Abstract

Graphs are a common way to represent multiple interactions between a set of agents, and such nature makes them suitable for various context. In particular, recently several research fields found useful to approach their problems via graphs, and this had a positive impact on the availability of graphs to analyse. Given a dataset in which the observations are graphs, it becomes useful to propose methodologies able to deal with such input, that can solve the classical tasks of machine learning. This study investigates the problem of applying regression models to graph structured input data and, in particular, predict intelligence scores from brain network connectivity analysis. This research field implies dealing with different problems: represent brain-network connections as graphs, substructure extraction and graph embedding and, finally, regression. In the first part of this work we provide an introduction to the state-of-the-art and a focus on the methods we take as reference. Specifically we use two recent papers based on functional analysis of brain networks: *Explainable Classification of Brain Networks via Contrast Subgraphs* proposed by Lanciano et al. [34] in 2020 and *Predicting full-scale and verbal intelligence scores from functional Connectomic data in individuals with autism Spectrum disorder* by Dryburgh et al. [22] (2019). In the former the authors propose a method of feature extraction based on a contrast subgraph algorithm that compares coupled sets of observations estimating subnetworks which are dense in one group and sparse for the other and viceversa, before deriving features used for the classification task. While, the latter method, consists in extracting substructure from the whole brain networks using significant correlations between brain connections and the target intelligence score and separately modelling the positive and negative correlation in a regression framework.

In the second part of this Master's thesis we firstly provide an application using the chosen approaches, adapting contrast subgraph for a regression problem framework, and then we propose some novelties based on the existing methods, still with an example on the same data set. Specifically, we use a real world dataset released by the Autism Brain Imagine Data Exchange (ABIDE) project [19], containing fMRI of patients having different age, sex, diagnostic category and other clinical variables. After adapting the method which was originally designed for a classification task in a regression-designed fashion, we train and test both state-of-the-art methods using a 5-fold cross-validated linear regression model. The proposed contributions consist in an extension of the study of the parameter space related to the contrast subgraph problem answering to diverse research questions. We also merge the features estimated using Connectome-based Predictive Model [22] and the new fashioned contrast subgraph based features. This last setting allows us to design a new features set that outperforms the existing methods in the prediction performance of the regression model.

Contents

Chapter 1

Introduction

In the last years in many fields such as finance, social media, biology, transportation systems, etc.. graph structured data have been widely used. This kind of data organization is extremely useful to understand connections in a network as well as to describe complex phenomenons in a more accessible way. Consequently, there are several different research developments in graph theory. Recent works on representation learning for graph structured data focus on learning distributed representations of graph substructures such as nodes and subgraphs [43]. The study of significant substructures in networks is crucial in the recent research employing different approaches of features selection or extraction. Another important field related to this data representation is graph embedding, namely the transformation of graphs into low dimensional vectors that preserve graph structures. This area is in turn framed on different approaches that are for example based on graph kernel, matrix factorization, edge reconstruction, deep learning or generative models [15]. The transformation of graphs into vectors allows researchers to reduce the complexity of wide networks, maintaining their features, and to use them as input for solving machine learning problems.
Therefore, the core idea of this work lies in the following statement: *"Is it possible to solve a regression task, given in input a dataset of graphs?"*
In order to answer, we devise a promising application that can test our intuition: brain networks. Brain networks are the result of a recent approach born in the context of neuroscience, for which a human brain can be mapped to a network, that brought researchers to make further findings with respect to those obtained with "traditional" neuroscience research activity. They are able to represent the map of the relationships between different regions of the brain, and consequently of the human behaviour. Building methods that are able to operate in this context is a great opportunity, that can support different research lines. In particular, solving a regression task in this context, would mean to be able to detect specific regions of the brain, responsible for the behaviour of a specific characteristic of the human brain. Specifically in the first part we apply two recently published methods and in the second we present a new approach of feature extraction enlarging the previously described methods. As regards the two recent methods we have: the former, proposed by Lanciano et al. in 2020 [34], that deals with an explainable classification problem using contrast subgraph to extract meaningful graph substructure in order to classify subjects affected by Autism Spectrum Disorder and Typically Developed. While in the latter, published by Dryburgh et al. in 2019 [22], the authors present an already existing connectome-based predictive model [49] to predict full-scale and verbal intelligence scores in individuals affected by Autism Spectrum Disorder.
Our purpose is to find an easily explainable method to extract relevant features from

brain networks in order to model the selected features and hence to estimate and predict target clinical variables such as, in our particular case study, individuals' IQ. Even if the method proposed by Lanciano et al. [34] is applied in a context of classification instead of regression, it is extremely explainable and flexible to be applied for our purposes and, for this reason, it is also our starting point for development and contributions we propose in the second part of this master thesis. Differently the study published by Dryburgh et al. [22] is already an application of predictive regression models to brain networks using different IQ scores as dependent variable and it also can be taken as an interesting reference for the proposed novelty.
Both existing methods and new contributions are provided with applications on a case study dataset provided by the Autism Brain Imagine Data Exchange (ABIDE) project [19] (more details are given in chapter 3.1).

This work was performed during a research internship at ISI Foundation under the supervision of Dr. Francesco Bonchi.

1.1 Background

First of all we need to give a brief introduction of the basics of the main concepts we deal with in this work, namely graphs and regression.
A graph $G(V, E)$ is defined over a set of nodes V and a set of edges E that are paired elements of V. Specifically $E \subseteq \{(s, v) \mid (s, v) \in V \text{ and } s \neq v\}$, where the vertices s and v are the endpoints of the edge (s, v) that is said to be incident to s and v. This kind of graph is called undirected graph, that differs from the directed graph in which edges have orientation. In this case a graph $G(V, E)$ is defined over a set of nodes V and a set of ordered pairs of edges, E.
A graph can be represented by an adjacency matrix A that is a squared matrix of dimensions $n \times n$, where n is equal to the cardinality of the set V, i.e. $n = |V|$. As we said, there is an edge between two vertices i and j if they are adjacent, in other words $(i, j) \in E$. Consequently, if this happens, $A_{ij} = 1$ while on the contrary $A_{ij} = 0$ if there is no connection between the two vertices. An undirected graph always has a symmetric adjacency matrix, that means $A_{ij} = A_{ji} \forall i, j \in V$.
In this summary we considered the general case of unweighted graph but the edges of any network can have different weights. However in this study we only deal with unweighted graphs.
Considering a subset of vertices $S \subseteq V$ we denote by $G[S]$ the subset of G induced by the set S, namely $G[S] = (S, E[S])$ where $E[S] = \{e = (u, v) \mid e \in E \text{ and } u, v \in S\}$.

Regression is used to model the effect of a set of explanatory variables, $x_1, x_2, ..., x_k$ on a variable y of primary interest. The set of variables are called *independent variables*, *covariates*, or *regressors*, while the variable of interest is called *dependent variable* or *response*. The various models differ for the types of y, that can be continuous, categorical, binary or counts; values that ca be assumed also by the covariates.
The main aspect of the regression model is that the relationship between the set of regressors and the variable of interest is not a deterministic function but it shows an error and this implies that the response is a random variable. We can only know the expected mean of y conditioned by the covariates, $E(y|x_1, x_2, ..., x_k)$. This can be written as follows:

$$y = E(y|x_1, ..., x_k) + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon.$$

The random deviation ϵ is called *error term*, *disturbance*, *random* or *stochastic component*. In regression analysis the data, $y_i, x_{i1}, ..., x_{ik}$ for $i = 1, ..., n$, is used to estimate the unknown parameters $\beta_1, ..., \beta_k$ and separate the systematic component from the stochastic one. The conditional mean of y is a combination of the covariates and, for all data, that means $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{ik} + \epsilon_i$, $i = 1, ..., n$. The most simple regression model is the linear regression model, where there are assumptions about the linearity of the relationship between regressors and dependent variable, about y that is preferred to be continuous and normally distributed and also about the errors that have to be independent and identically distributed (i.i.d.). More general models can be required for different response variables, i.e. binary variables, if the effect of covariates is nonlinear or if there are some heterogeneity that have to be taken into account [23]. However, we provide further details with respect to the applied models in the next sections.

1.2 Related works

In the recent years, starting from the mid 1990s, developments in the understanding of complex systems have led to the rise of network science. This growth is linked to the understanding that the behaviour of complex systems is produced by interactions among their constituent elements [14]. The wide range of available data sets has shown that different complex systems often share similar macroscopic behaviour. It is clear that the way which is chosen to analyse the connections in a network is crucial. Even if we focus our attention on studies based on graph theory there have been other recently realized methods to explore functional systems. These studies include mathematical models such as structural equation modelling [41][13], dynamic casual modelling [28] or Granger causality [10]. Their aim is to estimate "the causal influence that each element of a system exert on the behaviour of the other elements" ([14] p. 190).

1.2.1 Studies on brain networks connectivity

Depending on the type of network, the research has developed different approaches to detect and study the connection between elements. Brain connectivity can be described as: (i) structural, that denotes anatomical links, (ii) functional, that is statistical association or dependencies between elements, and (iii) effective, namely direct or causal relationship between elements [14]. The effective connectivity measures can be used to extract directed graphs that can be topologically described by graph theory. Functional network studies have been based on undirected graphs obtained from symmetrical statistical association between brain regions or simpler measures of functional connectivity. In general graph theory can be applied to an association matrix derived from either functional or effective connectivity measures, that can be used to extract undirected or directed graphs, respectively. But the neuroscientific studies up to date are based on functional connectivity [14] and in this work we focus on the study of functional connectivity in brain networks.

Studies on functional connectivity have been based on functional MRI (fMRI), electroencephalography (EEG), magnetoencephalography (MEG) or multielectrode array (MEA) data although functional connectivity based on resting state fMRI is one of the major techniques for brain functional studies. "Functional connectivity is defined as the temporal dependence of neuronal activity patterns of anatomically separated brain regions [1] [27] and studies have been shown the feasibility of examining functional connectivity between brain regions as the level of co-activation of functional MRI time-series measured during rest [39]" [54]. Biswal et al. [6] were the first demonstrating an high correlation between the fMRI BOLD time series of the left and right hemispheric regions of the primary motor network [6] [7]. After these pioneering results, several subsequent works showed high correlation not only between left and right hemispheric motor cortex but also between regions of "other known functional networks, like the primary visual network, auditory network and higher order cognitive networks ([18] [29] [17] [20] [21] [26] [39] [38] [53] [59])" [54]. These correlations between brain regions, as already said, can be represented with a graph structure, now we will see some of the state-of-the-Art techniques.

In the work proposed by Lanciano et al. [34] the authors compute pairwise Pearson correlation between the time series selecting them using a threshold corresponding to a certain percentile in the distribution of the correlations in absolute values. As already said, this approach is adopted also by Lord el al. [37] and Rubinov et al. [48]. There are other works in which positive and negative correlations are separated in order not to disregard relevant negative correlations ([22] [49] [25]). For example in Dryburgh et al. [22] they select the entries of the correlation matrices using the significance level of these matrices and the response variable. In this work the authors use a threshold with respect to the p-values as only way to select a subset of the functional connectivity and this is also a way to extract significant network substructures, however the solutions to the subnetworks selection problem are various. A widespread approach is to use different penalised regression. For example Beer at al. [4] propose a fused sparse group lasso with L_1 and L_2 norms to encourage structured and sparse solutions. A similar approach is also followed in Brown et al. [11] in which they proposed a regularized regression to identify anatomical subnetworks that are predictive of targeted clinical diagnoses or developmental outcomes, still using both absolute value and Euclidean norms. The use of the L_1 norm is seen also by Casanova et al. [16] in a LASSO regularized random forest classifier to find a sparse set of fMRI to classify adult males and female brains. In literature there are also examples of spatially regularized support vector machines to find a network substructure predictive of schizophrenia [56], a Laplacian-based regularization term which encourage subnetwork weights to smoothly vary between neighbour edges [36] [11], or the combination of an L_1 and Laplacian-based regularization term [30]. In Lanciano et al. [34] the problem of feature extraction is faced using a contrast subgraph technique that allows to extract substructures that are dense for some individuals, belonging to a class, and sparse for others, belonging to another class. The classes on which the contrast is based are the labels of the classification task.

1.2.2 Graph embeddings

Regarding the graph embedding issue there are several approaches to this problem. A possible and widely used procedure is matrix factorization. This is mainly used with graphs constructed from non-relational data for node embedding, typically watched as graph Laplacian eigenmap problems, and to embed homogeneous graphs [2] [45].

Deep learning based graph embedding applies deep learning models on graphs. These models are either a direct adoption from other fields or a new neural network model specifically designed for embedding graph data. The input is either paths sampled from a graph or the whole graph itself. Deep learning embedding can be based on random walk ([47] [24] [61]) or not ([12] [32] [33]), nonetheless both approaches are

made of numerous variants that are not object of our interest.
Edge reconstruction based optimization is another well known approach for graph embedding. This technique is applicable for most graph embedding settings a part from non-relational data and whole-graph embedding that have been not tried yet [15]. This third embedding technique directly optimizes an edge reconstruction based on objective functions by either maximizing edge-reconstruction probability or minimizing edge reconstruction.
Differently, in graph kernel approach the whole graph structure is represented as a vector consisting of the counts of elementary substructures that are decomposed from it [15]. There are three types of substructures defined in graph kernel: graphlet that is an induced and non-isomorphic subgraph of size-k [60], subtree patterns in which a graph is decomposed as its subtree patterns (an example is Weisfeiler-Lehman subtree [50]) and the third type that is based on random walks in which a graph is decomposed into random walks or paths and represented as the counts of these random walks [55] or paths [8].
The last of the main graph embedding approaches we consider in this concise literature review is the one based on generative models. "A generative model can be defined by specifying the joint distribution of the input features and the class labels, conditioned on a set of parameters [5]" [15]. This approach can be used for both node and edge embedding and, since it regards node semantics, is commonly used for heterogeneous graph [3] or a graph with auxiliary information [58].
These approaches are the ones present in the literature at the best of our knowledge and they regard graph embedding from a general point of view, hence they are related to all possible kind of data that are represented as graph. However, we will focus our attention on graph embedding techniques that are applied on brain networks and that are easily explainable and interpretable.

1.3 Outline

This thesis is organized as follows.

- In chapter 2 we present the issues faced in this work, then we describe in details the state-of-the-Art methods we chose as references and the applied models;
- In chapter 3 we illustrate the characteristic of the used data set ([19]) and we introduce the results of the application of the state-of-the-Art methods we picked: an explainable method for brain network classification based on contrast subgraph by Lanciano et al. [34] and a connectome-based predictive model for intelligence scores proposed by Dryburgh et al. [22].
- Then, in chapter 4 there is the description of the novelty we add to the already existing methods. We enlarged the research parameters space applying new measures to the existing methods and varying the prospective of feature selection. This chapter is constituted by two parts: the former in which we describe in detail our contributions and the latter in which we show application results.

Chapter 2

Problem and Methods

2.1 Problem Statement

The purpose of this work is to find a procedure to apply regression models on brain networks as input observations and clinical variables as variable of interest. In particular we want to test predictive models on brain network graphs using intelligence scores as dependent variable. This general task is not straightforward and it is developed in various passages. Therefore, imagine we are given N different subjects (i.e., the patients). Each one is associated to a graph(i.e., the brain network) in the set $G = \{G_1, ..., G_N\}$, each one defined over the same set of nodes V with $|V| = n$, and having a specific connectivity matrix each, according to the selected method: $G_i = (V, E_i)$ for $i = 1, .., N$. We are also given a vector $Y = \{y_1, ..., y_n\}$, that represent the response variable of the regression task (i.e., the intelligence score). Therefore, our problem requires to build a methodology able to, given a new patient, estimate the intelligence score based on its brain network. Hence, to solve this task, we need to carefully build a pipeline that is composed of two steps:

1. A feature engineering step, in which through specific network theory algorithms we are able to select the most informative part of the input graphs.

2. An embedding step, in which we synthesize in a vector the information previously obtained.

In this chapter we will go through these two specific steps, in particular giving a brief overview on the related literature and we will describe the methods and the procedures we chose and propose to solve our tasks.

2.2 Methods

2.2.1 Brain networks analysis using Contrast Subgraph

Having a set of graphs defined over the same set of nodes and identifying two groups in the population, the core idea in contrast subgraph is to find two sets of nodes that constitute a dense subgraph in the former group that is sparse in the latter and viceversa. Specifically, as explained in [34], having a dataset $\mathcal{D}$ where each observation $G_i = (V, E_i)$ is a graph defined over the same set of nodes, V, and considering two groups, the *condition group* $\mathcal{A} = \{G_1^A, ..., G_{r_A}^A\}$ and the *control group* $\mathcal{B} = \{G_1^B, ..., G_{r_B}^B\}$, we aggregate the information of the groups $\mathcal{A}$ and $\mathcal{B}$ in two *summary graphs*, $G^A = (V, w^A)$ and $G^B = (V, w^B)$, where $w^A, w^B : VxV \longrightarrow R_+$ are weights function that assign a value to each pair of vertices,u and v, that is:

$$w^A(u, v) = \frac{1}{r_A} \mid G_i^A \in \mathcal{A} \text{ s.t. } (u, v) \in E_i^A \mid,$$

namely the fraction of graphs $G_i^A \in A$ in which u is incident to v, and similarly for w^B.

In order to find the subset of vertices, $S \subseteq V$, whose induced subgraph is dense in the summary graph G^A and sparse G^B, we need to maximize the contrast-subgraph objective:

$$\delta(S) = e^{\mathcal{A}}(S) - e^{\mathcal{B}}(S) - \alpha \binom{|S|}{2} \tag{2.1}$$

where

$$e^{\mathcal{A}}(S) = \sum_{u,v \in S} w^A(u, v) \text{ and } e^{\mathcal{B}}(S) = \sum_{u,v \in S} w^B(u, v) \tag{2.2}$$

are the sum of edge weights in the subgraph of G^A and G^B, respectively, and induced by the vertex set S; α is a user-defined parameter that penalizes solutions of large size.[1]

At this point we need to define the way in which we generate the graphs for each subject that will be the input of our research work.

Considering the correlation matrix between the fMRI signals, $C_{116x116}$, we build an unweighted graph $G_i = (V, E_i)$ for the i^{th} patient that is defined over the set of nodes V and has edges between two nodes s and v if the correlation in absolute value is in top 20% of $|C_{116x116}|$, $E_i = \{e_{s,v} = 1 \Leftrightarrow |C_{s,v}| \geq q_{0.8} \quad \forall v, s \in V \text{ if } s \neq v\}$, giving as result an undirected and unweighted graph. "This approach is typical in litterature, see the works of Lord et al. [37] and Rubinov et al. [48]" [34].

Starting from these graphs, we use the estimated summary graphs to extract the features from each observation. Specifically, if we apply contrast on two groups, we consecutively have two sets of nodes resulting from the contrast subgraph algorithm: (i) V^A containing the nodes in the subgraph that is dense for the summary graph G^A and sparse for the summary graph G^B and (ii) V^B containing the nodes in the subgraph that is dense for the summary graph G^B and sparse for the summary graph G^A. After that, we compute the chosen metrics over the two subsets for each subject in the data set. As done in Lanciano et al. [34] we use the sum of edge weights in the subgraphs induced by V^A and V^B, that are $e^A(V^A)$ and $e^B(V^B)$, respectively. Since we consider unweighted graphs, the sum of the edge weights is nothing different from the number of edges in the induced subgraph: if we consider the formula in 2.2, $w^A(u, v)$ can be equal either to 1 or 0.

Hence, in this case in which we split the observations in two classes, we have two variables that we will use to model linear regression using the intelligence score as variable of interest. However, if we divide all the subjects in more classes the number of variables will increase as well as the number of contrasts. More precisely we will have a contrast for each possible pair of groups and, consecutively we will have a couple of variables for each of these contrasts.

Nonetheless we will discuss this point in more details in chapter 4.

[1] For more details see [34].

2.2.2 Connectome-based Predictive Model for intelligence scores

In the article of Dryburgh et al. published in 2019 [22], the authors propose the use of a connectome-based predictive model (CPM), developed by Shen et al. [49], to predict intelligence scores from functional connectome data that are derived from resting state fMRI (see figure 2.1).
They start from the connectivity matrices of each subject, namely the correlation matrix $C_{116\times116}$ of the fMRI signals. Applying a leave-one-out cross-validation where $N-1$ subjects are used for training and one for testing. After that the correlation coefficients are computed between these correlations and the IQ score of each subject, deriving also their statistical significance *p-value* from the training samples. These correlations are split into positive and negative matrices selecting the most significant ones applying a threshold of $p = 0.01$. This produce a negative matrix which will be used to build a negative predictive model and a positive matrix used to build a positive model building.
Since the number of significant entries in both the positive and the negative matrices can vary among all the subjects, the authors introduce at this point two summary values for each observation:

i) a positive summary value computing the sum of all the values in the positive subject-specific matrix;

ii) a negative summary value computed by summing all the negative values in the negative subject-specific matrix.

This allows to separately model positive and negative functional brain connection sets in two univariate linear regressions. Connectome-based predictive model assumes a linear relationship between the single-subject summary values and behavioral variables [22].
We need to give some remarks on this intelligible and coherent method. The authors decide to build a positive and a negative model for each population, keeping separate subjects with Autism Spectrum Disorder and Typically Developed. The reason is that they don't want to classify ASD and TD subjects nor to identify shared connections in the brain linked to intelligence. On the contrary, their aim is to identify the most relevant functional brain connections that can explain intelligence scores in healthy and autistic population, independently.
Another important remark regards the idea of teasing apart positive and negative connections. This is done to prevent bias when interpreting the connectome data and not to omit the inverse relationship that occur over two ROIs which can be important to understand brain connectivity [22].

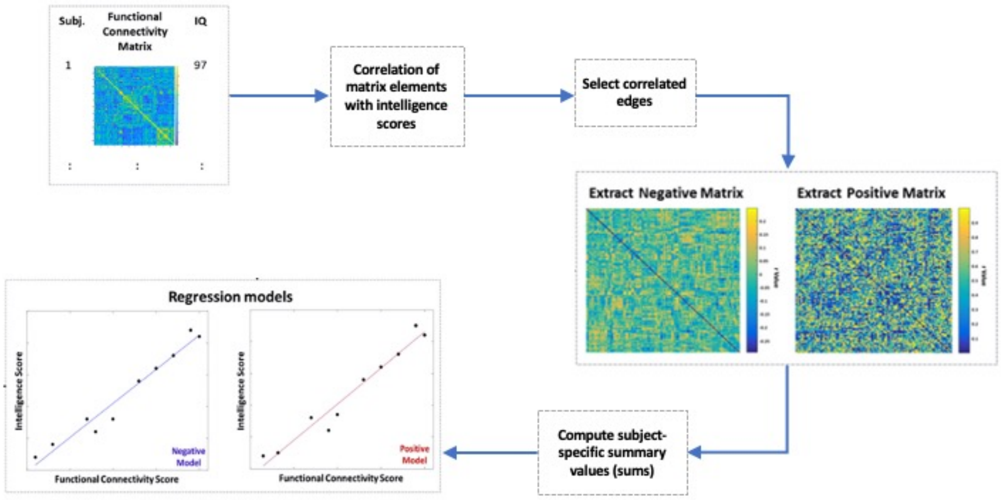

Subj.
Functional Connectivity Matrix
IQ
1
97
Correlation of matrix elements with intelligence scores
Select correlated edges
Extract Negative Matrix
Extract Positive Matrix
Regression models
Intelligence Score
Negative Model
Functional Connectivity Score
Intelligence Score
Positive Model
Functional Connectivity Score
Compute subject-specific summary values (sums)

Figure 2.1. Figure inspired to fig.1 of Dryburgh et al. [22]. The Connectome-based Predictive Model (CPM) (Shen et al. [49]) uses $C_{116\times116}$ connectivity matrix to first compute correlations between the entries of the matrix and the IQ score in a leave-one-out cross-validation setting. Then correlations are selected according to p-values below a predefined threshold ($p < 0.01$). Selected connections for each training subject are then split into two separates matrices: (i) significant positive correlations stored in a positive connectivity matrix (ii) significant negative correlations stored in a negative connectivity matrix. For each training subject all the positive and negative values in the matrices are summed obtaining two summary values for each subject, a positive and a negative summary score, respetively. Then pairs of linear regression models are trained: (i) a positive regression model mapping positive summary scores of the train set to the target intelligence score (ii) a negative regression model mapping negative summary scores of the train set to the target intelligence score. Finally, the left-out subject is used to test the previously trained models, for both positive and negative scenarios.

2.3 Models

The last point in our pipeline is the application of predictive models on the sets of features created in the previous steps. In chapter 3 we will consider only linear regression models, accordingly with the cited works that operate in a context of linear classification and linear regression, respectively. However in the contributions in chapter 4 we apply also different models to predict our response variables. In this section we give an introduction to all the models we use in this work.
We decide to apply three different models: linear regression, lasso regression and random forest. We use two models that infer linear correlation between dependent and independent variables while the third is used to understand which is the loss in using linear models with respect to a robust-non linear model such as random forest. It follows a brief overview of the selected models.

Multiple Linear Regression

Simple linear regression can handle only one independent variable X and a dependent variable Y, which is expressed as a linear function of X [42]. The value

y_i of variable Y, for every value x_i of variable X, is given by the equation:

$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

where the coefficients β_0 and β_1 represent the linear dependency of Y from the covariate X. The constant therm, called also intercept, represents the value of y for $x = 0$, while the slope of the straight line in the equation is the correlation between X and Y showing the change of Y when X has a unit increase. On the contrary, ϵ is the error, namely the difference between the estimated and true values of Y. In the case we have more than one independent variables, that is $X = \{x_{i1}, x_{i2}, ..., x_{ip}\}_{i=1}^{n}$, we have an extension of the simple linear regression for a generic observation i:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \epsilon_i = x_i^T \beta + \epsilon_i.$$

Considering the matrix notation: $Y = X\beta + \epsilon$ we have that the estimated values $\hat{Y} = X\beta = X(X'X)^{-1}X'Y$.However, in the multiple case there are some important assumptions that are needed to be verified [46]:

- *Linearity*: the response variable y should be linearly dependent to the explanatory variables X;
- *Independent and identically distributed errors*: The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms and they're supposed to have the same distribution;
- *Normally distributed errors*: $\epsilon_i \sim N(0, \sigma^2)$ with $i = 1, .., n$.
- *Homoscedasticity*: The variance of the residuals is the same for any value of X.

For interested readers we suggest machine learning manuals such as the one of Hastie at al. [31] or Wooldridge's masterpiece of econometric and cross section analysis [57].

Lasso Regression

The ordinary least squares (OLS) model simply minimizes the residual squared errors of the estimate. In the *lasso regression*, presented by Tibshirani in 1994 [51], we have an OLS with a shrinkage component that allows to literally shrink to zero non meaningful coefficients increasing the estimate on both sides of prediction accuracy and interpretation. Citing the author "There are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is prediction accuracy: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to zero some coefficients. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy. The second reason is interpretation. With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects." ([51] p. 1-2). Therefore we now present the definition of the *lasso model.*
Suppose that we have a set of couples (x^i, y_i), for $i = 1, ..., N$, with $x^i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ that is the independent variables and y_i is the response. Under the assumption that the observation are independent and the x_{ij} are standardize so that $\sum_i x_{ij}/N = 0$

and $\sum_i x_{ij}^2/N = 1$. Considering $\beta = (\beta_1, ..., \beta_p)^T$, the lasso estimate is defined as follows:

$$\begin{gathered}(\hat{\alpha}, \hat{\beta}) = argmin \sum_{i=1}^{N} \big(y_i - \alpha - \sum_j \beta_j x_{ij}\big)^2 \\ \text{subject to } \sum_j \mid \beta_j \mid \leq t\end{gathered} \tag{2.3}$$

$t \geq 0$ is a tuning parameter and, for all t, the estimate for α is $\hat{\alpha} = \bar{y}$. The problem in 2.3 is computationally a quadratic problem with linear constraint and we need to underline that the design matrix has to be full rank [51]. Neither going deep in the description of the lasso problem nor describing the methods to tune the parameter t are not the aim of this work. However we need to highlight the role of t, that is the amount of shrinkage that is applied to the estimate. Values of $t \leq t_0$ with $t_0 = \sum \mid \hat{\beta}_j^0 \mid$ "will cause shrinkage of the solutions towards zero, and some coefficients may be exactly equal to zero. For example, if $t = t_0/2$, the effect will be roughly similar to finding the best subset of size $p/2$."([51] p.3)

Random Forest

Breiman [9] was the first talking about this variant of bagging that averages a huge collection of de-correlated trees. The idea of bagging is to average many models with great variance but low bias. Trees can be used for bagging, since they are particularly noisy and, if grown sufficiently deep, they also have a low bias [31]. "Moreover, since each tree generated in bagging is identically distributed (i.d.), the expectation of an average of B such trees is the same as the expectation of any one of them. This means the bias of bagged trees is the same as that of the individual (bootstrap) trees, and the only hope of improvement is through variance reduction.[...] The idea in random forests (Algorithm 1) is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables. Specifically, when growing a tree on a bootstrapped dataset: *Before each split, select m of the input variables at random as candidates for splitting.* Typically values for m are $\sqrt{p}$ or even as low as 1.
After B such trees $\{T(x;\Theta_b)\}_1^B$ are grown, the random forest regression predictor is

$$\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$$

Θ_b characterizes the b^{th} random forest tree in terms of split variables, cutpoints at each node, and terminal-node values. Intuitively, reducing m will reduce the correlation between any pair of trees in the ensemble, and hence reduce the variance of the average"([31] p. 588-589).

Algorithm 1: Random Forest for Regression [31]

```
I. for $b$=1 to $B$ do
    1) Draw a bootstrap sample $Z^*$ of size $N$ from the training data.
    2) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively
    repeating the following steps for each terminal node of the tree, until the
    minimum node size $n_{min}$ is reached.
        (a) Select $m$ variables at random from the $p$ variables.
        (b) Pick the best variable/split-point among the $m$.
        (c) Split the node into two daughter nodes.
   end
II. Output the ensemble of trees $\{T_b\}_1^B$.
```

To make a prediction at a new point x:

$$\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Chapter 3

Experiments

3.1 Dataset

Before proceeding it is useful to give a brief but exhaustive background on the characteristics of our dataset, namely *resting-state fMRI* data.
When neurons of a certain brain area are activated, they are provided with more oxygen with an increasing blood flow by the adjacent capillaries, through a process called hemodynamic response. "This process results in a change in terms of the relative levels of oxyhemoglobin and deoxyhemoglobin that can be detected by MR imaging on the basis of their differential magnetic susceptibilities. This imaging approach is called blood oxygen level–dependent (BOLD) contrast imaging" ([40] p.1390). The BOLD signals change according to the change of the arterial partial pressure of oxygen and carbon dioxide [44] that is reflected in the frequencies of neural activity fluctuations measured by fMRI. By reason of the great amount of signals in MRI several techniques to reduce dimensions have been developed. Additionally, these signals have also a great amount of noise due to various factors that induce to apply preprocessing on the signal side (such as filtering or correction) or on the network-analysis side; see the exhaustive survey of Lang et al. [35] for more details.
In our study we use a real world and publicly avaliable dataset[1], released by the Autism Brain Imagine Data Exchange (ABIDE) project [19]. This dataset contains neuroimaging data of 1112 patients of different age, sex and diagnostic category[2], in particular 539 subjecs suffering from Autism Spectrum Disorder (ASD) and 573 typical controls. Following the example of Lanciano et al. in [34] we used data preprocessed pursuing the procedure denoted as DPARSF[3], followed by Band-Pass Filtering and Global Signal Regression. The results of the preprocessing part for each patient is a set of 116 time series of length 145, each of them obtained from one of the 116 ROIs produced by the brain partitioning done adopting the AAl atlas [52].
According to what is done in Lanciano et al. [34] and in Dryburgh et al. [22] we select a subset of the whole datasets choosing subjects with no significant demographic differences. We select 347 subjects that are males, under 30 years old, with no missing data or damaged signals in the fMRI exams and having the same type of test available for the intelligence score, namely the *WASI* test; 160 of them are ASD subjects while the other 187 are typically developed.

[1] http://preprocessed-connectomes-project.org/abide/index.html

[2] More details at the *ABIDE data legend* page.

[3] http://preprocessed-connectomes-project.org/abide/dparsf.html

3.2 Application settings

In this chapter we will train a linear regression model on different sets of features computed from input graphs following the two methods described in chapter 2. Specifically we have four different couples of observations' sets and response variables. These couples are constructed merging two groups of subjects, affected by Autism Spectrum Disorder (ASD) and Typically Developed (TD), respectively coupled with two response variables, FIQ and VIQ *WASI* intelligence scores. Obviously the features we compute from these sets of observations are different in the two approaches according to the procedures applied in each method. We now describe in details the operations and related choices of both approaches before presenting the results.

3.2.1 Contrast subgraph

The main idea of this technique, as we saw in chapter 2, is to estimate a subgraph that is dense for a group of observations and sparse for another and viceversa. Hence, the first point we need to illustrate is how we divide the subjects into different groups. In Lanciano et al. [34] the contrast algorithm is applied between the ASD and the TD subjects in order to classify if they show autism spectrum disorder or not. Following this idea, since our target variable is the intelligence score, we split the observations in *high* and *low* intelligence scores using the median, i.e. the 50^{th} percentile of the distribution, in order to obtain balanced groups. Considering that we use two intelligence scores and two sets of observations we will divide the considered data sets twice. In other words, before applying the contrast algorithm, we place the subjects with an intelligence score lower than the median in the *low IQ* group and the patients with an IQ higher than the median in the other group, namely the *high IQ*, and this is done for both FIQ and VIQ scores. In table 3.1 the median and other basic statistics of the intelligence scores distributions are shown. The analyses of this chapter focus in the simple case of comparison between two groups of observations that gives as result two estimated subgraph for each contrast. As we already said, the size of the estimated subset is controlled by the parameter α in the objective function 2.1 and, specifically, the relationship between this parameter and the number of nodes is inversely proportional. Since we do not know which is the optimal value of α we train a linear regression tuning this parameter to choose its value which corresponds with the best fitting for each data set.

	ASD data set (160 obs.)		TD data set (187 obs.)	
	FIQ	*VIQ*	*FIQ*	*VIQ*
min	76	55	79	67
median	**106**	**104**	**112**	**112**
max	148	149	144	141
mean	106.98	104.07	111.91	111.06
std	15.93	17.24	12.58	13.27

Table 3.1. In this table we report some statistics of the distributions of the intelligence scores we use as response variables to train the linear regression models in the following paragraphs. The median is used to divide the data sets into two balanced groups with *high* and *low* intelligence scores, respectively, which will be used to compute contrast subgraph algorithm and estimate the subgraph that is dense in one group and sparse in the other and viceversa.

Hence we exploit the contrast subgraph algorithm on the four data sets for values of $\alpha \in \{70, 75, 80, 85, 90\}$. The results, that are the lists of nodes that design the induced summary subgraphs, are shown in the appendix of this text (A.1-A.4). The intent of this work is not to analyze the brain subnetworks and the nature of brain connections estimated via contrast but it is mainly to find an algorithmic and explainable way to extract important features from brain networks and use them to predict quantitative variables. However, to give an idea of what the intermediate result of contrast subgraph algorithm can be, we give a visual example. In figures 3.1 and 3.2 we show the resulting sub-networks for the ASD data set, with FIQ as dependent variable and $\alpha = 80$, that are dense for the *low-IQ* group and sparse for the *high-IQ* group and viceversa, respectively. In this instance the *low-IQ* group is formed by patients that have an intelligence score in the FIQ-WASI test less or equal than 106, while subjects whose this test's result is in the interval $(106, 148]$ constitute the *high-IQ* group.

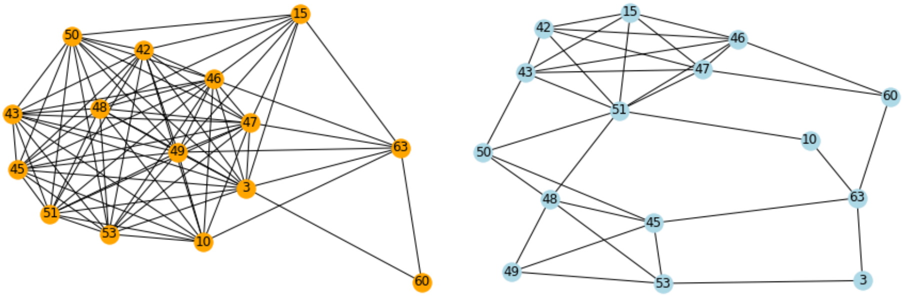


Figure 3.1. This figure shows an example of contrast subgraph result. The subjects taken into account show Autism Spectrum Disorder and the contrast is computed between patients that have an intelligence score resulting from the FIQ-WASI test in the lower fifteen percent of the sample and the ones that are in the upper bound, respectively. Here we have an example of the brain sub-network that is denser for subjects in the *low-IQ* class (left side) and sparse for the *high-IQ* group (right side) using an $\alpha = 80$.

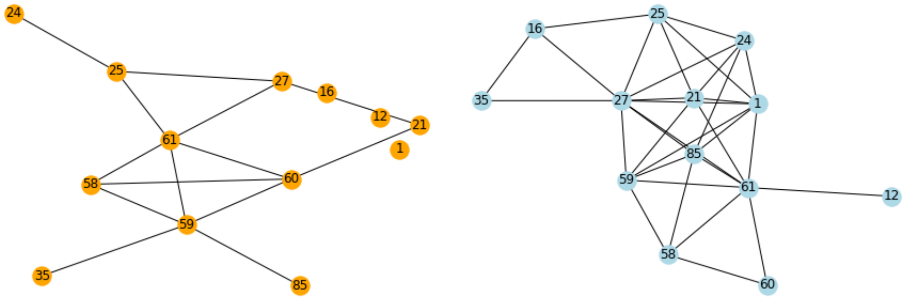


Figure 3.2. This illustration concerns the same case described in 3.1 computed on the ASD subjects divided into *low* and *high* IQ groups with respect to the 50^{th} percentile of the distribution of the FIQ-WASI test. But contrary to the previous figure it shows the subgraph that is dense for subjects belonging to the *high-IQ* group (in lightblue) and sparse for the *low-IQ* group (left side in orange).

Before proceeding in our analysis we need to exhibit another preliminary result regarding the tuning procedure of the α parameter. In the line plot below we show the average coefficient of determination obtained from a linear model tested in a

5-fold cross validation setting using the number of edges as dependent variables and the intelligence scores as response. The sum of the edges is computed over the sub-network resulting from the contrast subgraph algorithm applied with $\alpha = \{70, 75, 80, 85, 90\}$ and on all the data set-intelligence score couples we already mentioned. A part from the TD data set with the FIQ score as response variable that has the highest R^2 with $\alpha = 85$ in all the other settings the best value of the penalization parameter is 70. This is just an exploratory outcome that we use to select the best tuned parameter for each dataset. In paragraph 3.3 we will show the details of the model with the highest cross-validated coefficient of determination for each data set.

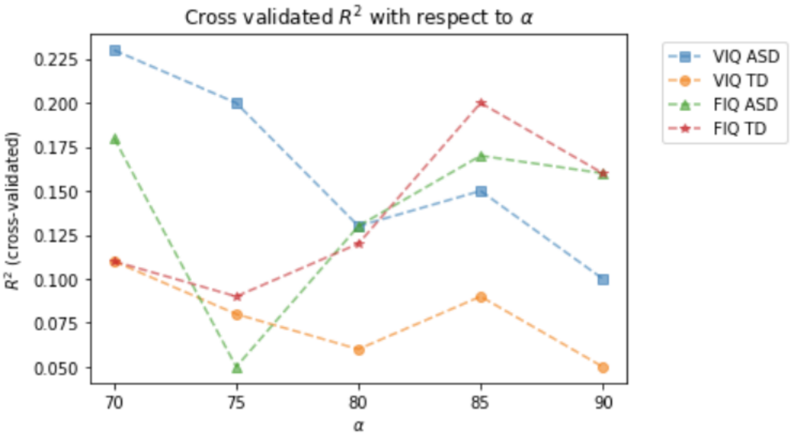


Figure 3.3. This line plot shows the R^2 of a linear regression model trained on all the data sets using 5-folds cross validation. The covariates used in the model are equal to the number of edges in the subgraph induced by the estimated subsets of the whole brain network via contrast subgraph. These subgraphs strictly depend on the value of the control parameter α reported on the x-axis of the plot.

3.2.2 CPM

As regards the Connectome-based Predictive Model (CPM) the procedure basically follows what we explained in 2.2.2. However, before presenting the results of our application we need clarify some issues regarding some adjustments we applied on the method proposed by Dryburgh et al. [22]. Basically in their article *Predicting full-scale and verbal intelligence scores from functional connectomic data in individuals with autism spectrum disorder* [22] they use an existing method implemented and described by Shen et al. [49]. In this work we mainly follow the original version of the CPM presented in *Using connectome-based predictive modeling to predict individual behavior from brain connectivity* [49]. Specifically, after computing the correlation matrices of the brain region of interest for each subject the CPM proceeds measuring the correlation between each entry of these matrices and the intelligence score in order to create two matrices for each subjects, the former containing the original correlation that are positively correlated with the IQ score and the latter with the correlations that are negatively correlated with the intelligence score, according with a predefined threshold of significance. Dryburgh et al. [22] use a robust linear regression to compute the relationship between ROI correlations and intelligence score, while we use the basic implementation proposed by Shen et al. [49] that uses Pearson's correlation.

An aspect in our CPM method application that differs from both the original implementation and Dryburgh et al.'s [22] is the choice of the cross validation method. After computing the positive and negative variables they use a leave-one-out cross validation in which, having N observations, a linear regression model is trained N times, each time leaving out a different subject for testing, hence at each round the model is firstly trained over $N-1$ subjects and then tested on the left-out one. Since the other techniques used in this work have an high computational cost and, as we will see, in the further sections we train several models, in this work we decide to train and test all the models in a 5-fold cross validation fashion, that is computationally less expensive even preventing overfitting.

3.3 Results

In this paragraph we present the outcomes of the preliminary analysis described in the previous session of this chapter. In table 3.2 we show a preview of the average coefficient of determinations for both test and train sets obtained from a 5-fold cross-validation fashion using a linear regression model. In the first row we have the R^2 related to the features extracted by contrast subgraph while in the others there are the coefficient of determinations of the Connectome-based Predictive Model. CPM-positive and negative models are simple linear regressions since we have a single independent variable, while the others have two covariates each.
In general the CPM-based approach outperforms the model with feature embedding performed via contrast. In particular we can see how the setting with both positive and negative variables reaches the highest average R^2 for all intelligence scores, having its peak in the case of ASD data set with an average value of 0.55 in the test sets. Another evidence we need to underline is the higher predictive power of the linear regression model in the case of subjects affected by autism spectrum disorder rather than in the typically developed data set. However, in the case of contrast subgraph related features for the FIQ-WASI test and typically developed patients we have a slightly higher score than for ASD data.

	ASD data set		TD data set	
Method - test (train) R^2	*FIQ*	*VIQ*	*FIQ*	*VIQ*
Contrast subgraph	0.17 (0.19)	0.25 (0.28)	0.20 (0.23)	0.14 (0.17)
CPM-positive	0.35 (0.4)	0.46 (0.49)	0.34 (0.37)	0.3 (0.35)
CPM-negative	0.37 (0.43)	0.44 (0.46)	0.35 (0.41)	0.31 (0.36)
CPM-positive and negative	0.55 (0.6)	0.55 (0.6)	0.43 (0.48)	0.37 (0.43)

Table 3.2. In this table we show the R^2 of a 5-fold cross-validated linear regression model on different settings of covariates and response variables in both test and train sets, outside and in brackets, respectively. By row we have the methods used to extract features from the estimated subgraph. These are the sum of sub-network's edges estimated via contrast subgraph and Connectome-based Predictive Model (CPM) in three different modalities: sum of correlations positively correlated with the intelligence scores, sum of correlations that are negatively correlated with the intelligence scores and both of them together. While by column we have the diverse response variables we use to train and test the model.

In the following pages we display the scatter plots of the predicted values against the true ones for both train and test sets separately for all cross-validation folds. However in the text we just show the plots related to the highest average test R^2 in contrast subgraph-based models and CPM ones, respectively, namely the ASD data set with VIQ intelligence score as y for both positive and negative features for the Connectome-based Predictive Model. The scatter plots of all the other models and settings are available in Appendix A (figures A.1-A.16).

It is noticeable the difference of fitting between the two setting of features in figures 3.4 and 3.5. Indeed in the CPM case all the test folds have an $R^2 > 0.5$ while the highest coefficient of determination in the contrast subgraph features setting is 0.43 and there is also an almost zero cross validated r-squared in the second fold. For CPM we reported also the correlation coefficient between the predicted and true values of the response variable, r, and its significance level p. This is done since in Dryburgh et al. [22] the linear regression model is evaluated using this measure, however we don't generally use it to appraise the models' prediction accuracy.

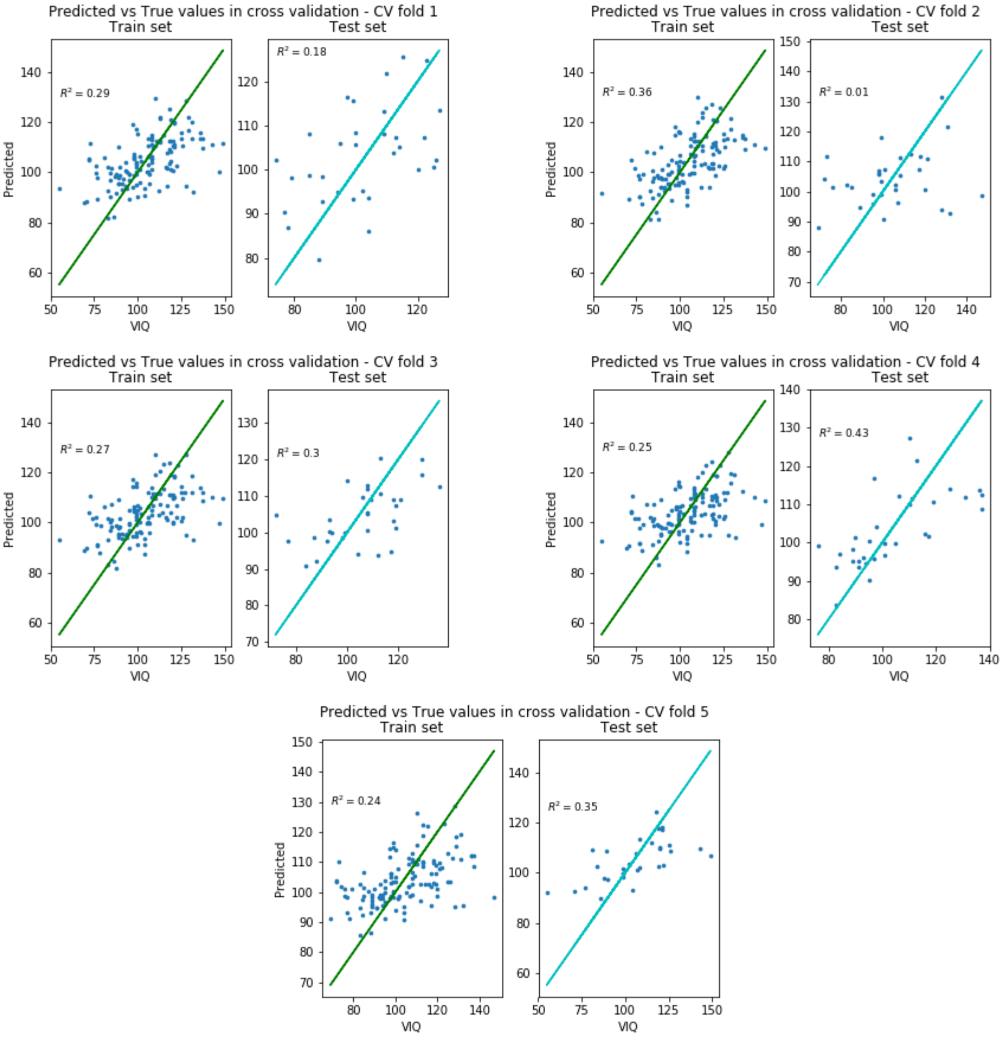


Figure 3.4. The plots in this figures are train-test paired scatter plots which have on the x-axis the values of the VIQ intelligence scores and on the y-axis the predicted values. Each couple refers to one of the five folds of the cross-validated linear regression model. In this case the independent features are computed using contrast subgraph technique with $\alpha = 70$.

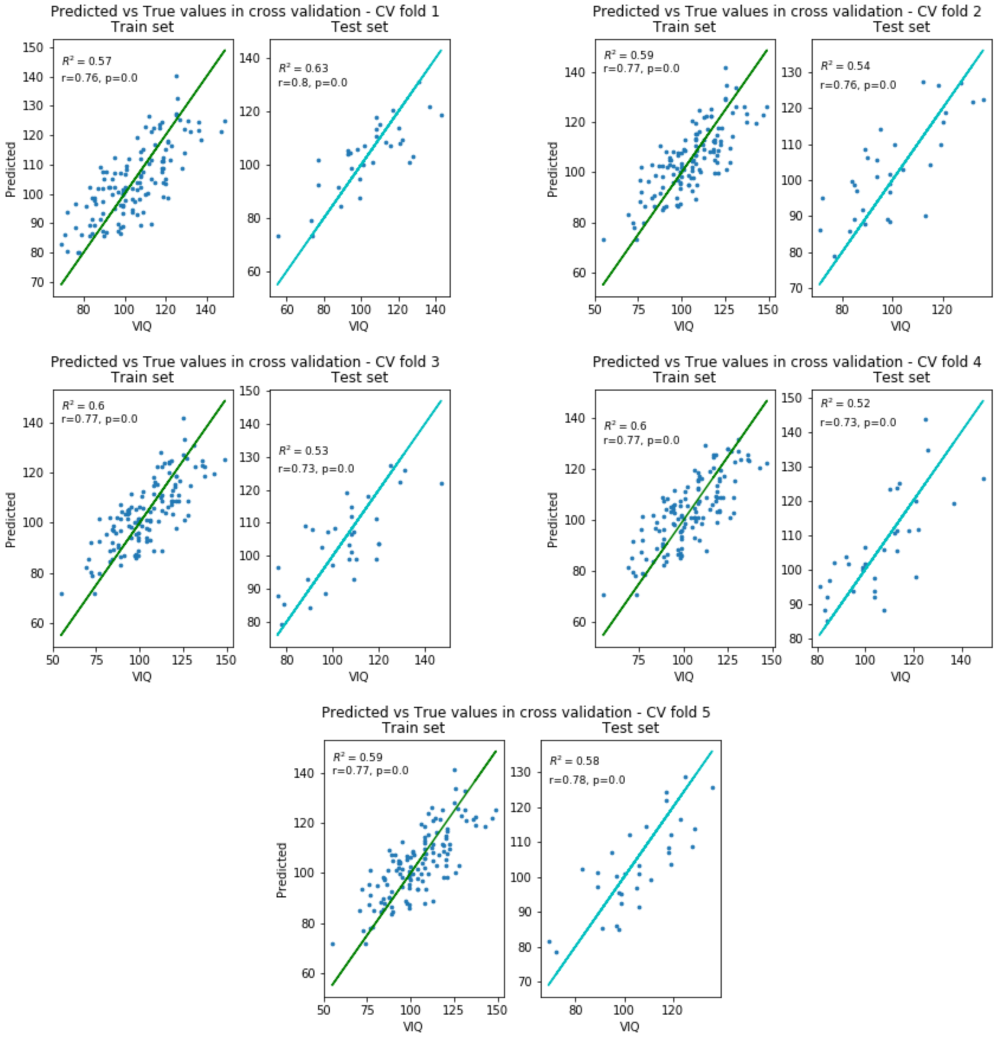


Figure 3.5. As in the previous figure we have coupled scatter plot of train-test cross-validated prediction with the intelligence score on the x-axis and the predicted values on the y-axis. However, differently from fig. 3.4, here we have the CPM estimate of the VIQ test score. Besides the coefficient of determination in the plots we reported also the correlation between the true and predicted values and its value of significance, r and p respectively.

Chapter 4

Contributions on the existing methods

In this chapter we introduce our research contributions to the cited methods. The starting point of our exploration is contrast subgraph method proposed by Lancino et al. [34] but we will also take advantage and inspiration from the Connectome-base Predictive Model studied by Dryburgh et al. [22]. We can summarize the object of our analysis in some key research questions:

- **RQ1**: *How different ways of dividing observations in contrast subgraph, on both criterion and number of groups sides, can affect the result?*
- **RQ2**: *There are any other metrics that can be used in features extraction process which performs better than the number of edges?*
- **RQ3**: *What is the impact of dealing with positive and negative correlations separately in the brain network construction?*
- **RQ4**: *How different regression algorithms perform in our context? And, in particular, does a robust non linear regression better perform than linear regression models?*
- **RQ5**: *is there any added value (in terms of performance) of the constrast subgraph-based features over the embeddings in the state of the art? Does conjoining the SOTA-features with our features provide better performance?*

In order to solve these problems we proposed different existing methods, models or metrics, depending on the specific task. First of all we will describe the applied methods and the choice we made for the research questions above, if needed, and then we will show the results of our experiments comparing them with the state of the art methods.

4.1 Proposed techniques

Specifically we need to describe the following points: the different criterions we used to divide the observations in different groups that will be paired and used as input for the contrast algorithm and the metrics applied to extract features from the estimated sub-networks. Actually we also applied different models in addition to the linear regression, namely a LASSO regression and a random forest. However, for a detailed description of the applied models we refer to section 2.3. As regards the control parameter α we chose to use the best values of chapter 3, namely 70 and 85.

4.1.1 Groups choice criterion (RQ1)

It is unequivocal that the way in which we split our dataset in order to compute the contrast subgraph algorithm is crucial as well as it is for the number of groups we choose. Indeed, the result of our research, and more precisely the feature extraction process, strongly depends on this aspect and, since we don't have any evidence on what the right choice could be, we approach the problem applying an extensive research based on different methods.
We call k the hyperparameter we need to infer, i.e. the number of groups in which we split data with respect to a variable $y \in N$, in our case the intelligence score. We take into account the following approaches:

- *intervals of the same length (E)*: having the distribution of the variable of interest y, we create k groups of size $\frac{y_{max}-y_{min}}{k}$;
- *intervals with respect to quantiles (Q)*: in this way we obtain k balanced groups. We consider as splitting points q_a, with $a \in [0, 1]$, as the quantiles of level a (for example the median is the quantile $q_{0.5}$). Hence we have, for each value of k, $k-1$ splitting points q_a with $a = \frac{p}{k}$ for $p = 1, ..., k-1$.
- *clustering (C)*: we run a non-hierachical clustering algorithm, namely the k-means algorithm, since we can choose the number of groups a priori, and then split the dataset according to the estimated labels.

In this work we vary the number of groups in $k = \{2, 3, 4\}$ for all the aforementioned settings. This choice is due to the exponential growth of the number of resulting contrasts, since it is equal to $D_{k,2} = \frac{k!}{(k-2)!}$ and, additionally, not to have empty groups increasing k too much with respect to the number of observation. Furthermore even having poorly filled sets or strongly unbalanced groups would negatively impact the analysis.

4.1.2 Metrics for sub-network embedding (RQ2)

In the article of Lanciano et al [34], dealing with classification, the authors consider one simple metric in the feature engineering part with respect to different contrast subgraphs, that is the number of edges for each induced subgraph. We decided to consider more global measures summarizing the whole graph $S \subseteq V$ like average clustering coefficient, diameter and number of edges as well. It follows a focus on the metrics we compute once completed the estimation of the subset for a single contrast, that are:

- **# of edges**

  $$e(S) = \sum_{u,v \in S} w(u, v)$$

  where $w(u, v)$ is the edge weight between nodes u and v that is either 1 or 0, since we consider undirected and unweighted graphs.

- **Average clustering coefficient**

  $$C = \frac{1}{n} \sum_{v \in S} c_v$$

with n equal to number of nodes in S and $c_v = \frac{2T(v)}{deg(v)(deg(v)-1)}$ is the clustering coefficient for the node v. For an unweighted graph, $T(v)$ is the number of triangles through the node v and $deg(v)$ is the degree of node v.

- **Diameter**
  that is the maximum of the shortest path between any couple of two nodes, u and v:
  $$d = max_{u,v} d(u,v) \quad \text{with } u, v \in S$$
  where the distance between any two nodes, $d(u,v)$, if we consider an unweighted graph, is equal to the number of edges in the shortest path. For semplicity, in the case of a disconnected graph, in which the diameter is infinite, we consider $d = 1000$.

Since these metrics are defined in different unit measures it can be appropriate to use some standardization before applying regression models. Having a matrix of data, X, composed by n features X^j, with $j = 1, ..., n$, then we have:

- *Scale*: Center to the mean and component wise scale to unit variance, i.e.:
  $$X^j_{scaled} = \frac{X^j - \bar{x}_j}{\sigma_j};$$
- *Min-Max scale*: Scale data from 0 to 1 with respect to max and min:
  $$X^j_{min\_max} = \frac{X^j - X^j_{min}}{X^j_{max} - X^j_{min}};$$
- *Logarithm*: logarithmic transformation:
  $$X^j_{log} = log(X^j + 1).$$
  using $X^j + 1$ in order to deal with nearly $0s$ entries.

4.2 Application and preliminary results

Before proceeding in our analyses and in order to give an exhaustive idea of our experiments, we need to define the space of our parameters and clarify all the steps in our pipeline. As a matter of fact what we have done with respect to the previous chapter is to enlarge the research parameter space so that we can detect which is the best setting to solve our problem. Indeed, we are dealing with a new and heterogeneous problem which has not been explored so far, that is predicting a target quantitative variable, such as the intelligence score, in a regression fashion using features extracted from complex networks.

4.2.1 Pipeline and parameter space

In this paragraph we provide a schematic view on the experiments' pipeline summarized in figure 4.1. Starting from our data, we split the observations into k subsets, according to different values of the picked IQ index and according to the methods described in subsection 4.1.1. After that, we compute contrasts subgraphs on these groups, obtaining a number of subsets for each observations equal to $D_{k,2} = \frac{k!}{(k-2)!}$, using two different values for the penalty parameter, which are $\alpha = \{70, 85\}$. The following step is to apply the described metrics (subsection 4.1.2) over the estimated brain sub-networks. The final step is to apply the chosen models (see section 2.3). We will also apply all the previously described preprocessing transformations for every set of features.
Giving a summary of the parameter space for the whole grid-search procedure, we have:

- **Dataset**: {FIQ-ASD, VIQ-ASD, FIQ-TD, VIQ-TD};
- **Criterion and number of groups for contrast**: {C,Q,E}×{2,3,4};
- **Control parameter**: $\alpha = \{70, 85\}$;
- **Metrics for subgraph embedding**: {# of edges, diameter, avg. clustering degree};
- **Models**: {linear reg., lasso reg., random forest}.

We study all the possible settings from these sets of parameters, that means the product of all possible combinations, i.e. a total of 648 models on different settings. Furthermore, if we consider that we apply to each resulting set of variables the three preprocessing transformation for graphs built on both positive and negative correlations separately, we have a total amount of 3888 trials.
As already said this is just a preliminary analysis, which we can consider as a parameters grid-search process that we use to find the best combination of different solutions for our problem. Every model is trained and tested using a 5-fold cross validation and we evaluate the goodness of the results using the test-folds average R^2 as reference measure. Below we will show the outcome of this consuming process which will be used in the application in the second part of this chapter.

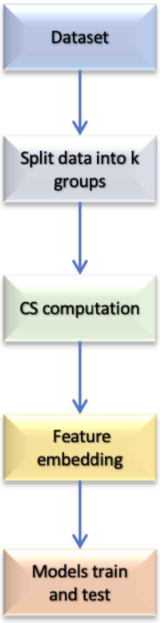


Figure 4.1. This figure shows in a schematic way the pipeline of our experiments. Starting from the picked dataset we first divide it into k groups according to the chosen criterion. After that contrast subgraph algorithm is applied to paired sets of observation. The following step consists in extracting features from the estimated sub-networks; these features are then used to train and test the models using cross-validation setting. For each of the described passages we proposed different solutions that are tuned in the searching procedure, for more details we refer to section 4.2.

4.2.2 First results

It is straightforward that, having this great amount of trials and many different settings of both data and tuned parameters, we cannot find a unique solution that is optimal for all the possible patterns. However the intent of this grid-search we made is not to find a global optimum but to explore the parameter space and find some common trends and solutions that can lead us to better performances. Hence, we will not identify optimal setting for each data set but an averaged global solution for each of the problems we previously presented. Specifically we examine the highest R^2 for each data set separately studied for each of the considered parameter (Figure 4.2) and we arbitrarily select an averaged value for each of them. Furthermore, as assumed in the research questions at the beginning of this chapter, we separately considered positive and negative correlations. This means that for each subject we build two brain networks, representing the positive and the negative correlations between brain regions, respectively. Then, we develop two parallel analysis for these two different conditions.
According to the results we obtained in this preliminary analysis we can summarize the chosen parameters as follows:

- **Group choice criterion**: Quantiles (Q);
- **Number of groups**: 4;

- **Control parameter**: $\alpha = 70$;
- **Metrics for embedding**: # of edges;
- **Model**: linear regression.

Making some remarks on these results we can say that referring to the first research question, namely which is the selected criterion for the observations splitting method and the number of groups, the answer is the quartiles of the intelligence scores distribution. The chosen metric (RQ2) is the *number of edges*, according to what is done in Lanciano et al. [34] and also in conformity of the objective function of the contrast subgraph algorithm (2.1) that maximizes the difference between the number of edges in the induced subgraphs of the two groups that are compared. About the third research question, that is *What is the impact of dealing with positive and negative correlations separately in the brain network construction?*, we can say, anticipating what we obtain in the final results (section 4.3), that we realize that we have a greater predictive power of negative correlations with respect to positive and setting in which we consider all the correlations in absolute value. However, merging the features of both positive and negative correlation we still have a better result compared to chapter 3. Regarding model selection issue (RQ4), we don't have any evidence that a non linear model performs better than a linear regression, and since we want our method as much simple and readable as possible, we continue working with an ordinary least squared regression model. We will discuss the solution to the fifth research question in the following paragraph.

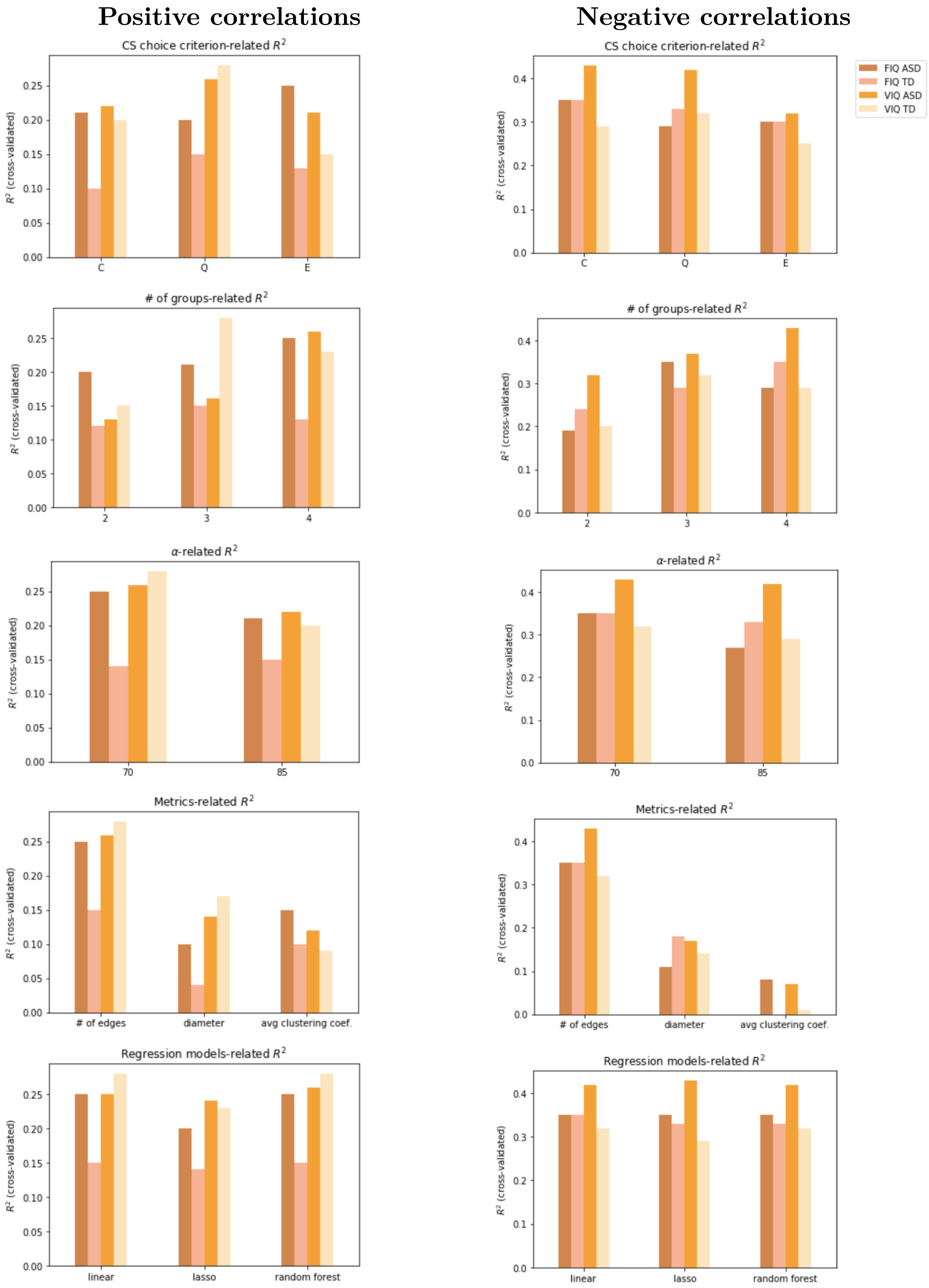


Figure 4.2. These bar plots show the highest cross-validated test sets R^2 reached in each of the four data sets (in legend) and related to each parameter separately considered in each row. This has been done for features computed using both positive and negative correlations, left and right columns respectively (see section 4.2 for more details).

4.3 Putting everything together

Now we use the selected parameters, features and model to compute dense and sparse sub-network via contrast, to extract features from the estimated brain substructure and then evaluate them in a regression context with the intelligence scores as response. As we have done in the previous sections of this work we train and test the models using a 5-fold cross-validation fashion, taking the average test and train R^2 to compare the goodness of fit of the different settings (all the results are reported in Table 4.1).
According to our research questions we separately perform our experiments on positive and negative brain connections. However we also merge the features extracted from the opposite brain correlations, as we have done for CPM in chapter 3, and we can see the results of these applications in the first three rows of table 4.1. We can see how the negative features have a larger predictive impact on the intelligence scores with respect to the variables computed from positive correlations. Moreover, they also have an higher coefficient of determination than the experiments we have done in chapter 3 (3.2). If we consider both train and test scores, we have a further overall increment of the goodness of prediction merging positive and negative features. If we focus on these three applications on the data set side we have that: in the subjects affected by autism spectrum disorder the intelligence score for which we observe a greater average R^2 is the VIQ-test while, on the contrary, for the typically developed subjects the test with higher performance is the FIQ-test, a part from the positive-feature case. In general, comparing these results with the ones obtained in chapter 3, we can say that a linear regression model better performs dealing with features obtained separating positive and negative correlation.

	ASD data set		TD data set	
Method - test (train) R^2	*FIQ*	*VIQ*	*FIQ*	*VIQ*
Contrast subgraph-positive	0.13 (0.29)	0.22 (0.38)	0.1 (0.22)	0.19 (0.33)
Contrast subgraph-negative	0.25 (0.4)	0.34 (0.44)	0.38 (0.45)	0.27 (0.37)
Contrast subgraph-positive and negative	0.23 (0.49)	0.4 (0.58)	0.39 (0.53)	0.26 (0.47)
Contrast+CPM	0.6 (0.73)	0.6 (0.76)	0.57 (0.69)	0.44 (0.63)
Contrast+CPM (with random forest)	0.53 (0.84)	0.57 (0.85)	0.55 (0.83)	0.53 (0.81)

Table 4.1. Here we report the average coefficients of determination of both test and train tests for different covariates-response settings. In the first three rows we have the contrast subgraph related dependent variables estimated for positive and negative correlation and the union of the two sets of variables. In the last two rows we considered a set of variables including both contrast subgraph and Connectome-based Predictive Model features trained with a linear regression model and a random forest regressor, respectively. By columns we have two different data sets, autism spectrum disorder and typically developed subjects and two response intelligence scores, FIQ and VIQ.

Referring to the research questions displayed at the beginning of this chapter, we still have to answer to the last question: *Does conjoining the state-of-the-art features with our features provide better performance?*
We therefore consider a last setting merging positive and negative features estimated using both contrast subgraph and Connectome-based Predictive model. Furthermore, in order to verify the predictive efficiency difference between linear and non-linear models we applied both linear regression and random forest regressor algorithms (last two rows of Table 4.1). Using this last setting of features we achieve a great increment of the average coefficients of determination, outperforming the results

obtained with the state of art methods we took into account. For the ASD data set, like the previous cases we exhibit an higher R^2 than the data set of typically developed patients, reaching an $R^2 = 0.6$ for both FIQ and VIQ intelligence scores. Additionally, a part from VIQ-TD data set, the linear regression outperforms the non-linear regressor overall. Indeed we have higher average coefficients of determination for the linear regression model in the test sets with respect to random forest for both ASD data sets and FIQ-TD. This evidence confirms the impression we had in the previous paragraph, namely that we do not lose much information using a linear model instead of a non-linear algorithm. In Appendix A we report the scatter plots of predicted against true values for each of the cross-validated folds related to the best fitting linear model.

Chapter 5

Conclusions

In this work we have challenged a recent problem coming from neuroscience, that is applying predictive regression models on brain networks. Facing this subject, in our research work we either applied two state-of-the-art techniques, and extended them with specific contributions to the pipelines.
In particular we selected two approaches of the literature based on functional connectivity analysis of brain networks, in the specific case correlations between ROIs. Even if the Connectome-based Predictive Model and the contrast subgraph-based approaches share the same perspective on the brain connectivity side, the former is originally employed to solve a linear regression problem while the latter is used in a classification framework. However one of the challenge we tackled in this work has been adapting a method designed for a classification task in a regression framework. Specifically we have seen that applying the contrast subgraph method on different classes of IQ, identified using the quartiles of the distribution as splitting points, and separately analyzing positive and negative correlations allows to reach a good prediction performance of the intelligence score. The results we achieved are comparable with the state-of-the-art CPM method that is designed for predicting the intelligence scores.
In addition, we discovered that the gained contrast subgraph features merged with the variables extracted using the Connectome-based Predictive Model produce a setting of features that, in a linear regression fashion, outperforms the SOTA methods. We also remark that, at least in our application, the use of non-linear regression algorithms does not give an increase of the predictive performance, suggesting that the selected features are effectively explanatory of the response variable, in a linear regression framework. Last but not least, we also noticed that in all the methods we considered have an higher predictive performance on subjects affected by Autism Spectrum Disorder: this could be due to the variety of data, given that the condition of Autism is not unique, but is represented by a wide spectrum.

However, these important results are obtained in a very specific application setting. Indeed, we worked with a predefined sets of observations, obtained selecting male patients in a certain age range and dividing subjects affected by Autism Spectrum Disorder and Typically Developed, in order to reduce as much as possible the noise in brain networks due to demographic characteristics. Hence, we plan to enlarge our research with future investigation generalizing our results. In particular in our on-going analysis we plan to:

- Work with other brain data sets and different scalar response variables for regression. Examples of interesting applications can be studying the predictable effect of brain networks on age or the impact of drugs;

- Analyze and compare the estimated substructure and the subnetworks embedding of the different methods, beyond regression performance;
- Study the subnetwork extracted from patients' whole brain networks, as well as the different embeddings, from a neuroscience standpoint and verify their connotation with the state-of-the-art.

Appendix A

Appendix

Chapter 3

α	***ASD data set - FIQ***
70	[3, 7, 9, 10, 13, 15, 22, 31, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 56, 57, 59, 63, 73, 93, 98, 103]; [0, 1, 6, 12, 16, 21, 23, 24, 25, 27, 34, 35, 40, 58, 59, 60, 61, 74, 82, 84, 85, 90, 99, 110]
75	[3, 9, 17, 22, 26, 28, 29, 62, 63, 64, 73, 76, 77, 79, 81, 92, 93]; [0, 1, 12, 13, 16, 17, 21, 24, 25, 26, 27, 34, 35, 58, 60, 61, 85, 110]
80	[3, 10, 15, 42, 43, 45, 46, 47, 48, 49, 50, 51, 53, 60, 63]; [1, 12, 16, 21, 24, 25, 27, 35, 58, 59, 60, 61, 85]
85	[42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 60]; [46, 54, 55, 64, 65, 75, 96, 97, 99, 111]
90	[42, 43, 46, 48, 49, 50, 51, 53, 60]; [46, 55, 64, 65, 96, 97, 99, 111]

Table A.1

α	***ASD data set - VIQ***
70	[3, 10, 13, 17, 29, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 59, 60, 62, 63, 73, 75, 77, 81, 87, 88, 93, 98, 101, 103]; [1, 5, 6, 11, 16, 17, 21, 23, 24, 25, 27, 34, 35, 38, 55, 57, 58, 61, 66, 67, 84, 85, 90, 97, 99]
75	[3, 10, 13, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 60, 62, 63, 73, 75, 77, 81, 88, 98]; [1, 6, 11, 16, 17, 21, 27, 38, 40, 46, 54, 55, 58, 64, 82, 84, 94, 96, 97, 99]
80	[13, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 60, 63, 88, 98]; [11, 21, 27, 38, 40, 46, 54, 55, 64, 84, 94, 96, 97, 99]
85	[13, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 60, 63, 98]; [1, 11, 16, 17, 21, 25, 27, 34, 35, 61]
90	[42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 60, 98]; [1, 16, 17, 25, 27, 34, 35]

Table A.2

α	***TD data set - FIQ***
70	[10, 11, 14, 22, 23, 26, 27, 31, 51, 57, 59, 64, 69, 82, 84, 89, 90, 101, 102, 103]; [18, 19, 27, 28, 29, 30, 33, 40, 42, 43, 72, 74, 83, 90, 91, 94, 98, 104, 105, 106, 108, 114]
75	[10, 11, 14, 22, 23, 26, 27, 59, 64, 69, 82, 84, 89, 90, 101, 102, 103]; [19, 28, 29, 33, 40, 43, 72, 74, 83, 90, 91, 94, 105, 106, 108, 114]
80	[10, 11, 14, 22, 23, 26, 27, 59, 64, 69, 82, 84, 89, 101, 102, 103]; [18, 19, 28, 29, 32, 33, 43, 72, 83, 105, 106, 114]
85	[10, 11, 22, 23, 64, 82, 84, 89, 101, 102, 103]; [18, 19, 29, 33, 43, 83, 91, 114]
90	[10, 11, 23, 64, 82, 84, 89, 101, 102, 103]; [17, 18, 29, 33, 43, 83]

Table A.3

α	***TD data set - VIQ***
70	[2, 3, 4, 6, 7, 8, 9, 10, 14, 15, 16, 17, 22, 23, 26, 54, 55, 64, 70, 72, 77, 78, 84, 88, 89, 95, 97, 98, 99, 111, 112]; [1, 16, 19, 21, 22, 23, 24, 25, 27, 28, 29, 32, 33, 35, 42, 43, 47, 50, 62, 63, 65, 80, 83, 93, 106, 107, 110, 114]
75	[2, 3, 4, 6, 8, 9, 14, 15, 16, 26, 46, 54, 55, 64, 72, 77, 78, 89, 97, 98, 99, 111, 112]; [1, 21, 22, 23, 24, 25, 27, 28, 29, 32, 33, 35, 43, 50, 62, 63, 65, 80, 83, 93, 110]
80	[2, 3, 4, 6, 8, 9, 14, 15, 26, 46, 54, 55, 64, 72, 77, 97, 98, 99, 111, 112]; [1, 21, 22, 23, 24, 25, 27, 29, 35, 43, 47, 50, 60, 62, 63, 110]
85	[2, 3, 4, 6, 8, 14, 26, 64, 72, 78, 97, 98, 99, 111, 112]; [1, 21, 22, 23, 24, 27, 29, 35, 43, 50, 60, 62, 63]
90	[2, 6, 8, 14, 47, 64, 98, 99, 111, 112]; [1, 22, 23, 27, 29, 35, 43, 50, 62, 63]

Table A.4

Results - Contrast Subgraph

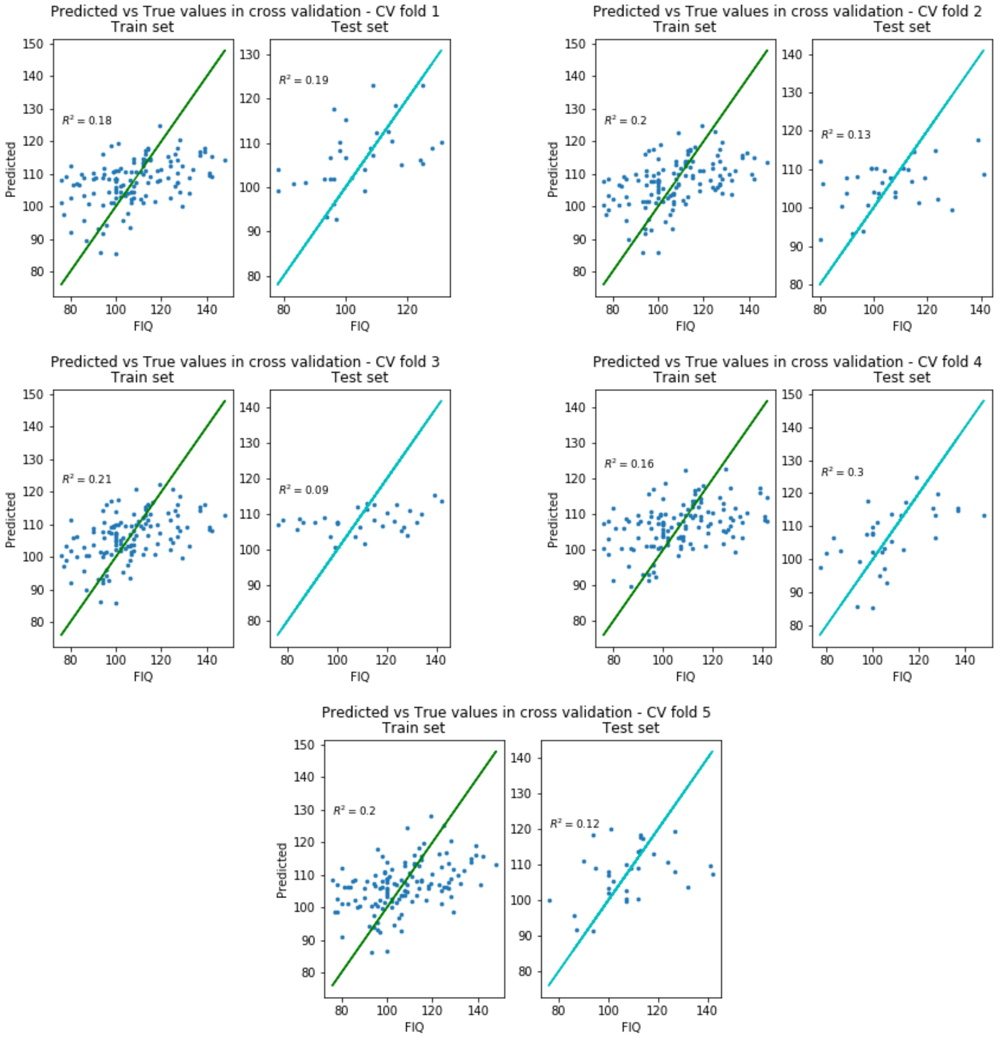


Figure A.1. Contrast subgraph FIQ ASD

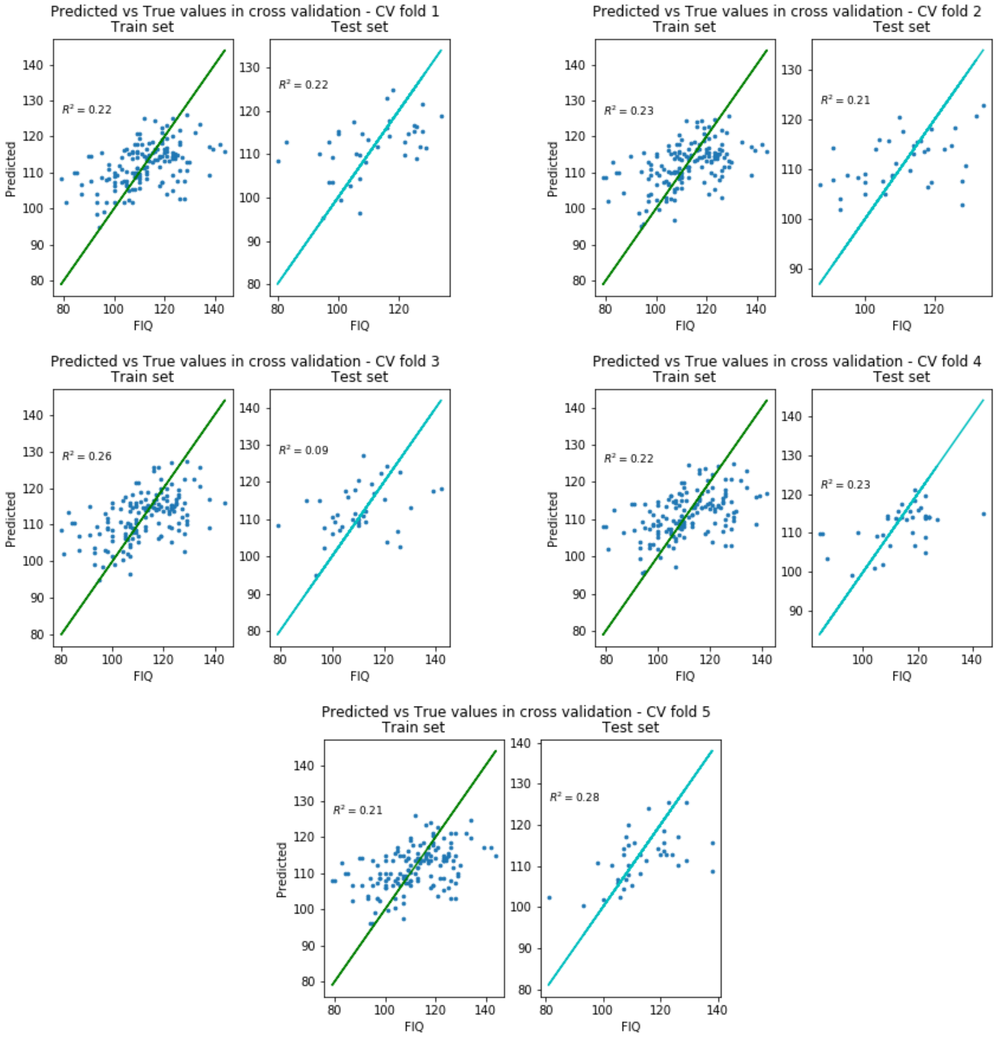


Figure A.2. Contrast subgraph FIQ TD.

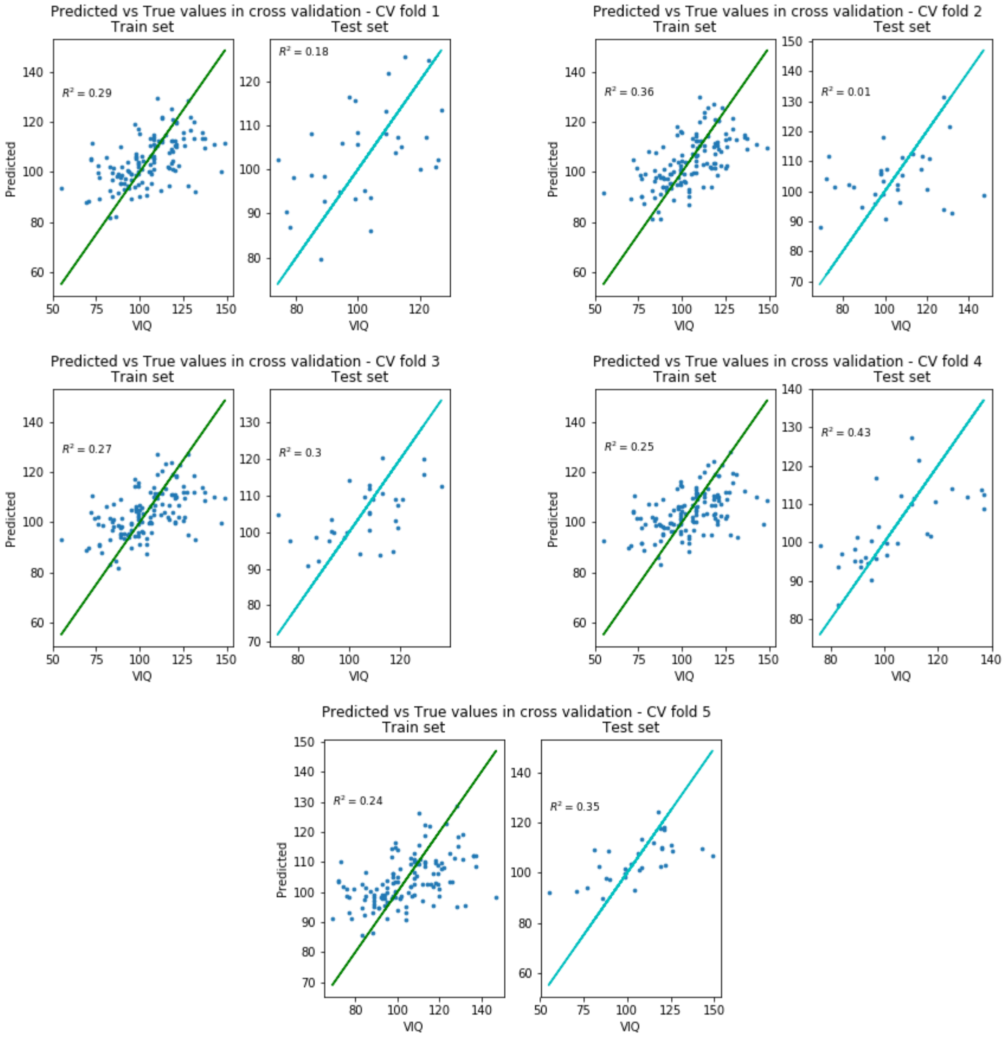


Figure A.3. Contrast subgraph VIQ ASD.

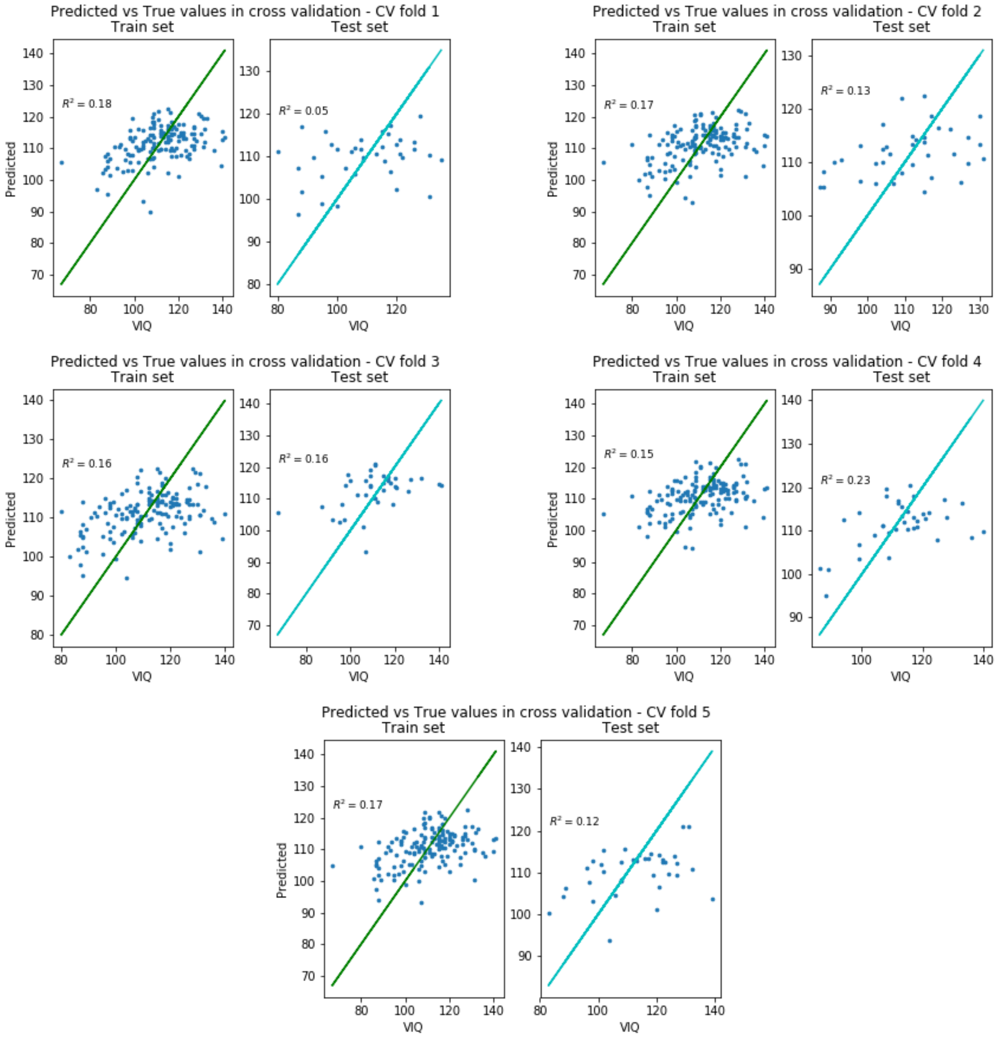


Figure A.4. Contrast subgraph VIQ TD.

Results - Connectome-based Predictive Model

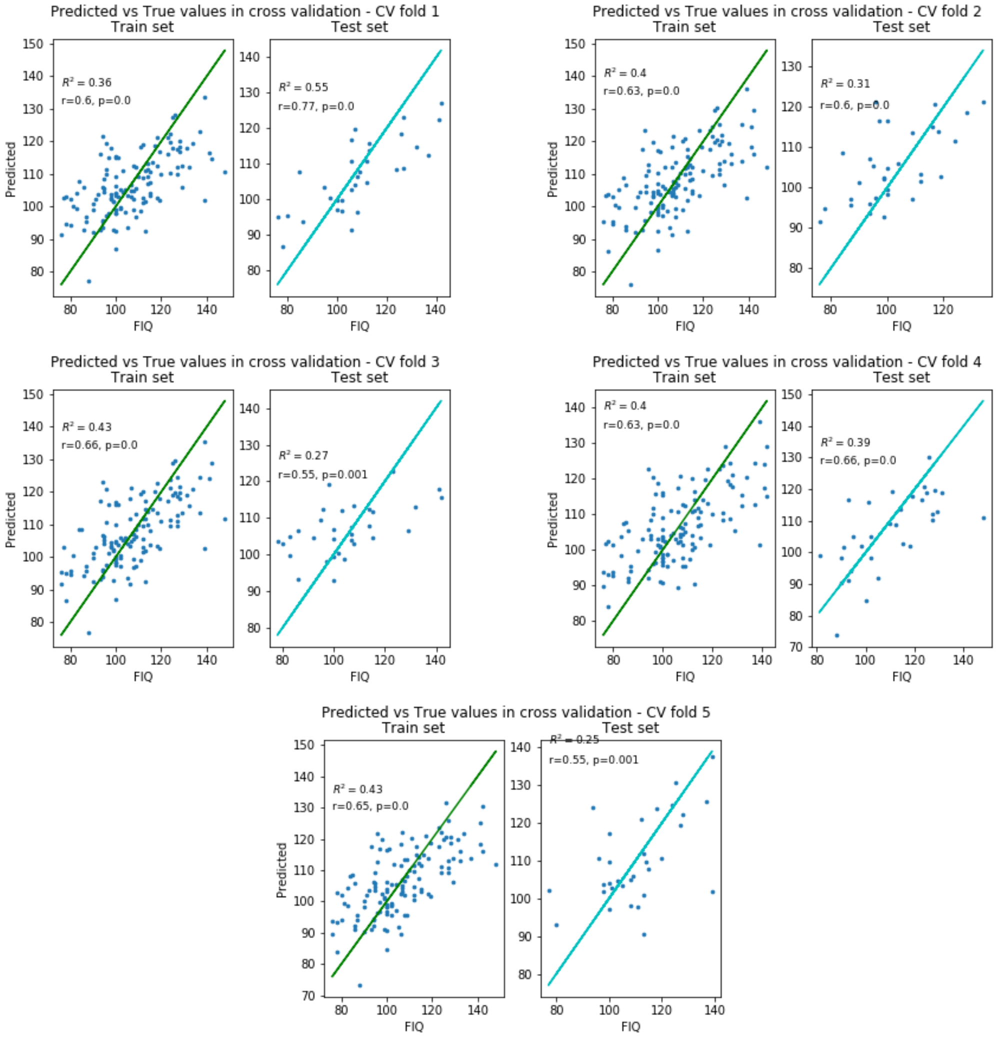


Figure A.5. Connectome-based Predictive Model FIQ ASD-positive.

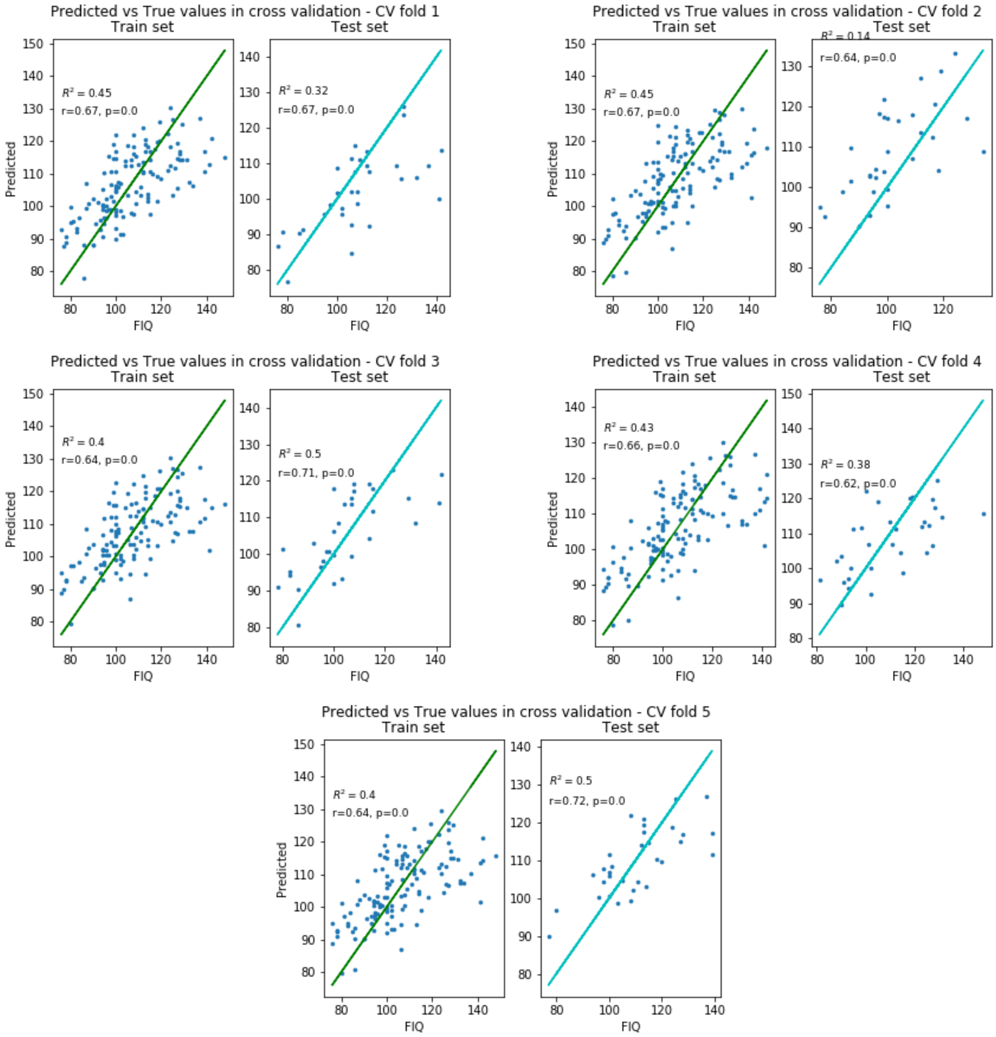


Figure A.6. Connectome-based Predictive Model FIQ ASD-negative.

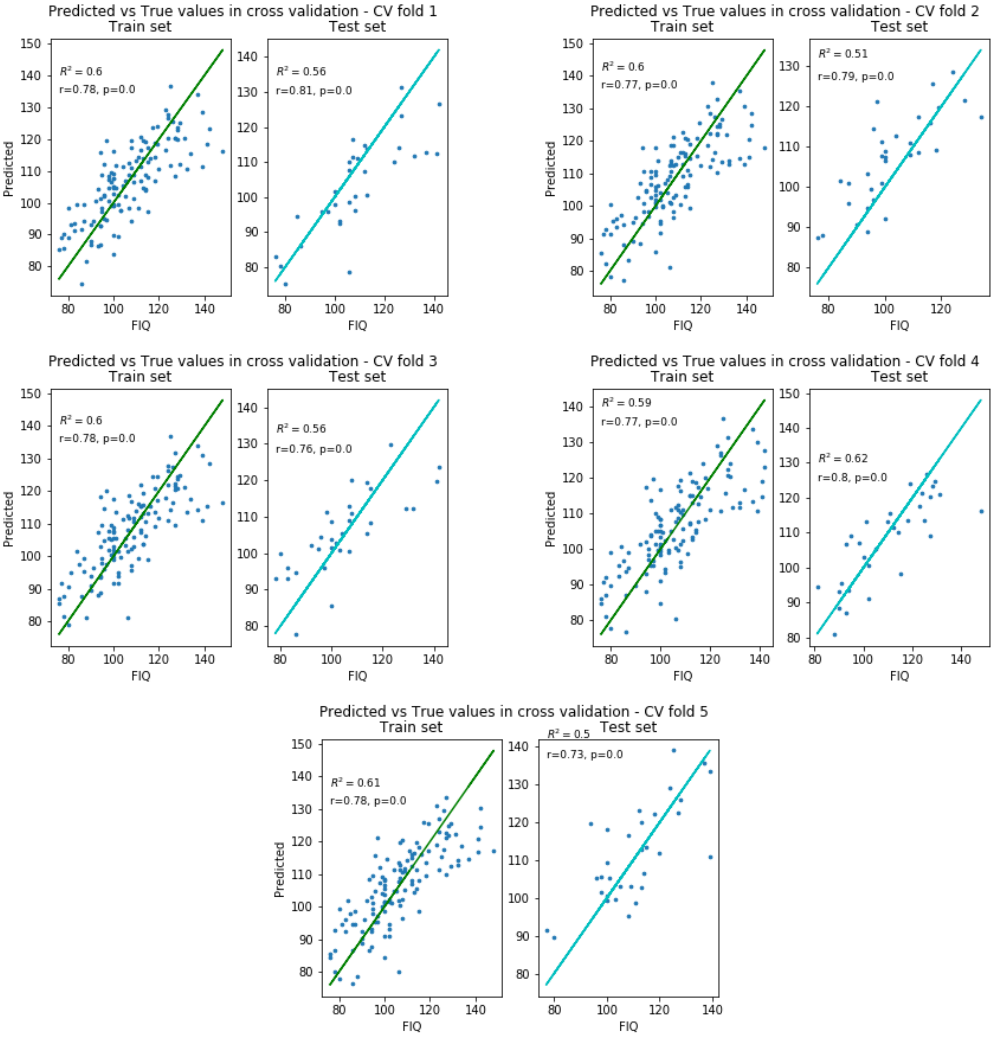


Figure A.7. Connectome-based Predictive Model FIQ ASD.

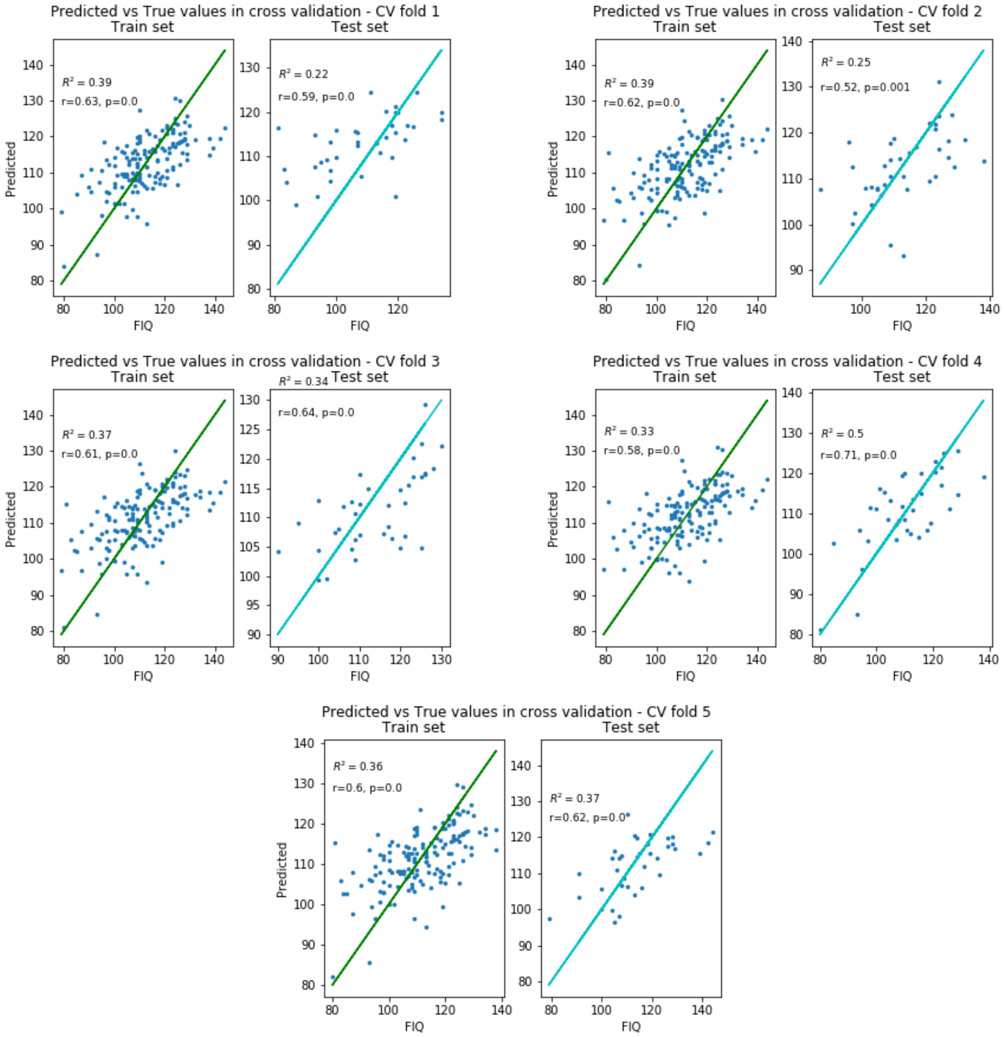


Figure A.8. Connectome-based Predictive Model FIQ TD-positive.

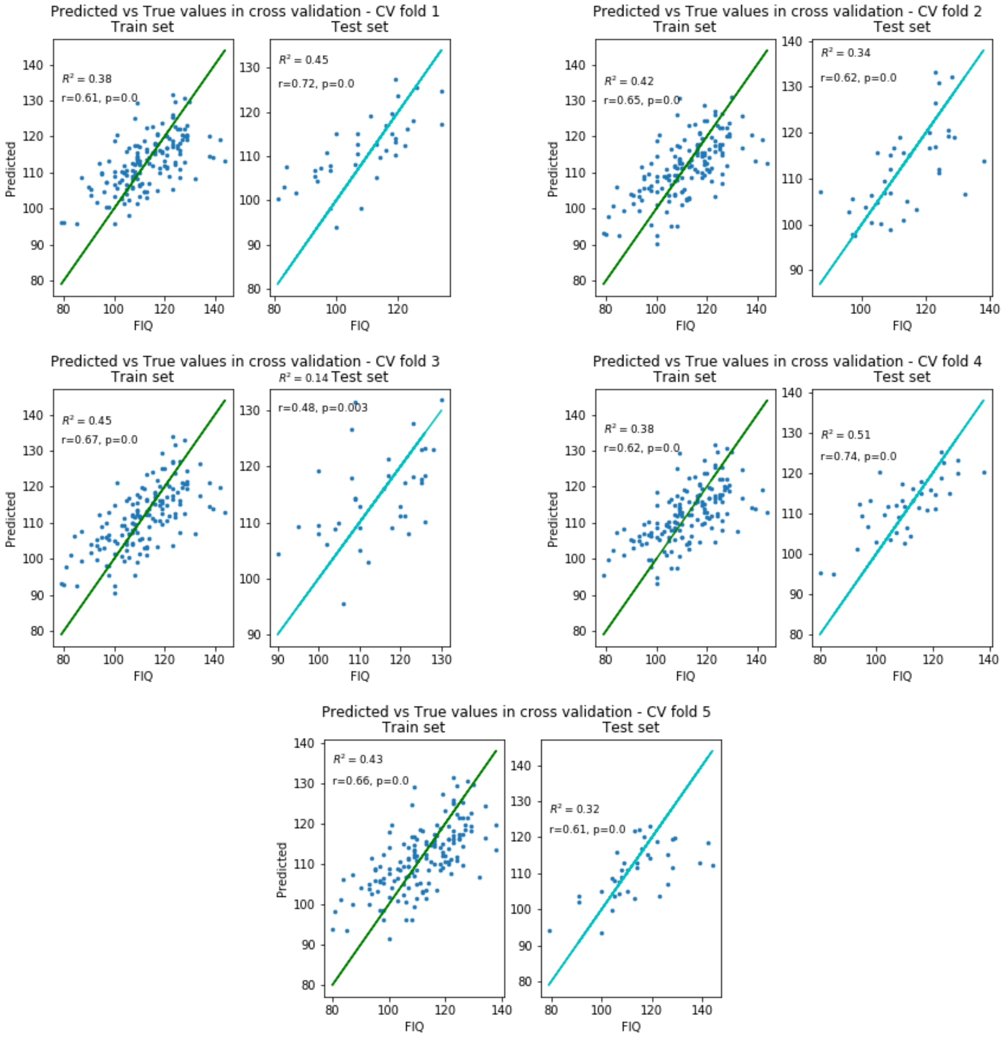


Figure A.9. Connectome-based Predictive Model FIQ TD-negative.

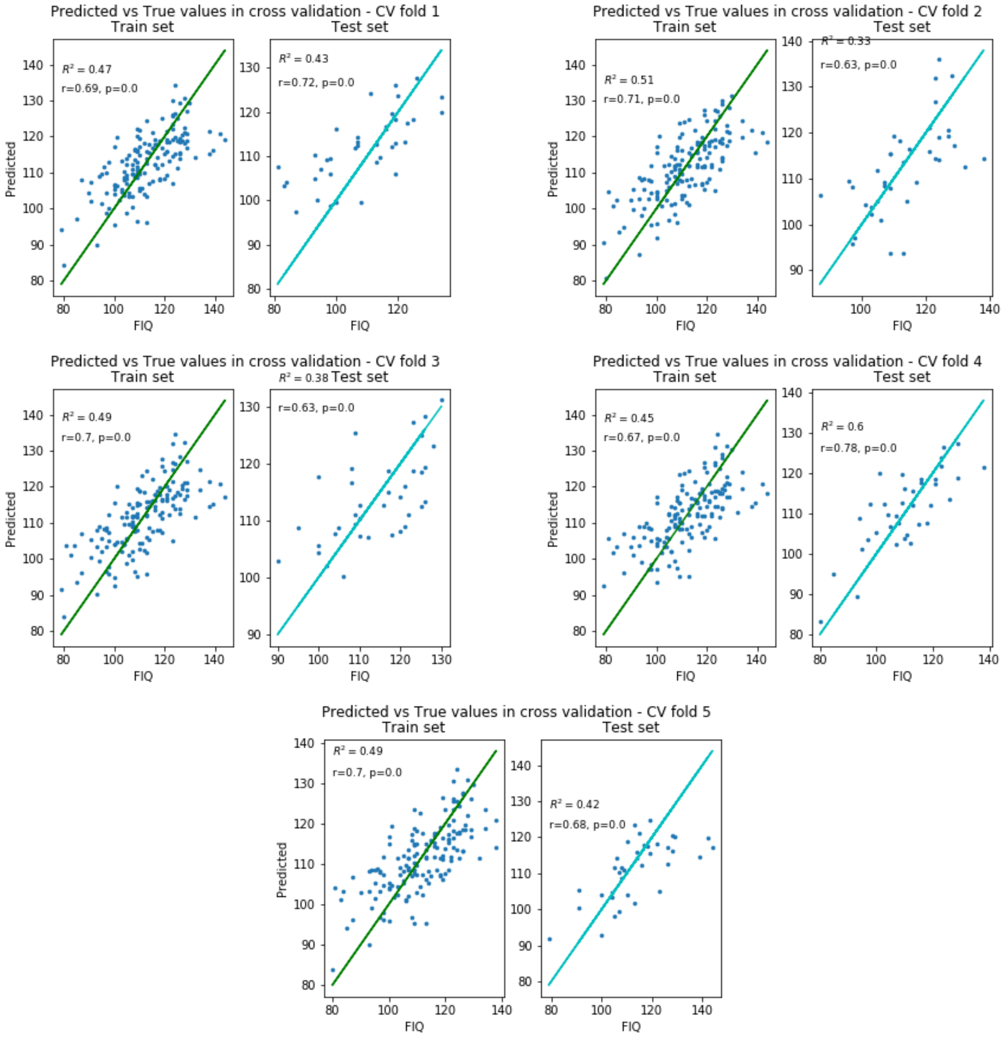


Figure A.10. Connectome-based Predictive Model FIQ TD.

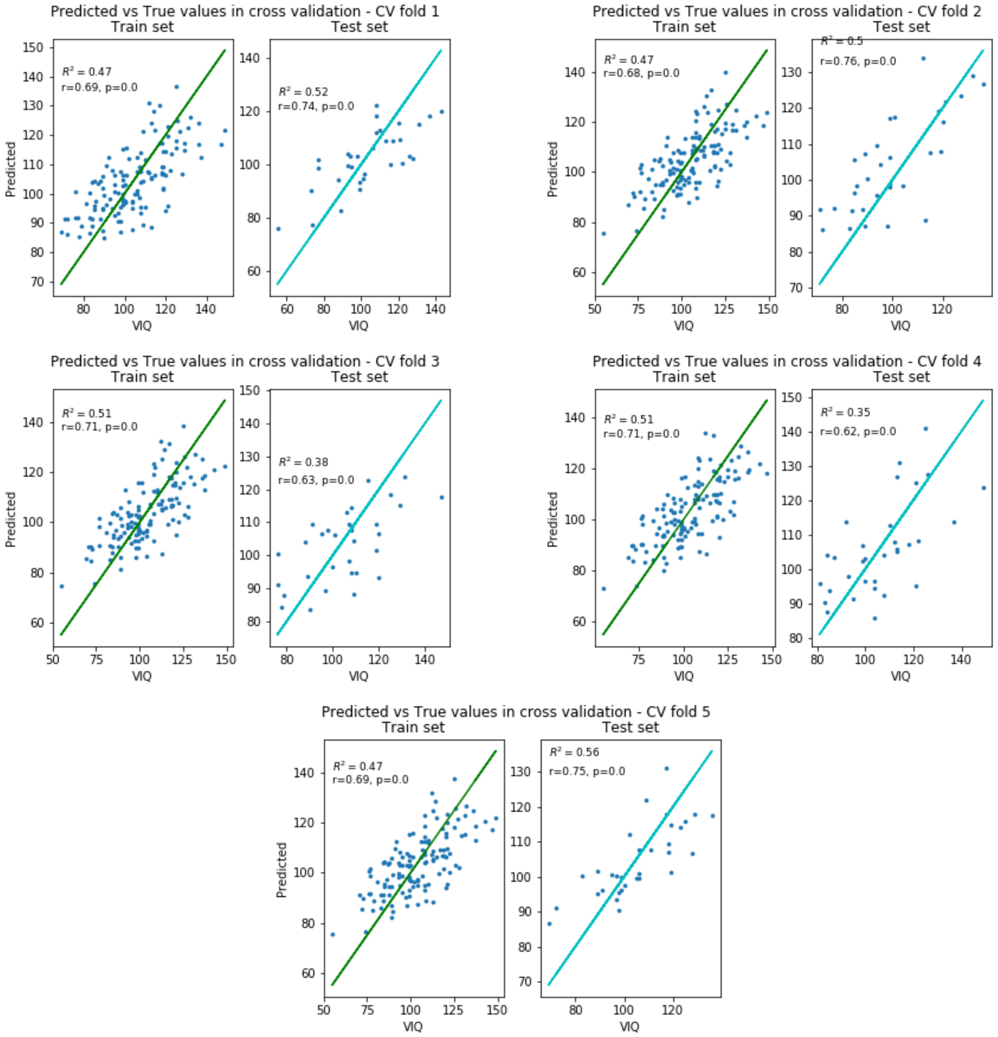


Figure A.11. Connectome-based Predictive Model VIQ ASD-positive.

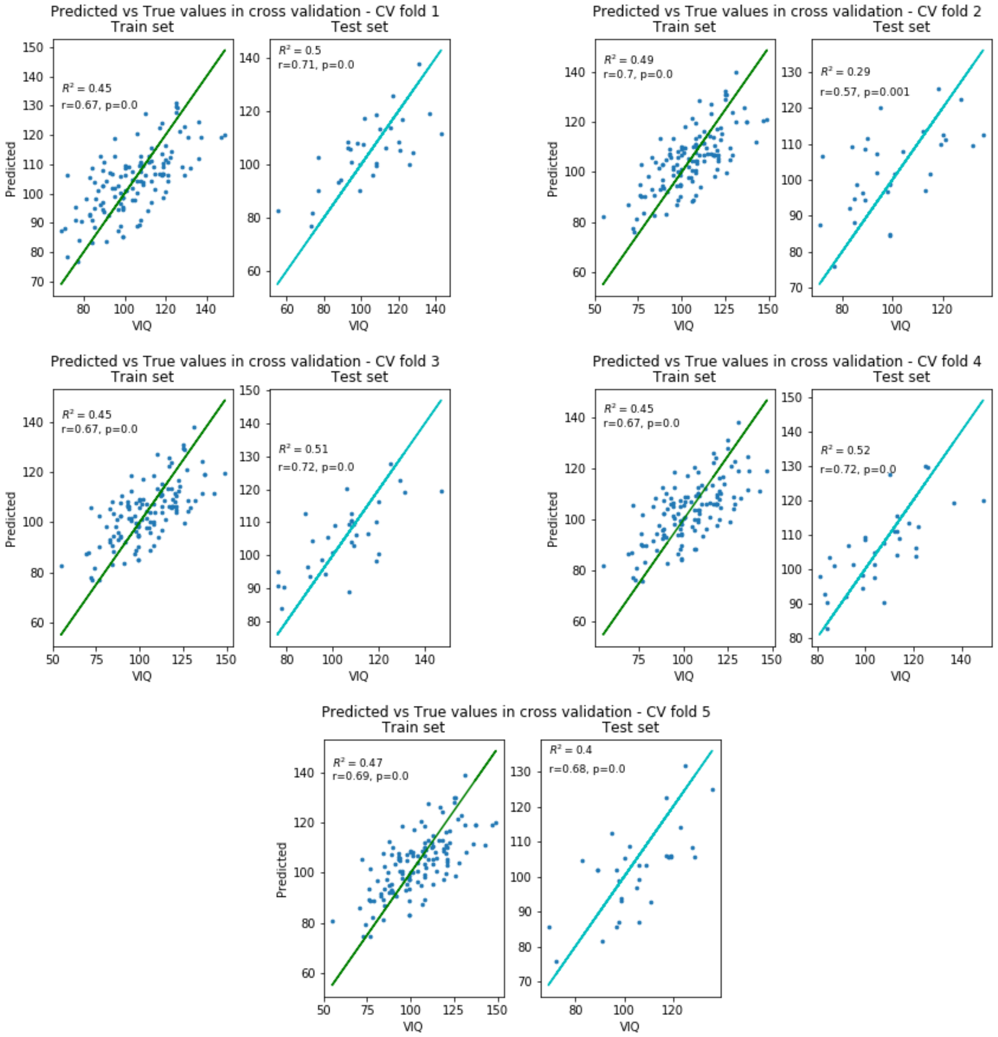


Figure A.12. Connectome-based Predictive Model VIQ ASD-negative.

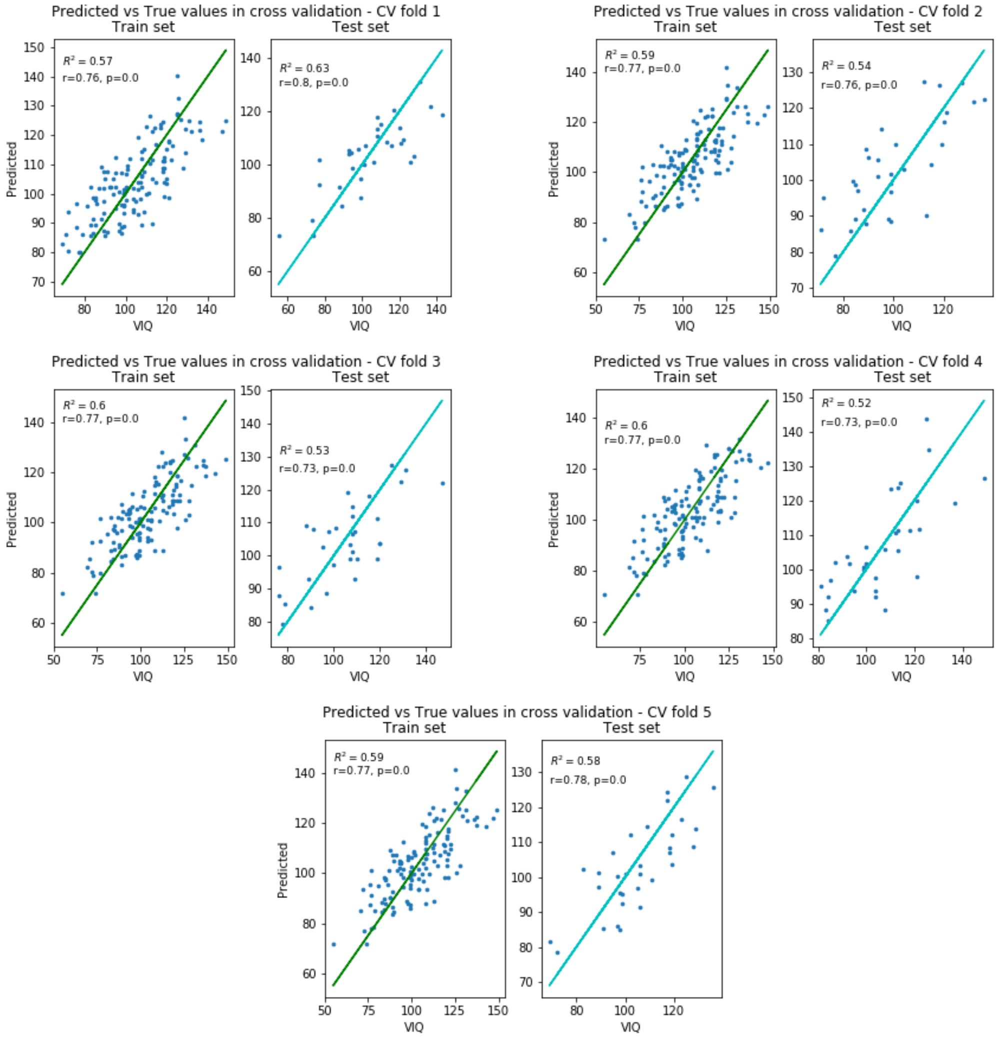


Figure A.13. Connectome-based Predictive Model VIQ ASD.

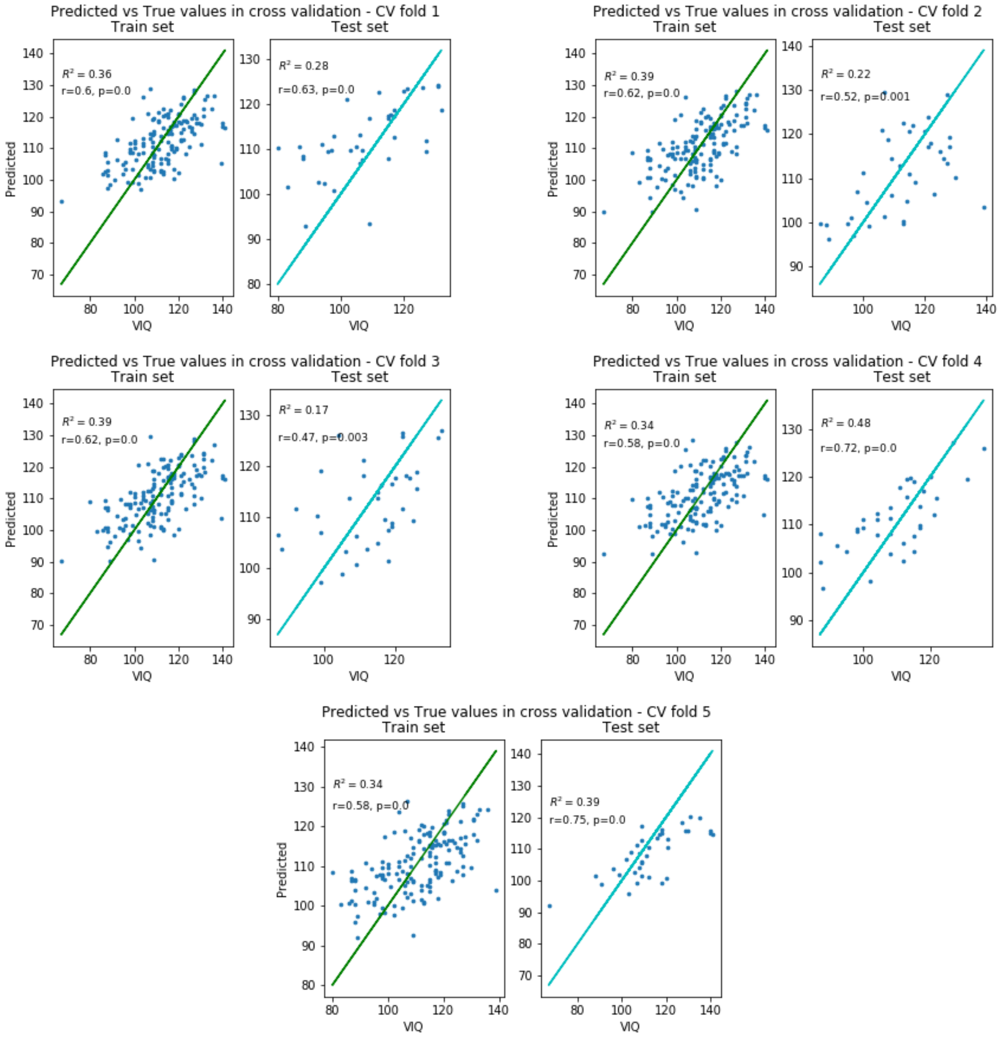


Figure A.14. Connectome-based Predictive Model VIQ TD-negative.

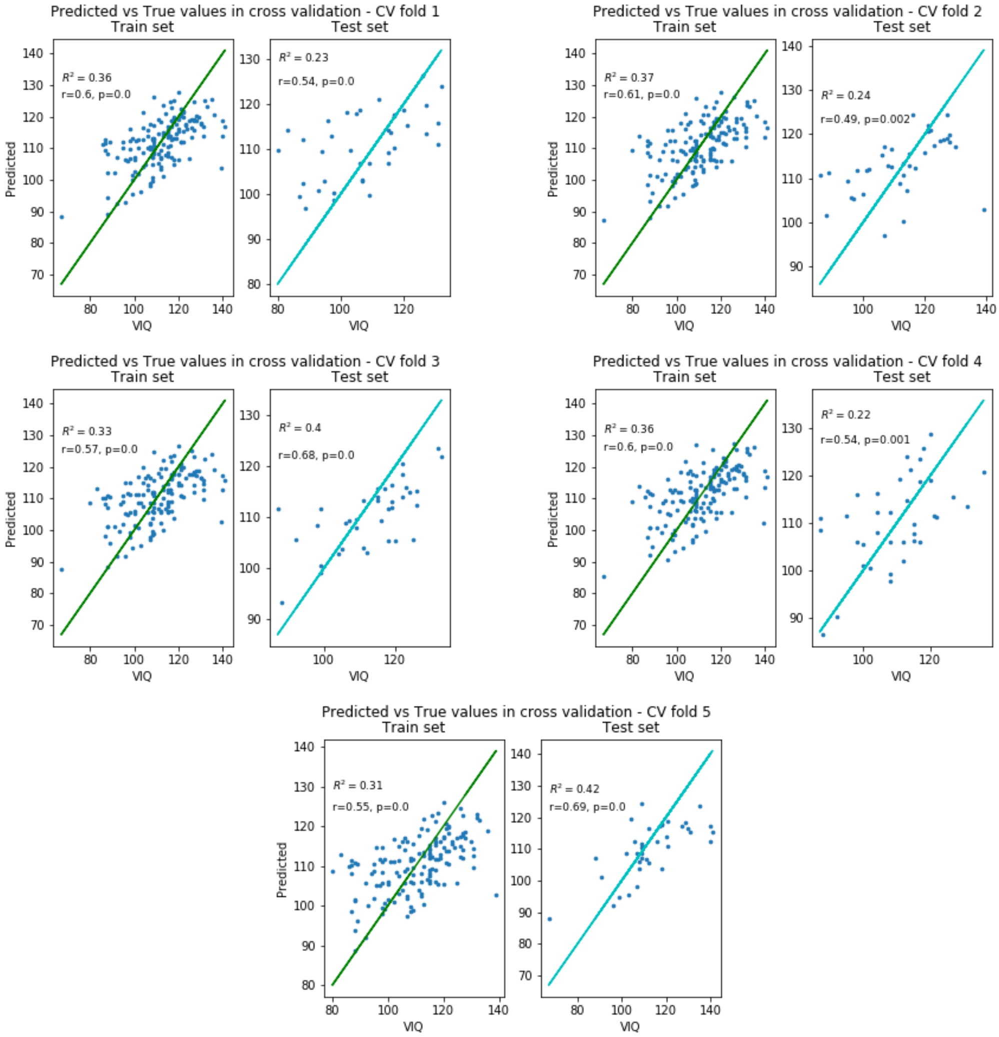
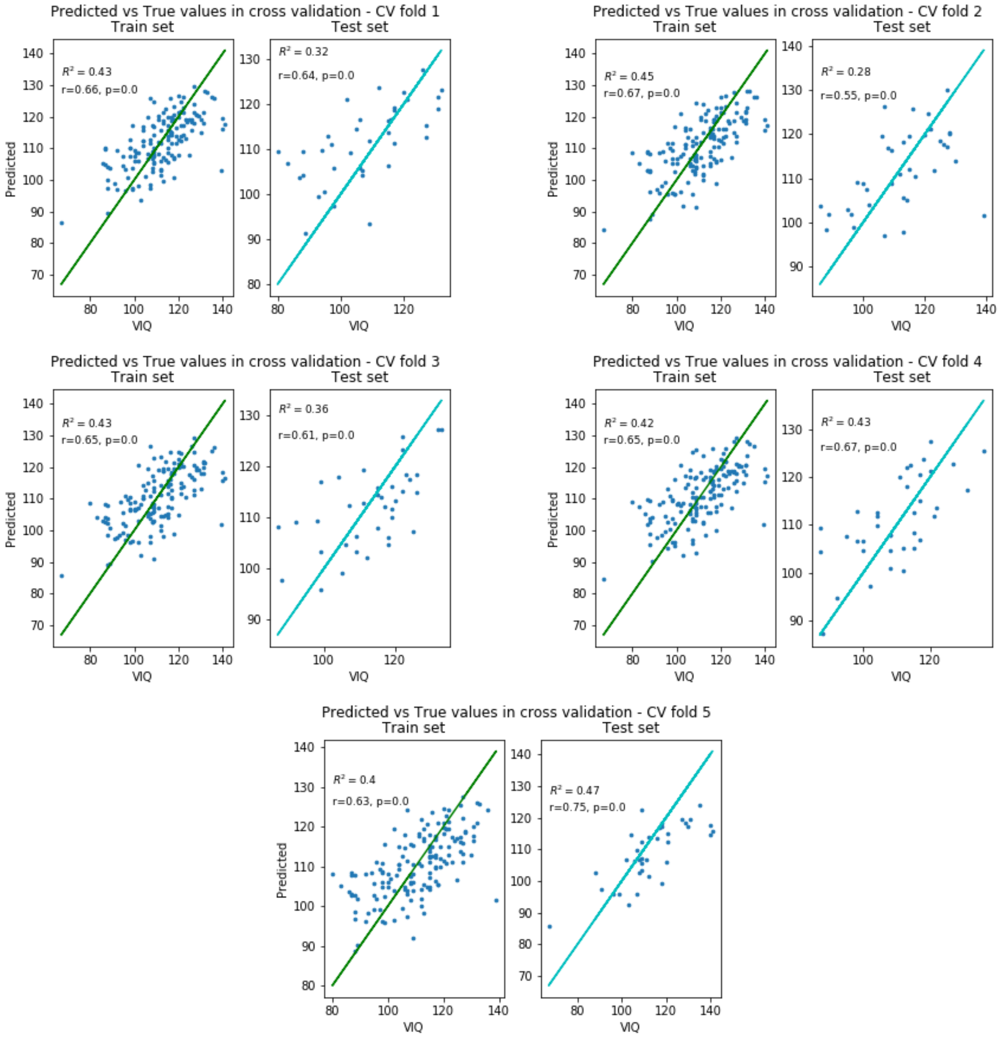


Figure A.15. Connectome-based Predictive Model VIQ TD-positive.



Predicted vs True values in cross validation - CV fold 1

Predicted vs True values in cross validation - CV fold 2

Predicted vs True values in cross validation - CV fold 3

Predicted vs True values in cross validation - CV fold 4

Predicted vs True values in cross validation - CV fold 5

Figure A.16. Connectome-based Predictive Model VIQ TD.

Chapter 4

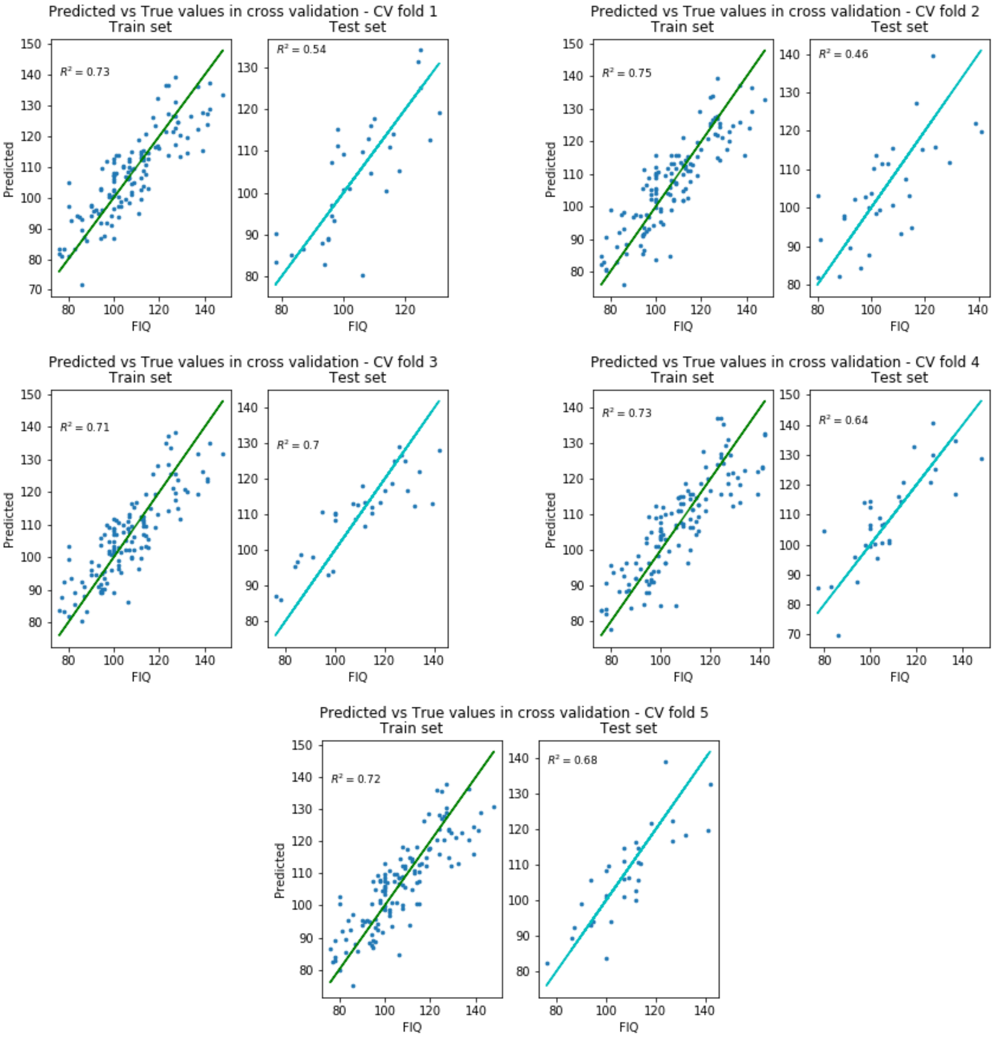


Figure A.17. Contrast subgraph and CPM features (both positive and negative), FIQ-ASD data set.

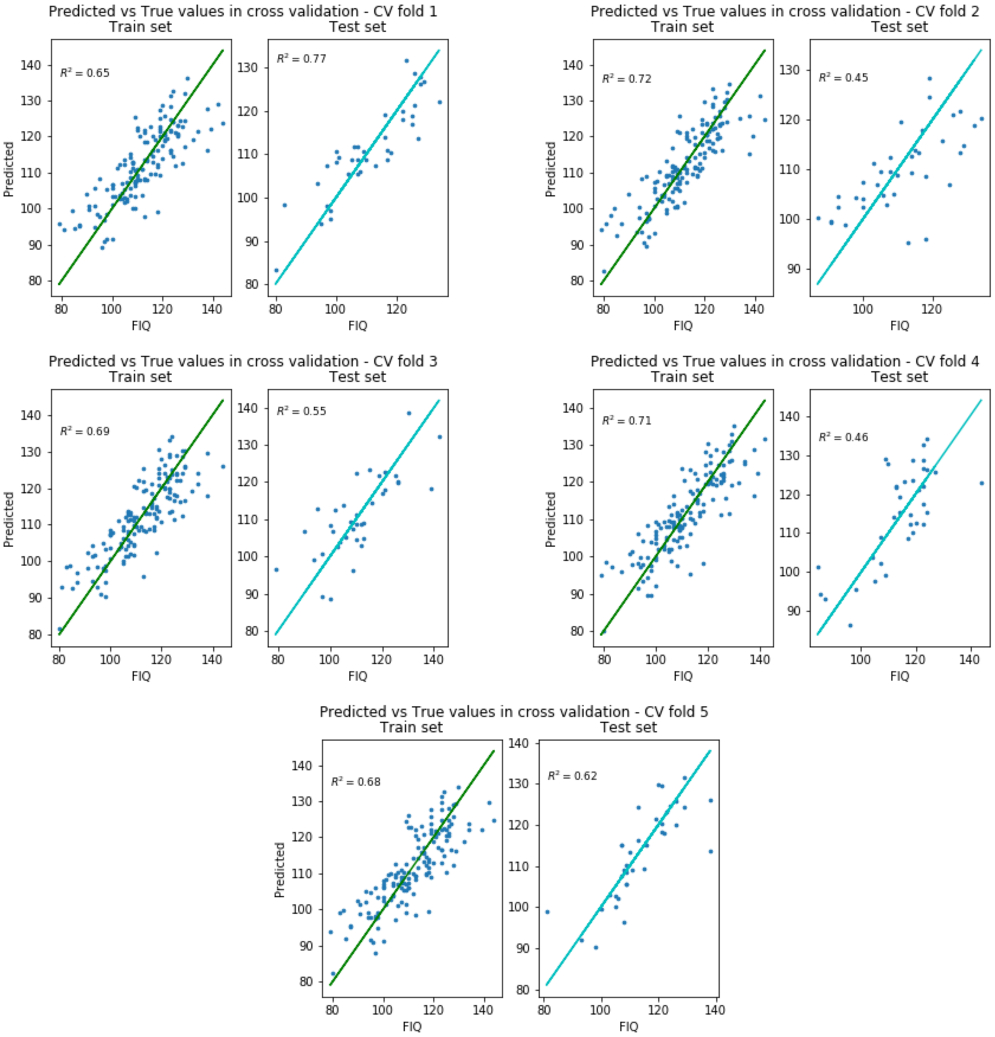


Figure A.18. Contrast subgraph and CPM features (both positive and negative), FIQ-TD data set.

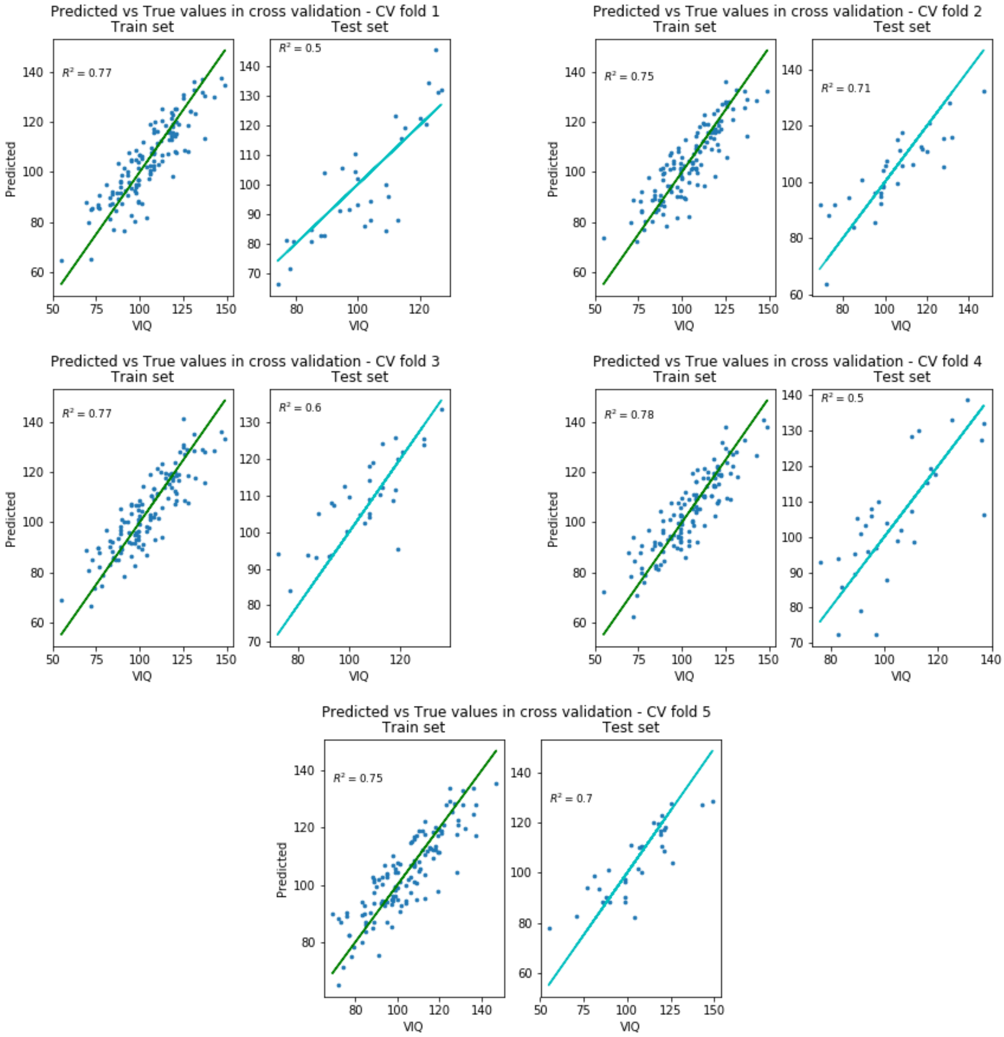


Figure A.19. Contrast subgraph and CPM features (both positive and negative), VIQ-ASD data set.

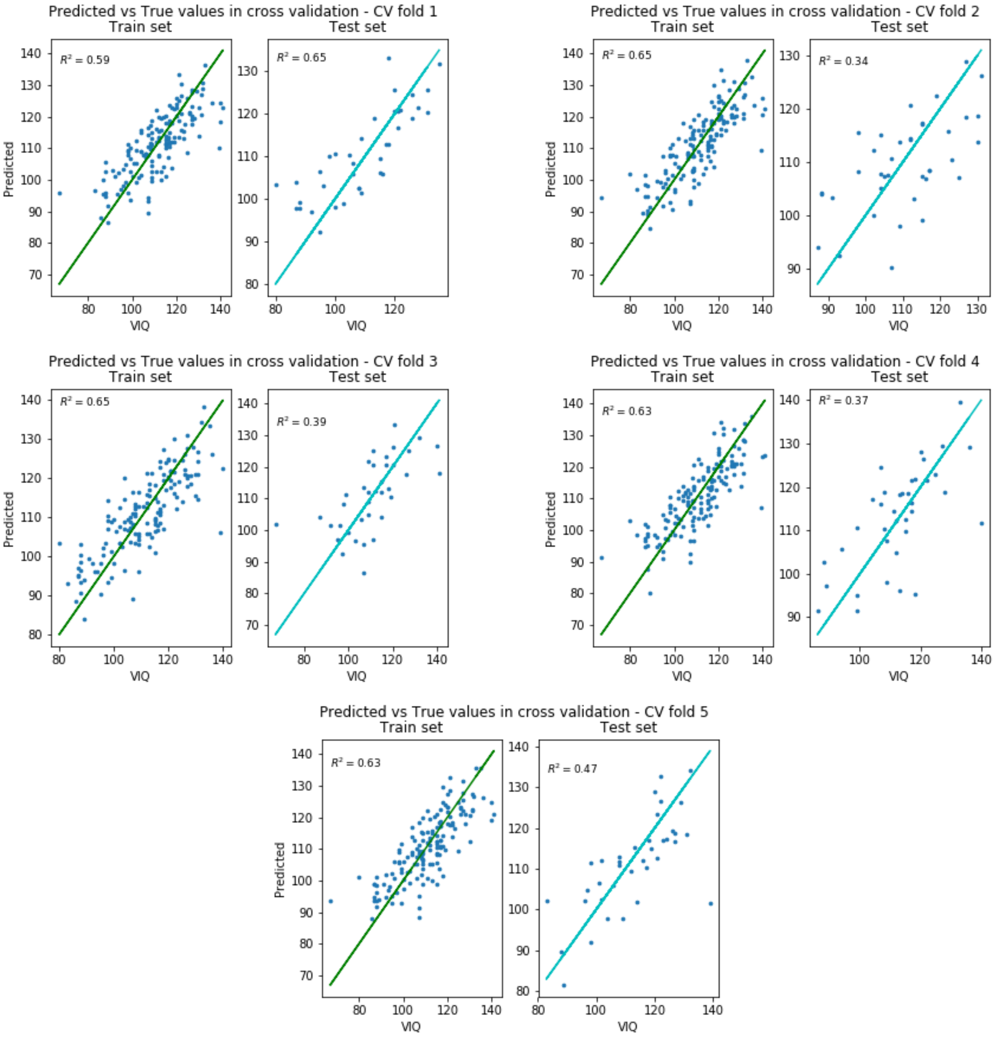


Figure A.20. Contrast subgraph and CPM features (both positive and negative), VIQ-TD data set.

Bibliography

[1] A. Aertsen, G. Gerstein, M. Habib, and G. Palm. Dynamics of neuronal firing correlation: modulation of" effective connectivity". *Journal of neurophysiology*, 61(5):900–917, 1989.

[2] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web*, pages 37–48, 2013.

[3] B. Alharbi and X. Zhang. Learning from your network of friends: A trajectory representation learning model based on online social ties. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 781–786. IEEE, 2016.

[4] J. C. Beer, H. J. Aizenstein, S. J. Anderson, and R. T. Krafty. Incorporating prior information with fused sparse group lasso: Application to prediction of clinical measures from neuroimages. *Biometrics*, 75(4):1299–1309, 2019.

[5] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8(3):3–24, 2007.

[6] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

[7] B. B. Biswal, J. V. Kylen, and J. S. Hyde. Simultaneous assessment of flow and bold signals in resting-state functional connectivity maps. *NMR in Biomedicine*, 10(4-5):165–170, 1997.

[8] K. M. Borgwardt and H.-P. Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[9] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences*, 101(26):9849–9854, 2004.

[11] C. J. Brown, S. P. Miller, B. G. Booth, J. G. Zwicker, R. E. Grunau, A. R. Synnes, V. Chau, and G. Hamarneh. Predictive connectome subnetwork extraction with anatomical and connectivity priors. *Computerized Medical Imaging and Graphics*, 71:67–78, 2019.

[12] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[13] E. Bullmore, B. Horwitz, G. Honey, M. Brammer, S. Williams, and T. Sharma. How good is good enough in path analysis of fmri data? *NeuroImage*, 11(4):289–301, 2000.

[14] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

[15] H. Cai, V. W. Zheng, and K. C.-C. Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

[16] R. Casanova, C. Whitlow, B. Wagner, M. Espeland, and J. Maldjian. Combining graph and machine learning methods to analyze differences in functional connectivity across sex. *The open neuroimaging journal*, 6:1, 2012.

[17] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, and K. Maravilla. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic resonance imaging*, 20(4):305–317, 2002.

[18] D. Cordes, V. M. Haughton, K. Arfanakis, G. J. Wendt, P. A. Turski, C. H. Moritz, M. A. Quigley, and M. E. Meyerand. Mapping functionally related regions of brain with functional connectivity mr imaging. *American journal of neuroradiology*, 21(9):1636–1644, 2000.

[19] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Neuroinformatics*, 4, 2013.

[20] J. S. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann. Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853, 2006.

[21] M. De Luca, S. Smith, N. De Stefano, A. Federico, and P. M. Matthews. Blood oxygenation level dependent contrast resting state networks are relevant to functional activity in the neocortical sensorimotor system. *Experimental brain research*, 167(4):587–594, 2005.

[22] E. Dryburgh, S. McKenna, and I. Rekik. Predicting full-scale and verbal intelligence scores from functional connectomic data in individuals with autism spectrum disorder. *Brain imaging and behavior*, pages 1–10, 2019.

[23] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression.* Springer, 2007.

[24] H. Fang, F. Wu, Z. Zhao, X. Duan, Y. Zhuang, and M. Ester. Community-based question answering via heterogeneous social network learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 122–128, 2016.

[25] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

[26] M. D. Fox and M. E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, 8(9):700–711, 2007.

[27] K. Friston, C. Frith, P. Liddle, and R. Frackowiak. Functional connectivity: the principal-component analysis of large (pet) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14, 1993.

[28] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[29] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.

[30] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor. Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, 72:304–321, 2013.

[31] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

[32] M. Henaff, J. Bruna, and Y. LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[33] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[34] T. Lanciano, F. Bonchi, and A. Gionis. Explainable classification of brain networks via contrast subgraphs. *arXiv preprint arXiv:2006.05176*, 2020.

[35] E. W. Lang, A. M. Tomé, I. R. Keck, J. Górriz-Sáez, and C. G. Puntonet. Brain connectivity analysis: A short survey. *Computational intelligence and neuroscience*, 2012.

[36] H. Li, Z. Xue, T. M. Ellmore, R. E. Frye, and S. T. Wong. Identification of faulty dti-based sub-networks in autism using network regularized svm. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 550–553. IEEE, 2012.

[37] A. Lord, D. Horn, M. Breakspear, and M. Walter. Changes in community structure of resting state functional connectivity in unipolar depression. *PloS one*, 7(8):e41282, 2012.

[38] M. Lowe, B. Mock, and J. Sorenson. Functional connectivity in single and multislice echoplanar imaging using resting-state fluctuations. *Neuroimage*, 7(2):119–132, 1998.

[39] M. J. Lowe, M. Dzemidzic, J. T. Lurito, V. P. Mathews, and M. D. Phillips. Correlations in low-frequency bold fluctuations reflect cortico-cortical connections. *Neuroimage*, 12(5):582–587, 2000.

[40] H. Lv, Z. Wang, E. Tong, L. M. Williams, G. Zaharchuk, M. Zeineh, A. N. Goldstein-Piekarski, T. M. Ball, C. Liao, and M. Wintermark. Resting-state functional mri: everything that nonexperts have always wanted to know. *American Journal of Neuroradiology*, 39(8):1390–1399, 2018.

[41] A. McIntosh, C. Grady, L. G. Ungerleider, J. Haxby, S. Rapoport, and B. Horwitz. Network analysis of cortical visual pathways mapped with pet. *Journal of Neuroscience*, 14(2):655–666, 1994.

[42] D. C. Montgomery and G. C. Runger. *Applied statistics and probability for engineers.* John Wiley & Sons, 2010.

[43] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.

[44] F. A. Nasrallah, L. Y. Yeow, B. Biswal, and K.-H. Chuang. Dependence of bold signal fluctuation on arterial blood co2 and o2: implication for resting-state functional connectivity. *Neuroimage*, 117:29–39, 2015.

[45] G. Nikolentzos, P. Meladianos, and M. Vazirgiannis. Matching node embeddings for graph similarity. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[46] J. W. Osborne and E. Waters. Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation*, 8(1):2, 2002.

[47] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[48] M. Rubinov, S. A. Knock, C. J. Stam, S. Micheloyannis, A. W. Harris, L. M. Williams, and M. Breakspear. Small-world properties of nonlinear brain activity in schizophrenia. *Human brain mapping*, 30(2):403–416, 2009.

[49] X. Shen, E. S. Finn, D. Scheinost, M. D. Rosenberg, M. M. Chun, X. Papademetris, and R. T. Constable. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *nature protocols*, 12(3):506–518, 2017.

[50] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

[51] R. Tibshirani. Regression shrinkage and selection via the lasso. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 58:267–288, 1994.

[52] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[53] M. Van Den Heuvel, R. Mandl, and H. H. Pol. Normalized cut group clustering of resting-state fmri data. *PloS one*, 3(4):e2001, 2008.

[54] M. P. Van Den Heuvel and H. E. H. Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[55] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

[56] T. Watanabe, D. Kessler, C. Scott, M. Angstadt, and C. Sripada. Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. *Neuroimage*, 96:183–202, 2014.

[57] J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[58] H. Xiao, M. Huang, L. Meng, and X. Zhu. Ssp: semantic space projection for knowledge graph embedding with text descriptions. In *Thirty-First AAAI conference on artificial intelligence*, 2017.

[59] J. Xiong, L. M. Parsons, J.-H. Gao, and P. T. Fox. Interregional connectivity to primary motor cortex revealed using mri resting state images. *Human brain mapping*, 8(2-3):151–156, 1999.

[60] P. Yanardag and S. Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374, 2015.

[61] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Expert finding for community-based question answering via ranking metric network learning. In *Ijcai*, volume 16, pages 3000–3006, 2016.