

The challenge of applying machine learning techniques to diagnose schizophrenia using multi-site fMRI data

by

Roberto Ivan Vega Romero

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

© Roberto Ivan Vega Romero, 2017

Abstract

One of the main challenges for the use of machine learning techniques in neuroimaging data is the *small n, large p* problem. Datasets usually contain only a few hundreds of instances (n), each of which is described using hundreds of thousands of features (p). In this dissertation, we explore the effects of reducing the number of features by analyzing 264 specific regions of interest of the brain, and increasing the number of instances by merging imaging data obtained from different scanning sites for distinguishing people with schizophrenia from healthy controls.

Empirical results show that, using features related to functional connectivity of the brain, we can achieve an accuracy above the chance level (over 70%), when using data from a single scanning site for both training and testing. However, this performance decreases when additional data from a different scanning site is used as part of the training process. We attribute the decrease in performance to *batch effects*: technical noise introduced at different scanning sites that confound the biological signal of interest.

Batch effects are often disregarded in association studies because there is often no statistically significant interaction between the scanning site and the variables being analyzed. In this work, we highlight important differences between association studies and prediction studies, and we argue that in the latter, batch effects matter. Our experiments reveal that not taking them into account reduces the performance of a learned classifier compared to using data from a single scanning site, even though this drastically reduces the size of the

training set. In addition, if we make the scanning site the target variable to predict, we can create a classifier that can distinguish *among sites* with an accuracy $> 80\%$.

We empirically show that if the same subjects are scanned in two different sites, then a neural network that maps the fMRI scan from one scanner into another is enough for correcting the batch effects. In more realistic situations, involving disjoint set of subjects, simple techniques like z-score normalization or whitening can remove batch effects caused by translations and scaling, or translations and rotations of the feature matrix. Both approaches proved successful in reducing the accuracy of scanning site classification to near chance level, but they were unable to improve the accuracy of schizophrenia diagnosis using multisite data. This is a strong indication that batch effects go beyond these simple linear transformations.

Finally, we explored the use of BECCA (batch effects correction using canonical correlation analysis) and approaches based on autoencoders for decreasing the influence of batch effects. These attempts were also unsuccessful under our test scenarios, suggesting that batch effects is a serious problem in prediction studies using fMRI data, and that more effort should be taken to understand their nature in order to reduce their influence.

To AnaLi,

For being my support, my inspiration, and the love of my life.

To my parents and brother,

For teaching me the values and attitudes that guide all my decisions.

Acknowledgements

These years at the University of Alberta have been an incredible experience. I have learned many things, not only academically speaking, but also in the at personal and cultural level. Many people and institutions have been part of this journey, and I cannot express in words my gratitude towards all of them.

I had the fortune of having two great mentors during the course of my studies: Russ Greiner and Matt Brown. They taught me not only how to frame and develop the ideas presented in this dissertation, but also the importance of collaboration, patience, and hard work. You can learn a lot by watching passionate people do what they love, and that has been my case with Russ and Matt. By observing them, I have a better understanding of what high quality research means. *What is the falsifiable claim?* is a question that will be present in all the research that I do from now on.

I want also to thank the professors Dale Schuurmans and Pierre Boulanger, who were part of the committee, and that took the time to read my dissertation and evaluate my work. I really appreciate your time and comments that improved the quality of this document.

I have met amazing people in these two years. They have shared with me their stories and experiences, and that has helped me to grow as a person. I have been living in Canada, but I feel like I have experienced China, India, Iran, Sri Lanka, Colombia, Bangladesh, U.S. and many other countries, all because of the people that I have met here. Special thanks go to my friends Kriti Khare, Andrew Hellmund, Shaiful Chowdhury, Fateme Bahri, Mohammad Salem, David Pizon and Paola Sanchez with whom I spent a wonderful time, even when we did not do research together. I also want to express my deepest gratitude to my friends: Bhaskar Sen, Neil Borle, Luke Kumar, Tanvir Sajed,

Mina Gheiratmand, and Negar Hassanpour. It was great to have those very interesting and fun discussions with you. Finally, as a TA, I learned a lot about many technical topics. I want to thank Junfeng Wen and Bernardo Avila for sharing part of their knowledge with me. I admire your hard work and technical knowledge.

My program started in September of 2014, but my path to the master program started much before. I want to thank the professors Gildardo Sanchez, Mauricio Antelis, Rita Fuentes and Alejandro Garcia for introducing me to research. You motivated me to pursue a graduate degree. Thanks also to Ken Bauer and Joaquin Campos, who encouraged me to apply to the University of Alberta, and to Leonardo Flores, who has always supported me through all my academic life.

Several institutions made possible that I studied here. I want to thank the Department of Computing Science of the University of Alberta for the support that they gave me by allowing me to be a teaching assistant. The Mexican Council of Science and Technology of Mexico (CONACYT) and the Public Secretariat of Education (SEP) gave very generous scholarships that allowed to cover my expenses in Canada. Without any of these institutions, I could not have been able to study my master degree.

My friends and family in Mexico have played a crucial role. I specially want to thank my parents, Roberto and Altagracia, and my brother, Alejandro, for all the support that they have given me through these years. Being part of a loving family is one of the greatest gift that I have received. Even when the distance is big, you are always with me.

Finally, I want to thank my loving wife, AnaLi. My life is happier and brighter since I met you. Thanks for all your encouragement, patience and support. We already have had incredible experiences together, and our journey is just starting. I can confidently say that I would not have done this without you.

Table of Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 1.1 | Problem definition | 5 |
| 1.2 | Contributions | 5 |
| 2 | Background | 7 |
| 2.1 | Basics of fMRI | 7 |
| 2.1.1 | Preprocessing pipeline | 9 |
| 2.2 | Functional connectivity | 10 |
| 2.3 | Association vs prediction studies | 11 |
| 3 | fMRI analysis from the machine learning perspective | 17 |
| 3.1 | Task description | 18 |
| 3.2 | Dataset | 19 |
| 3.3 | Feature extraction | 19 |
| 3.3.1 | Parcellation of the brain | 20 |
| 3.3.2 | Feature matrix | 21 |
| 3.4 | Support Vector Machine (SVM) | 22 |
| 3.5 | Learning algorithm and accuracy estimation | 24 |
| 4 | Classification results | 27 |
| 4.1 | Single site | 27 |
| 4.2 | Multiple sites | 30 |
| 4.3 | Batch effects | 32 |
| 4.3.1 | Scanning site classification | 33 |
| 4.3.2 | Traveling subject dataset | 36 |
| 4.3.3 | Solving the traveling subject problem | 37 |
| 5 | Reducing the influence of batch effects | 42 |
| 5.1 | Simple transformations I: Translation and scaling | 42 |
| 5.2 | Simple transformations II: Rotation and translation | 45 |
| 5.3 | BECCA | 48 |
| 5.4 | Non-linear transformations | 50 |
| 5.4.1 | Self-learning a feature representation | 50 |
| 5.4.2 | Bi-shifting autoencoders | 53 |
| 5.5 | Summary of methods | 55 |
| 6 | Conclusions | 57 |
| 6.1 | What is next? | 60 |
| 6.2 | Highlights | 60 |
| | Bibliography | 62 |
| | Appendices | 69 |

| | | |
|----------|--|-----------|
| A | Male versus female classification | 70 |
| A.1 | Gaussian Markov Random Fields | 71 |
| A.2 | Learning the models | 72 |
| A.3 | Dataset and results | 74 |
| B | Additional approaches | 76 |

List of Tables

| | | |
|-----|---|----|
| 3.1 | Number of participants in the dataset used for the experiments. Every participant was scanned 4 times, the total number of scans (instances) is 4 times the number of participants. | 20 |
| 4.1 | Single site dataset size and results using the 264 regions of interest. Statistically significant differences (using t-test, $p < 0.05$) between the mean accuracy and the baseline are marked in bold. | 28 |
| 4.2 | Single site dataset size and results using the 38 regions of interest corresponding to the fronto-parietal network and auditory network. Statistically significant differences between the mean accuracy and the baseline are marked in bold. | 29 |
| 4.3 | Classification accuracy for prediction of scanning site (binary classification) | 33 |
| 4.4 | Accuracy comparison in prediction of scanning site and participant ID before versus after batch effects correction. (Results on the hold-out set using the traveling subject data) | 41 |
| 5.1 | Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using z-score normalization. Values in bold indicate single site classification. | 44 |
| 5.2 | Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using whitening. Values in bold indicate single site classification. | 48 |
| 5.3 | Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using BECCA. Values in bold indicate single site classification. | 50 |
| 5.4 | Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using BECCA. Values in bold indicate single site classification. | 52 |
| 5.5 | Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using bishifting autoencoders. Values in bold indicate single site classification. | 55 |
| A.1 | Accuracy of Gaussian Markov Random Fields in the task of male versus female classification using the ADHD dataset (healthy controls) | 75 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Example of how machine learning can be used to distinguish between people with schizophrenia (SCZ) and healthy controls (HC) using fMRI data. | 2 |
| 2.1 | fMRI from a computational point of view. (a) Compute a 3-D matrix (corresponding to the BOLD signal at each voxel), every 2-3 seconds. (b) Every voxel has an associated time series. (c) Every voxel is associated with a particular location of the brain. Hence, this is an estimation of the level of activity at a particular location over time. | 8 |
| 2.2 | Regions of the brain with statistically significant differences between schizophrenia patients and healthy controls under the experiment of Walter et al. [63] (reproduced by permission of Oxford University Press) | 12 |
| 2.3 | Distribution of two hypothetical features across two groups of interest. The difference in the means of the activation in the frontal cortex (left) in both groups is statistically significant ($p < .01$), but that is not the case for the activation in the motor cortex (right) ($p = .63$) | 14 |
| 2.4 | Distribution of two hypothetical features across two groups of interest. Recall from Figure 2.3 that the difference in the means of the activation in the frontal cortex in both groups is statistically significant ($p < .01$), but that is not the case for the activation in the motor cortex ($p = .63$) | 15 |
| 2.5 | Distribution of two hypothetical features in two groups of interest. Even when both features have statistically significant differences between groups ($p < 0.01$), their predictive power is very limited. | 16 |
| 3.1 | Machine learning approach for classification problems | 17 |
| 3.2 | The two scenarios explored in this dissertations: a) Training set and test set come from the same scanning site, b) Add data from an external site to the training set. | 18 |
| 3.3 | (a) Mask used for extracting the time series of every region of interest. (b) Zero-mean time signals of all the voxel within the mask in a region of interest. | 21 |
| 3.4 | Every entry in the matrix on the left is the Pearson's correlation coefficient between two regions of interest. The correlation ($\rho = 0.71$) between the time series of ROI_1 and ROI_2 is shown as an example. Then, the upper triangular matrix (green) can be concatenated into a single vector, of length $l = \binom{264}{2}$ | 22 |

| | | |
|------|---|----|
| 3.5 | a) SVM creates a decision boundary that maximizes the margin between the decision boundary and the closest points to it. b) For the non-linearly separable case the slack variables ξ quantify the distance between the decision boundary and the points inside the margin, or in the wrong side of the decision surface. | 24 |
| 4.1 | Distribution of the accuracy on every scanning site after 30 experiments | 29 |
| 4.2 | Influence of the size of the training set on the performance in a hold out set. | 30 |
| 4.3 | Average accuracy after merging datasets from different scanning sites. | 31 |
| 4.4 | Mean accuracy after including data from all the scanning sites in the training set. | 32 |
| 4.5 | The feature vector for all subjects (both schizophrenia and healthy control) looks different across different scanning sites. | 34 |
| 4.6 | Analysis of the 2 networks used for prediction. Regions with high pairwise connectivity that are consistent across the different scanning site are shown in green. Note that the voxels corresponding to the auditory network are highly interconnected. | 35 |
| 4.7 | Physical location of the regions of interest that presented high connectivity across all the scanning sites. | 35 |
| 4.8 | Analysis of the 264 ROI. Regions consistent across the different scanning site are shown in green. Note that sensory/somatomotor, visual and auditory networks present high connectivity. | 36 |
| 4.9 | Projection of the feature matrix of the traveling subjects dataset into the first two principal components. Every point represents an fMRI scan, whose color represents a participant and whose shape identifies the scanning site. | 37 |
| 4.10 | Neural network architecture, which is essentially the same as the one used by autoencoders. The objective of the network is to produce an approximation of how data from scanning site A is represented in scanning site B. | 38 |
| 4.11 | Projection of the traveling subject dataset into the first two components after correcting the batch effects using a neural network. | 41 |
| 5.1 | Graphical representation of z-score removing translation and scaling. | 44 |
| 5.2 | Z-score normalization only corrects batch effects caused by translation and scaling of the data, but it is insufficient for other types of transformations. | 45 |
| 5.3 | Whitening can correct batch effects caused by rotations and translations of a dataset. | 47 |
| 5.4 | Intuition of the decomposition assumed by BECCA. Assume that there are $q = 3$ prototypes, and every participant can be represented as a linear combination of them. (Noise and batch effects are not represented in the figure.) | 49 |
| 5.5 | The stacked autoencoders train each layer independently: a) The raw inputs are mapped to themselves. b) The hidden layer of the previous autoencoder becomes the input for the next one. This process is repeated until the desired depth is achieved. The hidden layer of the last autoencoder can be used as the input to a learning algorithm. | 51 |

| | | |
|-----|--|----|
| 5.6 | Architecture of the neural network. The number of neurons of each layer is indicated at the top. | 52 |
| 5.7 | Bishifting Autoencoder. Every input from the one domain (target or source) is mapped to itself and to an approximation of how it would be represented in the other domain. | 54 |
| 5.8 | Bishifting Autoencoder. Every input from the one domain (target or source) is mapped to itself and to an approximation of how it would be represented in the other domain. | 56 |
| A.1 | Our methodology for classifying the resting-state fMRI scans as male or female. | 73 |

Chapter 1

Introduction

Over the last decades, many researchers have focused their careers on increasing our knowledge of the human brain and its disorders. There are more than 1,000 mental disorders of the central nervous system, and they cause more hospitalizations than any other disease group, including cancer and heart problems. They also represent a large economic burden, since their estimated cost is more than \$600 billion USD per year, only in the US [16]. Prompt diagnosis and prognosis can improve the quality of life of people with mental health problems, while saving millions of dollars in the process.

Unfortunately, making an accurate diagnosis is a challenging task, since many disorders have overlapping symptoms and there is no standard biologically-based clinical test yet [2]. This has triggered an increasing interest in the development of technological tools that can potentially help with the diagnosis or prognosis, such as the use of neuroimaging data. Among the different neuroimaging techniques, functional magnetic resonance imaging (fMRI) is one of the most promising [3].

fMRI is a non-invasive technique that measures the neuronal activity in the human brain [25], and it has become an important tool for studying the cognitive functions in healthy people as well as their changes in the presence of a mental disorder or illness [7]. However, an fMRI experiment produces a massive amount of data, in the range of tens of millions of real values for a single patient. It is impossible for a human being to analyze such amount of data, so the interest in building tools that perform this analysis automatically

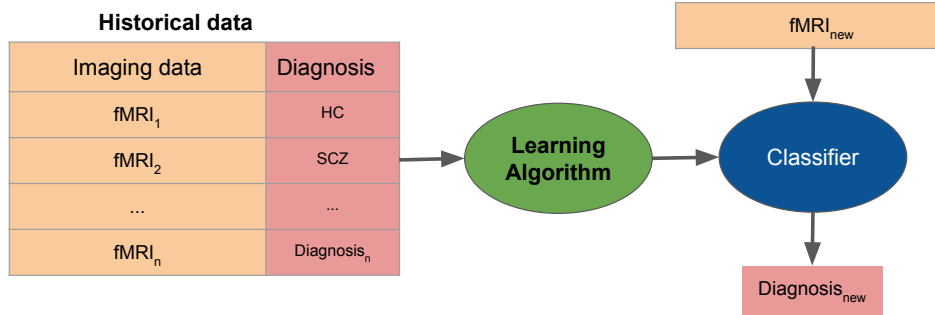


Figure 1.1: Example of how machine learning can be used to distinguish between people with schizophrenia (SCZ) and healthy controls (HC) using fMRI data.

is increasing. Machine learning is one of these tools.

Machine learning can be defined as a set of methods that identify patterns in historical data with the objective of using these patterns for making predictions in new, previously unseen data [41]. In the context of fMRI data and automatic diagnosis of mental illnesses, this would involve: collecting as many fMRI scans as possible of the groups of interest (for example, healthy controls and people with schizophrenia), learning a classifier by applying a learning algorithm for finding patterns in the collected data that are discriminative between the groups, and finally applying the resulting classifier to new instances to make predictions. Figure 1.1 depicts this approach. A detailed description of how to encode the fMRI scans into a feature vector that can be used by the learning algorithm to create the final classifier will be presented in detail in Section 3.

Several research groups have implemented machine learning approaches in the analysis of fMRI data in order to build predictors that can diagnose: attention deficit and hyperactivity disorders [6, 43], mild cognitive impairment and Alzheimer’s disease [34], schizophrenia [2], autism [69]; or classify people according to a variable of interest such as: gender [52], age group [62], or smoking status [44]. The reported accuracy of the different tasks varies from chance level to $> 85\%$, depending on the task, dataset, features, and learning algorithm used for creating the classifier.

Despite the differences in their objectives and performance, studies involv-

ing fMRI data face a common challenge: the high dimensionality of the data (in the range of tens of millions of features) and the relatively few number of instances (at most a few hundreds of them). This problem is known as *small n, large p*, where n refers to the number of instances and p refers to the number of features [22]. This situation is undesirable because machine learning approaches assume that the training sample is a good approximation of the real distribution of the data, which might not be the case with only a few instances in high dimensional space. At the same time, high dimensional data is likely to contain many redundant and irrelevant features that might obscure the patterns in the data, greatly reducing the performance of learning algorithms [70].

Two standard approaches for dealing with the *small n, large p* problem are: decrease the number of features, and increase the number of instances [61]. For solving the first problem, it is possible to use feature selection or dimensionality reduction algorithms. On the other side, one plausible way of increasing the number of instances is to simply merge fMRI data (of the same phenomenon) collected at different scanning sites into a single set, and then apply a learning algorithm to this new expanded dataset. Surprisingly, this naive approach does not work as expected. Despite having a bigger training sample, the accuracy of the predictions drops when using multi-site data relative to the accuracy obtained by applying the same algorithm on data from only a single site.

One of the reasons for this poor generalization is that machine learning algorithms assume that all the data X and their corresponding labels Y come from the same joint distribution $p(X, Y)$. Technical noise introduced at different scanning locations might confound the real biological signal in different ways, modifying the original distribution. This transformation makes the distribution of data obtained at two different scanning locations, a and b , different: $P(X, Y | a) \neq P(X, Y | b)$. This phenomenon is well known in genomic studies, and is called *batch effects* [37]. In fMRI studies, batch effects can be caused by a variety of factors including: field strength of the magnet, manufacturer and parameters of the MRI scanner, radiofrequency noise environments, differences in the scanning protocol, and the general experience of

the participants in the study [21].

While interscanner variability is a well known phenomenon in the neuroimaging community [19, 38, 73, 18, 21, 17], many researchers report that its effect are irrelevant for their studies, or can be corrected by including the scanning site as a variable in the model [53, 11, 60, 57, 9]. This differs from the empirical results presented by other groups, which show a decrease in the classification accuracy on multi-site data, or show that a model trained in data extracted from one scanning site does not generalize to data from a different site [69, 27, 43, 65]. An important difference distinguishes both groups: The former focuses on *association studies*, whose aim is to find statistical differences at the group level between two or more populations, while the second group focuses on *prediction studies*, whose aim is to make predictions at individual level. Most of the studies in neuroimaging fall in the first category, but there is also a growing interest in the use of machine learning techniques and neuroimaging data for the automatic classification of mental disorders [2]. For prediction studies, batch effects matter.

This dissertation is focused on the use of multi-site fMRI data for the diagnosis of schizophrenia, and is structured as follows: The rest of this chapter describes the problem being addressed, and summarizes the contributions of this work. Chapter 2 introduces the background required for working with fMRI data from a computational point of view, and describes the differences between association studies (common in neuroscience) versus prediction studies (common in machine learning). Chapter 3 describes the methods used for feature extraction, learning a classifier, and evaluating performance. Chapter 4 shows the results in single site and multi-site classification. Chapter 5 describes the techniques used to decrease batch effects. Chapter 6 presents the conclusions and future work. I include two appendices at the end of this dissertation. Appendix A describes our approach for sex classification using fMRI data, a related task to the diagnosis of schizophrenia that motivated the research presented in this dissertation. Finally, Appendix B lists other approaches that we used to classify between people with schizophrenia from healthy controls (or for sex classification), but whose results were inferior to

the ones presented in the main body of this document.

1.1 Problem definition

Given a training set $D = \{(S_1, y_1), (S_2, y_2), \dots, (S_n, y_n)\}$ with pairs of fMRI scans S_i and their respective labels $y_i \in \{C, D\}$ (for Control, Disease), extracted from a single scanning site, find a lower dimensional representation $D_{train} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ that can be given to a learning algorithm, to create a classifier that classifies a new instance, X , from a scan S as C or D, with an accuracy above chance, and as high as possible.

For the case of multi-site datasets, where $P(X, Y | a) \neq P(X, Y | b)$ for two scanning sites a and b , we assume that the discrepancy between the joint distributions is caused exclusively by batch effects. We further assume that there exist two functions, f_a and f_b , such that $P(f_a(X), Y | a) = P(f_b(X), Y | b)$. Our objective is to find the functions $\hat{f}_a(x)$ and $\hat{f}_b(x)$ that approximate f_a and f_b in order to create a new training set $D_{trainMulti} = \{(\hat{f}_a(x_1^a), y_1^a), \dots, (\hat{f}_a(x_n^a), y_n^a), (\hat{f}_b(x_1^b), y_1^b), \dots, (\hat{f}_b(x_m^b), y_m^b)\}$ which includes the n training instances from scanning site a , and the m training instances from scanning site b . Ideally, a classifier created using the same learning algorithm used for the single-site case, but fed with this expanded dataset, should achieve a better performance than the one fed with data from a single site.

1.2 Contributions

This dissertation makes the following specific contributions:

1. It empirically shows that one can learn a model using features related to the functional connectivity of the brain, that can distinguish between patients with schizophrenia versus healthy controls with an accuracy above chance level, and up to an average of 70% in single site classification. The number of features is effectively reduced by using domain specific information such as predefined regions of interest and information about the network topology of the brain. This step decreases the

computational cost of the learning process and increases the prediction accuracy relative to naively using all the available data.

2. It highlights important differences between association studies and predictions studies, and it empirically shows that batch effects are a serious problem that affect the latter (and most likely the former too). Naively incorporating fMRI data obtained from different scanning sites into the training set decreases the accuracy relative to using only instances extracted from a single site.
3. It proposes the use of a neural network, in a configuration similar to an autoencoder, to solve the batch effects when fMRI scans from the same participants are obtained at different scanning sites. It empirically shows the effectiveness of this approach by learning a model that can identify subjects with 100% accuracy, but identifies the scanning site at chance level (which is a strong indication that the batch effects were reduced)
4. It shows that simple methods, like z-score normalization can correct for translation and scaling in the data. Similarly, whitening can correct for translations and rotations. It empirically shows that these methodologies are not enough for solving our batch effects problem, in the sense that after applying the corrections, the classification accuracy when using multi-site data in the training process did not increase. This is a strong indication that batch effects go beyond these simple linear transformations.
5. It offers empirical evidence that the direct implementations of BECCA, stacked autoencoders, and bishifting autoencoders, which have been successfully used in other domain adaptation tasks, are not enough for solving the batch effects under our test scenarios related to the diagnosis of schizophrenia using fMRI data. This results suggest that more research effort is needed to understand the nature of batch effects and how to correct them.

Chapter 2

Background

2.1 Basics of fMRI

fMRI is an imaging technique that measures the changes in the oxygenation level of blood in the brain, a phenomenon known as the blood oxygen level dependent effect (BOLD effect) [56]. Since there is a coupling between the neuronal activity and the local control of blood flow and oxygenation in the brain (a process known as neurovascular coupling), it is possible to estimate the brain activity by measuring these changes [25]. An oversimplified description of the fMRI rationale is as follows:

1. Active regions in the brain require oxygen, which is delivered by the blood. This process changes the concentration of oxyhaemoglobin (oxygenated blood) and deoxyhaemoglobin (deoxygenated blood) in the blood vessels near the region of activity.
2. Oxyhaemoglobin and deoxyhaemoglobin have different magnetic properties, so the changes in their relative concentration in the blood can be detected by an MRI scan.
3. These changes are used as an estimation of the level of activity at certain location in the brain.

A typical fMRI scan obtains a 3-D volume of the brain every 2-3 seconds. Depending on the nature of the study, the collection of data from a single

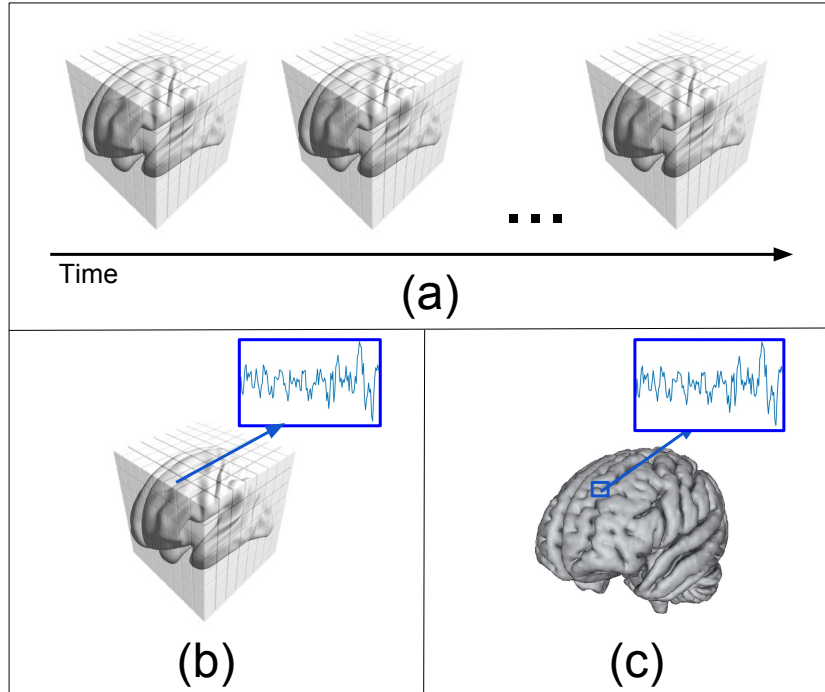


Figure 2.1: fMRI from a computational point of view. (a) Compute a 3-D matrix (corresponding to the BOLD signal at each voxel), every 2-3 seconds. (b) Every voxel has an associated time series. (c) Every voxel is associated with a particular location of the brain. Hence, this is an estimation of the level of activity at a particular location over time.

subject might take 5 – 30 minutes, which can be done in a single long run, or can be partitioned in few shorter runs [3].

From a computational point of view, every 3-D volume can be seen as a 3-D matrix, where every voxel represents a particular location in the brain, whose intensity value is related to the relative concentration of oxygenated blood at a particular moment. Since a volume is obtained every few seconds, the level of activity at any area of the brain is represented as the time series of its corresponding voxel (see Fig 2.1). As a typical fMRI 3-D volume contains approximately 500,000 voxels, an fMRI session over 8 minutes (involving 150 volumes), would have nearly 75,000,000 values for every single subject scanned. This amount of data poses a challenge for the analysis, since the number of instances (people scanned) is usually only a few hundreds.

2.1.1 Preprocessing pipeline

The analysis of fMRI data assumes that the same voxel ($[x,y,z]$ index in each 3-D scan) represents the same location in the brain for all the participants in an experiment. Unfortunately, raw fMRI data does not meet that criteria. First, there is variability in the brain size of different people, so the same voxel coordinates might represent two different brain regions in two people. Even with the scan from a single subject, there are different sources of noise that can affect the experiment, including: head motion, fluctuations in the electromagnetic field created by the scanner (thermal noise), or heart beating and breathing (physiological noise) [21]. At the same time, the expected fluctuations in the time signal of a particular region of the brain that is active is very small, usually only $\pm 5\%$, relative to its mean intensity value [25]. The combination of these two factors can severely damage the results of the analysis of fMRI data. In an effort to reduce their effect, a preprocessing step is applied before starting the analysis. The preprocessing of fMRI data is a well studied topic, which has been described in detail elsewhere [3]. Many tools, tutorials, books and software are available for performing this step. In our analysis we used the freely available FSL¹ package [28] and implemented the following preprocessing steps independently to every fMRI scan:

1. Motion Correction: When a subject moves, brain regions will move to different spatial locations in the scanner. In order to correct for this, we apply an affine transformation, which has 12 degrees of freedom (allows for translation, rotation, scaling and shearing along the three dimensions).
2. Coregistration: At the beginning of each fMRI experiment, usually a high resolution structural MRI image is acquired for every subject. This step involves aligning the fMRI brain volumes with their corresponding structural MRI image.
3. Spatial smoothing: In order to increase the signal to noise ratio, every

¹<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL>

volume of the fMRI data is smoothed by convolving it with a Gaussian kernel. This operation has the effect of substituting the intensity value of the voxel v_i in a volume by a weighted average of the intensity values of all the voxels within a mask centered at v_i . This preprocessing step is applied in an effort to decrease the effect of a fluctuating magnetic field in the MRI scan.

4. MNI normalization: In an effort to make every voxel map the same location of the brain across all the individuals, all the fMRI volumes are mapped to the ICBM152 template [15].

2.2 Functional connectivity

Functional connectivity is used to analyze the degree of synchrony among regions of the brain that are anatomically distant [49], which is defined as a statistical association between the time series associated with different parts of the brain [51].

Two types of methods are commonly employed for determining the functional connectivity of the brain: the ones based on regions of interest, and the ones based in independent component analysis (ICA). The first one extracts the time courses from each of a set of predefined regions of the brain, then determines the pairwise statistical association between regions by correlation [71], coherence [58], or conditional independence [51]. The output of this process is a graph where the regions of interest are represented by the nodes, and their statistical association is represented by the weights of the edges. ICA, on the other side, makes no assumption about the regions of the brain useful for the analysis and uses the time series from all the voxels to decompose the time signals into a set of statistically independent components. The output of this process is a set of spatial map that shows which parts of the brain are associated with the independent components, as well as their degree of activation. Note that this is a soft labeling, which means that a particular voxel can be associated with more than one statistical map (or with none of them).

The current belief in the neuroscience community is that the connectivity in the brain determines its computational properties [39]. There is also increasing evidence that differences in the functional connectivity are associated with alterations in cognitive and behavioral functions, suggesting that they also play a role in neurological and psychiatric disorders [4]. If the functional connectivity is *similar* among people with the same condition (*e.g.*, schizophrenia) and *different* from others with a different condition (*e.g.*, healthy controls), then we could use it to create a classifier that assigns new people to one of the groups, based on their functional connectivity. A detailed description of how we computed the functional connectivity for the experiments performed in this dissertation is given in Section 3.3

2.3 Association vs prediction studies

Traditionally, fMRI studies are used for analyzing the cognitive function of the brain, and its disruption in the presence of a mental disorder or mental illness. These studies typically seek explanatory models, in which a set of factors X are assumed to cause an underlying effect, which is measured by variable Y [54]. Their objective is to find these underlying factors (biomarkers) among all the features available in the data. For the specific case of fMRI data, the biomarker might take the form of regions of the brain that co-vary with a particular stimuli and that are statistically significantly different among the groups under study (*e.g.*, people with schizophrenia vs healthy controls) [7]. These type of studies are known as “association studies” [33].

Standard association studies require at least two stages: individual level analysis and group level analysis [3]. In the former one, the fMRI scan of every subject is analyzed individually, one voxel (or region of interest) at the time. The objective is to identify which voxels are associated with a time series that relates to the task being performed. The time series associated with every voxel (or region of interest) is modeled as $Y = X\beta + \epsilon$, where Y is the response associated with the explanatory variable X , based on their respective coefficients β ; here ϵ is the error term. Then a statistical test is

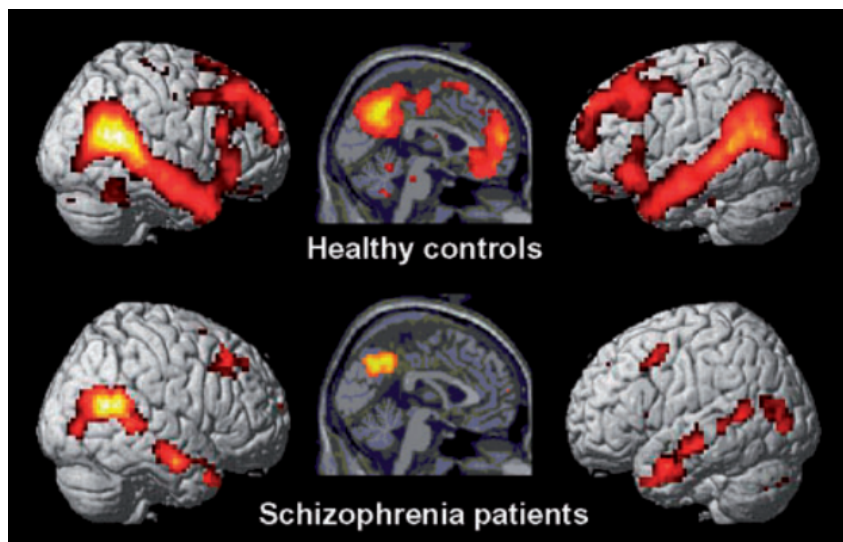


Figure 2.2: Regions of the brain with statistically significant differences between schizophrenia patients and healthy controls under the experiment of Walter et al. [63] (reproduced by permission of Oxford University Press)

applied to β (e.g., a t-test) to produce a statistical map, with a single value per voxel that represents its level of activation. Readers interested into the details of this process are referred to the work of Friston [20] and Ashby [3]. In the second phase, the individual statistical maps are combined into a single group map (e.g., one map for healthy people, and one map for people with schizophrenia) with the objective of finding statistically significant differences between groups.

A conclusion of an association study typically takes the form *people from group A presented more/less activation in regions X, Y, and Z compared with group B*, and then visually represents the regions in the brain that present differences. As a concrete example, Walter et al. [63] analyzed differences in the brain activation patterns between people with schizophrenia and healthy controls under activities with different degree of social interaction. They report that “...[patients with schizophrenia] showed less activation in three regions typically activated in theory of mind tasks, i.e., paracingulate cortex and bilateral temporo-parietal junctions...” and represent those differences in the image shown in Figure 2.2.

One of the most important characteristics of these studies is its explanatory

power: the goal is to build an *interpretable model* that brings insight about the nature of the problem under study, however, one caveat of explanatory models is that the research hypotheses are given in terms of theoretical constructs, which are descriptions of a phenomenon of interest [13], such as *functional connectivity*, rather than in terms of measurable variables [54]. Therefore, these models are difficult to evaluate numerically, so their reliability is typically measured by consistency (similar studies reporting similar results) and closeness to the known theoretical models of the problem under study.

Prediction studies, on the other side, focus on learning patterns using historical data with the objective of making predictions on new or future observations [41]. A typical prediction study follows the block diagram shown in Figure 1.1. Unlike association studies, the results are presented at individual level, rather than group level. Also, it is relatively easy to evaluate the performance of the methods used for learning the classifier, since there are well-defined metrics, such as accuracy, designed for this purpose.

Prediction studies are not as common as association studies in many fields, including neuroscience, because they might not help to produce an underlying theory [54]. However, there is a growing interest in these predictive models due to their potential to be applied in clinical settings [2]. As its name suggests, predictive studies focus on predictive power so, instead of finding a subset of all the features x that explain a phenomenon y using an univariate approach, they use a multivariate approach to find combinations of features that give the most probable outcome for a specific instance x , $\arg \max_y P(y|x)$. The caveat of these studies is that, since they might use hundreds or thousands of features, they are difficult to interpret. Besides, even if the number of features is low, the model learned might not be directly interpretable [23].

In many fields, models that focus on explanatory power are often assumed to have also a predictive power [54], but this is not necessarily the case. As a simple example, consider the following simulated scenario: A study measures the level of activation in the motor cortex and the prefrontal cortex in people with schizophrenia and healthy controls. The distributions of the features

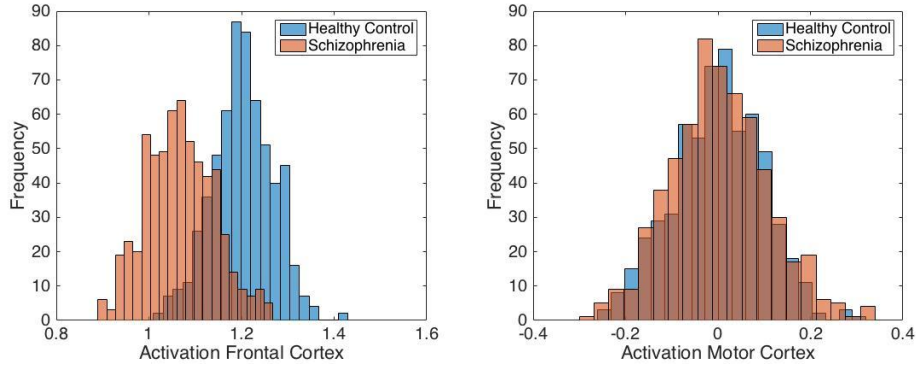


Figure 2.3: Distribution of two hypothetical features across two groups of interest. The difference in the means of the activation in the frontal cortex (left) in both groups is statistically significant ($p < .01$), but that is not the case for the activation in the motor cortex (right) ($p = .63$)

in both groups is shown in Figure 2.3. After running a t-test for finding statistically significant differences between both groups (association study), one could erroneously conclude that the activation in the motor cortex plays no role for distinguishing people with schizophrenia from healthy controls. Also, we observe that there is an overlap between the distribution of the activation of the frontal cortex in both groups.

On the other side, a prediction study would consider combinations of both features, seeking a separation surface that allows classification of new instances. Figure 2.4 shows the result of the prediction study where, because of the correlation structure of both variables, the activation of the motor cortex plays an important role in defining the decision boundary represented by the dotted line. In this particular case, if we encode the people with schizophrenia as $y_{scz} = -1$, healthy controls as $y_{hc} = 1$, activation in the frontal cortex as x_1 , and activation in the motor cortex as x_2 ; then the decision boundary is defined by the line $f(x_1, x_2) = 26.48x_1 - 14.39x_2 - 30.06 = 0$, and a new instance, $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$, would be classified as $y_i = \text{sign}(f(x_1^{(i)}, x_2^{(i)}))$. Note also that the overlap between the distribution decreases when both features are used together, in comparison with using only the features that are statistically significant in an association study; however, the interpretation becomes problematic since it is difficult to explain what a linear combination of the ac-

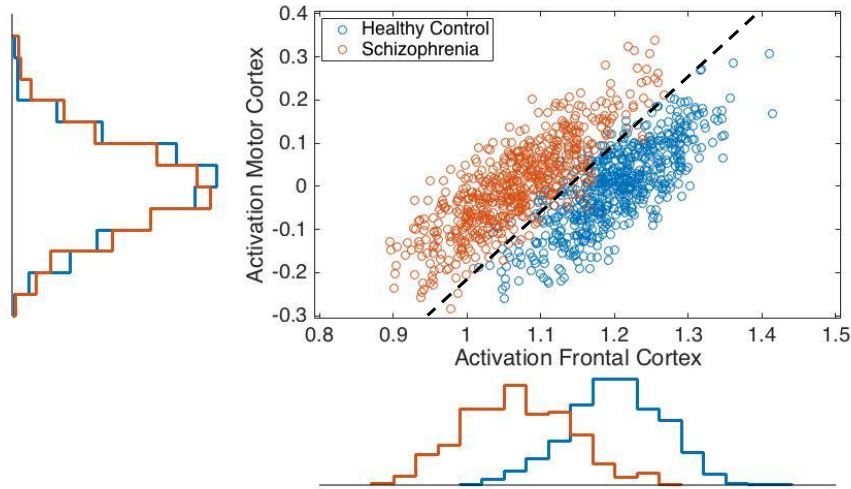


Figure 2.4: Distribution of two hypothetical features across two groups of interest. Recall from Figure 2.3 that the difference in the means of the activation in the frontal cortex in both groups is statistically significant ($p < .01$), but that is not the case for the activation in the motor cortex ($p = .63$)

tivation in two regions of the brain means. (This problem only increases with more dimensions: What does a linear combination of 500,000 voxels mean? What if we use basis functions to project the data into a different feature space?).

This simple example shows that features whose difference between groups is not statistically significant can still have predictive power and simply ignoring them might have a negative impact on the prediction accuracy. The opposite is also true: two features can have statistically significant differences between two groups without having any predicting power. This situation is illustrated in Figure 2.5.

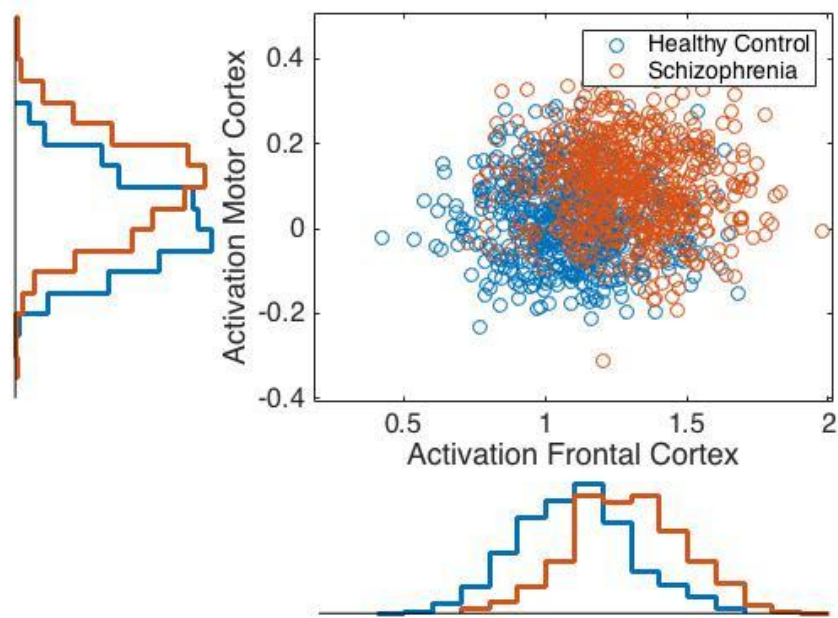


Figure 2.5: Distribution of two hypothetical features in two groups of interest. Even when both features have statistically significant differences between groups ($p < 0.01$), their predictive power is very limited.

Chapter 3

fMRI analysis from the machine learning perspective

In a classification problem using supervised machine learning, the objective is to learn a mapping from inputs x to outputs y , where $x \in \mathbb{R}^p$ is a p -dimensional vector containing the values of a set of features, and $y \in \{1, 2, \dots, C\}$, with C being the number of classes, indicates the class to which x belongs. This typically assumes that there exist an unknown function $y = f(x)$ that makes this mapping (or at least that $f(x)$ is a good approximation), and the goal is to apply a learning algorithm, $L(\cdot)$ on a labeled training set with n instances $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ to get an estimate of the function, $\hat{f} = L(D)$. It is then possible to make predictions on new instances $\hat{y}_{new} = \hat{f}(x_{new})$ [41]. This process is depicted on Figure 3.1, which is a generalization of Figure 1.1.

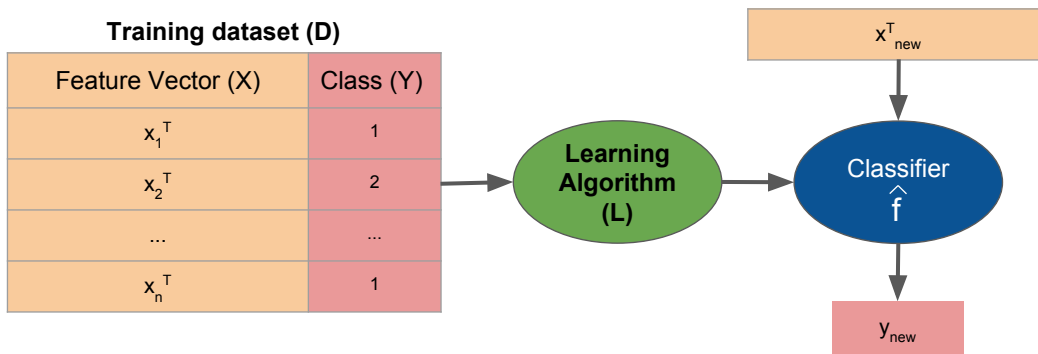


Figure 3.1: Machine learning approach for classification problems

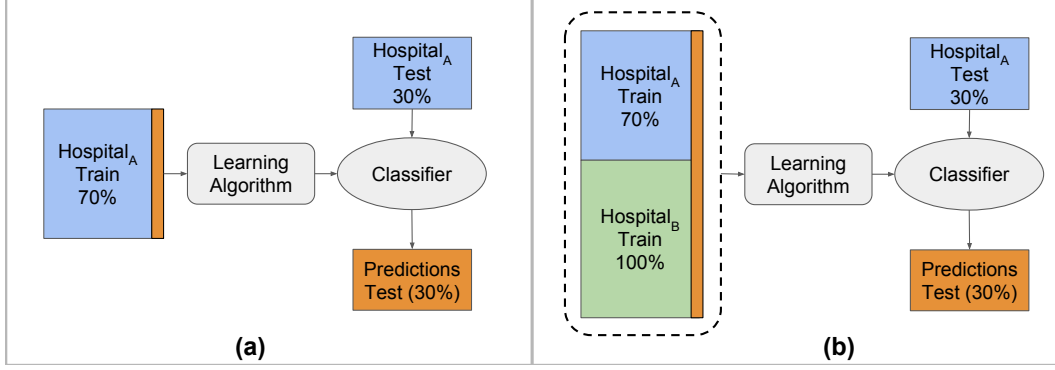


Figure 3.2: The two scenarios explored in this dissertations: a) Training set and test set come from the same scanning site, b) Add data from an external site to the training set.

3.1 Task description

Given a training set of labeled fMRI scans obtained from people with schizophrenia (SCZ) and demographically matched healthy controls (HC), create a classifier that makes predictions in a (disjoint) test set of fMRI scans with an accuracy above chance level when either:

1. The fMRI scans in the training set and test set where acquired in the same scanning site; or
2. fMRI scans from a different scanning site are added to the training set, but the fMRI scans in the test set are still from a single site. Since more data is available, the accuracy in scenario 2 should be higher than the accuracy obtained in scenario 1.

Both scenarios are depicted in Figure 3.2, and their performance will be evaluated computing the accuracy of the learned function $\hat{f}(\cdot)$ over a labeled test dataset D

$$\text{Acc}_{\hat{f}(\cdot)}(D) = \frac{1}{|D|} \sum_{[x_i, y_i] \in D} I(y_i = \hat{f}(x_i)) \quad (3.1)$$

where x_i represents the fMRI scan of the i th subject, $y_i \in \{SCZ, HC\}$ is his/her true class, $\hat{f}(x_i)$ is the predicted class, and $I(\cdot)$ is the indicator function.

3.2 Dataset

For our experiments, we used the FBIRN phase II dataset, which is a multisite study developed by the Function Biomedical Informatics Research Network (FBIRN). It contains data from 87 individuals with schizophrenia (59 males) and 85 healthy controls (70 males), both in the age range 18 – 70. Keator et al. provides a complete description of the study [31].

We used the data corresponding to the Auditory Oddball task, in which every participant is presented with a continuous sequence of two types of discrete stimuli: *Standards* and *Targets*. Each participant completes 4 experimental runs of 280 s each. Every run begins with a block of silence (15 s), followed by the presentation of Standard Tones (1000 Hz) that last 100 ms. Every 6 – 15 seconds, the Target tone (1200 Hz), of duration of 100 ms, is presented. While listening to the tones, the subjects are looking at a constant fixation cross in the middle of a screen, and they are instructed to press a button when they hear a Target tone.

After preprocessing the data, we eliminated the subjects who presented head movement greater than the size of one voxel at any point in time in any of the axis, a rotation displacement greater than 0.06 radians, or that did not pass a visual quality control assessment. The original released data contains scans extracted from 6 different scanning sites; however, we only used 4 of them. One of the sites was discarded because it lacked T1-weighted images, which were required as part of our preprocessing pipeline. The second site discarded contained only 6 subjects (5 with schizophrenia) after the quality control assessment, so it was not suitable for our analysis. Table 3.1 shows the number of instances that we used in the experiments.

3.3 Feature extraction

There are two common methods for computing the functional connectivity of the brain: Region of interest (ROI) based, and Independent Component Analysis (ICA). Literature in neuroscience that compares both methodologies

Table 3.1: Number of participants in the dataset used for the experiments. Every participant was scanned 4 times, the total number of scans (instances) is 4 times the number of participants.

| Original description | Site 3 | Site 9 | Site 10 | Site 18 |
|----------------------|--------|--------|---------|---------|
| Alias | Site 1 | Site 2 | Site 3 | Site 4 |
| # Healthy controls | 10 | 10 | 10 | 11 |
| # Schizophrenia | 11 | 12 | 13 | 12 |
| # Instances (m) | 84 | 88 | 92 | 92 |

report similar results when using one or the other [49]. For our experiments we decided to use a ROI approach for the following reasons:

- As shown in Figure 3.3, nearby voxels contain redundant information (they have very similar time series), so exploring all the voxels increases the computational cost without adding too much new information.
- ICA analyzes all the available data, which involves nearly 500,000 voxels per scan. On the other side, the number of regions of interest is usually just a few hundreds. Hence, a ROI approach dramatically reduces the number of features, which becomes a critical issue due to the limited number of instances in the training set.
- There exists extensive literature about the functional connectivity of the brain in schizophrenia. The information obtained from the association studies has found that the connectivity in some parts of the brain, like the prefrontal cortex, might be decreased in people with schizophrenia [7]. Since some regions of interest are deliberately located in these points of interest, we can take advantage of this domain-specific knowledge.

3.3.1 Parcellation of the brain

We used the regions of interest defined by Power et al. [45]. They parcellated the brain into 264 functional areas, with every area belonging to exactly one of 14 functional networks, and provided the coordinates (in the MNI space)

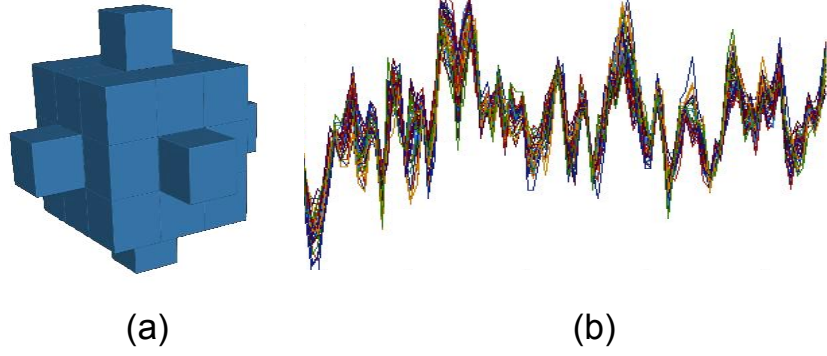


Figure 3.3: (a) Mask used for extracting the time series of every region of interest. (b) Zero-mean time signals of all the voxel within the mask in a region of interest.

of the center of every functional area. Every region was modeled as a 10 mm diameter spheres.

For determining the time series associated with every ROI, r , we approximated the 10 mm sphere with the mask of 33 voxels shown in Figure 3.3 centered at the coordinates of every region (every voxel is a $2 \times 2 \times 2$ mm³ cube). We then took the average of the zero-mean time signals of all the voxels within the mask independently for every subject s :

$$v(r, s) = \frac{1}{|N_r(s)|} \sum_{x_i \in N_r(s)} (x_i - \mu_i \mathbf{1}) \quad (3.2)$$

where x_i is the vector containing the time series of a single voxel, $\mu_i = \frac{1}{p} \sum_{j=1}^p x_i^{(j)}$ is the empirical average of all the p entries of the vector x_i , $|N_r(s)| = 33$ is the cardinality of the set of vectors $N_r(s)$, which represents the voxels within the mask centered at the coordinates of the region r for the subject s , and $\mathbf{1}$ is a vector of the same dimensions as x_i whose entries are all 1. The output of this process is a matrix $X \in \mathbb{R}^{p \times 264}$ for every subject that contains the time series associated to the 264 regions of interest.

3.3.2 Feature matrix

The functional connectivity of every subject can be estimated by computing the Pearson's correlation coefficient between the time series of every possible pair of regions of interest [47, 71]. The correlation coefficient between two

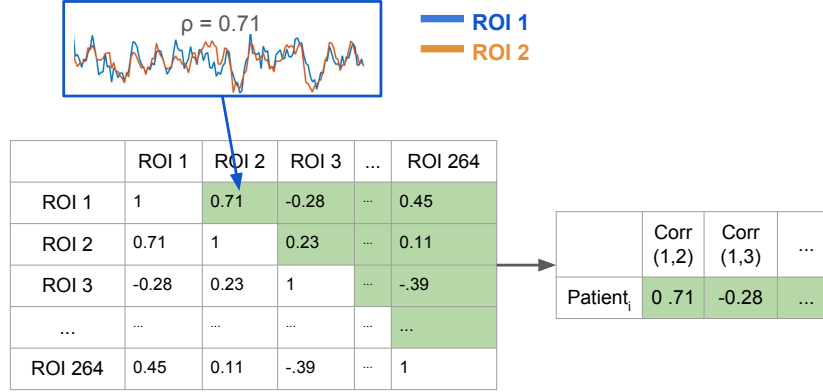


Figure 3.4: Every entry in the matrix on the left is the Pearson’s correlation coefficient between two regions of interest. The correlation ($\rho = 0.71$) between the time series of ROI_1 and ROI_2 is shown as an example. Then, the upper triangular matrix (green) can be concatenated into a single vector, of length $l = \binom{264}{2}$.

random variables, X and Y , is in the range $[-1, 1]$, and is defined as:

$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\text{Var}(X) \text{Var}(Y)} \quad (3.3)$$

where μ_X, μ_Y represent the mean of the random variables X and Y respectively, and $\text{Var}(X), \text{Var}(Y)$ represent their variances.

It is possible to represent the pairwise correlation of a single subject as a symmetric matrix R , where the entry $R_{i,j}$ contains the correlation coefficient between the regions of interest i and j . Note that all the elements in the diagonal of R are equal to 1. Therefore, we can characterize a single patient by extending the upper (or lower) triangular part of R into a feature vector. This process is exemplified on Figure 3.4. After repeating the same procedure to every fMRI scan, and creating a vector storing the label of every scan, we will have the labeled dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i , of length $l = \binom{264}{2}$, is the feature vector obtained from the i th scan, and $y_i \in \{\text{SCZ}, \text{HC}\}$ is its corresponding label.

3.4 Support Vector Machine (SVM)

Support Vector Machine is a learner that creates an optimal separating hyperplane for two classes that are linearly separable. It finds the hyperplane that

maximizes the margin between the training instances; see the linear boundary shown in the Figure 3.5 [22]. For the non-separable case, it is possible to add *slack variables*, ξ , that quantify the distance from points inside the margin, or in the wrong side of the boundary to the hyperplane. The decision boundary is then defined by the equation $f(x) = \omega^T x + b = 0$, where the weight vector ω is found by optimizing the following function (primal form of SVM):

$$\begin{aligned} \min_{\omega} \quad & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \\ & y_i(\omega^T x_i + b) \geq 1 - \xi_i, \quad (i = 1, 2, \dots, m) \\ & \xi_i \geq 0, \quad (i = 1, 2, \dots, m) \end{aligned} \tag{3.4}$$

The parameter C controls the width of the margin. It is possible to transform this constrained optimization problem into an unconstrained one by using Lagrangian multipliers and, after some mathematical manipulation along with the Karush-Kuhn-Tucker conditions, express its equivalent dual problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & \\ & 0 \leq \alpha_i \leq C, \quad (i = 1, 2, \dots, m) \\ & \sum_{i=1}^m \alpha_i y_i = 0 \end{aligned} \tag{3.5}$$

It is also possible create non-linear boundaries by transforming the data, using basis expansions, into a new feature space $\phi(x)$, and then compute a linear boundary on it. Note that if we substitute the basis functions $\phi(x_i), \phi(x_j)$ for the vectors x_i, x_j in Equation 3.5, the maximization problem is still in terms of the dot product of the basis function. This allows the use of the kernels (represented as $K(x_i, x_j)$) to efficiently create the linear boundary in the new feature space without having to explicitly compute $\phi(x)$.

The formulation of the dual problem present also an advantage when the number of features greatly exceeds the number of training instances (which is

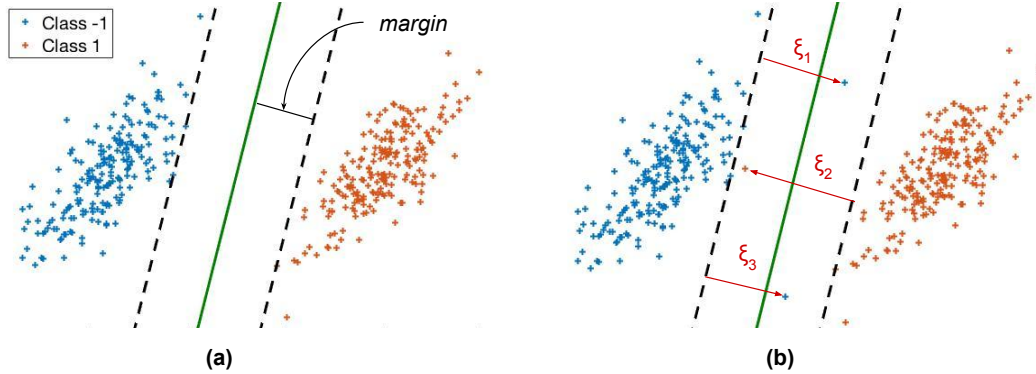


Figure 3.5: a) SVM creates a decision boundary that maximizes the margin between the decision boundary and the closest points to it. b) For the non-linearly separable case the slack variables ξ quantify the distance between the decision boundary and the points inside the margin, or in the wrong side of the decision surface.

the case with fMRI data). Note that in the primal formulation the variable to optimize is w , which means that it will estimate one coefficient per feature. The dual form, on the other side, optimizes α , so it will estimate a coefficient for every training instance.

Assuming a binary classification problem, where the label $y \in \{-1, 1\}$, the label of a new instance is computed by:

$$y_{new} = \text{sign} \left(b + \sum_i \alpha_i y_i K(x_{new}, x_i) \right) \quad (3.6)$$

with $b = y_k - \sum_i \alpha_i y_i K(x_k, x_i)$ for any k where $C > \alpha_k > 0$.

3.5 Learning algorithm and accuracy estimation

For learning a classifier that distinguishes between people with schizophrenia and healthy controls, we use a SVM with linear kernel¹ using the SVM library SVMLIB² [8]. For determining the best parameters for the classi-

¹We also experimented with the RBF kernel; however its high variance prevented it from generalizing to new data for this task. Therefore, we only report the results with the linear kernel.

²Software freely available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Algorithm 1 Learning algorithm

Input: D_{Train}, C **Output:** $Model$

```
1: procedure FINDBESTSVMMODEL
2:   Divide  $D_{Train}$  into 5 disjoint subsets  $\{D_1, D_2, \dots, D_5\}$ ; set  $D_{-i} = D - D_i$ 
3:   for  $c$  in  $C$  do
4:     for  $k = 1:5$  do
5:       Model  $\leftarrow$  trainSVM( $D_{-k}, c$ )
6:       Predictions  $\leftarrow$  predictSVM( $D_k, Model$ )
7:       TempAccuracy( $k$ )  $\leftarrow$  getAccuracy(Predictions, GroundTruth)
8:     end for
9:     Accuracy( $c$ )  $\leftarrow$  Average(TempAccuracy)
10:  end for
11:   $C_{best} \leftarrow \arg \max_c (Accuracy(c))$ 
12:  Model = trainSVM( $D, C_{best}$ )
13:  return Model
14: end procedure
```

fier, we used 5-fold cross validation and a grid search over the parameters $C = [2^{-5}, 2^{-3}, \dots, 2^{15}]$ as suggested by the authors of the library. The learning procedure is described in Algorithm 1. The subroutine *trainSVM* on line 5 refers to solving the problem described in Eq. 3.5, *predictSVM* refers to solving Eq. 3.6, while *getAccuracy* refers to solving Eq. 3.1. Also, when dividing the training set into 5 disjoint subsets (Line 2), we ensured that the proportion of elements of the different classes were essentially the same in every subset.

For estimating the accuracy of the learning algorithm, we followed to procedure described in Algorithm 2. Our objective was to estimate a distribution of accuracies for different partitions of a dataset, D , into a training set D_T and a hold out set D_H . The number of times that the experiment is repeated is set by the user in the variable *numExp*. For the purposes of this dissertation, we set *numExp* = 30. Note also that the output of Algorithm 2 is a vector containing the accuracy of every partition, instead of a single-point estimate of the accuracy. From this vector it is possible to estimate the mean expected accuracy and graph a histogram to get an estimate of the shape of the accuracy distribution.

As depicted in Figure 3.2, we partitioned the data into 70% for training set and 30% for the hold-out set in the case of single site classification. For the multisite case, we added 100% of the data obtained from a different scanning site to the training set.

Algorithm 2 Expected accuracy of the learning algorithm

Input: D , numExp

Output: Accuracy

```
1: procedure GETEXPECTEDACCURACY
2:   for  $i = 1:\text{numExp}$  do
3:     Divide  $D$  into 2 disjoint subsets  $D_T$  and  $D_H$ 
4:     Model  $\leftarrow$  findBestSVMModel( $D_T$ , C)
5:     Predictions  $\leftarrow$  predictSVM( $D_H$ , Model)
6:     Accuracy( $i$ )  $\leftarrow$  getAccuracy(Predictions, GroundTruth)
7:   end for
8:   return Accuracy
9: end procedure
```

Chapter 4

Classification results

4.1 Single site

For each single site classification experiment, we used data from one scanning site at a time for both training and testing, as depicted in Figure 3.2 (left). We computed the accuracy as described in the Algorithm 2. The data from every scanning site is represented as a matrix $D \in \mathbb{R}^{m \times p}$, with m being the number of scans and p being the number of features¹. Table 4.1 summarizes the results when using as features the pairwise correlation of the 264 ROI ($p = \binom{264}{2} = 34,716$ features) described in Section 3.3.2. Every participant in the dataset was scanned 4 times, and every scan counts as an individual instance; however, we split the data into a training set D_T and hold-out set D_H at a subject level – *i.e.*, all the 4 scans of a single subjects were part either of the training set or the hold-out set.

Note from Table 4.1 that only the data extracted from sites 2 and 4 have an accuracy that is statistically significant greater than the baseline, which is defined as the accuracy obtained by classifying all the instances as the majority class. One possible explanation for this result is that, even when the number of features was greatly reduced by the ROI analysis, the number of features is more than 350 times the number of instances, so the learning algorithm has problems identifying useful patterns.

It is possible to reduce the number of features even more by using domain-

¹We also performed the experiment using coherence, partial coherence, and the Fast Fourier Transform as features, but pairwise correlation achieved the best results.

Table 4.1: Single site dataset size and results using the 264 regions of interest. Statistically significant differences (using t-test, $p < 0.05$) between the mean accuracy and the baseline are marked in bold.

| | Site 1 | Site 2 | Site 3 | Site 4 |
|---------------------|--------|--------------|--------|---------------|
| # Participants | 21 | 22 | 23 | 23 |
| # Instances (m) | 84 | 88 | 92 | 92 |
| Baseline | 57.14% | 57.14% | 57.14% | 50 % |
| Accuracy | 61.67% | 67.5% | 53.69% | 62.91% |
| Std | 17.0% | 9.2% | 10.4% | 11.4% |

specific knowledge about schizophrenia. For example, previous studies have found that there are important differences in the functional connectivity of the prefrontal cortex [7]. By limiting the nodes to those belonging to the Fronto-Parietal network and the auditory network (since the participants are performing an auditory task), the number of ROI analyzed decreases to only 38, instead of the original 264. This reduces the actual number of features from $\sim 35,000$ to only 703. Table 4.2 shows that using this new approach the accuracy improved in all cases but site 3. Note this both increases the mean accuracy (among the 30 experiments) and also decreases the standard deviation of the results. Figure 4.1 shows the accuracy distribution estimated after applying Algorithm 2. Many prediction fMRI studies only report a single-point estimate of the accuracy. As shown in the figure, this is not a good estimator, since it can drastically change depending on the test set used. For example, an incorrectly computed single-point estimate for the accuracy in the data from site 1 could be $> 80\%$ when the distribution shows that, on average, the accuracy is only 64.33%.

A further analysis of the results shows that in all sites, but site 3, adding more data to the training set has a positive impact on the accuracy of the classifier produced by the learning algorithm; see Figure 4.2. The x-axis shows the percentage of the data that was used for creating a classifier, the rest was used as a hold-out set to estimate the accuracy (y-axis). These graphs suggest that adding more data in the training set should increase the expected accuracy of the learning algorithm. The assumption is that the data obtained

Table 4.2: Single site dataset size and results using the 38 regions of interest corresponding to the fronto-parietal network and auditory network. Statistically significant differences between the mean accuracy and the baseline are marked in bold.

| | Site 1 | Site 2 | Site 3 | Site 4 |
|---------------------|---------------|---------------|--------|---------------|
| # Participants | 21 | 22 | 23 | 23 |
| # Instances (m) | 84 | 88 | 92 | 92 |
| Baseline | 57.14% | 57.14% | 57.14% | 50 % |
| Accuracy | 63.33% | 70.35% | 55.0% | 70.20% |
| Std | 14.4% | 7.3% | 7.3% | 9.1% |

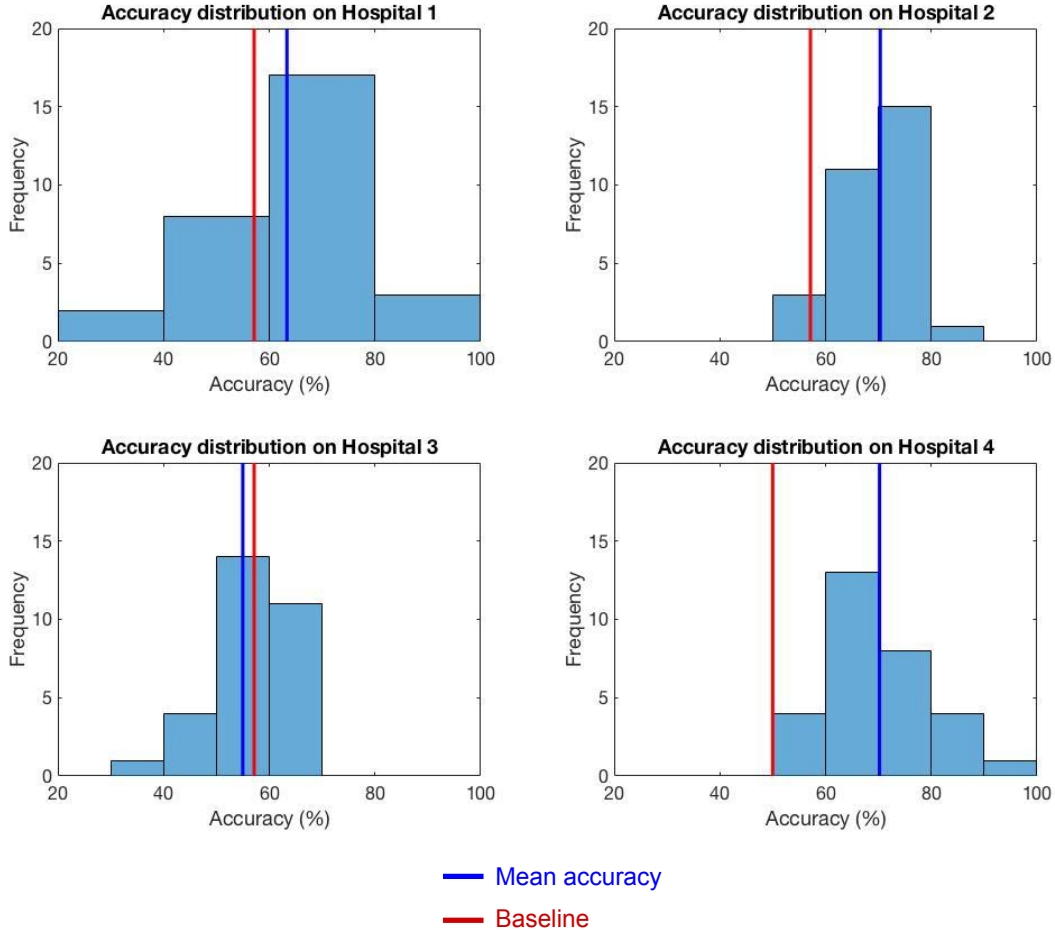


Figure 4.1: Distribution of the accuracy on every scanning site after 30 experiments

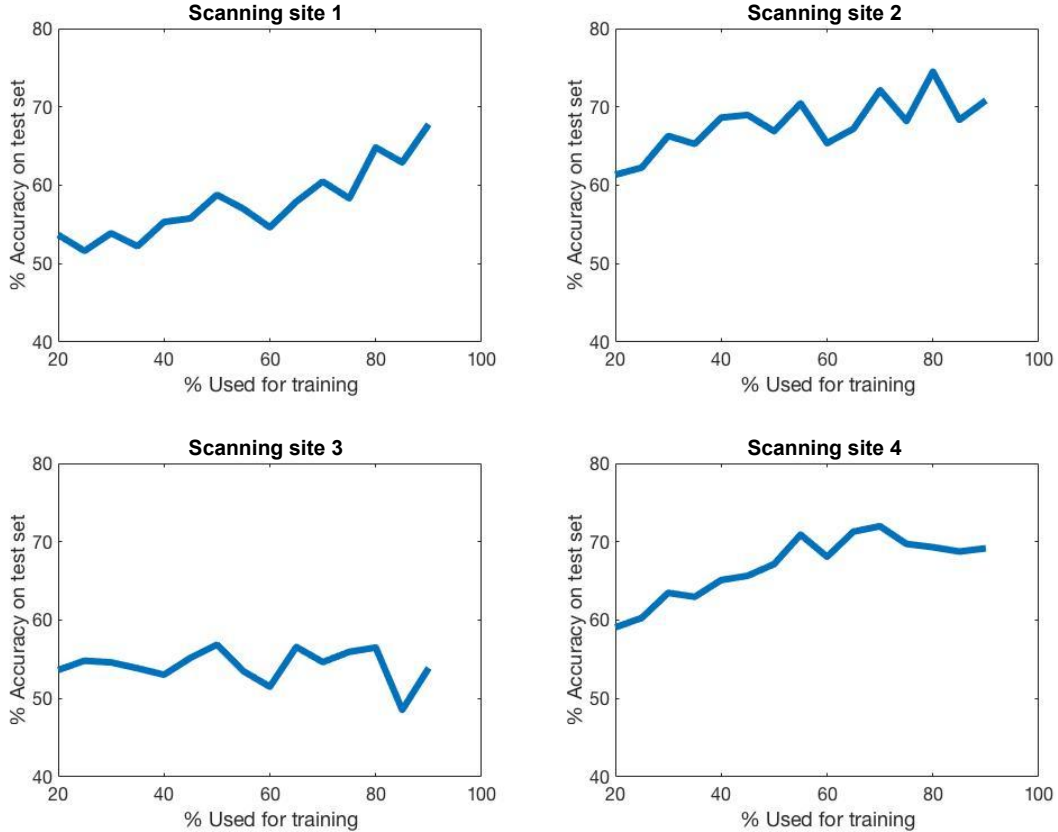


Figure 4.2: Influence of the size of the training set on the performance in a hold out set.

in the different scanning sites come from the same distribution. Therefore, we should be able to merge data from different sites in the training site in order to improve the performance of the classifier. Note that this is not the case for the scanning site 3, whose accuracy is at chance level regardless of the number of instances including in the training phase. These results suggest that the signal present in the data extracted from this scanning site is very weak, preventing the learning algorithm from detecting useful patterns for classification.

4.2 Multiple sites

The experiments using data extracted from multiple scanning sites follow the scheme depicted in Figure 3.2(b). Figure 4.3 shows the result of using data from two sites for training, while testing in data from only one site (we used only the 38 ROI for the multiple sites experiments.) The scanning site used

| | | Additional Train Set | | | |
|----------|------------|----------------------|----------------|----------------|----------------|
| | | Hospital 1 | Hospital 2 | Hospital 3 | Hospital 4 |
| Test set | Hospital 1 | 63.33 % | 64.45 % | 62.85 % | 64.64 % |
| | Hospital 2 | 72.73 % | 70.35 % | 67.14 % | 64.76 % |
| | Hospital 3 | 55.11 % | 65.35 % | 55.00 % | 56.78 % |
| | Hospital 4 | 76.35 % | 58.54 % | 63.02 % | 70.20 % |

Figure 4.3: Average accuracy after merging datasets from different scanning sites.

for test purposes is represented in the rows, while the scanning site added for training is represented in the columns. Ideally, the values in the diagonal should be the lowest in every row, since they include data from a single site; however, this is not the case. In most of the experiments, the changes in the accuracy were not statistically significant, even when Figure 4.2 suggested that more data should increase the accuracy. Only in 2 out of 12 cases did the accuracy increase significantly, but in 3 cases it significantly decreased. Besides, the behavior of the accuracy is not *symmetrical*. For example, adding data from the scanning site 1 to the data from scanning site 4 improved the accuracy over just using site 4; however, the “opposite” is not true: the accuracy of site 1 did not improve when adding data from site 4. Note that since we are using all the available data from a new site, the size of the augmented training set is more than double than the one in the experiments reported in Section 4.1.

Adding even more data to the training set does not correct this situation. If instead of adding data from one additional site, we could add the data from *all* the other sites into the training set; this leads to the accuracies shown on Figure 4.4. An interesting effect occurred in this case. Note that the accuracy of the scanning sites 1 and 3, which had a relatively low accuracy on the single site experiment, dramatically increased their average accuracy to 73.09% and 68.45%; however, the accuracy of the scanning sites 2 and 4 presented an important decrease in their accuracies relative to the single site experiment (although for the case of scanning site 4, the difference is not statistically

| | | Test Set | | | |
|-----------|-----------------|----------------|---------------|----------------|------------|
| | | Hospital 1 | Hospital 2 | Hospital 3 | Hospital 4 |
| Train Set | Single hospital | 63.33 % | 70.35 % | 55.00% | 70.20 % |
| | All hospitals | 73.09 % | 62.85% | 68.45 % | 65.83 % |

Figure 4.4: Mean accuracy after including data from all the scanning sites in the training set.

significant).

The results reported in Figures 4.3 and 4.4 reveal a big problem with multisite data. These results show that it is very difficult to determine if adding more data from a different site will have a positive, negative, or no effect in the classification accuracy. Unfortunately, this is not just a hypothetical scenario. If the objective is to develop classifiers that can be deployed for clinical use, the users will be interested in the prediction accuracy in their scanning site and typically have the option of incorporating data from different studies in order to improve their own performance. We see that there is no easy answer to this question.

4.3 Batch effects

In genomic studies, the noise added to the biological signals due to technical factors in the development of the experiments, and that differ from one experiment to another, is known as batch effects [37]. This phenomenon is also present in fMRI data, where we view it to be altering the joint distribution of the data in two different scanning sites, a and b , making $P(X, Y | a) \neq P(X, Y | b)$. The way in which these probabilities differ is still an open problem, but research suggest that it is influenced by a variety of factors including: field strength of the magnet, manufacturers and parameters of the MRI scanner, radiofrequency noise environments, differences in the scanning protocol, and the general experience of the participants in the study [21].

While several association studies claim that the effects introduced by the merging data extracted from different scanning sites plays no significant role

Table 4.3: Classification accuracy for prediction of scanning site (binary classification)

| | Site 2 | Site 3 | Site 4 |
|--------|--------------|--------------|--------------|
| Site 1 | 88.0% | 92.0% | 87.4% |
| Site 2 | - | 96.2% | 92.5% |
| Site 3 | - | - | 93.4% |

in their analysis [53, 11, 60, 57, 9], we found that it does play a very important role in prediction studies. Usually association studies include the scanning site as an extra feature in an effort to reduce its influence in their analysis, but this strategy did not improve the results in our study.

4.3.1 Scanning site classification

The batch effects can be further analyzed by changing the target variable in the classification task from $y \in \{\text{SCZ}, \text{HC}\}$ to $y \in \{\text{Site}_1, \text{Site}_2, \text{Site}_3, \text{Site}_4\}$ (the feature matrix is the same as the one used for learning a model to predict schizophrenia). As shown in Table 4.3, the accuracy is $> 88\%$ for binary classification (Site_i vs Site_j) in all cases. For the 4-class classification scenario the accuracy was, on average, $83.56\% \pm 5.0\%$, which is well above the chance level of $\sim 25\%$.

Scanning site classification results indicate that there are important differences in the feature vectors depending on the origin of the data. These differences are shown in Figure 4.5. Every row in the figure corresponds to a participant in the experiment, while every column represents a feature (correlation between two regions of interest). The color represent the value of the correlation coefficient. In hospital 1, there is a high correlation between all the features, across all the participants. On the other side, participants in the hospital 3 present much lower correlation values when compared with the other hospitals. This visualization reinforces the idea that the data from different scanning sites follow different joint distributions, and partially explains why simply merging the data for creating an augmented dataset does not achieve the expected results.

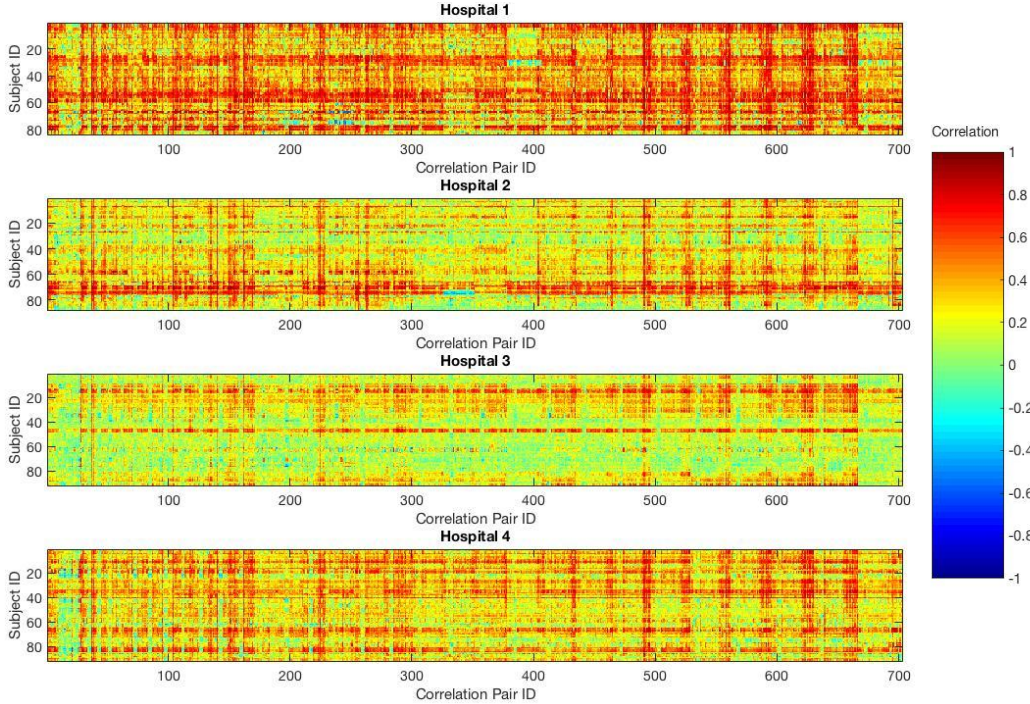


Figure 4.5: The feature vector for all subjects (both schizophrenia and healthy control) looks different across different scanning sites.

An interesting aspect of Figure 4.5 is that, despite the obvious differences across different hospitals, there are features that consistently present a high correlation value regardless of the scanning site. These features can be observed as *vertical lines* in the figure. A deeper analysis reveals that those features are mostly part of the auditory network – *i.e.*, the ROI inside this network are *highly connected* (see Figure 4.6). To address the concern that the high correlation values might be due to spatial proximity, we identify the physical location of the connected nodes; see Figure 4.7. The high connectivity occurs even if the ROI are in different brain hemispheres, eliminating the idea that the robustness across sites is due to an artifact, or spatial proximity. The visualization was performed using the package BrainNet Viewer [67].

Extending the analysis to include all the 264 ROI, we observe that the sensory/somatomotor network, auditory network and visual network present consistently high connectivity. These three networks are closely related to the fMRI task: watch a screen, hear sounds, and press a button (Figure 4.8). This is an interesting insight that suggests that not all the regions of interest

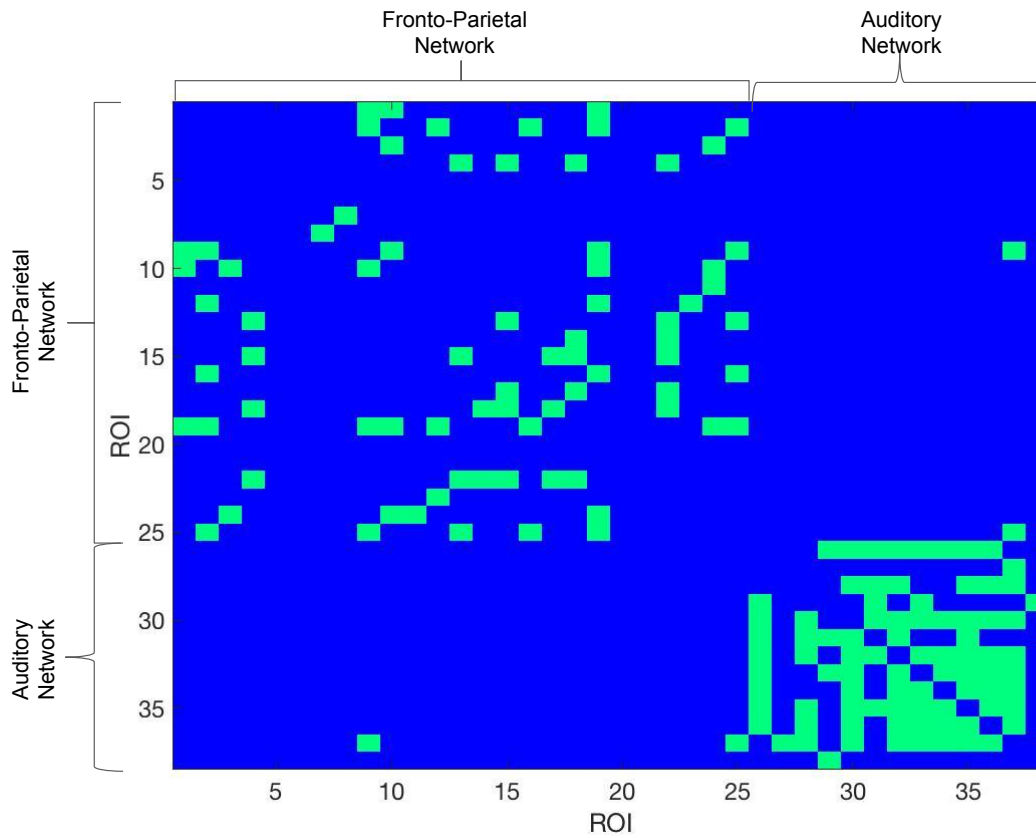


Figure 4.6: Analysis of the 2 networks used for prediction. Regions with high pairwise connectivity that are consistent across the different scanning site are shown in green. Note that the voxels corresponding to the auditory network are highly interconnected.

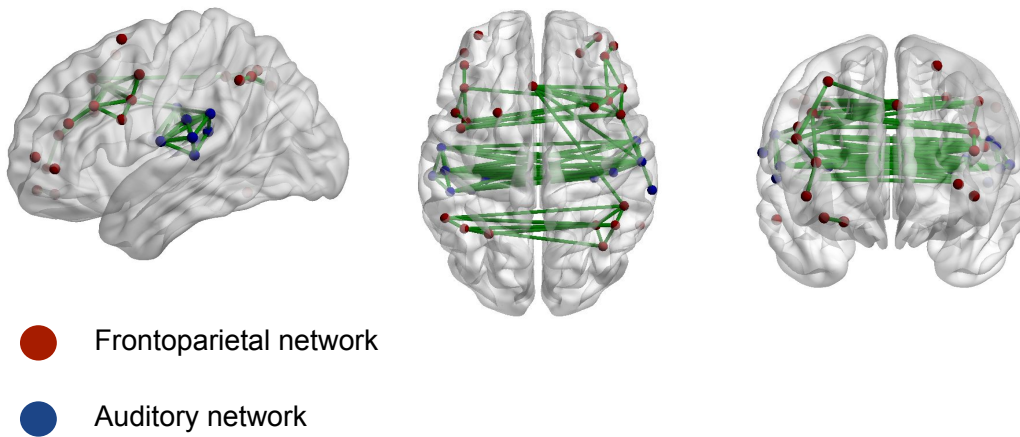


Figure 4.7: Physical location of the regions of interest that presented high connectivity across all the scanning sites.

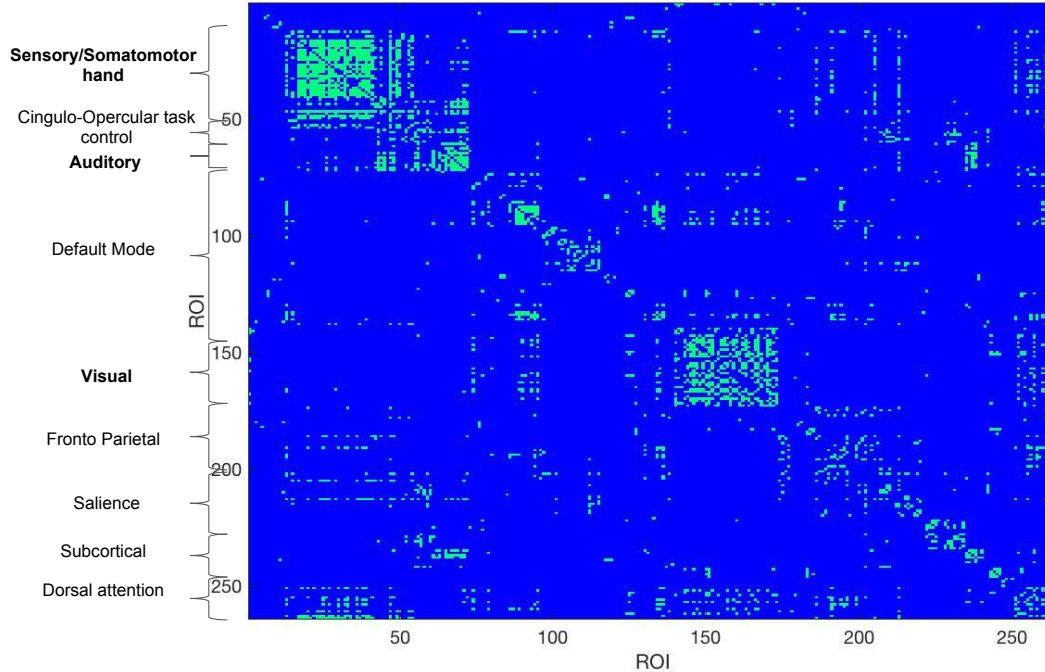


Figure 4.8: Analysis of the 264 ROI. Regions consistent across the different scanning site are shown in green. Note that sensory/somatomotor, visual and auditory networks present high connectivity.

are affected by batch effects in the same way. In particular, brain networks that are directly related to the task that participants are performing are more robust to noise than networks that are not directly related. Unfortunately, these high connectivity is not only consistent among scanning sites, but also between people with schizophrenia and healthy controls, which means that they have little predictive power for distinguishing between the two groups.

4.3.2 Traveling subject dataset

In an effort to better understand the effects that different scanning sites have on the data, the Biomedical Informatics Research Network (BIRN) designed an experiment in which 5 healthy participants traveled to 10 scanning sites, and were scanned 8 times in every site while performing the same set of activities [31]. To gain a visual intuition of these effects, it is possible to project the feature matrix into their principal components using Principal Component Analysis (PCA) and then plot the first two component to get a 2-D representation of the data. Figure 4.9 shows the result of this procedure. Every color

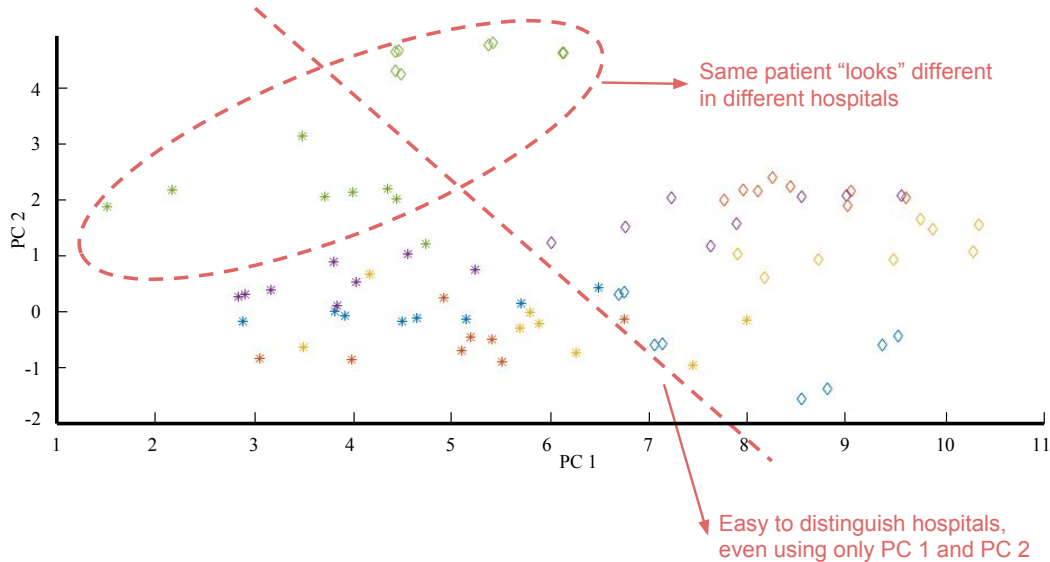


Figure 4.9: Projection of the feature matrix of the traveling subjects dataset into the first two principal components. Every point represents an fMRI scan, whose color represents a participant and whose shape identifies the scanning site.

in the figure represents a participant in the study, every shape represents a scanning site (we plotted only two sites for easiness of visualization), and every point represents an fMRI scan.

Figure 4.9 shows that the participants *look different* in different scanning sites. More important, there is a clear division between scans taken in one site versus the scans taken in the other one – *i.e.*, it is possible to create a linear decision boundary that separates both sites even in the reduced 2-D space.

4.3.3 Solving the traveling subject problem

An intuitive idea to deal with the batch effects is to assume that an unknown function, $g(X)$, maps the feature matrix X for a single patient from one scanning site A to a scanning site B , such that $X^B = g(X^A)$. When these input/output pairs are available, like in the case of the traveling subject dataset, a neural network with a single hidden layer is sufficient to compute an approximation of $g(x)$ for the input/output pairs used on the training set [24]. After computing this approximation, $\hat{g}(\cdot)$, we can concatenate X^B and $\hat{g}(X^A)$ into a single dataset, and use any learning algorithm to learn a classifier (SVM

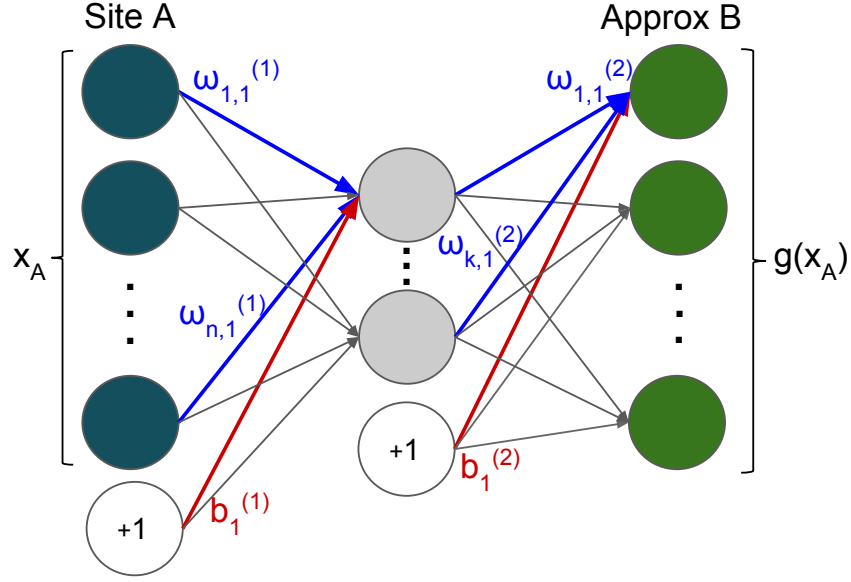


Figure 4.10: Neural network architecture, which is essentially the same as the one used by autoencoders. The objective of the network is to produce an approximation of how data from scanning site A is represented in scanning site B.

in this case). The architecture of the proposed neural network is shown in Figure 4.10. Note that is identical to the architecture to an autoencoder, with the difference that instead trying to copy its input into its output, it will map feature vectors from Site_A to Site_B .

The output of the neurons in the hidden and output layers is given by the function $f_{\omega,b}(x) = \sigma(\omega^T x + b)$, where $\sigma(x) = \frac{e^{2x}-1}{e^{2x}+1}$. The cost function optimized by the neural network is similar to the one used in sparse autoencoders, with the variation that the error is computed as the difference between the output of the neural network and the representation of the feature vector in a different site [42]:

$$J(\omega, b) = \frac{1}{m} \sum_{p=1}^m (\|x_p^B - g(x_p^A)\|^2) + \lambda \sum_{i,j,l} (\omega_{i,j}^{(l)})^2 + \beta \sum_{i=1}^k KL(\rho || \hat{\rho}_i) \quad (4.1)$$

where m is the number of participants whose feature vector are available for both site A and site B, x_p^B and x_p^A is the feature vector of obtained in scanning site B and A of the p th participant, $g(\cdot)$ is the output of the neural network,

$\omega_{i,j}^{(l)}$ is the weight that connects the i th neuron of the layer l with the j th neuron of layer $l + 1$, λ is a regularization term that controls the magnitude of the weights ω , β is a second regularization term that forces that the average activation ($\hat{\rho}_i$) of the i th neuron in the hidden layer to be close to a desired average activation ρ_i (usually a low number like 0.05), k is the number of neurons in the hidden layer, and $KL(p||q) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\frac{1-p}{1-q}$ is the Kullback-Leibler divergence between two Bernoulli random variables with mean p and q , respectively.

One way of measuring if this neural network is effective for decreasing the impact of batch effects in prediction studies is by creating two classifiers whose target variable are: (1) participant ID ($y \in \{1, 2, \dots, 5\}$), versus (2) scanning site ($y \in \{A, B\}$). Ideally, the first classifier should be successful, but the performance of the second task should be at chance level. In the dataset, every participant was scanned 8 times at each scanning site. Here, we used 4 scans for training the model (both, the neural network and the classifiers), and the other 4 were left as a hold-out set for estimating the performance. Defining $T_{i,j}^s$ as the matrix containing the time series corresponding to the j th scan of participant i in scanning site s , we apply Algorithm 3 to estimate the performance of the two tasks described above. Given the small number of instances available, we reduced the number of ROI to be analyzed. Since the participants were doing a task involving a visual activity, we analyzed only the visual network. Therefore $T_{i,j}^s \in \mathbb{R}^{31 \times 89}$ because the visual network involves 31 ROI, and the time series associated with every region has 89 timepoints.

The first step is to create the input/output pairs for the neural network. Every scan in the training set from the subject i obtained from the scanner A (input) will be mapped to the average scan (output) of the same subject obtained from scanner B in Algorithm 3. After that, the feature vector can be obtained for all the inputs/outputs. This feature vector is obtained by computing the pairwise correlation of the time series associated with each ROI as described in Section 3.3.2. The next step is to train the neural network by finding the values ω, b that minimize Equation 4.1. Then, using ω and b we can *translate* the feature vectors from site A to site B and learn two classifiers to

Algorithm 3 Traveling subject problem

Input: $\{T_{i,j}^s \mid i \in \{1, \dots, 5\}, j \in \{1 \dots 8\}, s \in \{A, B\}\}$ **Output:** $\text{Accuracy}_{\text{subject}}, \text{Accuracy}_{\text{site}}$

```
1: procedure PERFORMANCETRAVELINGSUBJECT
2: # Train a neural network to estimate how participants in site A look like
   in site B.
3:   for i in 1:5 do
4:      $\bar{T}_i^B \leftarrow \frac{1}{4} \sum_{j=1}^4 T_{i,j}^B$ 
5:      $X_i^B \leftarrow \text{getFeatureVector}(\bar{T}_i^B)$ 
6:     for j = 1:4 do
7:        $X_{i,j}^A \leftarrow \text{getFeatureVector}(T_{i,j}^A)$ 
8:       Generate input/output pairs  $\langle X_{i,j}^A, X_i^B \rangle$ 
9:     end for
10:  end for
11:  Concatenate the input/output pairs into a train set  $D_T^{(NN)}$ 
12:   $\text{Model}_{NN} \leftarrow \text{trainNeuralNetwork}(D_T^{(NN)})$ 
13:
14: # Estimate how participants in A look like in B
15:  for i = 1:5 do
16:    for j = 1:8 do
17:       $\hat{X}_{i,j}^B \leftarrow \text{Model}_{NN}(X_{i,j}^A)$ 
18:       $X_{i,j}^B \leftarrow \text{getFeatureVector}(T_{i,j}^B)$ 
19:    end for
20:  end for
21:  Merge  $\hat{X}_{i,j}^B$  and  $X_{i,j}^B$  into a training set  $D_T$  for  $i \in \{1, \dots, 5\}$ ,
    $j \in \{1, \dots, 4\}$ .
22:  Merge  $\hat{X}_{i,j}^B$  and  $X_{i,j}^B$  into a hold-out set  $D_H$  for  $i \in \{1, \dots, 5\}$ ,
    $j \in \{5, \dots, 8\}$ .
23:  Define the target vectors  $Y_T^{\text{site}}, Y_T^{\text{participant}}, Y_H^{\text{site}}, Y_H^{\text{participant}}$ 
24:
25: # Create the classifiers
26:   $\text{Model}_{SVM}^{\text{site}} \leftarrow \text{trainSVM}(D_T, Y_T^{\text{site}})$ 
27:   $\text{Model}_{SVM}^{\text{participant}} \leftarrow \text{trainSVM}(D_T, Y_T^{\text{participant}})$ 
28:
29:   $\text{Predictions}_{\text{site}} \leftarrow \text{Model}_{SVM}^{\text{site}}(D_H)$ 
30:   $\text{Predictions}_{\text{participant}} \leftarrow \text{Model}_{SVM}^{\text{participant}}(D_H)$ 
31:   $\text{Accuracy}_{\text{site}} = \text{getAccuracy}(\text{Predictions}_{\text{site}}, Y_H^{\text{site}})$ 
32:   $\text{Accuracy}_{\text{subject}} = \text{getAccuracy}(\text{Predictions}_{\text{subject}}, Y_H^{\text{subject}})$ 
33: end procedure
```

Table 4.4: Accuracy comparison in prediction of scanning site and participant ID before versus after batch effects correction. (Results on the hold-out set using the traveling subject data)

| Before correction | | After correction | |
|-------------------|-------------|------------------|-------------|
| Site | Participant | Site | Participant |
| 100% | 100% | 65% | 100% |

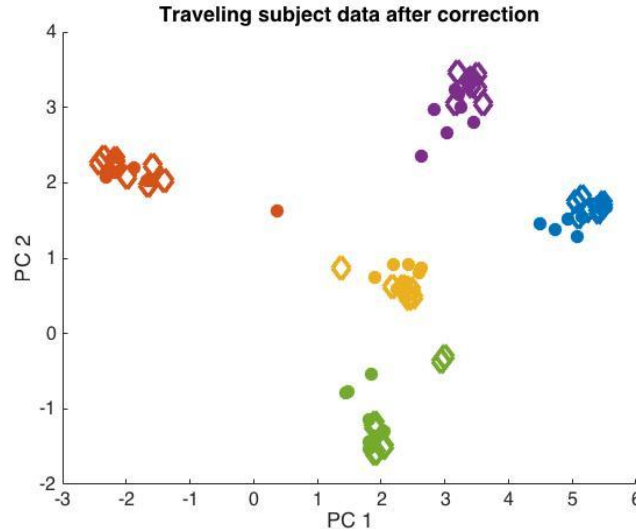


Figure 4.11: Projection of the traveling subject dataset into the first two components after correcting the batch effects using a neural network.

distinguish: (a) participant ID and (b) site. Finally, we test the performance of this algorithm in a hold-out set.

The results of applying Algorithm 3 to the traveling subject dataset are presented in Table 4.4. This table shows that this approach decreased the impact of batch effects, since it is not possible to distinguish between scanning site, but it is still possible to identify the subject perfectly. Figure 4.11 shows this results graphically. Note how, after projecting the output of the neural network into the two first principal components, the data is clustered by participant, but not by scanning site.

Chapter 5

Reducing the influence of batch effects

The solution proposed in Section 4.3.3 is effective when the same participants are scanned in both scanning sites; however, this is not the case in real scenarios. Without the matching input/output pairs, it is not possible to directly use a neural network to estimate the function $\hat{x}^B = g(x^A)$ that estimates the feature vector that would represent participants scanned in site A if they would have been scanned in site B. For this standard case, a different sets of techniques should be used for decreasing the batch effects.

5.1 Simple transformations I: Translation and scaling

An intuitive starting point is to use techniques that remove simple linear data transformations such as translations, rotations and scaling. The physical properties of a scanner influence the signal-to-noise ratio of the fMRI time signals [21]. This will in turn affect the correlation values between the time signals. If we assume that a scanner influences the correlation between the time signals of two regions of interest in the same way for all the patients, we can represent the relationship between every feature (pairwise correlation) in scanning sites A and B as:

$$X_i^B = \alpha_i X_i^A + \beta_i, \quad i = 1, 2, \dots, m \quad (5.1)$$

where X_i^B is a random variable representing the i th feature in the data ex-

tracted from scanning site B , X_i^A is its equivalent for scanning site A , α_i, β_i are the scaling and translation coefficients for the i th feature, and m is the number of features. Since we do not have pairs (x_i^A, x_i^B) for any subject, it is not possible to estimate the coefficients directly; however, we can remove their influence by using z-score normalization independently on data from each scanning site:

$$\bar{X}_i^A = \frac{X_i^A - E[X_i^A]}{\sqrt{\text{Var}(X_i^A)}} \quad (5.2)$$

$$\begin{aligned} \bar{X}_i^B &= \frac{X_i^B - E[X_i^B]}{\sqrt{\text{Var}(X_i^B)}} \\ &= \frac{\alpha_i X_i^A + \beta_i - E[\alpha_i X_i^A + \beta_i]}{\sqrt{\text{Var}(\alpha_i X_i^A + \beta_i)}} \\ &= \frac{\alpha_i (X_i^A - E[X_i^A])}{\sqrt{\alpha_i^2 \text{Var}(X_i^A)}} \\ &= \frac{X_i^A - E[X_i^A]}{\sqrt{\text{Var}(X_i^A)}}, \quad \text{for } \alpha_i > 0 \\ &= \bar{X}_i^A \end{aligned} \quad (5.3)$$

Note that even if we cannot recover the original values X_i^A and X_i^B , we can transform both variables into a new space \bar{X}_i^A and \bar{X}_i^B where they are equivalent. The effect of the translation is removed by subtracting $E[X]$, while the effect of a scaling by a positive number is removed by dividing by $\sqrt{\text{Var}(X)}$. Figure 5.1 shows this effect graphically for a simple example in two dimensions.

We applied z-score normalization independently to every dataset, and then repeated the experiments for the second scenario described in Section 3.1 (healthy controls versus people with schizophrenia in a multisite context). In addition, we attempted site classification, as described in Section 4.3.1, but using the fBIRN phase II dataset instead of the traveling subjects one. Table 5.1 shows the result for the schizophrenia versus healthy controls case. Ideally, the off-diagonal elements of the table should be higher than the diagonal elements (since the diagonal results use only information of a single

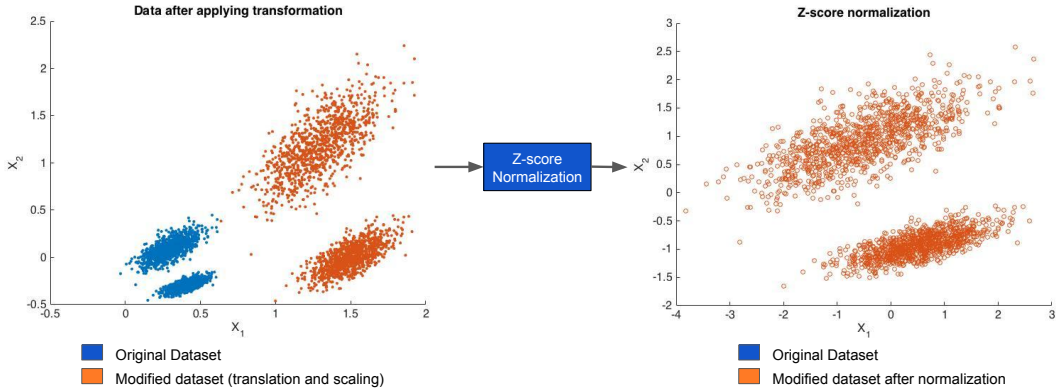


Figure 5.1: Graphical representation of z-score removing translation and scaling.

Table 5.1: Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using z-score normalization. Values in bold indicate single site classification.

| | | Additional train set | | | |
|----------|--------|----------------------|--------------|--------------|--------------|
| | | Site 1 | Site 2 | Site 3 | Site 4 |
| Test set | Site 1 | 63.3% | 63.5% | 56.9% | 62.5% |
| | Site 2 | 69.5% | 70.4% | 67.0% | 60.5% |
| | Site 3 | 45.7% | 50.7% | 55.0% | 60.5% |
| | Site 4 | 74.2% | 55.7% | 63.9% | 70.2% |

scanning site); however, this is not the case. Also, comparing the results of Table 5.1 with those shown in Figure 4.3, we can appreciate that the new results are even worse than naively merging the datasets! On the other side, it is no longer possible to predict the scanning site (accuracies at the chance level for all cases). If only the last test were performed, it would give the false intuition that z-score normalization is enough for removing the batch effects; however, as the performance of the task of diagnosing schizophrenia reveals, this is not the case.

At first sight, the results obtained after normalizing the data using z-score are counter-intuitive. The mean and standard deviation of every feature are now the same in datasets from different scanning sites, so they should be *more compatible*, which should be reflected as an increase in the prediction accuracy, not a decrease. A closer look into z-score normalization suggests why this

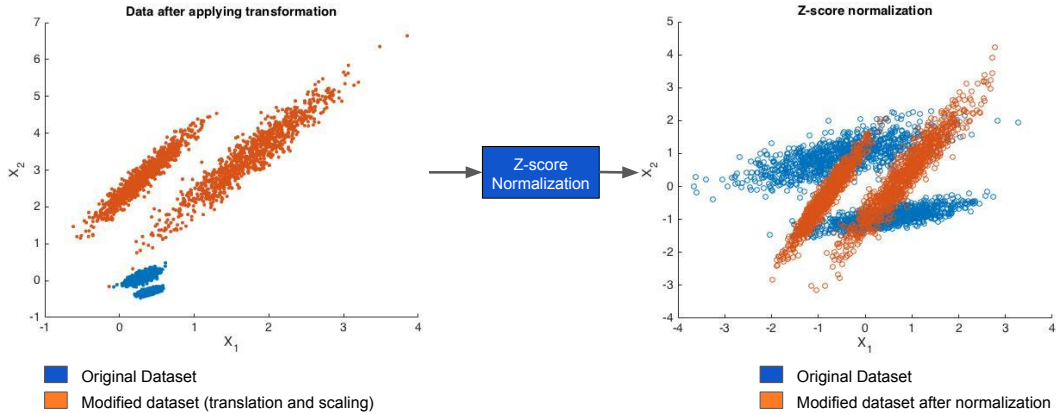


Figure 5.2: Z-score normalization only corrects batch effects caused by translation and scaling of the data, but it is insufficient for other types of transformations.

might the case. Consider $X^A \in \mathbb{R}^{n \times m}$ as a dataset extracted from scanning site A that contains the m -dimensional feature vector of n participants. Also, consider X^B as the result of applying a linear transformation to X^A :

$$X^B = X^A \alpha + \mathbf{1} \beta^T \quad \alpha \in \mathbb{R}^{m \times m}, \quad \beta \in \mathbb{R}^m \quad (5.4)$$

z-score normalization removes the effects of this linear transformation only when α is a diagonal matrix; however, it is insufficient when the off-diagonal elements $\alpha_{i,j} \neq 0$. This case is illustrated on Figure 5.2 and shows the limitation of any algorithm that exclusively removes the effects of scaling and translation.

5.2 Simple transformations II: Rotation and translation

Whitening is a linear transformation that can be viewed as a generalization of the z-score normalization. Besides making the mean of every feature equal to zero and its variance equal to one, it also removes the correlation between features by making its covariance matrix the identity matrix. One of the most common procedures to perform this process is *PCA Whitening* [32]. This transformation first rotates the data by projecting it into its principal components, and then it scales the rotated data by the square root of its eigenvalues

(which represent the variance of each new variable in the PCA space).

If the batch effects are caused by a rotation and translation of the datasets, then applying the whitening transformation to every dataset independently will remove the batch effects. To see why this is the case, consider the datasets X_A and X_B as defined in Equation 5.4. The zero-mean datasets, \bar{X}_A , can be obtained as:

$$\begin{aligned}\bar{X}_A &= X_A - \mathbf{1}E[X_A] \\ E[X_A] &= [E[X_A^1], E[X_A^2], \dots, E[X_A^m]]\end{aligned}\tag{5.5}$$

while for the case of \bar{X}_B :

$$\begin{aligned}\bar{X}_B &= X_A\alpha + \mathbf{1}\beta^T - \mathbf{1}E[X_A\alpha + \mathbf{1}\beta^T] \\ &= X_A\alpha + \mathbf{1}\beta^T - \mathbf{1}(E[X_A\alpha] - E[\mathbf{1}\beta^T]) \\ &= (X_A - \mathbf{1}[X_A])\alpha \\ &= \bar{X}_A\alpha\end{aligned}\tag{5.6}$$

The eigenvalues of the covariance matrix $\Sigma_A = \frac{1}{n-1}\bar{X}_A^T\bar{X}_A$ are obtained by solving the equation $\det(\Sigma_A - \lambda I) = 0$. For the special case when α is an orthogonal matrix with $\det(\alpha) = 1$ (which represents a rotation matrix)¹, $\alpha^T = \alpha^{-1}$ [10], the eigenvalues of the covariance matrix of \bar{X}_B :

$$\begin{aligned}\det\left(\frac{1}{n-1}(\bar{X}_A\alpha)^T(\bar{X}_A\alpha) - \lambda I\right) &= 0 \\ \det\left(\frac{1}{n-1}\alpha^T\bar{X}_A^T\bar{X}_A\alpha - \lambda I\right) &= 0 \\ \det(\alpha^T\Sigma_A\alpha - \lambda I) &= 0 \\ \det(\alpha^{-1}\Sigma_A\alpha - \alpha^{-1}\lambda I\alpha) &= 0 \\ \det(\alpha^{-1}(\Sigma_A - \lambda I)\alpha) &= 0 \\ \det(\alpha^{-1})\det(\Sigma_A - \lambda I)\det(\alpha) &= 0 \\ \det(\Sigma_A - \lambda I) &= 0\end{aligned}\tag{5.7}$$

¹All orthogonal matrices, α , have a determinant equal to +1, or -1. If it is positive, α is a rotation matrix. When the determinant is negative, it is a reflection matrix.

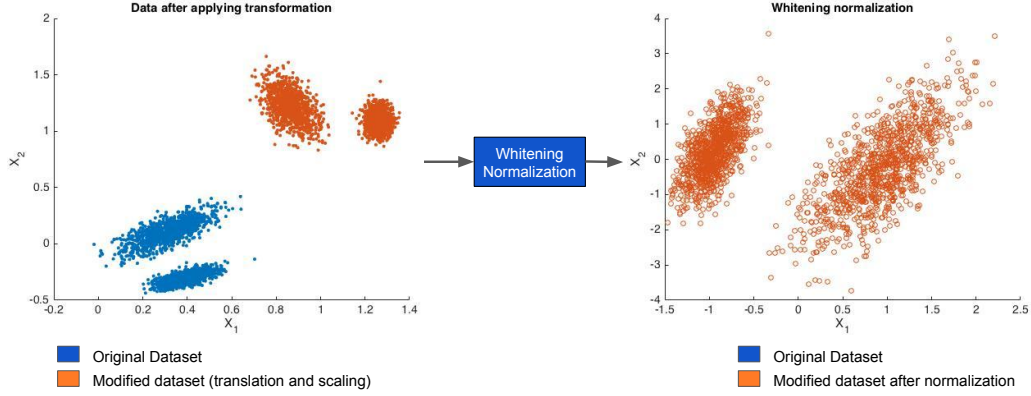


Figure 5.3: Whitening can correct batch effects caused by rotations and translations of a dataset.

As for the eigenvectors: if v is an eigenvector of Σ_A with an associated eigenvalue λ , then $\Sigma_A v = \lambda v$. Doing some mathematical manipulations:

$$\begin{aligned}
 \alpha \Sigma_A v &= \alpha \lambda v \\
 \alpha \Sigma_A I v &= \alpha \lambda v \\
 \alpha \Sigma_A \alpha^{-1} \alpha v &= \lambda \alpha v \\
 \Sigma_B(\alpha v) &= \lambda(\alpha v)
 \end{aligned} \tag{5.8}$$

Equations 5.7 and 5.8 show that, when the transformation matrix α is an orthogonal matrix with positive determinant, X_A and X_B will have the same eigenvalues, and the eigenvectors of X_B are just a rotation of the eigenvectors of X_A . Therefore, by projecting the data into those eigenvector, we obtain the exact same representation, removing the effects of translation and rotation. Figure 5.3 shows an example of this process with a 2-dimensional dataset.

Table 5.2 show the results of applying whitening to the problem of classifying schizophrenia versus healthy controls and site classification. Similar to z-score normalization, whitening reduces the performance of a scanning site classifier to chance level; however, it also decreased the accuracy of the classifier aimed to diagnose schizophrenia in a multisite context relative to naively merging the datasets.

Z-score normalization and whitening are common procedures applied to reduce the discrepancies between two datasets. In the first one, the marginal

Table 5.2: Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using whitening. Values in bold indicate single site classification.

| | | Additional train set | | | |
|----------|--------|----------------------|--------------|--------------|--------------|
| | | Site 1 | Site 2 | Site 3 | Site 4 |
| Test set | Site 1 | 63.3% | 39.4% | 44.6% | 66.1% |
| | Site 2 | 41.5% | 70.4% | 52.1% | 36.1% |
| | Site 3 | 47.4% | 45.3% | 55.0% | 47.5% |
| | Site 4 | 66.4% | 34.8% | 57.3% | 70.2% |

probabilities of every feature are the same for X_A and X_B ; however, it does not remove correlations present between features. Whitening, in addition, makes the covariance matrix equal to the identity matrix, making the features in the new space decorrelated. These techniques solve batch effects caused by translations and scaling, or rotation and translation of the dataset; however, they were insufficient for solving the batch effects in fMRI data, indicating that they are caused by more complex mechanisms that cannot be modeled by these simple linear transformations.

5.3 BECCA

BECCA (which stands for Batch Effect correction using Canonical Correlation Analysis) is a tool designed to remove batch effects caused by technical noise that confounds the true biological signal in the context of gene expression microarrays data [61]. It assumes that data extracted from two batches, X^A and X^B , are random samples of a common population, and that the two sets share a common biological signal shadowed by technical confounds that can be decomposed as:

$$\begin{aligned}
 X^A &= \alpha Y^A + \beta^A Z^A + \epsilon^A \\
 X^B &= \alpha Y^B + \beta^B Z^B + \epsilon^B
 \end{aligned}
 \tag{5.9}$$

where $X^i \in \mathbb{R}^{m \times n_i}$, which are matrices with m features and n^i instances, represent the observed data. αY^i , with $\alpha \in \mathbb{R}^{m \times q}$ and $Y^i \in \mathbb{R}^{q \times n^i}$, represents

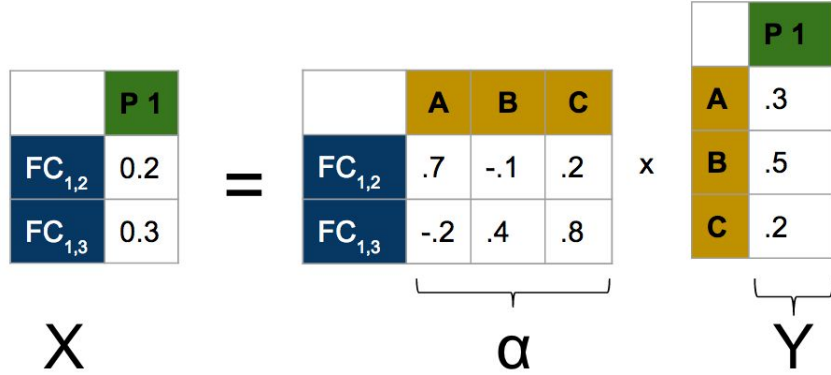


Figure 5.4: Intuition of the decomposition assumed by BECCA. Assume that there are $q = 3$ prototypes, and every participant can be represented as a linear combination of them. (Noise and batch effects are not represented in the figure.)

the biological component. $\beta^i Z^i$, with $\beta^i \in \mathbb{R}^{n \times r^i}$ and $Z^i \in \mathbb{R}^{r^i \times n^i}$, represents the influence of the batch effects, and ϵ^i represents noise in the measurements. Note that the matrix α is the same in both batches, indicating that they share a common biological signal.

To gain an intuition of this decomposition, imagine that the functional connectivity of every person is a function of their personality. For simplicity, assume that there are 3 *prototype personalities*: A, B and C, whose functional connectivity is known, and that the functional connectivity of a person is a linear combination of these 3 *prototypes*. In the context of Equation 5.9, α represents the functional connectivity of the prototypes, and Y_i represents the weights of every patient. A similar rationale follows for $\beta^i Z^i$. Figure 5.4 is a representation of this example in the absence of noise and batch effects.

In general, we can only observe X^A and X^B , while the other matrices are unknown. Under the assumption of this additive model, and that the batch effects are orthogonal to each other and to the signals of interest: ($\alpha^T \beta^A = 0$, $\alpha^T \beta^B = 0$, $(\beta^A)^T \beta^B = 0$), BECCA eliminates the noise and batch effect components of Equation 5.9 without explicitly modeling any of the matrices [61].

We applied BECCA² to the schizophrenia dataset in order to remove the

²Software freely available at: <https://sites.google.com/site/svaisipour/utilities>

Table 5.3: Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using BECCA. Values in bold indicate single site classification.

| | | Additional train set | | | |
|----------|--------|----------------------|--------------|--------------|--------------|
| | | Site 1 | Site 2 | Site 3 | Site 4 |
| Test set | Site 1 | 63.3% | 70.5% | 63.5% | 55.8% |
| | Site 2 | 68.1% | 70.4% | 65.9% | 64.3% |
| | Site 3 | 57.3% | 57.4% | 55.0% | 48.5% |
| | Site 4 | 73.6% | 68.4% | 56.5% | 70.2% |

batch effects, and then we used SVM to classify healthy controls versus people with schizophrenia; however, as shown in the Table 5.3, this approach was not successful either. Note that BECCA’s efficiency depends on how accurately it can estimate the covariance matrices of X^A and X^B . Unfortunately, the high dimensionality of the data, along with the small number of training instances and the high variance of the functional connectivity among the instances [14], might cause an inaccurate estimation of these matrices. Moreover, BECCA is not designed to optimize for classification accuracy; instead it is an unsupervised algorithm that attempts to remove the information that is not correlated between the two batches – so BECCA is not guaranteed to increase the prediction accuracy. On the other side, it is effective in removing signals that are not correlated between batches. When we applied BECCA to our data, the performance of the site prediction dropped to chance level, so it is no longer possible identify the scanning site using the extracted features.

5.4 Non-linear transformations

5.4.1 Self-learning a feature representation

A different approach to the problem is to use *representation learning algorithms* that attempt to transform the original data into a different space that facilitates supervised learning tasks. Self-taught learning is a paradigm that uses sparse coding to construct higher-level features in an unsupervised fashion [46]. The hope is that more abstract features of the data will be more

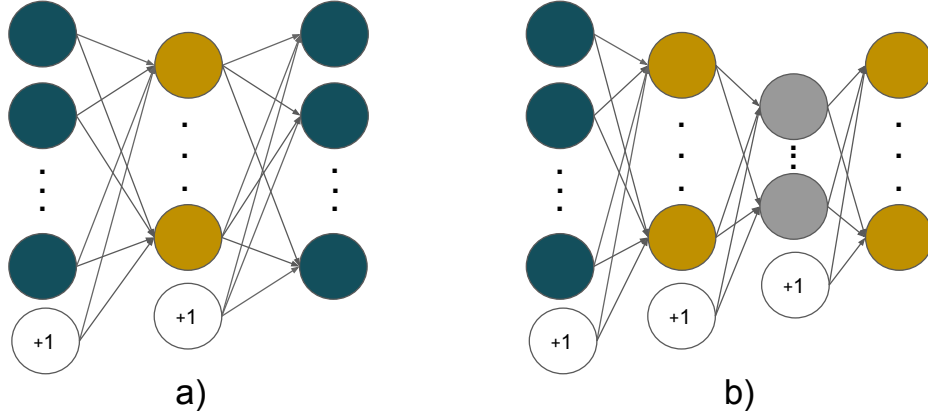


Figure 5.5: The stacked autoencoders train each layer independently: a) The raw inputs are mapped to themselves. b) The hidden layer of the previous autoencoder becomes the input for the next one. This process is repeated until the desired depth is achieved. The hidden layer of the last autoencoder can be used as the input to a learning algorithm.

likely to be shared among data extracted from different scanning sites [40].

One way of obtaining this representation is to use autoencoders as the basic building block, and then stack them into a deep architecture. The hidden layer of an autoencoder is used as the input to the next one (see Figure 5.5). Each autoencoder is trained independently in an unsupervised way, and the last hidden layer is the input to a supervised layer. Finally, a fine tuning of the network is performed to optimize all the weights. The idea is that the pre-training of the autoencoders will locate the parameters in a region parameter space that will reach a good local optimum [5].

Every autoencoder optimizes the cost function previously described in Equation 4.1, with the difference that now every input will map to itself. For the supervised learning task we used a softmax layer, which contains $K = 2$ output neurons, one for each class. Softmax regression minimizes the cost function [42]:

$$J(\omega) = - \sum_{i=1}^m \sum_{k=1}^K 1\{y_i = k\} \log \frac{\exp(\omega_k^T x_i)}{\sum_{j=1}^K \exp(\omega_j^T x_i)} \quad (5.10)$$

where y_i represents the class of the i -th instance, K represents the number of output neurons (classes), ω_j represents the weights connected to the j -th

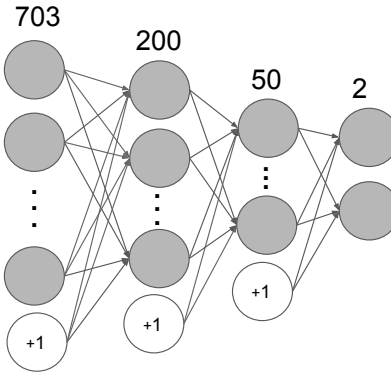


Figure 5.6: Architecture of the neural network. The number of neurons of each layer is indicated at the top.

Table 5.4: Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using BECCA. Values in bold indicate single site classification.

| | | Additional train set | | | |
|----------|--------|----------------------|--------------|--------------|--------------|
| | | Site 1 | Site 2 | Site 3 | Site 4 |
| Test set | Site 1 | 63.3% | 56.1% | 51.8% | 58.6% |
| | Site 2 | 65.7% | 70.4% | 52.5% | 62.9% |
| | Site 3 | 59.3% | 59.6% | 55.0% | 60.7% |
| | Site 4 | 56.9% | 52.8% | 54.4% | 70.2% |

output neuron, and x_i is the feature vector of the i -th training example.

For experimental purposes we used the architecture of the neural network shown in Figure 5.6. It consists in a network with 2 hidden layers. The first one had 200 hidden neurons and the second 50 (these parameters, along with the regularization coefficients were selected by cross-validation). The input layer consists of 703 neurons, while the output layer has 2 neurons (one for each class). The inputs to the network are the functional connectivity features extracted from fMRI data, new instances are classified to the class of the neuron with highest output. The results of using this neural networks for schizophrenia diagnosis in multisite data is shown in Table 5.4. As can be observed, this approach did not achieve the expected results.

Although disappointing, the results of applying neural networks to this problem is not unexpected. It was not even effective for decreasing the per-

formance of scanning site identification (the accuracy remains above 75% for most of the cases). Deep learning approaches usually require a vast number of training examples [46], which we do not have in this application. Note how the number of parameters to fit far exceed the number of training instances, which leads to overfitting (indeed, all the tested models achieved near 100% classification accuracy on the training set, but had a low accuracy on the test set). At the same time, these model have proven successful when there is structure in the data [42], such as images or audio; however, which might not be the case for pairwise correlation between regions of the brain.

5.4.2 Bi-shifting autoencoders

As seen in Section 4.3.3, using an approach similar to autoencoder could solve the batch effects problem for the traveling subject dataset. What prevents that approach for being used in the diagnosis of schizophrenia is that it requires fMRI scans from the same participants obtained in different scanning sites. Kan et al. proposed a bi-shifting autoencoder for unsupervised domain adaptation that is designed for cases when the training sample and test sample follow different distributions [30]. In this context, the discrepancy is not between training and test samples, but between data obtained from different scanning sites. The architecture of a bi-shifting autoencoder is depicted in Figure 5.7.

The main idea of the bi-shifting autoencoder is to have a set of weights W_c that transform the raw inputs x from the source or target domain into a common shared space by means of a non-linear transformation $z = f_c(x) = \sigma(W_c x + b_c)$, where b_c represents the bias term, and $\sigma(\cdot)$ is non-linear squashing function, such as the sigmoid function, or the hyperbolic tangent. Then a set of weights W_s, W_t, b_s and b_t , map z to itself, and to an estimation of how it would be represented in the other domain. For the source domain $g_s(z) = \sigma(W_s z + b_s)$, and $g_t(z) = \sigma(W_t z + b_t)$.

Since we only know the representation of the i -th instance in one of the domains, Kai et al. proposed to estimate its representation as linear combination of the instances in the other domain: $x_s^{(i)} = X_t \beta_i^T$ or $x_t^{(i)} = X_s \beta_i^T$, where

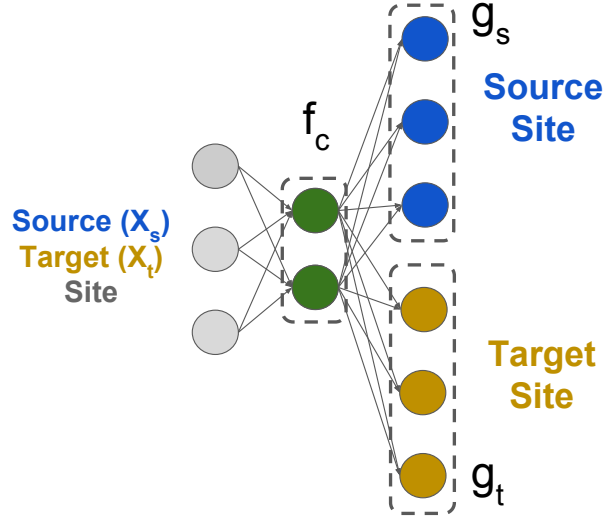


Figure 5.7: Bishifting Autoencoder. Every input from the one domain (target or source) is mapped to itself and to an approximation of how it would be represented in the other domain.

$X_t \in \mathbb{R}^{m \times n_t}$ is the matrix with the n_t m -dimensional vectors of the target domain, X_s is its analogue for the source domain, and β_i is the vector containing the coefficients of the linear combination of the instances. The cost function to be minimized is then:

$$\begin{aligned} & \arg \min_{W_c, b_c, W_s, b_s, W_t, b_t, B_s, B_t} \|X_s - g_s(z_s)\|_2^2 + \|X_t B_t - g_t(z_s)\|_2^2 \\ & + \|X_s B_s - g_s(z_t)\|_2^2 + \|X_t - g_t(z_t)\|_2^2 + \gamma \left(\sum_{i=1}^{n_s} |\beta_i^t|_1 + \sum_{i=1}^{n_t} |\beta_i^s|_1 \right) \quad (5.11) \end{aligned}$$

where $z_s = \sigma(W_c X_s + b_c)$, $z_t = \sigma(W_c X_t + b_c)$, $B_s \in \mathbb{R}^{n_s \times n_t}$ and $B_t \in \mathbb{R}^{n_t \times n_s}$ are matrices containing the β vectors that will reconstruct every instance of the source (target) domain as a linear combination of the instances in the target (source) domain, and γ is a regularization parameter that controls the sparsity in β .

The authors proposed an iterative method similar to expectation maximization to solve this problem [30]. They applied their idea to the task of face recognition and empirically showed that it improves the accuracy when using datasets from different distributions for the task of face recognition; how-

Table 5.5: Classification accuracy for the problem of healthy controls versus patients with schizophrenia after using bishifting autoencoders. Values in bold indicate single site classification.

| | | Additional train set | | | |
|----------|--------|----------------------|--------------|--------------|--------------|
| | | Site 1 | Site 2 | Site 3 | Site 4 |
| Test set | Site 1 | 63.3% | 65.6% | 59.5% | 64.7% |
| | Site 2 | 71.1% | 70.4% | 65.4% | 68.2% |
| | Site 3 | 53.8% | 60.0% | 55.0% | 58.9% |
| | Site 4 | 63.9% | 62.5% | 69.3% | 70.2% |

ever, they did not guarantee that their optimization methods will converge. The results of applying this methodology to our dataset reduced the discrepancy between the feature representation in different scanning sites. Scanning site classification decreased to chance level after using bishifting autoencoders. Table 5.5 presents the results for the schizophrenia diagnosis. Note that the accuracy when using multiple site data is slightly better than the one obtained when naively merging the data in most of the experiments (see Section 4.2) for most instances. The exceptions are the ones that involve site 3, which had a bad performance on its own; however the performance in the diagnosis of schizophrenia did not improve relative to using data from a single site, which is our main objective.

5.5 Summary of methods

Figure 5.8 summarizes the performance of the different methods used in an attempt to decrease the impact of batch effects for the task of separating patients with schizophrenia from healthy controls, and identification of scanning site. Each method made different assumptions about the data and the nature of batch effects. Z -score normalization (resp., whitening) are effective methods for removing translation and scaling, (resp., translation and rotation of the data); however, their lack of success in increasing the accuracy of schizophrenia diagnosis is a strong indication that batch effects in fMRI go beyond these simple linear transformations. BECCA assumes that the observed data is a











| Method | Decreases scanning site accuracy? | Increases SCZ vs HC accuracy? |
|-------------------------|---|---|
| z-score |  |  |
| Whitening |  |  |
| BECCA |  |  |
| Stacked autoencoders |  |  |
| Bishifting Autoencoders |  |  |

Figure 5.8: Bishifting Autoencoder. Every input from the one domain (target or source) is mapped to itself and to an approximation of how it would be represented in the other domain.

linear combination of some unknown *prototypes* plus technical noise introduced by unknown sources. It removes the signal from the batches (data from different scanning sites) that is not correlated with each other. Finally, stacked autoencoders and bishifting autoencoders use the *self-taught* [46] paradigm to learn a set of high-level features that extract commonalities between the dataset. Most of the approaches were successful in removing the signals in the datasets that allow the identification of the scanning site. Unfortunately, none of them were successful for increasing the accuracy of the classifier that separated people with schizophrenia from healthy controls.

Chapter 6

Conclusions

One of the biggest challenges for the application of machine learning algorithms into neuroimaging data is the tremendous disparity between the number of instances, usually in the range of few hundreds, and the number of features, usually in the range of tens of millions. To overcome this problem, two intuitive solutions are: (1) to reduce the number of features, and (2) to increase the number of instances. In this dissertation, we explored both approaches in the context of the diagnosis of schizophrenia. The objective was to build a classifier that could distinguish between people with schizophrenia and healthy controls using the FBIRN Phase II dataset. We tested the performance under two scenarios: (1) when the training and test data are obtained in a single scanning site and (2) when the training data is obtained from multiple sites, but the test data is a disjoint subset of only one of them.

For decreasing the number of features, we used the parcellation of the brain, proposed by Power et al. [45], that divides it in 264 regions of interest. These regions are also divided in 13 networks. Based on the nature of the task and literature about schizophrenia, we decided to extract the functional connectivity of the nodes corresponding to the auditory and fronto-parietal networks in order to learn a SVM classifier with linear kernel. This approach was successful, as it achieved an average accuracy of 63.33%, 70.35%, 55.0% and 70.20% in scanning sites 1, 2, 3 and 4 when using data from a single site. For sites 1, 2 and 4, the reported accuracy was statistically significant above the chance level. That was not the case for site 3. After analyzing

how an increase in the amount of data affected the performance of the learned classifier, we discover that for site 3, the accuracy was at chance level regardless of the increase in the size of the training data. This suggests that the signal extracted from data in this scanning site was too weak to be learned by the classifier.

For the multisite data scenario, we expected an increase of accuracy relative to using data from a single site in the training phase; however, naively merging the dataset into a single training set proved to be problematic. The accuracy went down, even when the size of the training set was double the size in the single site scenario. We attribute the decrease in the performance to the batch effects – *i.e.*, $P(X, Y | a) \neq P(X, Y | b)$ for two scanning sites a and b , where this difference is due to technical factors that confound the biological signal. The difference is such that, if we set as the target variable the scanning site, it is possible to achieve accuracies $> 88\%$ for the problem of binary site classification.

Multiple association studies have reported that the scanning site is a minor problem in neuroimaging, and that it does not have statistically significant interaction with the target variable (schizophrenia versus healthy control). In this work, we empirically show that this is not the case for prediction studies. For such prediction studies, the batch effect matters, and has a huge influence in the performance of the learning algorithms. We highlighted some important differences between association studies and prediction studies, and argued that not having a statistically significant effect does not mean that it is irrelevant. The empirical results that we present are consistent with this claim.

We empirically showed that if the fMRI scans of the same n participants are acquired at two sites a and b – *i.e.*, we have pairs (X_i^a, X_i^b) , $i = 1, \dots, n$ – then it is possible to learn a function $\hat{X}_i^b = f(X_i^a)$ that decreases the impact of the batch effects. We successfully used neural networks, with an architecture similar to autoencoders, in the traveling subject dataset and achieved an accuracy of 100% in the task of subject identification, but only 65% for the task of scanning site identification – as desired.

For the case of the schizophrenia dataset, in which every participant was

scanned only at a single site, we attempted to decrease the influence of the batch effects in the performance of the classifier using techniques that correct for translation, rotation or scaling in the data, such as z-score normalization and whitening. The negative results obtained suggest that batch effects are a more complex phenomenon than these simple linear transformations. We then tried BECCA, an algorithm that has successfully been used for batch effect corrections in microarray data by removing the data that is not correlated between two sites. However, it did not improve the accuracy in schizophrenia diagnosis using fMRI data. Functional connectivity has been reported to have high variance between subjects, even in studies using only healthy people. This high variance, in addition to the low number of instances, can be problematic for the estimation of the covariance matrices of the data, whose correct estimation are essential for BECCA to work. Finally, we used stacked autoencoders and bishifting autoencoders. Both have been successful in the problem of domain adaptation in image classification tasks, but were not successful in fMRI data using functional connectivity as features. Researchers in the community that use deep learning approaches agree that the number of training instances, and structure in the data, are important factors for the success of these approaches; unfortunately, we have a very small number of instances (relative to the number of instances available for imaging tasks) and it is not clear what kind of structure is present in a functional connectivity matrix. These negative results suggest that batch effects is not a trivial problem, and that more information about their nature is required in order to successfully apply machine learning algorithms to fMRI data.

These results also indicate that special careful should be taken when using any these methodologies for doing association studies. Note that, in absence of the prediction task of schizophrenia versus healthy controls, someone might incorrectly assume that these procedures are enough for removing the effects introduced by using data extracted from different scanning sites. It is true that after applying these methodologies, the data has more common characteristics: same mean and standard deviation for z-score and whitening, a higher correlation between batches after using BECCA, or a set of common high-level

features after using approaches based on autoencoders; however, the different datasets are still *not compatible* in the sense that a classifier learned from sites A and B will be less accurate in predicting the class of new instances from site A , versus a classifier learned using data only from site A , even if the size of the first training set is twice as large of for the second.

6.1 What is next?

The work presented in this dissertation is a first step towards directly removing the batch effect in fMRI data for classification purposes. The negative results obtained are an indication that further analysis is needed to better understand this problem. The next steps in our research in this field includes explicitly including in the model the different sources of noise that are known to distort the fMRI signals, such as the magnetic field inhomogeneity, or the drift in the time signals. These parameters can be estimated with the use of phantoms [21]. A second approach would be the create a classifier for a single scanning site, but using as priors the information learned from the rest of the scanning sites. Some groups are already researching the application of multitask learning to achieve this objective, and they are showing promising results, suggesting that is a direction worth pursuing [65].

6.2 Highlights

This dissertation focused on the challenges of applying machine learning algorithms to multisite fMRI data. Its main contributions were:

- It offered empirical evidence that batch effects is a problem that negatively impacts the task of schizophrenia diagnosis using fMRI data and machine learning techniques. Many association studies report that interscanner variability does not interfere with their analysis, which is not consistent with our results. We highlighted the differences between association studies and prediction studies (like ours), and argue that batch effects play an important role in the latter.

- It showed that z-score normalization or whitening are sufficient to solve batch effect problems caused by translations and scaling, or rotation and scaling, respectively. Our experiments suggest that batch effects for the diagnosis of schizophrenia go beyond these linear transformations. It also empirically showed that a neural network can solve the batch effects problems when the same participants are scanned in different scanning sites, achieving an accuracy of 100% in subject identification, and near chance accuracy in scanning site classification, as desired.
- It empirically showed that BECCA, stacked autoencoders, and bishifting autoencoders, which have been successful in similar problems in different fields, cannot be directly applied to solve our batch effects problem. This strongly suggest that more research effort is needed to solve this problem that is usually overlooked by most of the recent studies that use multi-site fMRI data.
- It empirically showed that, using features related to functional connectivity, it is possible to learn a classifier that distinguishes between people with schizophrenia and healthy controls with 70% accuracy using data extracted from a single site. Incorporating domain specific knowledge, like parcellating the brain in 264 regions of interest, and limiting the analysis to brain networks known to be related to schizophrenia improved the accuracy of the classifiers and decreased the computational cost of the learning algorithm relative to not using domain specific knowledge.

Bibliography

- [1] The adhd-200 sample http://fcon_1000.projects.nitrc.org/indi/adhd200/. [Online; accessed 17-March-2015].
- [2] Mohammad R Arbabshirani, Kent A Kiehl, Godfrey D Pearlson, and Vince D Calhoun. Classification of schizophrenia patients based on resting-state functional network connectivity. *Frontiers In Neuroscience*, 7:133, 2013.
- [3] F Gregory Ashby. *Statistical Analysis of fMRI Data*. The MIT Press, 2011.
- [4] Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, Dan Nolan, Edward Bryant, Tucker Hartley, Owen Footer, James M Bjork, Russ Poldrack, Steve Smith, Heidi Johansen-Berg, Abraham Z Snyder, and David C Van Essen. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169 – 189, 2013.
- [5] Yoshua Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.
- [6] Matthew R G Brown, Gagan S Sidhu, Russell Greiner, Nasimeh Asgarian, Meysam Bastani, Peter H Silverstone, Andrew J Greenshaw, and Serdar M Dursun. ADHD-200 global competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Frontiers In Systems Neuroscience*, 6:69, 2012.
- [7] Vince D Calhoun, Tom Eichele, and Godfrey Pearlson. Functional brain networks in schizophrenia: a review. *Frontiers In Human Neuroscience*, 3:17, 2009.
- [8] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- [9] Jiayu Chen, Jingyu Liu, Vince D Calhoun, Alejandro Arias-Vasquez, Marcel P Zwiers, Cota Navin Gupta, Barbara Franke, and Jessica A Turner. Exploration of scanning effects in multi-site structural mri studies. *Journal Of Neuroscience Methods*, 230:37 – 50, 2014.
- [10] David Cherney, Tom Denton, Rohit Thomas, and Andrew Waldron. *Linear algebra*. Freely available at: <https://www.math.ucdavis.edu/~linear/>, 2016.

- [11] Christian Dansereau, Yassine Benhajali, Celine Risterucci, Emilio Merlo Pich, Pierre Orban, Douglas Arnold, and Pierre Bellec. Statistical power and measurement bias in multisite resting-state fMRI connectivity. 2016.
- [12] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection for microarray gene expression data. *Journal of Bioinformatics & Computational Biology*, 3(2):185 – 205, 2005.
- [13] J R Edwards and R P Bagozzi. On the nature and direction of relationships between constructs and measures. *Psychological methods*, 5(2):155 – 174, 2000.
- [14] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664 – 1671, 2015.
- [15] Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, and D Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313 – 327, 2011.
- [16] Society for neuroscience. *Brain Facts. A primer on the brain and nervous system*. Available online at: <http://www.brainfacts.org/book>, 2012.
- [17] Lee Friedman and Gary H Glover. Report on a multicenter fMRI quality assurance protocol. *Journal Of Magnetic Resonance Imaging: JMRI*, 23(6):827 – 839, 2006.
- [18] Lee Friedman, Gary H. Glover, Diana Krenz, and Vince Magnotta. Reducing inter-scanner variability of activation in a multicenter fMRI study: Role of smoothness equalization. *NeuroImage*, 32(4):1656 – 1668, 2006.
- [19] Lee Friedman, Hal Stern, Gregory G Brown, Daniel H Mathalon, Jessica Turner, Gary H Glover, Randy L Gollub, John Lauriello, Kelvin O Lim, Tyrone Cannon, Douglas N Greve, Henry Jeremy Bockholt, Ayse-nil Belger, Bryon Mueller, Michael J Doty, Jianchun He, William Wells, Padhraic Smyth, Steve Pieper, Seyoung Kim, Marek Kubicki, Mark Vangel, and Steven G Potkin. Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping*, 29(8):958 – 972, 2008.
- [20] K J Friston, J Ashburner, S J Kiebel, T E Nichols, and W D Penny, editors. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.
- [21] Douglas N Greve, Gregory G Brown, Bryon A Mueller, Gary Glover, and Thomas T Liu. A survey of the sources of noise in fMRI. *Psychometrika*, 78(3):396 – 416, 2013.
- [22] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009.
- [23] Stefan Haufe, Frank Meinecke, Kai Grgen, Sven Dhne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96 – 110, 2014.

- [24] Simon Haykin. *Neural networks and learning machines*. Pearson Education Inc, Upper Saddle River, third edition edition, 2009.
- [25] David J Heeger and David Ress. What does fmri tell us about neuronal activity?. *Nature Reviews Neuroscience*, 3(2):142 – 151, 2002.
- [26] Helene Hjelmervik, Markus Hausmann, Berge Osnes, Ren Westerhausen, and Karsten Specht. Resting states are resting traits an fmri study of sex differences and menstrual cycle effects in resting state cognitive control networks. *PLoS ONE*, 9(7):1 – 10, 2014.
- [27] Wolfgang Huf, Klaudius Kalcher, Roland N Boubela, Georg Rath, Andreas Vecsei, Peter Filzmoser, and Ewald Moser. On the generalizability of resting-state fMRI machine learning classifiers. *Frontiers In Human Neuroscience*, 8:502, 2014.
- [28] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *NeuroImage*, 62(2):782 – 790, 2012.
- [29] Daphna Joel, Zohar Berman, Ido Tavor, Nadav Wexler, Olga Gaber, Yaniv Stein, Nisan Shefi, Jared Pool, Sebastian Urchs, Daniel S Margulies, Franziskus Liem, Jrgen Hnggi, Lutz Jncke, and Yaniv Assaf. Sex beyond the genitalia: The human brain mosaic. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 112(50):15468 – 15473, 2015.
- [30] Meina Kan, Shiguang Shan, and Xilin Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [31] David B. Keator, Theo G.M. van Erp, Jessica A. Turner, Gary H. Glover, Bryon A. Mueller, Thomas T. Liu, James T. Voyvodic, Jerod Rasmussen, Vince D. Calhoun, Hyo Jong Lee, Arthur W. Toga, Sarah McEwen, Judith M. Ford, Daniel H. Mathalon, Michele Diaz, Daniel S. O’Leary, H. Jeremy Bockholt, Syam Gadde, Adrian Preda, Cynthia G. Wible, Hal S. Stern, Aysenil Belger, Gregory McCarthy, Burak Ozyurt, and Steven G. Potkin. The function biomedical informatics research network data repository. *NeuroImage*, 124, Part B:1074 – 1079, 2016. Sharing the wealth: Brain Imaging Repositories in 2015.
- [32] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. 2015.
- [33] Sheehan Khan. *The Budgeted Biomarker Discovery Problem*. PhD thesis, 2015.
- [34] Ali Khazaei, Ata Ebrahimzadeh, and Abbas Babajani-Feremi. Application of advanced machine learning methods on resting-state fmri network for identification of mild cognitive impairment and alzheimer’s disease. *Brain Imaging and Behavior*, 10(3):799–817, 2016.
- [35] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.
- [36] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

- [37] Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y. Weiss-Sols, Robin Duque, Hugues Bersini, and Ann Now. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics*, 14(4):469 – 490, 2013.
- [38] V A Magnotta and L Friedman. Measurement of signal-to-noise and contrast-to-noise in the fBIRN multicenter imaging study. *Journal of Digital Imaging*, 19(2):140 – 147, 2006.
- [39] T Medkour, A T Walden, and A Burgess. Graphical modelling for brain connectivity via partial coherence. *Journal Of Neuroscience Methods*, 180(2):374 – 383, 2009.
- [40] Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian J. Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, Pascal Vincent, Aaron Courville, and James Bergstra. Unsupervised and transfer learning challenge: a deep learning approach. In Isabelle Guyon, G. Dror, V Lemaire, G. Taylor, and D. Silver, editors, *JMLR W&CP: Proceedings of the Unsupervised and Transfer Learning challenge and workshop*, volume 27, pages 97–110, 2012.
- [41] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [42] Andrew Ng, Jiquan Ngiam, Chuan Y. Foo, Yifan Mai, and Caroline Suen. UFLDL Tutorial. http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial, 2010.
- [43] Emanuele Olivetti, Susanne Greiner, and Paolo Avesani. ADHD diagnosis from multiple data sources with batch effects. *Frontiers In Systems Neuroscience*, 6:70, 2012.
- [44] Vani Pariyadath, Elliot A Stein, and Thomas J Ross. Machine learning classification of resting state functional connectivity predicts smoking status. *Frontiers In Human Neuroscience*, 8:425, 2014.
- [45] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and Steven E Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665 – 678, 2011.
- [46] R Raina, A Battle, H Lee, B Packer, and A Y Ng. Self-taught learning: Transfer learning from unlabeled data. In *Machine learning -International workshop then conference*, International conference on machine learning (ICML 2007), pages 759 – 766, 2007.
- [47] J Richiardi, S Achard, H Bunke, and D Van De Ville. Machine learning with brain graphs: Predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal Processing Magazine*, 30(3):58 – 70, 2013.
- [48] Irina Rish and Genady Ya Grabarnik. *Sparse modeling : theory, algorithms, and applications*. Chapman & Hall/CRC machine learning & pattern recognition series. Boca Raton, FL : CRC Press : Taylor & Francis Group, 2015., 2015.

- [49] Cristina Rosazza and Ludovico Minati. Resting-state brain networks: literature review and clinical applications. *Neurological Sciences*, 32(5):773 – 785, 2011.
- [50] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059 – 1069, 2010.
- [51] Raymond Salvador, John Suckling, Christian Schwarzbauer, and Ed Bullmore. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions: Biological Sciences*, (1457):937, 2005.
- [52] Theodore D Satterthwaite, Daniel H Wolf, David R Roalf, Kosha Ruparel, Guray Erus, Simon Vandekar, Efstathios D Gennatas, Mark A Elliott, Alex Smith, Hakon Hakonarson, Ragini Verma, Christos Davatzikos, Raquel E Gur, and Ruben C Gur. Linked sex differences in cognition and functional connectivity in youth. *Cerebral Cortex (New York, N.Y.: 1991)*, 25(9):2383 – 2394, 2015.
- [53] J M Segall, J A Turner, T G M van Erp, T White, H J Bockholt, R L Gollub, B C Ho, V Magnotta, R E Jung, R W McCarley, S C Schulz, J Lauriello, V P Clark, J T Voyvodic, M T Diaz, and V D Calhoun. Voxel-based morphometric multisite collaborative study on schizophrenia. *Schizophrenia Bulletin*, 35(1):82 – 95, 2009.
- [54] Galit Shmueli. To explain or to predict?. *Statistical Science*, (3):289, 2010.
- [55] Robert H. Shumway and David S. Stoffer. *Time series analysis and its applications : with R examples*. Springer texts in statistics. Springer, New York, 2011.
- [56] N B Smith and A Webb. *Introduction to Medical Imaging: Physics, Engineering and Clinical Applications*. Cambridge Texts in Biomedical Engineering. Cambridge University Press, 2010.
- [57] Cynthia M Stonnington, Geoffrey Tan, Stefan Klppel, Carlton Chu, Bogdan Draganski, Jr Jack, Clifford R, Kewei Chen, John Ashburner, and Richard S J Frackowiak. Interpreting scan data acquired from multiple scanners: a study with alzheimer’s disease. *Neuroimage*, 39(3):1180 – 1185, 2008.
- [58] Felice T Sun, Lee M Miller, and Mark D’Esposito. Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *NeuroImage*, 21(2):647, 2004.
- [59] Lixia Tian, Jinhui Wang, Chaogan Yan, and Yong He. Hemisphere- and gender-related differences in small-world brain networks: A resting-state functional mri study. *NeuroImage*, 54(1):191 – 202, 2011.
- [60] Jessica A Turner, Eswar Damaraju, Theo G M van Erp, Daniel H Mathalon, Judith M Ford, James Voyvodic, Bryon A Mueller, Aysenil Belger, Juan Bustillo, Sarah McEwen, Steven G Potkin, Fbirn, and Vince D Calhoun. A multi-site resting state fMRI study on the amplitude of low frequency fluctuations in schizophrenia. *Frontiers In Neuroscience*, 7:137, 2013.

- [61] Saman Vaisipour. *Detecting, correcting, and preventing the batch effects in multi-site data, with a focus on gene expression Microarrays*. PhD thesis, 2014.
- [62] Svyatoslav Vergun, Alok S Deshpande, Timothy B Meier, Jie Song, Dana L Tudorascu, Veena A Nair, Vikas Singh, Bharat B Biswal, M Elizabeth Meyerand, Rasmus M Birn, and Vivek Prabhakaran. Characterizing functional connectivity differences in aging adults using machine learning on resting state fmri data. *Frontiers In Computational Neuroscience*, 7:38, 2013.
- [63] Henrik Walter, Angela Ciaramidaro, Mauro Adenzato, Nenad Vasic, Rita Bianca Ardito, Susanne Erk, and Bruno G. Bara. Dysfunction of the social brain in schizophrenia is modulated by intention type: An fMRI study. *Social Cognitive & Affective Neuroscience*, 4(2):166 – 176, 2009.
- [64] Lubin Wang, Hui Shen, Feng Tang, Yufeng Zang, and Dewen Hu. Combined structural and resting-state functional mri analysis of sexual dimorphism in the young adult human brain: an mvpa approach. *Neuroimage*, 61(4):931 – 940, 2012.
- [65] Takanori Watanabe, Daniel Kessler, Clayton Scott, and Chandra Sripada. Multisite disease classification with functional connectomes via multitask structured sparse SVM, year = 2014,. *Sparsity Techniques in Medical Imaging (STMI)*.
- [66] Kai Wu, Yasuyuki Taki, Kazunori Sato, Hiroshi Hashizume, Yuko Sassa, Hikaru Takeuchi, Benjamin Thyreau, Yong He, Alan C. Evans, Xiaobo Li, Ryuta Kawashima, and Hiroshi Fukuda. Topological organization of functional brain networks in healthy children: Differences in relation to age, sex, and intelligence. *PLoS ONE*, 8(2):1 – 14, 2013.
- [67] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: A network visualization tool for human brain connectomics. *PLoS ONE*, 8(7):1 – 15, 2013.
- [68] Chunsheng Xu, Chuanfu Li, Hongli Wu, Yuanyuan Wu, Sheng Hu, Yifang Zhu, Wei Zhang, Linying Wang, Senhua Zhu, Junping Liu, Qingping Zhang, Jun Yang, and Xiaochu Zhang. Gender differences in cerebral regional homogeneity of adult healthy volunteers: A resting-state fmri study. *BioMed Research International*, 2015:1 – 8, 2015.
- [69] Noriaki Yahata, Jun Morimoto, Ryuichiro Hashimoto, Giuseppe Lisi, Kazuhisa Shibata, Yuki Kawakubo, Hitoshi Kuwabara, Miho Kuroda, Takashi Yamada, Fukuda Megumi, Hiroshi Imamizu, Sr Nez, Jos E, Hidehiko Takahashi, Yasumasa Okamoto, Kiyoto Kasai, Nobumasa Kato, Yuka Sasaki, Takeo Watanabe, and Mitsuo Kawato. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature Communications*, 7:11254, 2016.
- [70] L Yu and H Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Machine learning -International workshop then conference-*, volume 2 of *International conference machine learning*, pages 856 – 863, 2003.

- [71] Andrew Zalesky, Alex Fornito, and Ed Bullmore. On the use of correlation as a measure of network connectivity. *Neuroimage*, 60(4):2096 – 2106, 2012.
- [72] Yufeng Zang, Tianzi Jiang, Yingli Lu, Yong He, and Lixia Tian. Regional homogeneity approach to fmri data analysis. *Neuroimage*, 22(1):394 – 400, 2004.
- [73] Kelly H Zou, Douglas N Greve, Meng Wang, Steven D Pieper, Simon K Warfield, Nathan S White, Sanjay Manandhar, Gregory G Brown, Mark G Vangel, Ron Kikinis, and 3rd Wells, William M. Reproducibility of functional mr imaging: preliminary results of prospective multi-institutional study performed by biomedical informatics research network. *Radiology*, 237(3):781 – 789, 2005.

Appendices

Appendix A

Male versus female classification

Besides the *large n, small p* problem that is present in the analysis of most fMRI datasets, and the problem of batch effects in multi-site analysis, there is an extra layer of complexity that complicates the task of creating an automatic tool for diagnosis of schizophrenia: the reliability of the labels used for learning the classifier. As discussed on Chapter 3, supervised learning techniques apply a learning algorithm L to a labeled dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ to learn a function, $\hat{f} = L(D)$, in order to make predictions $\hat{y}_{new} = \hat{f}(x_{new})$, for x_{new} instances. In consequence, if the labels $\{y_1, \dots, y_n\}$ are misleading –i.e. some people with schizophrenia are erroneously labeled as healthy control (or vice versa), the learning algorithm might not find the patterns that distinguishes one group from the other. Unfortunately, this might be the case for the case of diagnosis in psychiatry. Since there is no standard biologically-based clinical test yet [2] for the identification of mental diseases, different psychiatrists might make different diagnoses.

In order to test the performance of the learning algorithm without the ambiguity of the labels, we learned a classifier to distinguish between males and females. Unlike the case of diagnosis of schizophrenia, for the task of sex classification Gaussian Markov Random Fields achieved a better accuracy than Support Vector Machines. Surprisingly, this task was more difficult than expected, and the accuracy achieved was just slightly above chance level (see Table A.1 in section A.3). Despite that many research articles report important differences in the structure and function of the brain in males and females [26, 66, 68, 59],

many of these studies are association studies. This means that they find group differences, which are not necessarily predictive [29]. A few studies report near 70 % accuracy in sex classification in large, single site, studies [52, 64]; however, we were unable to replicate their results in our data. Besides, we noticed that using multi-site data did not increase the accuracy relative to using data from only a single site. These results motivated the research presented in this dissertation. The rest of this appendix presents the techniques used and the experiments performed for the task of sex classification.

A.1 Gaussian Markov Random Fields

A Gaussian Markov Random Field (GMRF) is similar to Markov Random Fields, with the difference that now the random variables take continuous values [35]. A multivariate Gaussian distribution over the random variables X_1, X_2, \dots, X_n is parametrized by an n dimensional mean vector μ and an $n \times n$ covariance matrix Σ . The density function is then defined as:

$$p(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$

Any Gaussian distribution can be represented as a pairwise Markov Network with quadratic node and edge potentials. These kind of networks are known as Gaussian Markov Random Fields [35]. One advantage of these models is that two Gaussian variables, X_i and X_j are conditionally independent given the rest of the variables if and only if their corresponding entries $\Sigma_{i,j}^{-1} = \Sigma_{j,i}^{-1} = 0$ [36]. Therefore, learning the structure of a GMRF reduces to the problem of finding zero entries on the inverse of the covariance matrix Σ^{-1} , which is also known as the precision matrix $\Omega = \Sigma^{-1}$. Using this approach, we generated a functional connectivity graph (parameterized by Ω) for males (Ω_m) and another for females (Ω_f). Given the fMRI scan of a new participant, x_{new} , we computed $P(x_{new} | \Omega_m)$ and $P(x_{new} | \Omega_f)$, and assigned x_{new} to the class with highest probability.

A common approach when finding probabilistic graphical models is to chose

the simplest one that adequately explains the data [48]. One metric that can be used for evaluating the model fit is the likelihood function, which measures the probability of the data given the model. Since a fully connected model will output the highest likelihood function on the training data, a regularization term for penalizing complex models is required.

For a dataset D with n independent and identically distributed samples x_1, x_2, \dots, x_n , where each variable in x_i follows a Gaussian distribution with zero mean, its log-likelihood function can be expressed as:

$$L(D) = \frac{n}{2} [\log(|\Omega|) - \text{tr}(A\Omega)] + \text{const} \tag{A.1}$$

$$A = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$$

Since the constant term is independent of the mean and covariance, it can be ignored in the optimization function. The objective is then to find the model that maximizes the penalized likelihood function with a restriction in the number of parameters, which is known as the *sparse inverse covariance selection problem*:

$$\max_{\Omega} \log(|\Omega|) - \text{tr}(X^T X \Omega) - \lambda \|\Omega\|_1 \tag{A.2}$$

where X is an $n \times m$ matrix. n is the number of instances in the dataset, m is the number of random variables, λ is the regularization term, and $\|\Omega\|_1$ is the l_1 norm of the precision matrix. For our specific task of sex classification using fMRI data, $m = 264$ is the number of regions of interest, while n is the number of timepoints in the time series of a particular region multiplied by the number of participants included in the training set.

A.2 Learning the models

Our methodology involves steps depicted in Figure A.1. In general terms, it involves creating a graphical model for every class that we are interested in classifying. Then, for every subject to be classified, we will compute the

likelihood of the data given a particular model. The subject will be classified with the class whose model has the highest likelihood.

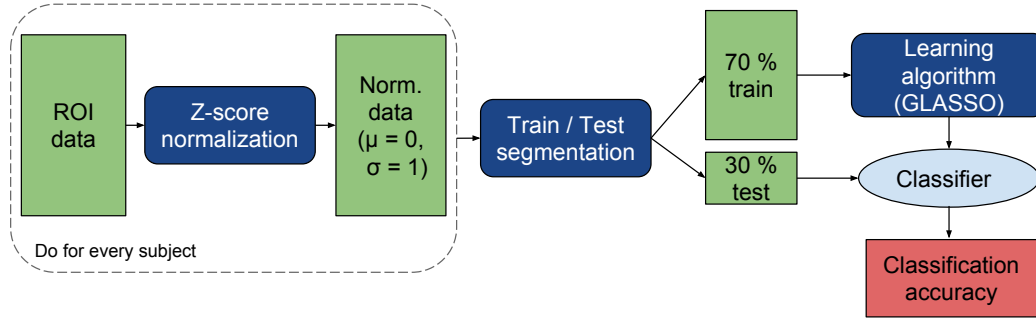


Figure A.1: Our methodology for classifying the resting-state fMRI scans as male or female.

1. *Extract regions of interest.* Since every time series consists of 91 time points, every subject is represented by a matrix, x , of 91×264 .
2. *Normalize data.* We normalized the data of every region of interest of every subject independently. Every time series (column of x) was normalized to mean $\mu = 0$ and standard deviation $\sigma = 1$.
3. *Separate data into train and test set.* Separate the dataset in groups of interest. For this particular case the classes are male or female. Then randomly select the data of 70% of the subjects in each class for training and use the remaining 30% for testing purposes.
4. *Concatenate time series.* Steps 5 and 6 are implemented using 5-fold cross validation and the training set from the last step. Using the 4 subsets of each round, concatenate the time series corresponding to the same region of interest. This will result in a matrix X of $n * 91 \times 264$ per class, where n is the number of subjects present in the subsets.
5. *Construct the model.* Using Eq. A.2, construct a model for each class. Then, test the performance of the model on the remaining subset of data (of the cross validation process). Performance was measured as

the percentage of samples correctly classified on the subset not used for training. To classify a new instance x we simply compute the likelihood of x given each model using Eq. A.1, and assign x to the class whose model gave the highest likelihood. Since the constant term is independent of the model, it can be ignored for the classification step. We experimented with the following values for the regularization term: $\lambda = [0.001, 0.003, 0.01, 0.03, 0.1, 0.3]$, and selected the value of λ with the highest (internal) cross-validation accuracy.

6. *Classify new data.* After selecting the best model for each class (steps 5 and 6), test the performance of the model on the test set created on step 4. Report the accuracy on this test set.

A.3 Dataset and results

We used the ADHD-200 [1] dataset for our experiments. It contains 973 resting-state fMRI scans from both healthy controls and people diagnosed with attention deficit and hyperactivity disorder (ADHD), collected across 8 independent imaging sites. The age range of the entire sample is 7-21 years. The ADHD-200 dataset was preprocessed for Brown et al. [6], and we used these preprocessed data for our experiments.

Out of the 973 scans, we used only the ones corresponding to healthy subjects. We performed two sets of experiments, one using single-site data (using only the data from each of the 3 scanning sites with the higher number of instances), and another one using multi-site data, using data from all the healthy subjects across the 8 scanning sites. Table A.1 shows the results of these experiments in the hold out set, as well of the number of instances available for every class. From the total of instances, 70% were used for training purposes. The remaining 30% formed the hold-out set. We repeated the experiments 30 times, with different 70/30 splits, and the reported accuracy is the average over the 30 experiments. As can be seen, the accuracy is slightly above chance level (baseline) for all the cases, but the improvement is very small. It was also surprising to see that the accuracy did not improved by a large margin in the

Table A.1: Accuracy of Gaussian Markov Random Fields in the task of male versus female classification using the ADHD dataset (healthy controls)

| Data | # Male | # Female | Baseline | Accuracy |
|------------|--------|----------|----------|----------|
| Site 1 | 84 | 59 | 58.6% | 64.7% |
| Site 5 | 55 | 53 | 50.9% | 58.3% |
| Site 7 | 48 | 42 | 55.5% | 66.9% |
| Multi-site | 301 | 259 | 54.0% | 61.2% |

multi-site scenario, even when the amount of the available data is more than triple than in single site experiments. These results motivated the research about batch effects presented in this dissertation.

Appendix B

Additional approaches

Section 3 discussed how we extracted features from the fMRI datasets using functional connectivity, which was estimated by taking the pairwise correlation between the 264 regions of interest defined by Power et al. [45]. After extracting these features, we used support vector machines to learn a classifier that produced the results shown in Section 4. In addition to the results reported in the main body of this dissertation, we also implemented several approaches to learn a classifier that could distinguish between people with schizophrenia and healthy controls, or between males and females (in the case of the sex classification problem). Since this approaches achieved a lower accuracy than the one achieved using pairwise correlation, they are not included in the main text; however, we list them in this appendix along with references for the interested reader:

- Adding site as feature: This is the most naive approach to try to solve the batch effects problem; however, the problem was not solved by including scanning site as one of the features.
- Gaussian Markov Random Fields: This approach was explained in Section A.1. In this approach, every point in a time series is considered as an independent and identically distributed instance, which is evidently not true. We removed the temporal autocorrelation in an effort to reduce the impact of this deviation of the assumption.
- Fast Fourier Transform: We concatenated the power spectrum of the

time series corresponding to the 264 regions of interest into a single vector for every patient. Details about the Fast Fourier Transform can be consulted in the book of Shumway and Stoffer [55].

- Coherence: Similar to correlation, coherence is a measure of linear relationship between two time series, with the difference that coherence focuses on the frequency domain, while correlation does it on the time domain [3].
- Regional homogeneity: This approach consists in using a metric, such as the Kendall's coefficient concordance, to measure the similarity of a given voxel with its nearest neighbors. This similarity is condensed into a single number per every voxel, resulting in a regional homogeneity map that can be used as a feature for further classification [72]
- Partial correlation/coherence: In some cases, two random variables X and Y are correlated because of a common third variables Z ; however, they might be uncorrelated given Z . Pairwise correlation (coherence) is unable to distinguish this particular scenario. Partial correlation (coherence) identifies an additional lineal association between X and Y after removing the effects of Z [58, 39].
- Graph statistics: After computing a functional connectivity graph for every participant in the experiment, it is possible to extract some additional measures of brain connectivity by analyzing the graph. Rubinov and Sporns provide a good overview of some of the relevant graph metrics for neuroscience [50]. We used node strength and node degree in our experiments.
- Feature selection: In addition to the feature selection described in this dissertation, we also used the feature selection algorithms: minimum redundancy maximum relevance (MRMR) [12] and fast correlation based filter (FCBF) [70].