algorithm both Resting-State and Task-Based fMRI data.

6.2.1 Description of the dataset

In this section we will use a private clinical dataset we call MST which is described later. The analysis of this dataset was the actual trigger for this project[1] in which we wanted to find physiological indicators of a good recovery in patients treated with a music supported therapy (therefore MST data) after suffering a stroke in the brain. All the subjects of the analysis suffered of a loss of mobility in one of their upper extremities, and all of them were treated with an experimental technique in which the patients combine a standard therapy routine with

[1]The analysis of this data using a more classical approach to temporal inference have been included in this report in the Appendix A. The interested reader can safely skip this description and read the appendix instead

sessions of playing and hearing music. It have been proven that such routines can improve the performance of the therapy in some patients [36].

The objective of the problem assessed with this dataset is to characterise the brain activation of those patients while listening to music showing an improved performance in the therapy in opposition to the patients showing a standard performance in the treatment. This characterisation is expected to be implemented in the form of a connection between parts of the brain, as the therapy is constructed in the basis of an interaction between the motor and auditory cortices of the brain [37]. Therefore, we need to analyse the whole brain at once, and not only a part of it.

This dataset is really limited, because we need to deal only with patients suffering from similar lesions and being treated with the same rare experimental procedure. This limitation is translated in a dataset with just nine patients. The recordings have 200 volumes each, which is translated into 1800 instances.

We have two groups in the dataset: improved performance (4 patients) and normal performance (5 patients). Each instance had about 110000 voxels, mapped in the MNI standard of $4mm \times 4mm \times 4mm$.

The data was appropriately preprocessed in the standard terms in the field, which includes several steps: normalisation (correcting differences among the brains of the patients and correction of the possible movements during the experiment), noise reduction, bandpass filtering and some Gaussian blur to refine possible mistakes in the motion correction [7] [8].

The nature of the problem and the characteristic of the data are introduced in much more depth in Section A.2

6.2.2 Single-Subject Preprocessing

Our analysis started with all the appropriate fMRI preprocessed as explained in the previous section. However, we still needed to perform some high-level preprocessing over the data before using it in out algorithm.

In the first place, we need to perform some Task-Based specific transformations to take into account the structure of a Task-Based fMRI recording. These kind of recordings are composed by a series of task periods separated by resting periods. This resting periods are then subtracted to the task periods to filter secondary activations and artefacts. Usually, the final signal is described as the task signal in which we have removed the mean of the adjacent resting period and we have normalised to that same variance. In that way, all voxels are normalised according to their mean activation during the resting periods. This representation is a variant of the usual z-score used in statistics, in which we characterise the standard behaviour of each attribute by looking at the resting period [7].

After this correction, we preserve approximately one half of the volumes of the original recording (the other half correspond to resting periods). A second step is then done to reduce the cost of the algorithm by reducing a little bit more the number of volumes considered for each subject.

This reduction is simply performed with a PCA (with the same architecture as in Section 3.3.1) run over the volumes belonging to the subject. Note that, as the linear combinations are performed over the volumes of the same subject no class information is destroyed during the process.

At the end of the single-subject preprocessing we preserve between 20 and 50 volumes for each patient (the precise number depends on the eigenvalues of the PCA). After that, the data is gathered in a single dataset and fed to the algorithm. The treated dataset had a final total of 280 instances.

6.2.3 Results

We run the algorithm to find four Components, all of them were quite similar, but complementary to each other. One of this solutions is shown in Figure 6.7.

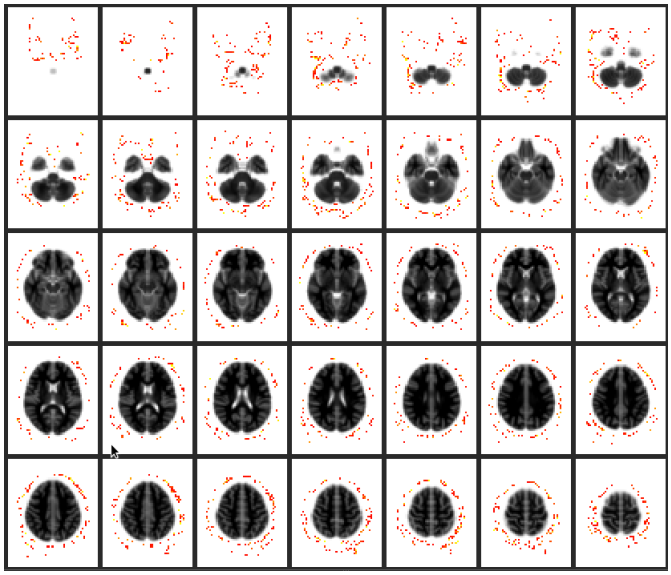

Figure 6.7: A representation of the first Independent Component found by our algorithm. The component was thresholded and superimposed to a standard of the brain to appreciate the position of the activation pattern within the brain

In comparison, we offer the Independent Component (extracted using Group-ICA over this same data with another 26 ICs as described in Section A.3.3) showing the greatest similarity with the solution of our algorithm. This component is shown in Figure 6.8.

This is what is usually called a *noisy network* in the fMRI context, an Independent Component produced by a noisy source (probably due to some not-corrected movement of the patient or some artefact from the recorder[2]). These kind of networks always appear when performing an ICA over this kind of data and they are usually localised in the outside region of the brain. The

[2]Another plausible explanation is that we were just looking a noise resulting from a bad choice of the threshold. The choice of the thresholding point for the components is an open issue in our algorithm. This subject will be discussed again the next section and later in the discussion in Sections 7.1.1 and 7.2.1

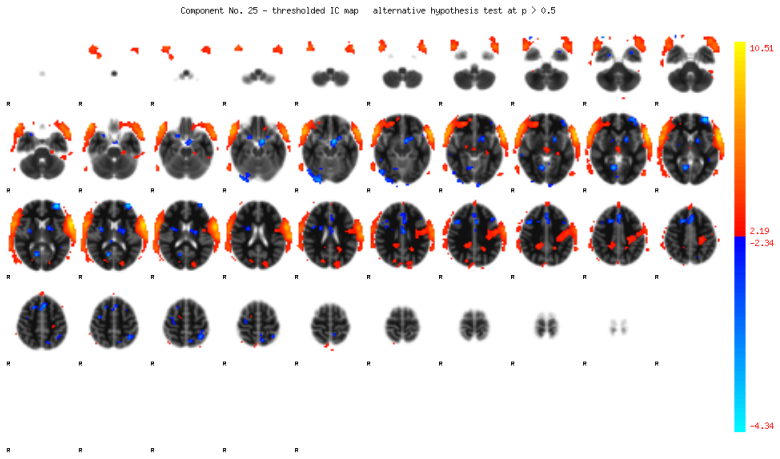

Figure 6.8: One of the networks found using ICA over the dataset. The Independent Component is represented here by the red points of the image. The black and white background is a standard representation of the brain. The image was automatically produced by FSL-MELODIC [29]

problem with this kind of networks is that, when dealing with a small sample, they can be very discriminative. They have a lot of randomly activated voxels and therefore a perfect target for overfitting.

To clarify this effect we run an alternative procedure for analysing this kind of data over the dataset consisting in to first perform an ICA, compute the time courses using a GLM and making inference with such time courses. The most discriminative network according to this procedure was the same network we found before, the one exposed in Figure 6.8. The discriminative power was measured using linear SVMs over the time courses corresponding to each network. The cited IC had the best performance in a 3000 Cross-Validation procedure, showing a mean performance of 0.68. The following best-scored network showed a performance of 0.66, so this value has to be considered carefully. Still, that was the most discriminant network.

6.2.4 Discussion

The results for this dataset were inconclusive. Indeed, we can see that our algorithm found something very close to the most discriminative network as defined by the classical procedures. The visual differences between the images showing both results (the one from ICA and the one form BD-DICA) should not be taken very seriously. Our algorithm is finding a slightly different representation than an Independent Component (i.e. a solution for our algorithm is not a solution for ICA).

However, the result is not satisfactory at all, as it shows that our algorithm cannot avoid to overfit the data when a candidate network exist. This problem is usually overcome in the classical approach by removing the noisy components from the set of candidates by visual inspection (performed by experts in the field) before starting the inference. This is, of course, not an option in our case, as we do not have access to $\mathcal{S}$ beforehand.

In exchange, our procedure was extremely faster. The algorithm took only one minute to find the first discriminant network, and the whole process, including the Single-Subject preprocessing took less than 20 minutes. The same procedure using the classical approach took 30 minutes only to find the discriminative networks (to that we have to add the the processing time of the GLM and the inference analysis). These times are even more promising if we take into account that the code of our algorithm has not been optimised and it is written in MATLAB, whereas the classical approach were run using the software package FSL [29], a serious tool implemented in C widely used in the fMRI community.

6.3 Resting-State dataset

Resting-State fMRI data can be analysed in a more direct fashion than Task-Based fMRI, but the set of tools applicable to this problem is smaller (for instance, Tensor-ICA is not directly applicable). In addition, Resting-State analysis is performed in the whole brain in most of applications. For all these reasons, this kind of data is a more desirable target for ourselves algorithm than Task-Based fMRI data.

6.3.1 Description of the dataset

For the Resting-State experimentation we used a very extensive public dataset called ADHD200 [38].

The data deals with the clinical problem of characterising ADHD, a mental disorder affecting between 5% and 10% of scholarised children which is very hard to diagnose with the classical psychological techniques, based in (often subjective) behavioural indicators. An fMRI characterisation for ADHD is very interesting, not only because it can help to better understand the disorder, but also because it would make it much easier to diagnose and threat.

This dataset is offered already preprocessed following the standard fMRI Resting-State procedures [39]. The whole dataset includes a large number of subjects, but we restricted ourselves to a subset including instances from the Neuroimage sample [40] and the Oregon Health and Science University sample [41].

Our subsample included 37 controls and 37 patients suffering from Combined ADHD, one of the three modalities of the disorder. The instances have about 130000 voxels.

6.3.2 Single-Subject Preprocessing

The original recordings have, for each subject, 257 volumes. As in the previous case, it is convenient to perform a PCA over the volumes of each subject to reduce the computational cost of the algorithm. However, the computation of the z-score representation is much easier this time, as we normalise the data according to its own distribution, taking as reference the whole resting-state signal. This process is actually a part of the PCA process itself, so it does not require any additional step.

The final subjects had about 40 volumes each one, which makes a total of about 2900 instances. In comparison with the 280 from the Task-Based case, a much larger data sample.

6.3.3 Results

As in the Task-Based case, we look for four components. Only the two first ones show a good score in the objective function. These two networks are printed in Figure 6.9 and, as it can be seen, they seem to have a very well defined structure. The representation with the same threshold for the other two found networks is almost empty, self for a small number of voxels uniformly scattered around the image. We also have to stress the fact that the threshold for this components was much larger than the threshold imposed in the results from Section 6.2[3].

Even when we still do not have a good way to measure the meaning of that threshold in significance terms, it is clear that whereas the maps found in the Task-Based experimentation would not pass the statistical tests, these new networks would pass. Specifically, the components are normalised to be demeaned and have unit variance before printing them into an fMRI volume. The network of Figure 6.7 was thresholded so that only the voxels with values over $v > 0.2$ were displayed. In contrast, the images from Figure 6.9 were thresholded so that $v > 2$. Both thresholds are rather arbitrary, as no formal statistical correspondence have been defined in any case. However, this last threshold is the one usually employed by fMRI scientists to visualise raw normalised ICA outputs.

6.3.4 Discussion

The resulting networks are much better defined than the results obtained in the Task-Based data. We believe that this improvement in the quality of the result might have two sources: first, it could be a consequence of the increase of the size of the sample respect with the Task-Based case; second, this could reflect the fact that our assumptions are correctly fulfilled by Resting-State data but they are not by Task-Based data.

In any of the cases, it seems that the results are conclusive in this case. The two obtained networks are very complementary and they are actually part of the same Resting-State registered network [38]. This network is shown in Figure 6.10

To check if this network is actually the most discriminant for our subset of subjects we used the extracted time courses for the network in Figure 6.10 and nine extra well known Resting-State networks. These networks were selected for the team gathering the data [38] and the time courses for each network and patient were offered as part of the dataset.

To perform inference (i.e. to find out what is the most discriminant IC for our subsample of the dataset) we characterised each of the 74 instances with the mean of the square of their time courses (the linear term was removed from the experimentation because the time attached

[3]As before, we refer the reader to Sections 7.1.1 and 7.2.1 for a deepest discussion about this issue

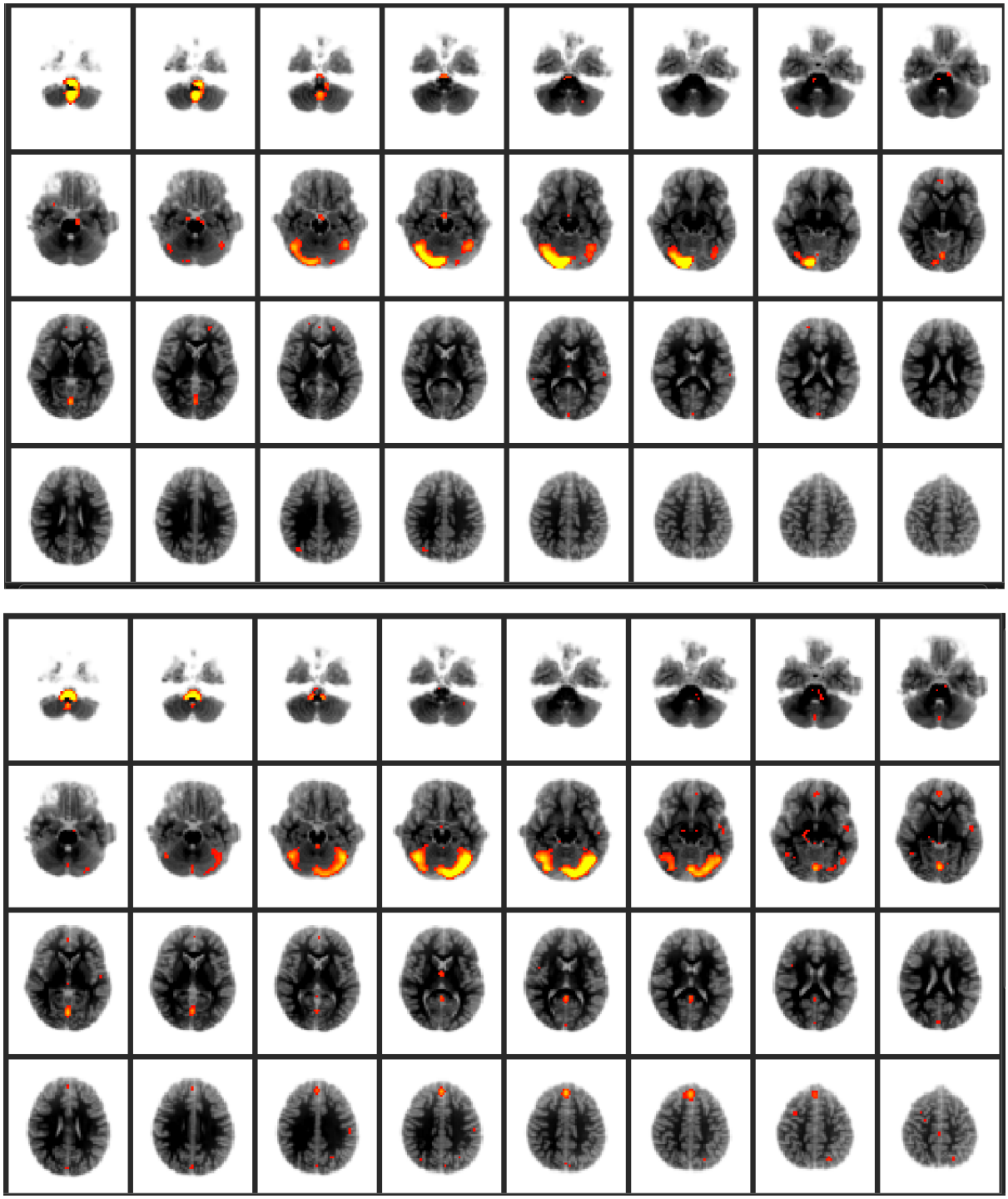

Figure 6.9: A representation of the first two components found by our algorithm for the ADHD200 dataset. The component was thresholded and superimposed to a standard of the brain to appreciate the position of the network.

Figure 6.10: A representation of one of the standard Resting State networks, showing a great similarity with the found network of Figure 6.9. Image by [38]

to the data had been demeaned). The inference was performed using linear SVMs with cross-validation as we did with the MST data (this process is explained in more detail in Section A.3.5).

The results of the experimentation performed using only the extracted time courses show that the network in Figure 6.10 was actually the most discriminant one when using a linear classifier[4].

This result adds some empirical evidence over the validity of our assumptions and the performance of our algorithm. However, it should not be forgotten that this dataset hold a very large sample with respect to usual fMRI studies.

[4]This result confirms that the output of our method is correct in the sense that it found something really similar to what the classical algorithm would found with the same data. However, it should not be taken as a valid result in an fMRI data analysis context. Indeed, a more detailed analysis shows that the result do not survive to any test of significance performed over the classical approach and therefore it is not conclusive as a neuroscientific result.

Chapter 7

Discussion

During the closing chapter of this thesis we will briefly discuss the results of our development, exposing its strong points and its weak spots. During this process, will discuss the lacking parts of the algorithm and later, as Future Work, we will try to draw some solutions together with some possible extensions for the algorithm. In this section, we will also introduce an architecture to use BD-DICA as a Feature Extraction technique.

7.1 Conclusions

We have presented a reliable, relatively efficient and robust algorithm to find independent non-Gaussian discriminant patterns characterising a dataset. This algorithm is capable of dealing with data of high dimensionality presenting several strategies to not fall into overfitted solutions. However, our succeed is very limited in practical terms. In one hand, the results over the validity of the algorithm for dealing with Task-Based data are completely inconclusive, and they show that our algorithm can overfit to a given source when the sample is small. In the other hand, we could not provide a significance test, a key point in data characterisation, for the results of our algorithm

During this section we will draw in more depth those conclusions, regarding the use of this algorithm in both fMRI and general data.

7.1.1 BD-DICA for fMRI data

In our opinion, the experiments performed with the Resting State dataset show that this algorithm has the potential to become a powerful tool in the field of fMRI data characterisation. In the other hand, the experiments performed with the Task-Based dataset show that the algorithm has not reach the required maturity to face complex problems in which the addition of prior information can be crucial.

Strong points

The advantages of this approach with respect to the classical ones have been widely described during this whole document. We are presenting a single-step procedure, much faster than the complex staged processes, based in formalised assumptions that have been widely tested and proved by the community during the last decade.

In addition, our algorithm does not require an exact value for the number of sources N. We just need an estimation for the size of the reduced set $\mathcal{Z}$, that does not need to be precise at all. This parameter is automatically found in our implementation of the algorithm by the PCA function we used (implemented by J. Hurri as part of the FastICA library [42]). In comparison, the number of sources is a crucial parameter in most of the classical approaches using ICA to perform inference (specially in those using infomax), and the results can depend very hardly in the exactitude of this number.

Another minor advantage of our algorithm with respect to the classical approaches is the possibility of selecting the amount of components we want to extract in an on-line manner. We can monitor the search and, after a number of networks have been extracted, decide if we want to go for another one or we want to stop the process. However, this is not a great advantage in the case of fMRI characterisation in which, in general, one or two networks are enough to characterise a given behaviour.

Weak spots

The disadvantages are also clear: our algorithm cannot include additional prior information about the data, which is crucial in most of the Task-Based whole brain experiments. In this prior information we include the capacity of rejecting noisy networks, a disadvantage that have been made clear during our experiments.

More importantly, there is no procedure defined in our algorithm to deal with significance tests, which eventually leads to the lack of a formally valid thresholding point for the extracted component. This problem is described in more detail bellow.

We have also failed to extract any conclusive result regarding the validity of our assumptions in Task-Based data. Further experiments could provide a better insight on whether Task-Based data is a valid candidate for this algorithm but, for the time being, we think that BD-DICA should only be used over Resting-State data.

Several possible solutions to the problems described in this section are sketched in Section 7.2.

The thresholding problem In general, when a result is extracted from the data, it is important to know how significant your result is. In the case of fMRI and our algorithm this is translated in knowing, for a given component extracted from the data, how likely is for each of the voxels to belong to the given component. In other words, the probability of 0those voxels of being also activated in the component if there were no real structure behind the class information.

The way of representing this kind of information in Group-ICA is to perform some kind of significance test and use as intensity of each voxel of the IC the significance of the hypothesis *this voxel belongs to this IC*. This kind of representation is often called t-map, as it is spatial map of t-statistics for the given Independent Component.

Moreover, dealing with spatial differences between the same component for different groups of subjects, another kind of t-map can be extracted in which now the hypothesis is that the voxel is activated for one group but not for the other.

This t-map representation is very useful when visualising the networks, as we can choose a threshold as if it were a p-vale (i.e. we choose to display the voxels belonging to such network under a certain degree of significance).

The most important missing part of our algorithm is precisely the capacity to print that kind of information. This presents two problems: first, we do not have a formal measure to know where to threshold our ICs; second, we cannot know how reliable the result is (i.e. how likely it would be to extract that result with no label information over the data).

7.1.2 BD-DICA for general data

It is worth to remark that, even when this algorithm has been constructed for and tested with fMRI data, it could surely work as well with any other dataset whose problem satisfies Definition1. Moreover, the properties of the algorithm make it perfect to work with high dimensional data with a small number of instances in comparison (i.e. $D \gg M$). This is often the case of a lot of Computer Vision applications, where the whole input of the sensor is shown to the classifier during the learning phase.

Specifically, the algorithm present to major advantages with respect to other traditional approaches to Dimensionality Reduction (an adaptation for the algorithm to Feature Extraction is presented in Section 7.2.4) like PCA, LDA or FE-ICA.

Scalability of the algorithm

One of the main problems facing a data characterisation problem is that often it is necessary to work with the whole data representation instead of a reduced version of the instances. This is a key issue in ICA, in which the distribution of each of the instances along its features has to be present at all points of the algorithm to correctly measure the non-Gaussianity of the projection.

The architecture of the BD-DICA is built in such a way that such large dimensionalities are manageable in an efficient way by the algorithm. Actually, the computational cost of the algorithm is linear in both, dimensionality and number of instances. Empirical evidence of this property is provided in Figure 7.1.

This scalability is a great advantage in comparison with other methods like LDA. Even performing some Feature Extraction (such as PCA) before LDA presents really hard scalability

problems not only in time but also in space (the covariance matrix needed to perform a PCA in the direction of the features in an fMRI set of volumes could need up to 8GB of RAM).

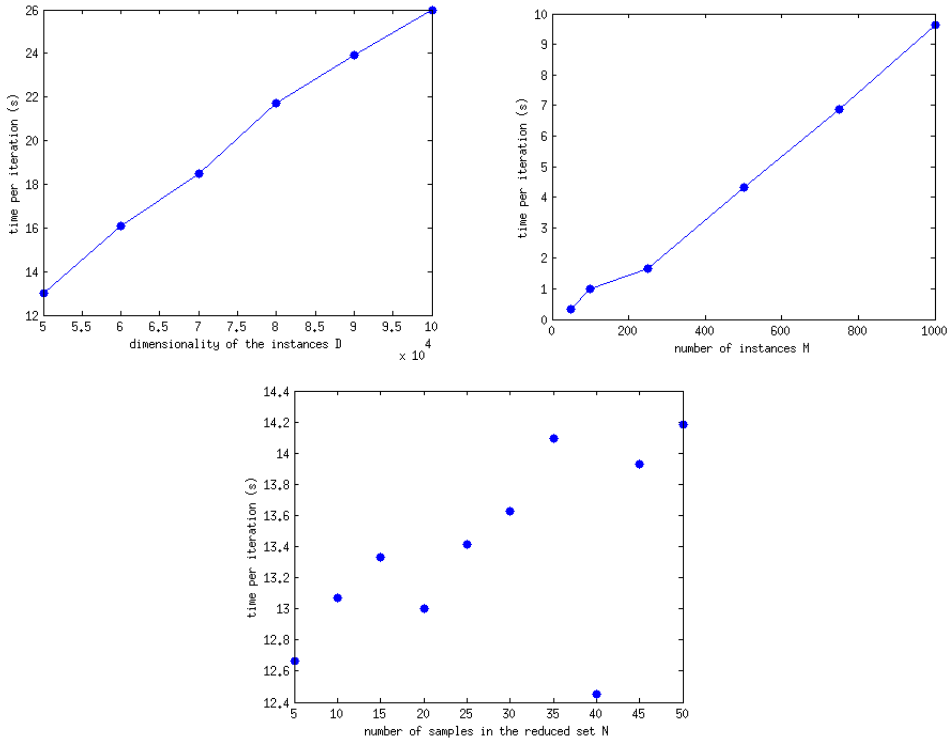

Figure 7.1: Time dependency of the algorithm shown empirically using different synthetic datasets. Note that the dimension of the instances is scaled by a 10^4 factor.

In the other hand, we can see that the dependence on the size of the reduced dataset $\mathcal{Z}$ in the interval used by the algorithm is negligible.

Avoiding overfitting

Our algorithm have been constructed keeping in mind the usual problems carried by a large dimensionality of the data. We have already talked about the scalability of the algorithm but there is another risk usually associated with high dimensional data: the risk of overfitting.

To reduce this risk, we have two main strategies. First of all, the use of a reduced dataset $\mathcal{Z}$ to span the search space for the projector, which makes it much difficult for the FLD to find configurations taking advantage of the noise in the data to find highly separable projections. The number of vectors in that dataset is a key parameter in this strategy: the lower, the less the possibilities of overfitting.

More importantly, the use of the Negentropy term in the objective functions makes the algorithm to not accept insubstantial solutions with no structural meaning, which are usually related with

a poor generalisation capacity. In the fMRI context, we say that the incentive of the optimisation process to achieve highly separable projections are as great as the incentives for this projector to have physiological significance. This same affirmation is valid for other data in different contexts.

This also represents a great advance with respect to other methods like LDA or classifiers dealing directly with the volumes (like k-NN), which suffer heavily the Course of Dimensionality.

7.1.3 Using the algorithm: some guidance

We have seen a direct application of the algorithm to some dataset in Chapter 6. However, we have not reveal what are the best parameters to successfully find discriminant ICs in a new dataset. We will discuss now the effect of the different parameters of the method.

The parameter κ

This is probably the most important parameter of our algorithm. It fixes the proportion of discriminative power and Negentropy we want in our solution. The bigger the κ, the more important the BDFLD term is in the algorithm.

In the ideal case in which all our assumptions are fulfilled and there exist no noise, a run with $\kappa = 1$ (its bigger value) would yield quickly to the correct solution, as the most discriminant projection is, after all, a source vector. The Negentropy term helps to avoid overfitting, distancing the search from those noisy projectors that take advantage of the random variation of the data. Therefore, κ should be decreased as the noise gets bigger and bigger in relation with the number of instances we have.

However, a extremely low value of κ could make the algorithm to choose the most non-Gaussian network before the most discriminative one. So careful must be taken in the choice of this parameter.

Our choice for the analysis of the real fMRI data was $\kappa = 0.2$, which yield to good solutions in the case of the synthetic data, but $\kappa = 0.5$ worked better in the cases with low noise.

Samples in the reduced dataset

The other adjustable parameter of the algorithm is the size of the reduced dataset $\mathcal{Z}$. As we saw previously in this section, it is convenient to keep this number small to decrease the chances of overfitting. However, a extremely low value could yield to a subspace not containing all the sources vectors.

Probably the best way of choosing this parameter is to inspect the eigenvalues of the PCA, if that is the way in which the samples are being reduced. In general, the greater variability will be observed in the vector space containing the source vectors, whilst the variations provoked by small changes in the shape of the networks for different observations and other artefacts will

be smaller in comparison, if good noise-reduction preprocessing have been applied to the data beforehand.

Therefore, a tipping point in the eigenvalues of PCA could imply that we are changing the kind of basis vector from those responsible from the large variations to those responsible for the small irrelevant ones.

7.1.4 The assumptions, revisited

At the beginning of this work we presented some assumptions. It is time now to look backwards and revise them briefly.

The strongest assumption we have made in this work is the orthogonalisation of the basis sources $\mathbf{s}_i \in \mathcal{S}$. This orthogonalisation was an hypothesis maintained to construct the BDFLD but it has not been forced during the search of components. Actually, we cannot completely assume that the networks are orthogonal to each other, because in that case we would break the assumption of statistical independence (i.e. we could extract information of a vector by looking at another one).

The results with the synthetic data show that the partial compliance of the data with this assumption is enough to produce satisfactory results in much of the times. At the same time, the results obtained with the Resting-State dataset demonstrated that there are complementary interesting networks overlapping with each other. Therefore, our strategy seems to work with the data we have inspected.

This does not necessarily hold, however, for the Task-Based data. Indeed, one of the things we saw when performing ICA over our Task-Based dataset was that there were classical Resting-State networks coexisting with the more clustered task-based related regions of the brain. Unfortunately, the set of such regions overlap, in general, greatly with the Resting-State networks as described by Biswal [13]. Therefore, we cannot conclude that this algorithm would work with Task-Based fMRI data.

Another assumption was the linear separability, later extended to quadratic plus linear separability. This assumption has demonstrated to work reasonably well for the synthetic and Resting-State datasets, and there is nothing indicating that this would stop working in any other fMRI dataset, where the mechanisms of the mixing of the source vectors are similar.

Finally, we think that the non-Gaussianity and independence assumptions do not require a review. They have been established based on a large set of empirical evidence present in the literature [10] and everything seemed to work in that sense in our algorithm.

7.2 Future Work

During this section we will show some ideas we have to further improve the algorithm and partially fix some of the weak spots we have found during the journey. This section represents

all the things we wanted to do but we could not implement on time, and show the possible future directions of our research in Neuroimage.

7.2.1 Statistical significance tests

A possible path to create a statistical significance test for the components extracted by our algorithm is to follow the classical strategy of the Fisher's Permutation test [43]. This kind of test draws a baseline with which the extracted solution can be compared by running the algorithm several times using different random permutations of the labels, destroying the real information contained in the classes but preserving the distribution of the data. In theory, if the algorithm cannot find something solid using mislabelled data, a solid result should be interpreted as a real characterisation of an existing structure. However, if the algorithm is capable of finding acceptable similar solutions even with mislabelled data, the results for the real dataset should be tagged as inconclusive.

This procedure could be used to produce a sample of the distribution of the final values for the objective function $\mathcal{J}$ or the score of the BDFLD, which could help to interpret the significance of the obtained network.

However, performing such tests requires a huge number of permutations, which is prohibitive with our algorithm taking up to 15 minutes in finding a component. Even in the best case scenario, a duration of 1 minute of analysis would yield to 35 hours of analysis to have a reasonable amount of samples in our distribution. Therefore, we need a way of improving performance of the algorithm before implementing those kind of tests.

Another different problem is the extraction of the t-maps of the networks. This problem could be solve using a probabilistic approach to ICA [44]. We are, however, still far from the designing of such improvement in our algorithm.

7.2.2 Improving performance with the representation

The most expensive step in our algorithm is to compute all the projections of the data over the proposed projector ξ in every iteration. In this section, we will propose an approach to reduce this kind of projections to a fixed number. This solution can be used to perform quick random permutation tests or to improve the computational performance of the algorithm in situations requiring a large number of iterations.

The main idea of this approach is that we can decompose all the samples in the dataset $\mathcal{X}$ in to the basis spanned by $\mathcal{Z}$, which in theory corresponds to the original vector space $\mathcal{V}$.

In this context, an observation $\mathbf{x}_i \in \mathcal{X}$ can be expressed as $\mathbf{x}_i = {r_i}^j \, \mathbf{z}_j$, where of course the ${r_i}^j$ are the coefficients of $\mathbf{x}_i$ in this new basis.

Now, consider the projector $\boldsymbol{\xi}$, which is actually obtained in our algorithm as a linear combination of the vectors in the reduced sample: $\boldsymbol{\xi} = w^j \, \mathbf{z}_j$.

The projection of the sample $\mathbf{x}_i \in \mathcal{X}$ over the projector ξ can then be written in the following way:

$$\begin{aligned} \langle \boldsymbol{\xi}, \mathbf{x}_i \rangle &= \xi_j \, {x_i}^j & (7.1) \\ &= w_k \, {\mathbf{z}^k}_j \, {r_i}^l \, {\mathbf{z}_l}^j \\ &= w_k \, {r_i}^l \, \delta_l^k K(k) \\ &= w_k \, {r_i}^k \, K(k) & (7.2) \end{aligned}$$

where we have assumed that the samples in $\mathcal{Z}$ are orthogonal to each other (this is actually the case if we extract $\mathcal{Z}$ using PCA) and $K(k) = \|\mathbf{z}_k\|^2$.

Therefore, to characterise a given projection, all we need to know are the coefficients ${r_i}^k$ and the constants $K(k)$. All this numbers can be extracted at the same cost of N iterations, where N is the size of the reduced set $\mathcal{Z}$. Of course, the operations in Equation 7.2 are much cheaper as the operations in Equation 7.1. The first one is actually equivalent to a problem with dimension $D = N$. Note that, in the fMRI context, $N \sim 20$, $D \sim 10^5$ and the number of iterations required to extract a single component can oscillate between 10 and 30.

Unfortunately, we have not had the time to implement this approach and perform the appropriate experiments to empirically check the predicted theoretical performance.

7.2.3 A Fixed-Point algorithm for BD-DICA

FAST-ICA is an implementation of ICA developed by Hyvarinen et al. [30] [42] in which instead of a Gradient Ascend approach to find the maxima of the objective function a Fixed-Point algorithm is used. This version of ICA works much faster than the Gradient Ascend versions.

A Fixed-Point optimisation algorithm does not travel through the search space looking for a maxima, but tries to reach a point satisfying the conditions of the maximum. More specifically, a maximum is characterised by a null derivative, therefore, we can try to find a point satisfying $\nabla \mathcal{J} = 0$ to reach the solution. This can be achieved by using Newton-Raphson over the previous equation, as Hyvarinen did for the case of BD-ICA [3].

We would still need to compute derivatives, but the number of iterations could be much smaller. Unfortunately, we did not have enough time to derive the expression for such algorithm and perform the corresponding experiments to check if this approach could lead to a faster version of BD-DICA.

7.2.4 BD-DICA as a Feature Extraction technique

Even when this algorithm is clearly oriented towards data characterisation, the scalability with the dimensionality of the instances and its strategies to avoid overfitting make it a nice candidate to Feature Extraction.

Now, as we have already said this is a Basis-Decomposition algorithm and its adaptation to Feature Extraction is not direct. Our proposal for such adaptation is, however, relatively straightforward: as the procedure tries to optimise the parameters for the representation obtained by the BD-Transformation, it seems reasonable to use this same transformation to represent the data using the resulting parameters.

This approach presents, however, two possibilities, depending on the interpretation of the quadratic term of the BD-Transformation we choose. To see that, consider that we have extracted n components $\mathbf{y}_i$. We can represent the sample $\mathbf{x} \in \mathcal{X}$ in this way:

$$\mathbf{x} \longrightarrow \left(\langle \mathbf{y}_1, \mathbf{x}\rangle, \langle \mathbf{y}_1, \mathbf{x}\rangle^2, \langle \mathbf{y}_2, \mathbf{x}\rangle, \langle \mathbf{y}_2, \mathbf{x}\rangle^2, \dots \langle \mathbf{y}_n, \mathbf{x}\rangle, \langle \mathbf{y}_n, \mathbf{x}\rangle^2\right) \tag{7.3}$$

or in this other way:

$$\mathbf{x} \longrightarrow \left(\langle \mathbf{y}_1, \mathbf{x}\rangle, \langle \mathbf{y}_2, \mathbf{x}\rangle, \dots \langle \mathbf{y}_n, \mathbf{x}\rangle, \|\mathbf{x}_\perp\|^2\right) \tag{7.4}$$

Equation 7.3 assumes that the quadratic term is important by itself, whereas Equation 7.4 assumes that the quadratic term is only important is it represents the perpendicular projection of the sample with respect to the hyperplane described by the projectors.

Both representations are equally valid for a general dataset. Probably they perform different in different situations. Neither of them is better than the other even when the BDFDL optimises a representation much similar to Equation 7.3, since the optimisation is performed over each component at each time.

As in the previous case, we had not the opportunity to appropriately test these representations in a real data experimentation, and therefore they are left as future work.

7.2.5 Further extensions

Some further extensions can be added to the algorithm easily by adding a term depending on parameters that can be encode as a w-application in the algorithm's objective function.

A possible extension of this kind is, for example, a term penalising highly scattered components. This can be useful to avoid noisy sources as discussed in Section 7.1.1.

A simpler extension can be made to penalise the presence of the component over desired features (e.g. voxels in regions belonging to non-interesting areas of the volumes, as the zones without brain or as in the previous example of the auditory and visual cortices).

The development of such extensions could be actually very easy. They are, however, out from the scope of this work which does not intend to develop a sophisticated algorithm, but to set the grounds of an approach to Basis Decomposition D-ICA.

7.2.6 Orthogonal Basis Decomposition

As a last idea in this gathering of possible future directions, we want to introduce the idea of Orthogonal Basis Decomposition Feature Extraction. We believe that the orthogonality assumption can be further extend to develop new techniques oriented towards Data Characterisation.

As an example, let us introduce an experiment we performed in this direction using Deep Networks. Deep Neural Networks (DNNs) [45] are a set of ANN architectures considering a large number of hidden layers. Those kinds of architectures allow to represent different abstractions of the data in different layers, and they are therefore a nice instrument for data characterisation.

In our experiment, we constructed a DNN with D inputs, a very small number of neurons in the hidden layers (4-5 units) and a single unit in the output layer. Then we fed our noiseless synthetic data to the network and, once the learning was done, we inspected the weights connecting the input layer with the first hidden layer. Note that we have D weights for each unit in the hidden layer and that they are ordered, so they have the same structure as the instances of our dataset.

Indeed, consider the input z of one unit of the first hidden layer when we show the sample $\mathbf{x} \in \mathcal{X}$ to the network: $z = w_i\, x^i$ where w_i are the weights of that layer. Now, if $\mathbf{x}$ is actually a linear combinations of source vectors $\mathbf{s}_i$, the input z will represent the mixing coefficient for that sample if the weight vector $\mathbf{w}$ is one of the source vectors. Therefore, if the network optimises its weights to maximise accuracy, it is reasonable to expect to find source vectors in the weights of the first hidden layer.

The complicated operations needed to extract information from the coefficients can be implemented automatically by the network in the rest of the hidden units.

The experiment we performed confirms that hypothesis. We were able to find the discriminant network in two of the five hidden units of the first hidden layer with a similarity of sim ~ 0.78.

Of course, this is just an example, but this kind of approach can be taken to find more efficient ways of characterising data satisfying those conditions.

7.3 Summary

During this work, we have presented the Basis-Decomposition Discriminant Independent Component Analysis (BD-DICA) algorithm. To do that, we have first proposed a new framework called Basis-Decomposition to express operations made with the instances of the data instead of with the features and we have reformulated the two already existing architectures of ICA in the terms of our framework.

Later, we have developed a generalised version of the FLD to deal with arbitrary parametrisable transformations and we have specifically developed the Basis-Decomposition transformation, building in this way a Basis Decomposition version of the famous discriminant.

Then we have used the two Basis-Decomposition algorithm to built a Basis-Decomposition version of Discriminant ICA, which was our first objective.

We have tested the algorithm with synthetic and real data. We have obtained satisfactory results in the synthetic and one of the real datasets, dealing with Resting State fMRI, but inconclusive results with the other dataset, dealing with Task-Based fMRI.

Finally, we have presented some conclusions and some future directions to further develop the algorithm and extend the scope of the Basis-Decomposition framework.

The end.

Appendix A

A case of study in fMRI

In this appendix we will introduce which was the first objective of this thesis in its preliminary versions: to analyse the MST problem described in Section A.2. This work was done using classical Machine Learning techniques for temporal inference after an ICA. This project was carried along with a parallel project in Resting-State that triggered what finally became the main topic of this Master's Thesis, the BD-DICA. However, our work with the MST data allowed to us to understand in a much deeper sense the structure behind the fMRI data, and we think that the exposition of this analysis could also help the reader to get more familiar with the characteristics of this special kind of signal. In our opinion, the results obtained with this dataset justify partially some of the design decisions we took for the BD-DICA project, whereas give a more deep explanation about the causes behind the weak points of the algorithm.

A.1 About fMRI data and the BOLD signal

A.1.1 The BOLD signal

Functional Magnetic Resonance Image (fMRI) is a Neuroimaging technique to scan blood-oxygenated levels in the human brain used since 1992.

The technique utilises the differences between the magnetic properties of oxygen-rich and oxygen-poor blood to register evidence of brain activity. Specifically, physiological studies of the brain cells seem to indicate that brain activation requires an increase of the levels of oxygenated blood in the regions of activation [46]. This incoming flow of oxygenated blood displaces the de-oxygenated blood, presenting a gradient in the oxygenation levels of the blood in the activating brain areas. This gradient shows up about 2 seconds after the activation [7]. After this increasing, the cells consume the oxygen and the levels go back to its previous state.

This phenomenon allows us to characterise large-scale neural activity as a change in the oxygen levels in the blood in the area of activation. We say *large scale* because the spatial resolution of the technique is highly limited. The blood flow irrigates relatively large areas of the brain, so we cannot characterise activation of small clusters of neurons (to not say nothing about

individual ones). Therefore, fMRI is a technique that allows us to study correlations between the activation of brain areas and certain behaviours or mental conditions.

The resolution of an fMRI machine is measured in voxels. The voxels are the equivalent to pixels in the 3-Dimensional space of the brain. The smallest the voxel in proportion to the brain, the larger the resolution of the recordings performed by the machine. Current fMRI present voxels with edges between 0.5mm and 2mm.

But returning to the previous discussion, we still need to explain how can we measure this changes on the blood flow without cutting the skulk of the subject in two halves. This is possible because, whereas de-oxygenated haemoglobin (dHb) is paramagnetic, the oxygenated haemoglobin (Hb) is diamagnetic [7]. Therefore, we just need to find a method to measure the gradients of magnetic responses in the brain.

As the reader might have already guessed, this is done by means of magnetic resonance. This phenomena occurs when we expose magnetically active materials to a oscillating magnetic field. When we apply a magnetic field to such materials, the spin of their unpaired electrons tends to align with the applied field. This alignment creates by itself an induced magnetic field, that can be measured from outside the material. As the external field oscillates, this phenomenon is produced constantly in the material.

The twist is that whereas paramagnetic materials present an induced field in the same direction than the external applied field, diamagnetic materials present an induced field in the opposite direction of the external field. Therefore, Magnetic Resonance can be used to detect the amounts of Hb and dHb within the brain in a not invasive way.

However, electrons usually need some time to adjust to the changes in the external field, so the oscillations have to be low. The fMRI machine register a snapshot of the brain for each of those oscillations, allowing us to observe the state of the brain each time the external magnetic field have been reversed. The time separating those snapshots fixes the temporal resolutions of the recordings, which is usually of about 2 seconds [7].

The signal described in this section, produced by this changes in the blood magnetism in the brain is usually called Blood-oxygen-level dependent (BOLD) signal.

A.1.2 fMRI data description and preprocessing

fMRI data constitutes of a series of 3-Dimensional snapshots called volumes. Each of those, contains a large number of voxels, representing each of the areas of the brain in which we are measuring the levels of Hb and dHb.

A problem arises however when visualising volumes: the position of the head in the 3-Dimensional space in which the signal has been measured usually changes due to unavoidable movements of the patient. This could be problematic if we want to keep track of a determined voxel. Movement correction techniques should be applicable to remove this effect.

A much serious problem arises when comparing activations of different subjects. Not only the positions inside the machine tend to be different, but also the brain size may differ. It could

be even possible that both patients present slightly different sizes in some brain areas. Making inference about the participation of an area of the brain in different subjects is, therefore, difficult.

To overcome this difficulty it is necessary to project all the volumes of all patients into a common pattern usually defined by an external standard. This procedures are computationally expensive and difficult, involving transformations up to 2000 degrees of freedom and they are usually performed with the help of structural MRI images form the patients, showing a higher resolution [8].

The standardised spaces are well defined and widely used in the fMRI community. There exists an standard for each resolution (i.e. for voxel edges of 1, 2, 4 and 8mm). To draw an idea of the size of the volumes, the MNI (the standard template proposed by the International Consortium of Brain Mapping) space of 4mm has 110000 voxels.

After those normalisations, the outer side of the brain is usually removed. This includes not only the skull and other not-related tissues, but also the empty space out of the head which is usually full of noise.

Other usual preprocessing steps include noise reduction, bandpass filtering and spatial smoothing. The noise reduction step is usually performed with very sophisticated techniques varying by a simple filter to an intelligent system removing artefacts from major arteries which always carry oxygenated blood. The bandpass filtering is crucial to remove physiological rhythms not related with brain activity (as the beating of the heart) that can arise in the recording as well as other artefacts produced by the machine itself. The spatial smoothing is usually the last step of the preprocessing and it is intended to polish the work of the normalisation and motion correction steps, which in most of the times have to interpolate the values of the voxels in the new representation.

For most of applications involving whole-brain analysis, independently on the original spatial resolution of the recording, it is convenient to resample the volumes to a 4mm standard to reduce the computational complexity of the analysis (this is actually a requisite in ICA, which needs to store in memory whole volumes).

A.1.3 fMRI experiments

Human experiments with fMRI are performed in a relatively non-invasive way. The subject is asked to hold still during the experiment inside the machine. The recordings usually take about 8 minutes or less but sometimes the experimentation has different stages. The magnetic fields are believed to be harmless to the tissues, but the machine have been described as claustrophobic and the generators of the magnetic field produce a very loud noise.

The set of the experimentation is therefore sometimes difficult, as the subject have a limited capacity to perform tasks.

The length of an fMRI recording is usually of about 240 volumes (with a TR = $2s$ exactly 8 minutes, as we said before).

A.2 The Music-Supported-Treatment Problem

During this appendix, we will try to find predictors of the success of a recovery therapy called Music Supported Therapy (MST) believed to be related with the audio-motor coupling of the human brain. We will use of course fMRI data to find those indicators.

In the case of study, we use data from patients that have suffered an stroke in the brain, resulting in a lost of mobility in one of their arms. The physiological reason of this effect is the death of the cells in the affected brain area, in this case responsible of the movement of the affected arm. This patients are usually treated with conventional and constraint induced therapy, two methods that, by inducing a phenomenon of plasticity in the brain, are capable of assigning the responsibilities of the dead brain area to another one (usually the symmetric one located in the opposite hemisphere).

Music Supported Therapy has been recently developed to improve this plasticity, and therefore to improve the performance of the treatment in this kind of patients [47]. MST combines the strategies of the classical treatments with the audition and execution of music. This strange effect is believed to be connected with the audio-motor coupling of the brain, a series of mechanisms that allow us to react rapidly to sudden noises from the environment, and that it is ultimately responsible of the capacity of musicians to adapt its execution to the sound of their instrument [37].

Along several studies [36], MST has demonstrated to show a better performance than conventional and constraint induced therapy in the recover of the patients. However, still a non negligible variance has been observed between patients, in which differentiate two groups: a group of patients showing a standard recovery in the same way than patients treated with classical therapies and another group showing a clearly improved performance.

The aim of this problem is to try to characterise this variation in the success of the MST therapy using fMRI data collected from a group of patients with similar affections in upper extremities. Several studies explained the success of the MST therapy with a well documented effect called audio-motor coupling [37], a connection in brain activity from auditory and motor regions believed to be responsible of the plasticity phenomena responsible of the recovery of the patients [48]. Therefore, it would be perfect if we were able to find patterns involving those cortices in the characterisation of our phenomenon.

Our study is part of a much larger project [47] devoted to characterise in every possible aspect the Music Supported Therapy. For that purpose lots of recordings have been taken from the set of affected patients, including recordings during the performance of motor and auditory tasks previous, in the middle and after the treatment.

Our objective is to characterise the performance of the recovery using only auditory data taken before the therapy. We want to use only auditory data to avoid to learn the gravity of the lesion, greatly characterised by the activation during the motor tasks. We want to use only data previous to the therapy for two reasons: first, it would be too easy to learn how well the subject have been recovered by just looking at the aftermath images, in which the auditorium coupling is known to be strengthened; second, from a clinical point of view, it would be very interesting to be able to predict how good is going to perform a subject before assigning her a

determined therapy. This second objective is, however, just theoretical. We have a quite reduced sample of patients and we do not expect to obtain large significances, but a sign justifying a larger study.

A.3 Experimentation

A.3.1 Recording the data

As we said, we will centre ourselves in recordings of auditory tasks performed before the therapy. Those recordings were taken just once for each patient and they last for 194 volumes (i.e. about 6.5 minutes).

The first 12 seconds are left for the patient to get used to the noise of the measure and then a series of melodies are played. The patient is asked to focus in the sounds and not to think at anything else. Each melody lasts about 15 seconds. Between the melodies, a resting period of other 15 seconds is recorded during which the patient does not hear nothing and it is asked to keep the mind in blank.

There are four kind of melodies played this way. Two of them are whole melodies, and the other two are just sequences of tones. Those four kind of melodies are played three times in total during the experiment, yielding to a total of 12 playings and 13 resting periods.

Before the analysis, the first 6 volumes (i.e. the first 12 seconds corresponding to the adaptation time of the patient) are removed from the recording.

A.3.2 Preprocessing and description of the data

The preprocessing of the data was performed thoroughly by the research group conducting the whole project [47]. The preprocessing of this kind of data is actually very hard, as the normalisation process involves patients affected by a stroke, having completely deactivated a whole part of the brain. To complete the normalisation, masks should be constructed for each patient to cover this affected parts.

In any case, we received all the data perfectly preprocessed and registered in an MNI standard of $4mm \times 4mm \times 4mm$, which totals about 110000 voxels per volume.

We have two groups in the dataset: improved performance (4 patients) and normal performance (5 patients). The construction of those groups was made based in the ARAT index, a quantitatively behavioural indicator describing the gravity of the reduction of mobility of the arms of the patients.

We had access to the ARAT score of the patients before and after the treatment. To separate the patients in two groups we just needed to split the distribution of the pairs of those indices into two separate independent distributions. This was done by assuming a linear dependence between the pre and post ARATs with a different linear coefficient for each of the groups. The

split of the subjects was chosen such that the linear regression in the two groups presented the best possible fit. After this process, some patients should need to be removed from the study because they show similar affinities to the linear fits of both groups. This patients correspond to cases with high pre ARAT scores, in which any possible variance in the recover performance cannot be appreciated because of the small margin of recover of the patient.

A.3.3 Finding the Independent Components

Our analysis is performed in three steps following the guidance exposed in Section 2.1.3. First, we will extract the Independent Components from the data using Group-ICA (see Section 2.1.2). Then, we will extract a representation of the time courses of those components and finally we will perform inference using that representation.

To extract the Independent Components, we feed the data into Spatial Group-ICA algorithm (we used FSL MELODIC [29]). The number of components is estimated automatically by the software and chosen to be 27. Additionally, MELODIC performs a significance test and prints the Independent Components (ICs) into NIFTI volumes in which the value of a voxels represents the t-stat of that voxel belonging to a given IC.

These representations are then thresholded and examined. Some of them are easily identifiable as noise and they are discarded from the analysis, whereas other components are directly associated with the lesions and also discarded. Only 11 of the initial 27 components survived to this filter.

An example of both, filtered (because of noise presence) and kept components are shown in Figure A.1. An additional example of a noisy network from this analysis is shown in Figure 6.8

A.3.4 Constructing the descriptors

Once the ICs have been identified, we need to represent each of them as a vector ready to be fed to a classifier.

Our descriptor for each patient and each network has a representation of the previously described four kind of melodies. Actually, since we have only 9 subjects, we decided to increase the number of samples by taking each of the three repetitions as an observation by itself. Each of these samples is represented by a vector of four numbers, one for each of the melodies.

To obtain those numbers, we first used a Generalised Linear Model (GLM) to estimate the time courses associated with each of the networks. This regression finds the best mixing coefficients associated to each of the ICs to reconstruct the original fMRI signal. Therefore, the series of all those coefficients along time are a good representation of the relative activation of each of the ICs during the experiment.

These coefficients are then separated to correspond to each of the iterations and each of the stages. The coefficients of each of the tasks are normalised using the values of the adjacent

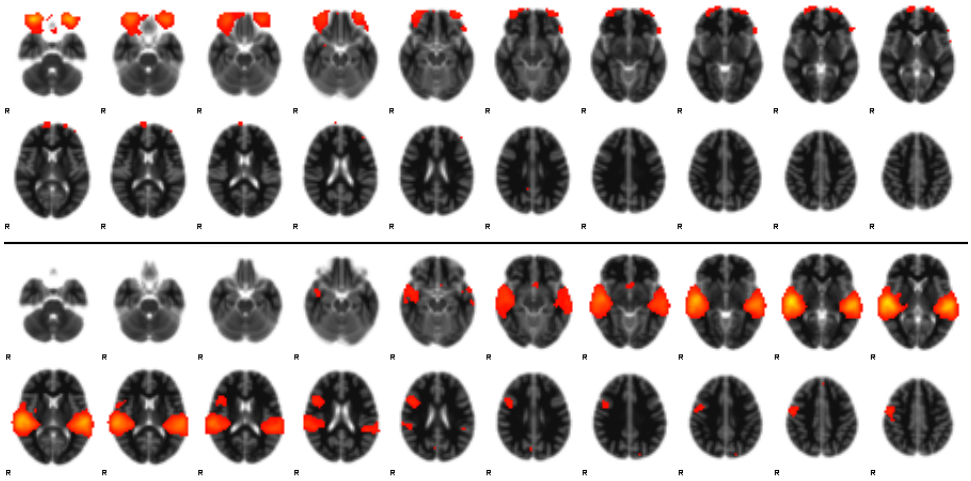

Figure A.1: Two ICs found by Group-ICA in the MST data. Up, an irrelevant network corresponding to either noise or not correctly removed tissues (perhaps the eyes?). Down, a component activated in the auditory cortex. The Independent Components are represented here by the red points of the image. The black and white background is the MNI standard used in the normalisation process. The images were automatically produced by FSL-MELODIC [29].

resting period by subtracting the mean of the resting and dividing by its variance. The number representing that stage in this iteration is, finally, the mean of the normalised coefficients.

After this process, we have 27 observations with 4 features for each of the 11 Independent Components. The observations are labelled with the class of the patient (we have just two classes) but we also take record of which observation belong to each patient since we will need it before to make the permutation tests, since the observations are not fully independent from each other.

Alternatively to this representation, we also built a similar vector using the square of the time courses instead (i.e. each feature corresponds to the mean of the square of the time courses). We performed the inference stage separately for this two representations.

A.3.5 Inference

In this stage of the process we use linear Support Vector Machines (SVMs) to classify the observations in the two existing groups. The learning/testing procedure is performed using n-fold Cross-Validation, with 4 samples for the testing in each iteration and using 2000 iterations for each test. The discriminative power of each of the ICs (or combinations of ICs) is then measured with the performance obtained in the cross validation procedure.

We decided to test each of the networks and, furthermore, combinations of two and three networks. To test a given combination, we just put together the four descriptors of each network involved in the combination making a $4 \times n$ feature vector (with $n = 1, 2, 3$).

The tests were run following a greedy strategy. Firstly, we test all the networks, one by one. The best performing one is then selected as one of the outputs of the procedure. Then we test the possible combinations of two ICs involving the one found in the previous step (therefore, we have 11 combinations). The best performing combination of two ICs is then held and we repeat again the same process with combinations of three networks involving the two best performing ones found before. In this way, we just need to perform 33 tests instead of the $11^3 + 11^2 + 11$ that would be necessary to test all the possible combinations.

A.4 Results

A.4.1 Measuring statistical significances

Once we have found our discriminant ICs, we need to obtain a measure of the statistical significance of the result. For that, we use a non-parametric method called Permutation Test. The main idea behind this technique is to execute the inference algorithm several times feeding into the algorithm different permutations of the labels of the observations. In this way, we remove all the information contained in the data without altering the distribution of the features. The result is the distribution characterising the expected output of the inference process (in this case, the expected obtained performance) when there is no class structure behind the data.

Our first statistical test (test-1) is for checking if we could have obtained the same performance in our chosen combination of ICs using random data with the same distribution than ours. We performed an additional second test, to check if we could have obtained that performance following the whole process. This second test (test-2) is connected with the so called corrected[1] *p-value*s and measures the significance of our result involving the whole supervised part of the experiment.

In other words, while the first test tries to answer if we could find the same results in a given combination of ICs if there were no structure behind the data, the second test tries to answer if we could find a (combination of) IC(s) showing the performance of our results with random data.

For both tests, we first prepared all the possible combinations of the labels of the observations. To ensure that the classifier is not learning directly from the individuals, we forced the observation belonging to the same patient to have the same class label. Mathematically, we needed to prepare all the different permutations of a vector of nine binary values (two classes). In addition, we also removed the permutations showing the weaker changes (i.e. when only two different labels have been permuted).

After all that process, we had about 120 permutations of the labels. For the first test, we just recorded the performance of the given permutation using the data with all the built combinations of the labels in the same manner than in the real experiment (i.e. 4-folded Cross-Validation with 2000 its.). For the second test, we repeated the whole process than before for each of the label permutations. The best performances of each of the possible permutations were recorded.

[1] The result is however not *corrected* in any sense in our case, as we compute the significances in an exact way.

IC(s)	Performance	test-1	test-2
22	0.66	1.49	0
22, 14	0.68	1.45	0
22, 14, 13	0.77	2.37	0

Table A.1: Main results of our experimentation for the MST data using the plain representation (i.e. using directly the mean of the time courses). Please note that any negative value has been represented with a zero.

IC(s)	Performance	test-1	test-2
5	0.76	2.30	0.99
5, 14	0.80	2.89	0.89
5, 14, 13	0.81	3.15	0.88

Table A.2: Main results of our experimentation for the MST data using the squared representation. Please note that any negative value has been represented with a zero.

These recordings show the probability distribution that we would obtain using the same procedure of a dataset with no real information about the classes. With that distribution, we can easily measure the probability of finding our result if our data would not have class information at all.

As the founded probability distributions were Gaussian-shaped, we measured their mean/variance to find the significance of our results. Let σ be the variance of that distribution, μ its mean and p the performance of the given combination of ICs. Then, our parameter of significance can be written as $t = (p - \mu)/\sigma^2$.

A.4.2 Results of the experimentation

The results of best combinations for 1, 2 and 3 ICs, including the significance values founded with our two permutation tests are summarised in Tables A.1 and A.2.

The four networks included in the results (i.e. ICs 5, 13, 14 and 22) are displayed in Figure A.2

A.4.3 Conclusions

The results using simply the mean of the time courses for representing each instance is not significant at any level for the *test-2* test of significance. Note that even the results for the t in the *test-1* are large enough, we are not considering this as a valid significant test.

The results for the other representation, including the mean of the squares of the time courses are, however, more significant. Specifically, the t of the *test-2* shows that the results correspond to a $p < 0.2$, with the hypothesis of that similar results could have been obtained by chance. This

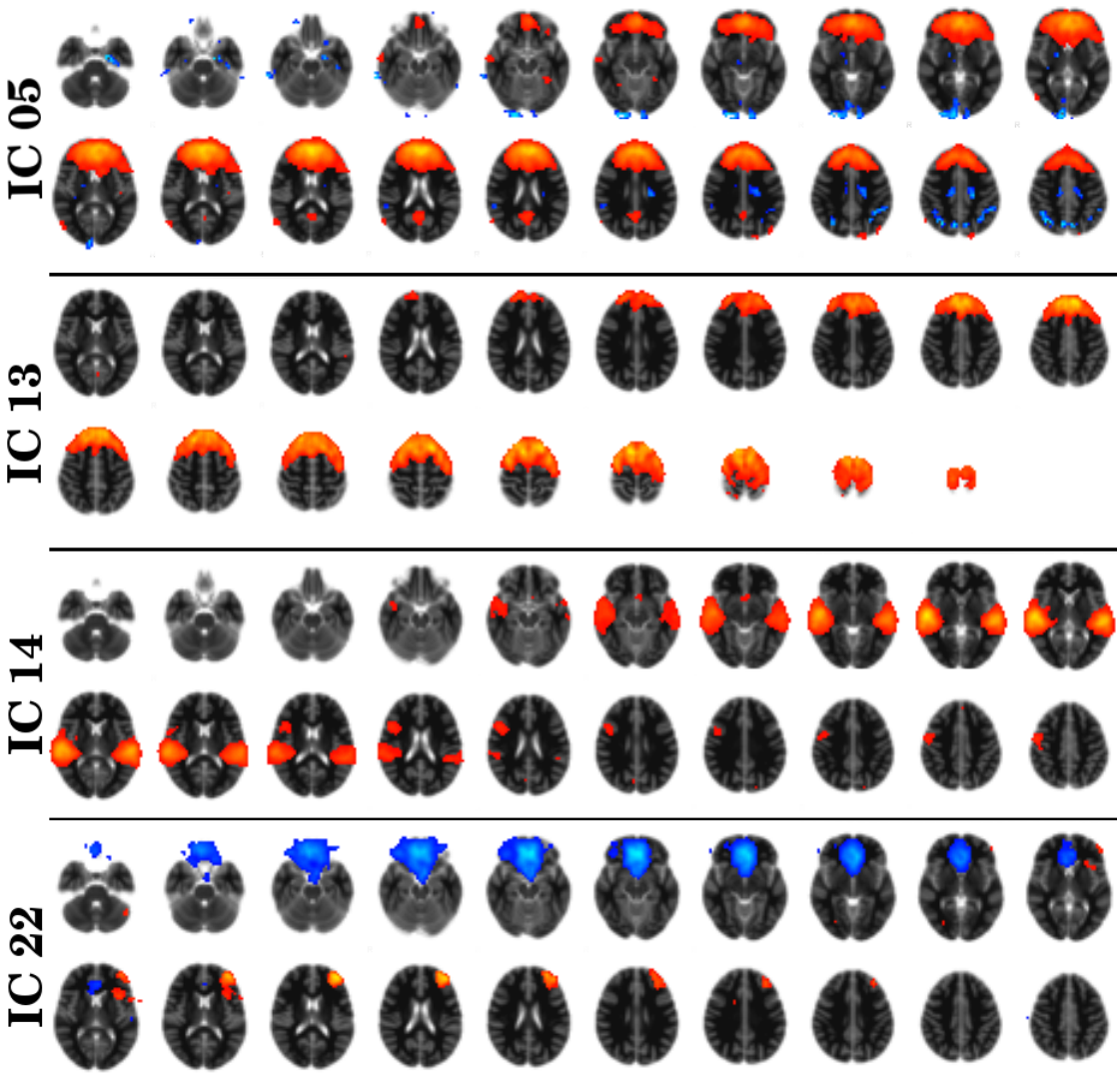

Figure A.2: From top to bottom, ICs 5, 13, 14 and 22, as found by Group-ICA over the MST data. This four networks show some discriminant properties. The Independent Components are represented here by the red an blue points of the image (this last one present significant negative values instead of positive). The black and white background is the MNI standard used in the normalisation process. The images were automatically produced by FSL-MELODIC [29].

is not a very good result, but at least shows significance at some level. This is an encouraging result, if we take into account that we have only nine samples.

In addition, the obtained networks seem to have all the sense from the point of view of MST. Let us consider the most significant results, including only the ICs 22 and 14[2] (see Figure A.2. The IC 22 show a clear frontal presence, in zones related with cognition and intelligence, that has been connected with recovery performances in different treatments [49]. This connection has been empirically proven, but in addition it has a very intuitive explanation: people with a better cognitive abilities usually show a better performance on the recovering from brain lesions. Because of that, we believe that this IC is not really characterisation the performance with the MST treatment, but the performance with any treatment, including MST.

The IC 14 has, however, a much direct implication in MST. This network shows two contributions. In one hand, we have the two large clusters in both hemispheres at the middle of the section which are known to correspond to the auditory cortex [46]. In the other hand, the smaller cluster in the left top side in the sections correspond to sensi-motor regions. This kind of network was the same observed in [47] with the same set of patients. In this study, it is observed that the network shown here in Figure A.2 presents a giant change during the treatment in a single-subject study. Indeed, this network present a much larger activation after the treatment and it is believed to be the one responsible from the induction of plasticity in MST. It would be reasonable to argue that the activation of that network before the treatment could be a predictor of the success of the treatment.

Having said that, it is important to note that our *p-value* is not small enough to consider these results as a real scientific discovery. However, as we said, we believe that they are very encouraging and we plan to resume this project once the set of patients has grown and we have a sample large enough to resolve the effect we want to characterise.

[2]The other two networks play a much unimportant role in the set of results, but we believe that their presence in the frontal lobe could indicate that their discriminant power comes from the same source as the IC 22.

Appendix B

Details on the BD-DICA algorithm

We have strategically skipped all the implementation details of the algorithm for two reasons: first, it is convenient to gather all information regarding the implementation in the same place, in case some reader wants to make its own code to execute the algorithm; second, some expressions will take a large portion of space to be derived, and their inclusion in the main discussion could divert the attention of the reader from the more complex theoretical discussion.

In this appendix, we will derive the missing parts of the algorithms developed along this work: the generalised LDA algorithm, the BDFLD algorithm and the BD-DICA algorithm. As the reader might guess, each of them is supported in the previous one.

Finally, we will expose the dynamics of the BD-DICA algorithm, which, in our opinion, are not entirely trivial.

B.1 Gradient Ascend algorithm for generalised LDA

As in the following two sections, we will deal with the basis of the already draw algorithm, in this case Algorithm 4.1, in which we have only skip the derivatives.

In this case, we need to derive the following gradient:

$$\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\Phi(\mathcal{X},\mathcal{D},\mathcal{T}_{\boldsymbol{\alpha}}) = \boldsymbol{\nabla}_{\boldsymbol{\alpha}}\left(\frac{\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{\sigma_1^2+\sigma_2^2}\right) \tag{B.1}$$

where we are differentiating with respect to the set of parameters $\boldsymbol{\alpha}$, and $\boldsymbol{\mu}_i$ and σ_i are defined as in Equations 4.14 and 4.15.

Now note that output of the transformation $\mathcal{T}_{\boldsymbol{\alpha}}$ is, in general, a vector. Let us define

$$\mathbf{u}_i(\alpha) \equiv \mathcal{T}_{\boldsymbol{\alpha}}(\mathbf{x}_i) \tag{B.2}$$

and the sets

$$\mathcal{U}_{(c)} \equiv \{\mathbf{u}_i : \mathbf{x}_i \in \mathcal{X}_{(c)}\} \tag{B.3}$$

Note that we have omitted the explicit dependence with the parameters of the transformation $\boldsymbol{\alpha}$ for simplicity.

We can rewrite Equations 4.14 and 4.15 in the following way:

$$\boldsymbol{\mu}_c = \frac{1}{M_c} \sum_{\mathbf{u} \in \mathcal{U}_{(c)}} \mathbf{u} \tag{B.4}$$

$$\sigma_c = \sum_{\mathbf{u} \in \mathcal{U}_{(c)}} \|\boldsymbol{\mu}_c - \mathbf{u}\|^2 \tag{B.5}$$

Expanding the expression in its components, this expressions are written as:

$$\mu_c{}^i = \frac{1}{M_c} \sum_{\mathbf{u} \in \mathcal{U}_{(c)}} u^i \tag{B.6}$$

$$\sigma_c = \sum_{\mathbf{u} \in \mathcal{U}_{(c)}} (\mu_c{}^i - u^i)(\mu^c{}_i - u_i) \tag{B.7}$$

Now, we can also write $\boldsymbol{\alpha}$ in components α^i. With the spirit of making the derivation simpler we will compute the derivatives of the generalised FLD with respect to one of the components of $\boldsymbol{\alpha}$, rewriting the gradient operator in the following way:

$$\boldsymbol{\nabla} = \left(\frac{\partial}{\partial \alpha^1}, \frac{\partial}{\partial \alpha^2}, ... \right) \equiv (\partial_1, \partial_2, ...) \tag{B.8}$$

From now on we will use that notation to write the gradient. Specifically, for the components of the $\boldsymbol{\alpha}$ vector we will use ∂_β whereas for differentiating with respect to the components of $\mathbf{u}$ we will use ∂_i.

Now, consider the derivative with respect to an arbitrary component of the parameter vector $\boldsymbol{\alpha}$ of the generalised FLD:

$$\partial_\beta \Phi = \frac{1}{\sigma_1^2 + \sigma_2^2} \left(\partial_\beta \left(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \right) \left(\sigma_1^2 + \sigma_2^2 \right) + \left(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \right) \partial_\beta \left(\sigma_1^2 + \sigma_2^2 \right) \right) \tag{B.9}$$

We can compute separately now the two derivatives of the last expression:

$$\begin{aligned}\partial_\beta \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \partial_\beta (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\, \partial_\beta (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\, \partial_\beta (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\, (\partial_\beta \boldsymbol{\mu}_1 - \partial_\beta \boldsymbol{\mu}_2) \end{aligned} \tag{B.10}$$

where we have used that the transposition commutes with the derivative, and

$$\partial_\beta \left(\sigma_1^2 - \sigma_2^2\right) = 2\, \partial_\beta \sigma_1 + 2\, \partial_\beta \sigma_2 \tag{B.11}$$

The derivatives of the mean can be written as the derivatives of its components:

$$\partial_\beta {\mu_c}^i = \frac{1}{M_c} \sum_{\mathbf{u} \in \mathcal{U}_{(c)}} \partial_\beta u^i \tag{B.12}$$

And the derivatives of the σ as:

$$\partial_\beta \sigma_c = \sum_{\mathbf{u} \in \mathcal{U}_{(c)}} (\partial_\beta {\mu_c}^i - \partial_\beta u^i)(\partial_\beta {\mu^c}_i - \partial_\beta u_i) \tag{B.13}$$

This reduces the problem to find the derivatives of the transformed instances ${u_i}^j$. Now, if we call $\mathcal{T}_{\boldsymbol{\alpha}}^j(\mathbf{x}_i)$ to the jth component of the transformation $\mathcal{T}_{\boldsymbol{\alpha}}$ [1], we can formally define the components of the $\mathbf{u}$ instances:

$${\mathbf{u}_i}^j(\boldsymbol{\alpha}) \equiv \mathcal{T}_{\boldsymbol{\alpha}}^j(\mathbf{x}_i) \tag{B.14}$$

So we can write the derivatives in the following way:

$$\partial_\beta {\mathbf{u}_i}^j(\alpha) = \partial_\beta \mathcal{T}_{\boldsymbol{\alpha}}^j(\mathbf{x}_i) \tag{B.15}$$

This ends the derivation of the gradient of the generalised FLD as the last expression depends only on the chosen transformation $\mathcal{T}_{\boldsymbol{\alpha}}$.

[1] Actually, we could picture each of the $\mathcal{T}_{\boldsymbol{\alpha}}^j$ as a transformation by its own, and interpret $\mathcal{T}_{\boldsymbol{\alpha}}$ as a vector representation of the gathering of all those transforms.

B.2 Gradient Ascend algorithm for the BDFLD

The Basis Decomposition FLD is just a generalised FLD considering the Basis-Decomposition Transformation as defined in Definition 5.12. Therefore, we only need to compute the derivatives of the transformation.

To simplify the notation, as the BD Transformation has only two components, we will slightly change the notation for the components of the transformed instances in the following way:

$$u_i = \mathcal{T}^1_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{\|w_j\ \mathbf{z}^j\|}\ \langle w^j\ \mathbf{z}_j, \mathbf{x}_i\rangle \tag{B.16}$$

$$v_i = \mathcal{T}^2_{\mathbf{w}}(\mathbf{x}_i) = u_i^2 \tag{B.17}$$

where, remember, we are assuming that the instances $\mathbf{x} \in \mathcal{X}$ are normalised. Note that we have renamed the generic parametrisation $\boldsymbol{\alpha}$ to fit the parameters of the BD Transformation $\mathbf{w}$. We will still use, however, Greek letters to derive with respect to the parameters of the transformation (i.e. $\partial_\beta \equiv \partial/\partial w^\beta$).

The derivatives for the second component of the transformation can be trivially reduced in terms of the derivatives of the first component:

$$\partial_\beta v_i = 2u_i\ \partial_\beta u_i \tag{B.18}$$

Therefore, we just need to compute the derivatives for the u_i to complete the derivation:

$$\begin{aligned}\partial_\beta u_i &= \partial_\beta\langle\boldsymbol{\xi}, \mathbf{x}_i\rangle \\ &= \frac{1}{\|w^j\ \mathbf{z}_j\|}\ \partial_\beta\langle w^j\ \mathbf{z}_j, \mathbf{x}_i\rangle + \partial_\beta\left(\frac{1}{\|w^j\ \mathbf{z}_j\|}\right)\langle w^j\ \mathbf{z}_j, \mathbf{x}_i\rangle\end{aligned} \tag{B.19}$$

As before, we will compute separately those two derivatives:

$$\begin{aligned}\partial_\beta\langle w^j\ \mathbf{z}_j, \mathbf{x}_i\rangle &= \partial_\beta(w^j)\langle\mathbf{z}_j, \mathbf{x}_i\rangle \\ &= \delta^j_\beta\ \langle\mathbf{z}_j, \mathbf{x}_i\rangle \\ &= \langle\mathbf{z}_\beta, \mathbf{x}_i\rangle\end{aligned} \tag{B.20}$$

$$\begin{aligned}
\partial_\beta \left(\frac{1}{\|w^j \ \mathbf{z}_j\|} \right) &= \frac{-1}{2 \ \|w^j \ \mathbf{z}_j\|^3} \partial_\beta \left(\|w^j \ \mathbf{z}_j\| \right) \\
&= -\frac{1}{\|w^j \ \mathbf{z}_j\|^3} \langle w^j \ \mathbf{z}_j, \partial_\beta w^j \mathbf{z}_j \rangle \\
&= -\frac{1}{\|w^j \ \mathbf{z}_j\|^2} \langle \boldsymbol{\xi}, \delta^j_\beta \mathbf{z}_j \rangle \\
&= -\frac{1}{\|w^j \ \mathbf{z}_j\|^2} \langle \boldsymbol{\xi}, \mathbf{z}_\beta \rangle \\
&= -\frac{1}{\|w^j \ \mathbf{z}_j\|^2} u_\beta
\end{aligned} \tag{B.21}$$

Putting altogether Equations B.19, B.20 and B.21 we finish our derivation:

$$\begin{aligned}
\partial_\beta u_i &= \frac{1}{\|w^j \ \mathbf{z}_j\|} \ \langle \mathbf{z}_\beta, \mathbf{x}_i \rangle - \frac{1}{\|w^j \ \mathbf{z}_j\|^2} \ u_\beta \ m \langle w^j \ \mathbf{z}_j, \mathbf{x}_i \rangle \\
&= \frac{1}{\|w^j \ \mathbf{z}_j\|} \ \langle \mathbf{z}_\beta, \mathbf{x}_i \rangle - \frac{1}{\|w^j \ \mathbf{z}_j\|} \ u_\beta \ \langle \boldsymbol{\xi}, \mathbf{x}_i \rangle \\
&= \frac{1}{\|w^j \ \mathbf{z}_j\|} \left(\langle \mathbf{z}_\beta, \mathbf{x}_i \rangle - u_\beta \ u_i \right)
\end{aligned} \tag{B.22}$$

B.2.1 Dealing with non-normalised instances

In some situations, the instances set cannot be normalised. This kind of constraint can arise from adding more terms into the objective function. For instance, when adding the BD-ICA contribution later, if we would not use the reduced set of instances $\mathcal{Z}$ but the original set $\mathcal{X}$ to construct the candidates $\boldsymbol{\xi}$, it would be convenient to force the instances to be demeaned and to have unit variance, as we did with $\mathbf{z} \in \mathcal{Z}$. Of course, this last constraint is incompatible with the norm normalisation. We could still maintain two versions of the instances, but it is customary to be thrifty in this sense when dealing with fMRI. In these kind of cases, it could be convenient to have a discriminant working with non-normalised instances. As we said before, such discriminant can be built with just substituting $\mathbf{x}$ by $\mathbf{x}/\|\mathbf{x}\|$. In terms of the discriminant, we only need to change the transformation in Equation 5.13 by the alternative transformation in Equation B.23, described as follows:

$$\mathcal{T}^{\mathrm{BD}}_{\mathbf{w}}(\mathbf{x}) \equiv \begin{pmatrix} \frac{1}{\|\mathbf{x}\|} \langle \boldsymbol{\xi}, \mathbf{x} \rangle \\ \frac{1}{\|\mathbf{x}\|^2} \langle \boldsymbol{\xi}, \mathbf{x} \rangle^2 \end{pmatrix}$$

This change is linearly propagated to the derivatives. This means that instead of Equation B.22 we need to use Equation B.23:

$$\partial_\beta u_i = \frac{1}{\|w^j\ \mathbf{z}_j\|}\left(\frac{1}{\|\mathbf{x}_i\|}\langle \mathbf{z}_\beta, \mathbf{x}_i\rangle - u_\beta\ u_i\right) \tag{B.23}$$

and instead of Equation B.18, we have to use Equation B.24:

$$\partial_\beta v_i = \frac{2}{\|\mathbf{x}\|} u_i\ \partial_\beta u_i \tag{B.24}$$

B.3 Gradient Ascend algorithm for BD-DICA

We are now in position to show the final derivatives for BD-DICA. As before, we will compute the derivative with respect to an arbitrary component of $\mathbf{w}$ instead of the gradient of Equation 5.19.

Consider the following expression:

$$\partial_\beta \mathcal{J}(\mathbf{w}) = (1-\kappa)\ \partial_\beta J(\mathbf{w}) + \kappa\ \partial_\beta \Phi(\mathbf{w}) \tag{B.25}$$

The second term on the right side of B.25 has been already solved along the previous two sections. We will therefore centre ourselves in the first term, corresponding to the objective function of BD-ICA in Equation 3.5. During this section, we will use the notation $\mathbf{y} = w^j\ \mathbf{z}_j$ to save some ink.

$$\begin{aligned} \partial_\beta J(\mathbf{w}) &= \partial_\beta (E[G(y)] - E[G(\nu)])^2 \\ &= 2(E[G(y)] - E[G(\nu)])\ \partial_\beta E[G(y)] \end{aligned} \tag{B.26}$$

where we have assumed that $E[G(\nu)]$ is a constant, for the reasons exposed before in Section 3.3.1. Actually, as all distributions will be demeaned and will have unit variance, μ is actually a normal distribution. That means that, for our choice of $G(y)$ described in Equation 3.7, it can be derive [5] that:

$$E[G(\nu)] = -\frac{1}{\sqrt{2}} \tag{B.27}$$

Following our derivation, we need still to compute the flowing derivative:

$$\begin{aligned} \partial_\beta E[G(y)] &= \frac{1}{N}\sum_i^N \partial_\beta G(y^i) \\ &= \frac{1}{N}\sum_i^N \partial_{y_i}(G(y^i))\ \partial_\beta y^i \end{aligned} \tag{B.28}$$

Now, the derivatives of $G(y)$ and $y^i = w^l {z_l}^i$:

$$\partial_y G(y) = -\,\partial_y \left(e^{-\frac{y^2}{2}} \right) = y\, e^{-\frac{y^2}{2}} = y\, G(y) \tag{B.29}$$

$$\partial_\beta\, y = \partial_\beta\, w^l\, {z_l}^i = \delta^l_\beta\, {z_l}^i = {z_\beta}^i \tag{B.30}$$

Grouping the terms:

$$\partial_\beta J(\mathbf{w}) = 2 \left(E[G(y)] + \frac{1}{\sqrt{2}} \right)\, \frac{1}{D} \sum_i^D y^i\, G(y^i)\, {z_\beta}^i \tag{B.31}$$

B.4 Dynamics of the Gradient Ascend algorithms

As the reader might already noticed, the objective function of BD-DICA has a series of spurious maxima we cannot avoid. They are caused by the Negentropy term, which have maxima in all the projections aligned with the source vectors $\mathbf{s} \in \mathcal{S}$. Therefore, it would be interesting to count with an implementation of Gradient Ascend capable of skipping small maxima.

To deal with this problem, we have implemented two different strategies. First of all, the first steps of the algorithm are conducted with a different (larger) κ, to seed the starting point of the actual process to a position with a high score in the discriminant.

In addition, inspired in the implementation of several algorithms for back propagation in Neural Networks (see [50]) we added a term of inertia to the Hill Climbing process which could help to avoid falling in too shallow maxima. The implementation of this term is quite easy. Suppose that the variation in $\mathbf{w}$ of the kth iteration is Δ^k. Then, in the following iteration:

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \boldsymbol{\nabla} \mathcal{J}(\mathbf{w}) + m\Delta^k \tag{B.32}$$

where m is the mass term, which manages the amount of inertia of the algorithm.

The resulting algorithm is highly configurable. We have found a combination of such parameters that work relatively fine with fMIR data: the number of steps in the initialisation (set to 5), the κ value for that part of the process ($\kappa_{init} = \kappa + 0.2$), the mass term ($m = 0.4$), the learning rate ($\eta = 0.15$), the tolerance for convergence ($\theta = 10^{-5}$), a possible number of max iterations (we chose MaxIt = 99). The value for κ will depend on the amount of noise and the size of the dataset, but a value of $\kappa \sim 0.4$ has demonstrated to work nicely.

If another combination has to be performed to deal with new data, we can offer some advices. We have seen that a too high distances between the κs can make the algorithm start too far away from the optimal condition, which is translated in a large number of iterations. Each of those iterations can take from 2 to 40 seconds, so this number should be taken seriously.

A $m > 0.5$ might cause the algorithm to oscillates indefinitely when a reasonable learning rate is selected and it is therefore not recommended.

Within the parameters of the initialisation it is possible to select a different η for this part of the process. This can help to jump quickly to a point with good solutions and spare some iterations in the rest of the process if $\eta_{\text{init}} > \eta$. We have hard coded $\eta_{\text{init}} = 4\eta$. This combination seems to work very good. However, be careful when increasing this number as the Hill Climbing of the rest of the process inherits a term of inertia form the initialisation.

An additional implemented possibility is to repeat the initialisation for several random seeds and select the best one for the rest of the process. This can help to avoid local maxima when dealing with situations with lots of spurious solutions. In the case of fMRI, however, it proved to be of no use at all.

An implementation of this algorithm as described in this appendix will be soon publicly available in www.github/qtabs/BDDICA/ running under MATLAB. The code could theoretically works perfectly under Octave self for an external library implementing PCA. In addition, SPM is needed to import fMRI images.

Bibliography

[1] J. Hernández Orallo, M. J. Ramírez Quintana, and C. Ferri Ramírez, *Introducción a la Minería de Datos*. Pearson Educación, 2004.

[2] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fmri data," *Trends in cognitive sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications and Control Series, Wiley, 2004.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, vol. 10. Wiley, 2001.

[5] C. S. Dhir and S.-Y. Lee, "Discriminant independent component analysis," *Trans. Neur. Netw.*, vol. 22, pp. 845–857, June 2011.

[6] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data," *Neuroimage*, vol. 45, no. 1, pp. S163–S172, 2009.

[7] R. Buxton, *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*. Cambridge University Press, 2002.

[8] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, eds., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.

[9] V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. J. Pekar, "ICA of functional MRI data: An overview," in *Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, (Nara, Japan), pp. 281–288, April 2003.

[10] B. B. Biswal, "Resting state fmri: a personal history," *Neuroimage*, vol. 62, no. 2, pp. 938–944, 2012.

[11] S. A. Harrison and F. Tong, "Decoding reveals the contents of visual working memory in early visual areas," *Nature*, vol. 458, no. 7238, pp. 632–635, 2009.

[12] D. Saur, O. Ronneberger, D. Kümmerer, I. Mader, C. Weiller, and S. Klöppel, "Early functional magnetic resonance imaging activations predict language outcome after stroke," *Brain*, vol. 133, no. 4, pp. 1252–1264, 2010.

[13] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, *et al.*, "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4734–4739, 2010.

[14] V. Calhoun, T. Adali, G. Pearlson, and J. Pekar, "A method for making group inferences from functional mri data using independent component analysis," *Human brain mapping*, vol. 14, no. 3, pp. 140–151, 2001.

[15] J. Stone, "Independent component analysis: a tutorial introduction, 2004."

[16] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fmri: a tutorial overview," *Neuroimage*, vol. 45, no. 1, pp. S199–S209, 2009.

[17] Z. Yu-Feng, H. Yong, Z. Chao-Zhe, C. Qing-Jiu, S. Man-Qiu, L. Meng, T. Li-Xia, J. Tian-Zi, and W. Yu-Feng, "Altered baseline brain activity in children with adhd revealed by resting-state functional mri," *Brain and Development*, vol. 29, no. 2, pp. 83–91, 2007.

[18] D. Cordes, V. M. Haughton, K. Arfanakis, J. D. Carew, P. A. Turski, C. H. Moritz, M. A. Quigley, and M. E. Meyerand, "Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data," *American Journal of Neuroradiology*, vol. 22, no. 7, pp. 1326–1333, 2001.

[19] Y. Du and Y. Fan, "Group information guided ica for fmri data analysis.," *NeuroImage*, vol. 69, pp. 157–197, 2013.

[20] C. F. Beckmann, C. E. Mackay, N. Filippini, and S. M. Smith, "Group comparison of resting-state fmri data using multi-subject ica and dual regression," *Neuroimage*, vol. 47, p. S148, 2009.

[21] Y. Fan, Y. Liu, H. Wu, Y. Hao, H. Liu, Z. Liu, and T. Jiang, "Discriminant analysis of functional connectivity patterns on grassmann manifold," *Neuroimage*, vol. 56, no. 4, pp. 2058–2067, 2011.

[22] J. Sui, T. Adali, G. D. Pearlson, and V. D. Calhoun, "An ica-based method for the identification of optimal fmri features and components using combined group-discriminative techniques," *Neuroimage*, vol. 46, no. 1, pp. 73–86, 2009.

[23] W. Lu and J. C. Rajapakse, "Approach and applications of constrained ica," *Neural Networks, IEEE Transactions on*, vol. 16, no. 1, pp. 203–212, 2005.

[24] E. Kreyszig, *Differential geometry*. Dover Publications, 1991.

[25] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[26] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[27] C. C. M. Grinstead and J. L. Snell, *Introduction to probability*. American Mathematical Soc., 1997.

[28] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[29] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[30] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[31] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[32] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 41–48, Aug. 1999.

[33] T. Li, S. Zhu, and M. Ogihara, "Using discriminant analysis for multi-class classification: an experimental investigation," *Knowledge and information systems*, vol. 10, no. 4, pp. 453–472, 2006.

[34] J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *Neural Networks, IEEE Transactions on*, vol. 18, no. 5, pp. 1433–1441, 2007.

[35] J. Marsden and M. Hoffman, *Basic Complex Analysis*. W. H. Freeman, 1999.

[36] S. Schneider, T. Münte, A. Rodriguez-Fornells, M. Sailer, and E. Altenmüller, "Music-supported training is more efficient than functional motor training for recovery of fine motor skills in stroke patients," *Music Perception: An Interdisciplinary Journal*, vol. 27, no. 4, pp. 271–280, 2010.

[37] A. Rodriguez-Fornells, N. Rojo, J. L. Amengual, P. Ripollés, E. Altenmüller, and T. F. Münte, "The involvement of audio–motor coupling in the music-supported therapy applied to stroke patients," *Annals of the New York Academy of Sciences*, vol. 1252, no. 1, pp. 282–293, 2012.

[38] M. R. G. Brown, G. S. Sidhu, R. Greiner, N. Asgarian, M. Bastani, P. H. Silverstone, A. J. Greenshaw, and S. M. Dursun, "Adhd-200 global competition: Diagnosing adhd using personal characteristic data can outperform resting state fmri measurements," *Frontiers in Systems Neuroscience*, vol. 6, no. 69, 2012.

[39] C. Craddock, T. N. Bureau, T. A. . consortium, and V. T. ARC, "The athena preprocessing for the adhd200 dataset," 2012.

[40] J. Buitelaar, J. Sergeant, R. Mindera, A. A. Vasquéz, R. Cools, S. V. Faraone, B. Franke, C. Hartman, P. H. D. Heslenfeld, D. Norris, J. Oosterlaan, N. Rommelse, D. Slaats-Willemse, and M. Zwiers, "Adhd200, neuroimage sample," 2012.

[41] D. Fair, J, Nigg, B. Nagela, and D. Bathula, "Adhd200, ohsu sample," 2012.

[42] J. S. H. Gävert, J. Hurri and A. Hyvärinen, "Fastica for matlab 7.x and 6.x v2.5," 2005.

[43] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Human brain mapping*, vol. 15, no. 1, pp. 1–25, 2002.

[44] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 2, pp. 137–152, 2004.

[45] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[46] B. Kolb and I. Whishaw, *Introduction to Brain and Behavior with CDROM.* Worth Publishers, 2001.

[47] N. Rojo, J. Amengual, M. Juncadella, F. Rubio, E. Camara, J. Marco-Pallares, S. Schneider, M. Veciana, J. Montero, B. Mohammadi, E. Altenmüller, C. Grau, T. F. Münte, and A. Rodriguez-Fornells, "Music-supported therapy induces plasticity in the sensorimotor cortex in chronic stroke: a single-case study using multimodal imaging (fMRI-TMS).," *Brain injury : [BI]*, vol. 25, no. 7-8, pp. 787–793, 2011.

[48] E. Altenmüller, J. Marco-Pallares, T. Münte, and S. Schneider, "Neural reorganization underlies improvement in stroke-induced motor dysfunction by music-supported therapy," *Annals of the New York Academy of Sciences*, vol. 1169, no. 1, pp. 395–405, 2009.

[49] S. A. Huettel, P. B. Mack, and G. McCarthy, "Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex," *Nature neuroscience*, vol. 5, no. 5, pp. 485–490, 2002.

[50] T. M. Mitchell, *Machine Learning.* New York, NY, USA: McGraw-Hill, Inc., 1 ed., 1997.