AN INFORMATION THEORETIC REPRESENTATION OF BRAIN CONNECTIVITY FOR COGNITIVE STATE CLASSIFICATION USING FUNCTIONAL MAGNETIC RESONANCE IMAGING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

ITIR ÖNAL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2013

Approval of the thesis:

AN INFORMATION THEORETIC REPRESENTATION OF BRAIN CONNECTIVITY FOR COGNITIVE STATE CLASSIFICATION USING FUNCTIONAL MAGNETIC RESONANCE IMAGING

submitted by **ITIR ÖNAL** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering**

Prof. Dr. Fatoş T. Yarman Vural
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Göktürk Üçoluk
Computer Engineering Department, METU

Prof. Dr. Fatoş T. Yarman Vural
Computer Engineering Department, METU

Prof. Dr. İ. Hakkı Toroslu
Computer Engineering Department, METU

Assist. Prof. Dr. Sinan Kalkan
Computer Engineering Department, METU

Dr. Onur Pekcan
Civil Engineering Department, METU

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ITIR ÖNAL

Signature :

ABSTRACT

AN INFORMATION THEORETIC REPRESENTATION OF BRAIN CONNECTIVITY FOR COGNITIVE STATE CLASSIFICATION USING FUNCTIONAL MAGNETIC RESONANCE IMAGING

Önal, Itır
M.S., Department of Computer Engineering
Supervisor : Prof. Dr. Fatoş T. Yarman Vural

September 2013, 83 pages

In this study, a new method for analyzing and representing the discriminative information, distributed in functional Magnetic Resonance Imaging (fMRI) data, is proposed. For this purpose, a local mesh with varying size is formed around each voxel, called the seed voxel. The relationships among each seed voxel and its neighbors are estimated using a linear regression equation by minimizing the expectation of the squared error. This squared error coming from linear regression is used to calculate various information theoretic criteria. Then, the optimal mesh size, which represents the connections among a voxel and its neighbors, is estimated by minimizing these information theoretic criteria with respect to mesh size. The optimal mesh size is used to represent the degree of connectivity such that if the optimal mesh size is small, then the voxel is assumed to be connected with a small number of neighbors. On the other hand, high optimal mesh size indicates that voxels are massively connected. The proposed method shows that the local mesh size with the highest discriminative power depends on the participants, samples in the experiment, and voxels. The results indicate that the local mesh model with optimal mesh size can successfully represent discriminative information.

Keywords: Local Mesh Model, Information Theoretic Criteria, Model Order Selection, fMRI

ÖZ

FONKSİYONEL MANYETİK REZONANS GÖRÜNTÜLEME İLE BİLİŞSEL SÜREÇ SINIFLANDIRMASI İÇİN BEYİNDEKİ BAĞLANIRLIĞIN BİLGİ TEORETİK TEMSİLİ

Önal, Itır
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Fatoş T. Yarman Vural

Eylül 2013 , 83 sayfa

Bu çalışmada, fonksiyonel Manyetik Rezonans Görüntüleme (fMRG) verisinde dağılmış ayrımsayıcı bilginin analizi ve ifadesi için yeni bir metot önerilmiştir. Bu amaçla, her vokselin (tohum vokselin) çevresinde değişken boyutlarda bir yerel örgü oluşturulur. Her tohum voksel ile komşuları arasındaki ilişki, bir doğrusal bağlanım denklemi kullanılarak, hatanın karesi minimize edilecek şekilde kestirilir. Doğrusal bağlanımdan elde edilen hata karesi çeşitli bilgi teoretik kriterlerin hesaplanmasında kullanılır. Daha sonra, voksel ve komşuları arasındaki ilişkileri temsil eden ideal örgü boyutu, bilgi teoretik kriterlerin örgü boutuna göre minimize edilmesiyle kestirilir. İdeal örgü boyutu, bağlanırlık derecesinin ifade edilmesinde kullanılır, öyle ki eğer ideal örgü boyutu küçükse vokselin az sayıda komşusuna bağlandığı varsayılır. Öte yandan, yüksek ideal örgü boyutu voksellerin yoğun olarak bağlantılı olduğunu gösterir. Önerilen metot en yüksek ayrımsayıcı güce sahip yerel örgü boyutunun katılımcılara, deneydeki örneklere ve voksellere bağlı olduğunu göstermektedir. Sonuçlar, ideal örgü boyutuna sahip yerel örgü modelinin, ayrımsayıcı bilgiyi başarıyla temsil ettiğini gösterir.

Anahtar Kelimeler: Yerel Örgü Modeli, Bilgi Teoretik Kriterler, Model Derecesi Seçme, fMRG

To all beloved

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Prof. Dr. Fatoş Tünay Yarman Vural not only for her guidance, unfailing support and encouragement throughout the research, but also for her valuable advices that shape my future. She was the architect of this work and source of inspiration for me with her extensive knowledge, creative thinking and vision.

I wish to offer my special thanks to Assist. Prof. Dr. İlke Öztekin for her generous support and suggestions. I am especially indebted to my colleagues Dr. Mete Özay, Orhan Fırat, Ömer Ekmekçi and Dr. Gülşah Tümüklü Özyer for the illuminating discussions, their valuable support, and cooperation throughout the courses of this investigation.

I appreciate the feedback offered by my thesis committee members Prof. Dr. Göktürk Üçoluk, Prof. Dr. İsmail Hakkı Toroslu, Assist. Prof. Dr. Sinan Kalkan and Dr. Onur Pekcan.

I owe special thanks to all my dearest friends Beybin İlhan, Ece Bulut, Gökçe Aköz, Sinem Balıkçıoğlu, Gülfem İnaner, Derya Erdem and Cemre Artan for their motivation and kind friendship. I should also thank my colleagues at Computer Engineering Department of METU, especially Aybike Şimşek, İlkcan Keleş and Utku Şirin for their support and friendship during my academic research.

I would like to extend my deepest gratitude to my family, without whom I couldn't accomplish this far. I am grateful to my dearest mom, Deniz and my dearest dad, Mustafa for their abiding love, never-ending encouragement and confidence. I also wish to thank my brother Mert, my sister-in-law Figen since no matter how far they are, I always felt their support and motivation with me. I am also grateful for my precious little nephew Orhun, since he gave me the best motivation by learning to call me aunt with his cutest smile during my research period.

Last but surely not the least, my heartfelt appreciation goes to my dearly beloved Mert for his presence beside me each and every single day of the thesis period. Without his endless encouragement, unwavering motivation, joyful and helpful accompany and warm love, this

thesis would not have been completed. I am also deeply grateful for him since he turned my life into an adventurous journey, in which he accompanied me to the far end of the world.

I acknowledge the support of TÜBİTAK (The Scientific and Technological Research Council of Turkey) BİDEB 2210 graduate student fellowship during my M.Sc. education.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

NOMENCLATURE

$\bar{s}_j$	Voxel coordinates
$\epsilon(\tau)$	The error at time τ
$\hat{c}_i$	Estimated class label of sample at t_i
$\hat{C}^{te}$	Estimated class label vector of test data
$\hat{E}_\rho$	Expected value of squared error estimated for participant ρ
$\hat{E}_{cl,\rho}$	Expected value of squared error estimated for class *cl* and participant ρ
$\hat{E}_{i,\rho}$	Expected value of squared error estimated for sample at time instant t_i and for participant ρ
$\hat{E}_{j,\rho}$	Expected value of squared error estimated for voxel at coordinates $\bar{s}_j$ and for participant ρ
$\hat{p}_\rho^{AIC}$	Optimal mesh size estimated for participant ρ using AIC
$\hat{p}_\rho^{BIC}$	Optimal mesh size estimated for participant ρ using BIC
$\hat{p}_\rho^{FPE}$	Optimal mesh size estimated for participant ρ using FPE
$\hat{p}_\rho^{MDL}$	Optimal mesh size estimated for participant ρ using MDL
$\hat{p}_{cl,\rho}^{AIC}$	Optimal mesh size estimated for class *cl* and participant ρ using AIC
$\hat{p}_{cl,\rho}^{BIC}$	Optimal mesh size estimated for class *cl* and participant ρ using BIC
$\hat{p}_{cl,\rho}^{FPE}$	Optimal mesh size estimated for class *cl* and participant ρ using FPE
$\hat{p}_{cl,\rho}^{MDL}$	Optimal mesh size estimated for class *cl* and participant ρ using MDL
$\hat{p}_{i,\rho}^{AIC}$	Optimal mesh size estimated for sample at t_i and participant ρ using AIC
$\hat{p}_{i,\rho}^{BIC}$	Optimal mesh size estimated for sample at t_i and participant ρ using BIC

$\hat{p}^{FPE}_{i,\rho}$	Optimal mesh size estimated for sample at t_i and participant ρ using FPE
$\hat{p}^{MDL}_{i,\rho}$	Optimal mesh size estimated for sample at t_i and participant ρ using MDL
$\hat{p}^{AIC}_{j,\rho}$	Optimal mesh size estimated for voxel at $\bar{s}_j$ and participant ρ using AIC
$\hat{p}^{BIC}_{j,\rho}$	Optimal mesh size estimated for voxel at $\bar{s}_j$ and participant ρ using BIC
$\hat{p}^{FPE}_{j,\rho}$	Optimal mesh size estimated for voxel at $\bar{s}_j$ and participant ρ using FPE
$\hat{p}^{MDL}_{j,\rho}$	Optimal mesh size estimated for voxel at $\bar{s}_j$ and participant ρ using MDL
μ^{IC}_{ρ}	Mean of optimal mesh sizes for participant ρ using one of the criteria IC
ρ	Participant
σ^2_p	Mean squared error and also the maximum likelihood estimate of error variance
σ^{IC}_{ρ}	Standard deviation of optimal mesh sizes for participant ρ using one of the criteria IC
θ	A real valued parameter vector of linear regression equation
$\varepsilon_{i,j,\rho}$	Error obtained from linear regression equation at time t_i for the voxel at coordinates $\bar{s}_j$ belonging to participant ρ
$\varepsilon_{i,j}$	Error obtained from linear regression equation at time t_i for the voxel at coordinates $\bar{s}_j$
$\vartheta_j = \{v(t_i, \bar{s}_j)\}_{i=1}^{N}$	A time series of voxel at coordinates $\bar{s}_j$
$a_{i,j,k}$	Arc weight of local mesh model at time t_i, between the seed voxel at coordinates $\bar{s}_j$ and the neighboring voxel at $\bar{s}_k$
acc	Accuracy of the classifier
$AIC(p)$	Akaike Information Criterion estimated for mesh size p
$AIC_{\rho}(p)$	AIC estimated for mesh size p for participant ρ
$AIC_{cl,\rho}(p)$	AIC estimated for mesh size p for class cl and for participant ρ

$AIC_{i,\rho}(p)$	AIC estimated for mesh size p for sample at t_i and for participant ρ
$AIC_{j,\rho}(p)$	AIC estimated for mesh size p for voxel at $\bar{s}_j$ and for participant ρ
$BIC(p)$	Bayesian Information Criterion estimated for mesh size p
$BIC_{\rho}(p)$	BIC estimated for mesh size p for participant ρ
$BIC_{cl,\rho}(p)$	BIC estimated for mesh size p for class cl and for participant ρ
$BIC_{i,\rho}(p)$	BIC estimated for mesh size p for sample at t_i and for participant ρ
$BIC_{j,\rho}(p)$	BIC estimated for mesh size p for voxel at $\bar{s}_j$ and for participant ρ
$corr_{jk}$	Correlation coefficient between time series of voxels ϑ_j and ϑ_k
$cov_{jk}(\vartheta_j, \vartheta_k)$	Covariance of the signals ϑ_j and ϑ_k
D^{te}	Test dataset
D^{te}	Test feature matrix
D^{tr}	Training dataset
D^{tr}	Training feature matrix
$dist(ins_1, ins_2)$	Distance between two instances
$E^{cl}_{i,j}(.)$	Expectation on all time instants t_i belonging to class cl and all voxels at $\bar{s}_j$
$E_{i,j}(.)$	Expectation on all possible time instants t_i and all voxels at $\bar{s}_j$
$E_i(.)$	Expectation on all time instants t_i for the same voxel
$E_j(.)$	Expectation on all voxels at $\bar{s}_j$ for the same time instant
$f_r(ins)$	r^{th} attribute of instance ins
$FPE(p)$	Final Prediction Error estimated for mesh size p
$FPE_{\rho}(p)$	FPE estimated for mesh size p for participant ρ
$FPE_{cl,\rho}(p)$	FPE estimated for mesh size p for class cl and for participant ρ
$FPE_{i,\rho}(p)$	FPE estimated for mesh size p for sample at t_i and for participant ρ

$FPE_{j,\rho}(p)$	FPE estimated for mesh size p for voxel at $\bar{s}_j$ and for participant ρ
$G = [g(1), ..g(M)]$	A vector of time series data having size M
$g(\tau)$	The value of time series function at time instant τ
M	Number of voxels
$MDL(p)$	Rissanen's Minimum Description Length for ARMA processes estimated for mesh size p
$MDL_{\rho}(p)$	MDL estimated for mesh size p for participant ρ
$MDL_{cl,\rho}(p)$	MDL estimated for mesh size p for class cl and for participant ρ
$MDL_{i,\rho}(p)$	MDL estimated for mesh size p for sample at t_i and for participant ρ
$MDL_{j,\rho}(p)$	MDL estimated for mesh size p for voxel at $\bar{s}_j$ and for participant ρ
N	Number of samples
N^{te}	Number of test samples
N^{tr}	Number of training samples
p	Order of linear regression model (mesh size)
$RIS(p)$	Rissanen's Minimum Description Length estimated for mesh size p
t_i	Time instant
$v(t_i, \bar{s}_j)$	Intensity value of a voxel at time instant t_i and coordinates $\bar{s}_j$
$v_{\rho}(t_i, \bar{s}_j)$	Intensity value of a voxel belonging to participant ρ at time instant t_i and coordinates $\bar{s}_j$
$var_j(\vartheta_j)$	Variance of the signal ϑ_j
${\mu_{vox_\rho}}^{IC}$	Mean of optimal mesh sizes estimated for a voxel, for participant ρ using one of the criteria IC
${\sigma_{vox_\rho}}^{IC}$	Standard deviation of optimal mesh sizes estimated for a voxel, for participant ρ using one of the criteria IC

LIST OF ABBREVIATIONS

FPE	Final Prediction Error
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
MDL	Minimum Description Length
IR	Item Recognition
JOR	Judgment of Recency
IC	Information Criteria
MVPA	Multi-voxel Pattern Analysis
fMRI	functional Magnetic Resonance Imaging

CHAPTER 1

INTRODUCTION

1.1 Problem Definition

Mankind has always wondered how the human brain functions. In order to answer this unacknowledged question, people developed many theories. In 17th Century, Rene Descartes claimed that the effect of pineal gland on the surrounding ventricles is the main factor of activity in the brain. One century later, Emanuel Swedenbord assumed that the cortex itself was responsible for the cognition and cortex was divided into parts, each responsible for different cognition like thought, vision etc. Later in 19th century, the idea that different brain regions represent different aspects of human mind was popular. Moreover, during these days the amount of brain tissue, responsible for a cognitive function, is believed to determine the influence on behavior. Since the amount of tissue could not be measured directly, the bumps and the flattenings in the skull are believed to reflect the volume of different regions. Therefore, a child devoted to his pet was believed to have bumps on the skull over the area responsible for representing love. On the other hand, a liar would have flattening in the skull over the area responsible for honesty.

Later on, scientists gave up the idea of examining bumps in the skulls and started to examine the changes in brain physiology. However, the experiments were invasive and based on damaging the brain of animals. The measurements obtained from such experiments were useful to relate brain functions with brain regions. However, many aspects of cognition remained unacknowledged since such invasive experiments could not be conducted on humans. Years later, technology evolved so that with the functional Magnetic Resonance Imaging (fMRI), scientists began taking pictures of activations in the brain. Due to its noninvasive nature, fMRI is systematically used in the experiments on humans. Now, we live in the decade in which fMRI is the dominant technique used to understand how human brain functions.

The ultimate goal of all these studies throughout the centuries is to understand how brain functions, or with a different perspective "mind reading". People have always been curious about what others are thinking, feeling, dreaming etc. In many science-fiction themes, there were supernatural characters who has the ability to read other’s mind. A system in the science-fiction classic Brainstorm, was able to record and play back the experiences of people so that people can live the experiences of others. Although the current technology is years away from such innovations, using fMRI measurements in mind reading is popular and results of experiments are promising.

1.2 Proposed Cognitive Model

The proposed methods in this thesis aim to contribute to the innovations in mind reading using fMRI. In mind reading experiments, voxel (smallest unit of fMRI data) intensity values are recorded during a memory encoding and retrieval process using fMRI. During these experiments, samples presented to the user belongs to different classes. As a first step of mind reading, the aim is to classify these samples using the recorded activations of brain. Therefore, as in the classical machine learning techniques, first feature vectors are formed from the fMRI measurements using a training set and test set. Then, these features are used to train and test a k-NN classifier.

In this study, a local mesh model is used as a backbone to model the relationships among voxel intensities. In their study Ozay et. al, [47] showed that relationships among voxels are more discriminative than the voxel intensity values. Therefore, in this thesis rather than voxel intensity values, the arc weights of local mesh model, which represent the relationships among voxels, are used as features. In local mesh model, each voxel is represented as a linear combination of its nearest neighbors. Therefore, the arc weights of the local mesh are estimated using a linear regression equation by minimizing the squared error.

In previous studies, the number of neighbors to form a mesh around voxels was not determined and usually for mesh sizes in an interval, the classification results were listed. Yet, in this study, the squared error obtained from the linear regression equation is used in the selection of optimal number of neighbors. The optimal mesh size is estimated by minimizing the information theoretic criteria, which are Final Prediction Error (FPE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Rissanen's Minimum Description Length (MDL). Among them, FPE makes a trade-off between the complexity (determined by the mesh size) and degree of fit (determined by the error term) to estimate the optimal mesh size. On the other hand, AIC assumes that there is an unknown information distribution in the brain and estimates the optimal mesh size as the one that best approximates to this unknown distribution. Unlike these two, BIC uses a Bayesian approach and estimates the optimal mesh size as the one that leads to true model among other candidates. Finally, MDL considers representing the information in a compressed form and with this manner, it estimates the optimal mesh size as the one that best represents the information.

The major purpose of this study is to estimate an optimal mesh size which may vary depending on participant, sample or voxel. In this study, it is assumed that while one voxel is densely connected to others, another one may be connected to a few number of its neighbors. If the optimal mesh size is small, it is assumed that voxels make few connections with their neighbors. On the other hand a large mesh size implies that voxels are massively interconnected.

How the optimal mesh size varies based on the participant, class, sample and voxel is analysed in this study.

In this study, data obtained from a working memory experiment, in which the samples belong to either one of the item recognition (IR) task or judgment of recency (JOR) task are used. By forming a local mesh around each voxel and using the arc weights as features, samples belonging to two different cognitive tasks (IR or JOR) are aimed to be classified. Therefore, from training and test data, training and test feature vectors are formed respectively using the arc weights of the local mesh model. Using a k-NN classifier, the 2-class classification task is performed.

Briefly, a local mesh is formed around a seed voxel and the seed voxel is represented in terms of its nearest neighbors using a linear regression equation. The error term, obtained from this regression equation is used to calculate the information theoretic criteria separately for each participant, class, sample and voxel. Then the mesh size minimizing either one of these criteria is selected as the optimal mesh size so that corresponding seed voxel is represented as a linear combination of its nearest neighbors having optimal mesh size. The arc weighs of the mesh having optimal mesh size are used to form feature vectors. Finally, a classifier is trained using the feature vector formed using the training data and is tested with the feature vectors formed using the test data.

1.3 Contributions

- In the literature, information theoretic criteria are mainly used for model order selection of autoregressive functions. In other words, information criteria are used to answer how many previous values of a time series should be used to estimate the output itself. In this study, this information theoretic approach is adopted to a spatial data. Here, information theoretic criteria are used to estimate the number of spatially nearest neighbors. The results indicate that, the information theoretic criteria can also be used to estimate the model order of spatially distributed data.

- In the previous studies of local mesh model, which is the backbone of this study, the focus was not on the mesh size. In those studies [19, 47], for some mesh size in an interval, the classification process was performed and classification performances were listed. Moreover, around each voxel, local mesh of same size was formed and they did not provide an optimal mesh size. Unlike previous studies, the main focus is on the selection of optimal mesh size in this thesis and how the optimal mesh size differs for a participant, class, sample and voxel is analyzed.

- In the literature, generally, a parameter is selected using cross validation on training data. Here, we propose a method to classify cognitive states in which the parameter (the mesh size), is estimated independent of the training data.

- The major observation of this study is that, proposed methods give a better performance compared to the classical methods reported in the literature. Therefore, proposed methods in this study are promising to be used in the classification of various cognitive states.

- Finally, in this thesis a new hypothetical brain connectivity called "local relational connectivity" is defined. Although this thesis does not prove the existence of such connectivity network, the high classification performances achieved using the edges of this network as features to the classifier give us the intuition of the existence of the suggested connectivity model.

The work presented in this thesis has appeared in the following publications:

- I. Onal, M. Ozay, O. Firat, I. Oztekin, F. T. Yarman Vural, "An Information Theoretic Approach to Classify Cognitive States Using fMRI", 13th IEEE International Conference on BioInformatics and BioEngineering (BIBE), 2013
- I. Onal, M. Ozay, O. Firat, I. Oztekin, F. T. Yarman Vural, "Analyzing the Information Distribution in the fMRI measurements by estimating the degree of locality",35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), 2013
- I. Onal, M. Ozay, O. Firat, I. Oztekin, F. T. Yarman Vural, "Information Distribution Analysis in the fMRI measurements with Degree of Locality Estimation",IEEE 21th Conference on Signal Processing and Communications Applications (SIU), 2013

1.4 Outline of the Thesis

In Chapter 2, a brief literature survey on fMRI and brain connectivity are provided. Moreover, the popular MVPA methods aiming to classify cognitive states of brain are overviewed. Then, local mesh model, on which this thesis is built, is introduced and the information theoretic criteria to select model order are surveyed. Finally, a brief information about how k-NN classification algorithm works is presented.

Chapter 3 introduces the proposed methods to classify cognitive states. Here, a hypothetical local relational connectivity network based on the local mesh model is constructed and the importance of error term, obtained from the linear regression equation in the local mesh model, is explained. Mainly this chapter focuses on how to select the number of neighbors to form a mesh by combining the error term and information theoretic criteria. Moreover, how the feature vectors to be used in the classification are formed using the information theoretic approach are explained in detail.

Chapter 4 represents the analysis of the proposed methods in Chapter 3. In this chapter, results of the experiments and performance comparisons of proposed methods and the available methods in the literature are provided. Furthermore, how the optimal mesh size varies based on participant, class, sample and voxel is analyzed.

In the final chapter, Chapter 5, outcomes of overall study are discussed and the possible directions of this work are pointed out.

CHAPTER 2

AN OVERVIEW ON BRAIN CONNECTIVITY BASED ON FMRI DATA

In this chapter, for the purpose of providing background to the reader, one of the current neuroimaging technologies, fMRI, and various representations of brain connectivity are surveyed. For the reader to get an impression on how cognitive states are classified, current multi voxel pattern analysis methods (MVPA) are presented as related work. Moreover, as the backbone of this thesis, local mesh model is introduced and how it is used to classify cognitive states is explained. Finally, various information theoretic criteria, used as model order selection methods are overviewed.

2.1 Data Acquisition: Functional Magnetic Resonance Imaging (fMRI)

The ability to visualize how the human brain functions is one of the most remarkable developments in 20th century, because the brain images reflect a subtle information about the hidden structures in the brain. After the discovery of neuroimaging, functional neuroimaging methods have been used to detect the active brain regions while the subject performs a cognitive task. Usually, in these methods, a subject is exposed to cognitive task and during the task, the activated parts of the brain are revealed with the help of neuroimaging technology. Hence, functional neuroimaging methods serve as a powerful tool to understand the mappings between the brain regions and cognitive functions.

Functional Magnetic Resonance Imaging (fMRI) is one of the neuroimaging techniques in which MRI scanners are used to measure changes in brain activations. As its name implies, MRI scanner has three main items namely "magnetic", "resonance" and "imaging" [30].

- **Magnetic:** The first item "magnetic" means the static magnetic field created by the scanner with the aim of aligning nuclei of atoms in the human body. Since the body contains a lot of water, MRI machines make use of nuclei of hydrogen (^{1}H) atoms, called protons. As a first step, MRI scanner creates a static magnetic field with a powerful electro-magnet in it. The magnetic field, created inside the MRI scanner is usually about 3 - 4 teslas(T), which is 50.000 times greater than the field created by Earth and this field has the capacity to affect the magnetic nuclei of hydrogen. In the absence of such significant magnetic field, the protons point randomly in different directions. However, in the MRI scanner, they are aligned in the direction of this strong magnetic field and this is called an equilibrium state.

- **Resonance:** The second term is "resonance". After the alignment of nuclei with the magnetic field, electromagnetic waves that resonate at a particular frequency are emitted by the radiofrequency coils of MRI scanner to disturb the nuclei of atoms and perturb the equilibrium [30]. This process is named "resonance". In this phase, atoms are excited and absorb the energy emitted by the radiofrequency pulse. Then the radiofrequency pulse is turned off so that the hydrogen atoms can return to the equilibrium state and release the energy. As it can be seen, there is a continuous static magnetic field in the MRI scanner whereas the radiofrequency fields are created for a short time and then turned off. The released energy can be detected by the radiofrequency coils and defined as the MR signals. However, since this MR signal carries no spatial information, it can not be directly used for imaging.

- **Imaging:** The third term of MRI, "imaging" represents the phase where MR signals are turned into brain images. Current MRI scanners adopt the pioneering work of Lauterber et. al. [36], in which three orthogonal gradients are used to generate 2D and 3D MRI images. During the imaging phase, additional magnetic fields are created by the gradient coils so that nuclei of atoms at different locations wobble in different speeds. Using Fourier analysis, the spatial information can be recovered from the signal.

Functional Magnetic Resonance Imaging (fMRI) is an MRI procedure that measures the changes in brain function over time. Although fMRI scans and MRI scans both use the same principles of atomic physics, MRI visualizes anatomical structure of the brain (Fig. 2.1) while fMRI visualizes activity of the brain (Fig. 2.2). Therefore, images of MRI scans represent the anatomical structure of the brain whereas images of fMRI scans represent the activity within the anatomic structure of the brain [2]. However, fMRI does not directly image neuronal activity instead, it visualizes physiological changes correlated with neuronal activity that occur in the brain [30].

In 1936, Pauling and Coryell [49] discovered that the magnetic characteristics of hemoglobin depends on whether it is bound to oxygen or not. Oxygenated hemoglobin (Hb) is diamagnetic, meaning that it does not have any unpaired electrons, so, its magnetic moment is zero. On the other hand, deoxygenated hemoglobin (dHb) is paramagnetic, that is, it has unpaired electrons and dHb has strong magnetic moment. Arterials containing oxygenated blood cause little or no distortion to the magnetic field in the surrounding tissue, whereas capillaries and veins containing blood that is partially deoxygenated distort the magnetic field in their neighborhood [4, 43]. By distorting the surrounding magnetic field, dHb cause nuclei there to lose magnetization faster. Hence, higher level of MR signal intensity is produced where blood is oxygenated compared to the locations where the blood is deoxgyenated. In 1990, Ogawa et. al. [42, 44] discovered that by detecting the changes in blood oxygenation, the activated areas

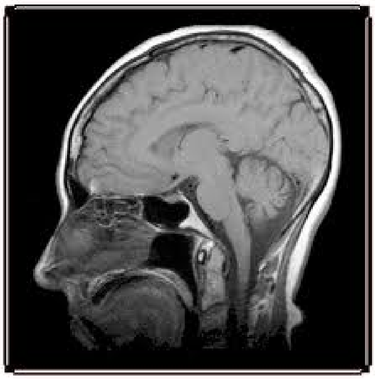


Figure 2.1: An MRI image of head [3].

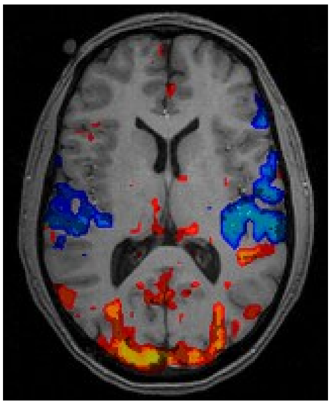


Figure 2.2: An fMRI image showing brain activations [1].

in the brain can be acquired using MRI procedure. Moreover, he stated that change in the level of oxygen in the blood, in other words, the change in the strength of MR signal caused by the paramagnetic property of deoxygenated hemoglobin, determines the blood-oxygenation-level-dependent (BOLD) contrast.

When a neuron becomes active, it needs energy to return to its original state. Since the source of energy, glucose, is not stored in the brain, it must be supplied with blood flow. Therefore, blood flows to the active area to transport glucose and oxygen in order for neurons to return their original state. While oxygen bounds to deoxygenated hemoglobin as a consequence of a neural activity, dHb is replaced with Hb and the ratio of oxygenated to deoxygenated blood increases [56]. Due to the decrease in dHB, MR signal increases in the active area. As a result, by measuring and imaging the oxgyenation level in blood, fMRI based on BOLD contrast gives information about the neural activity indirectly. The early studies employing BOLD contrast based fMRI started in 1992 [10, 34, 45] and BOLD contrast fMRI is still used as a powerful tool to measure brain activity.

The increase in MR signal as a response to the neural activation is called Hemodynamic Response (HR) and is parametrized by a Hemodynamic Response Function (HRF) [14] (Fig. 2.4). As it can be seen from Fig. 2.4, in BOLD HR, there occurs an *initial dip* for a short time following the onset of a neuronal activity. This dip may result from the initial oxygen extraction before the later overcompensatory response [30]. The maximum value in BOLD HR signal is named as *peak* and it is achieved about 4 to 6s after the stimulus. After neural activity stops, the BOLD HR signal falls below the original level, called the undershoot. Then, the signal recovers to the original level in time. In this thesis, only the peak values corresponding to given stimulus are taken into account.

During each fMRI scan, BOLD signal measurements are recorded to form 2D slices of brain. Then, all slices across the brain are combined to form 3D brain images. Since brain images are three dimensional, they are discreticized into *voxels*, which are volumetrix pixels [14]. Therefore, *voxel* is the smallest spatial unit of fMRI data which consists of about thousands of brain cells.

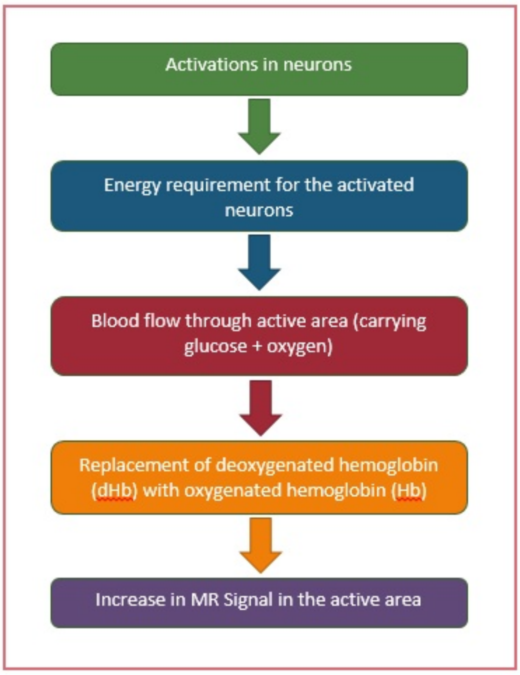

Activations in neurons
Energy requirement for the activated neurons
Blood flow through active area (carrying glucose + oxygen)
Replacement of deoxygenated hemoglobin (dHb) with oxygenated hemoglobin (Hb)
Increase in MR Signal in the active area

Figure 2.3: An overview of the physiological changes in the brain that lead to BOLD fMRI data.

.

2.2 Brain Connectivity

Brain connectivity refers to the patterns of connections between units (individual neurons at microscale, neuronal populations at mesoscale or brain regions at macroscale) of brain. It can be classified into three major categories based on the type of connections. If these connections correspond to anatomical links, the connectivity becomes *anatomical connectivity*. On the other hand, if they are patterns of statistical dependencies or causal interactions, then the connectivities are named as *functional connectivity* or *effective connectivity*, respectively [62].

2.2.1 Anatomical Connectivity

The connectivity of brain units (neurons or regions) that are physically or structurally linked to one another is called anatomical connectivity. Generally, in order to visualize and investigate the anatomical connectivity of brain *in vivo*, Diffusion Tensor Imaging (DTI) which is a non-invasive MRI method is used [35]. The pattern of anatomical connectivity is rather stable for short time scales like seconds or minutes. However, for longer time scales like hours to days, some alterations can be detected in the anatomical connectivity.

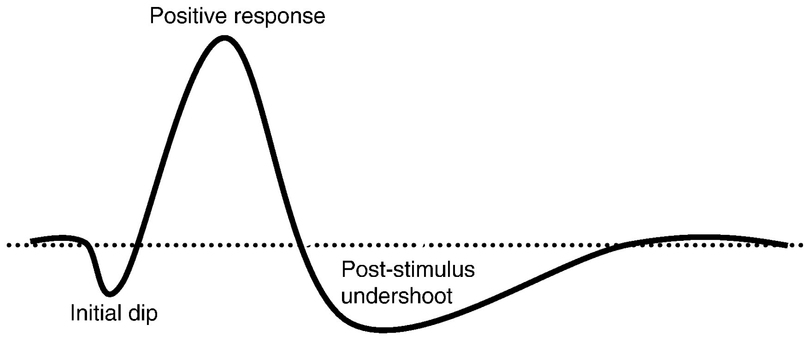


Figure 2.4: Hemodynamic Response Function (HRF), from [29].

.

2.2.2 Functional Connectivity

Neurons or in a rough scale, voxels do not function alone. Rather, their interactions with other such elements enable cognitive tasks to be performed in an orchestrated manner [28]. Recent studies focus on how different parts of the brain connect and coordinate with each other to perform a particular cognitive function rather than identifying activated brain regions under a cognitive task [37]. Friston et. al. [21] defines functional connectivity as the temporal correlations between spatially remote neurophysiological events. Even if there is no statistical link between two units of brain, the functional connectivity among them may be calculated [62].

Suppose that a time series of voxel at coordinates $\bar{s}_j$ is represented with $\vartheta_j = \{v(t_i, \bar{s}_j)\}_{i=1}^{N}$. If the functional connectivity is defined as temporal correlations between voxels, zero-order correlation coefficient between time series of voxels ϑ_j and ϑ_k is calculated by,

$$corr_{jk} = \frac{cov_{jk}(\vartheta_j, \vartheta_k)}{\sqrt{var_j(\vartheta_j) var_k(\vartheta_k)}}\ , \tag{2.1}$$

where $corr_{jk}$ represents the zero-order correlation coefficient between time series of voxels ϑ_j and ϑ_k, $cov_{jk}(\vartheta_j, \vartheta_k)$ is the covariance of the signals and var_j is the variance of the signal ϑ_j.

Functional connectivity is a statistical concept in which statistical dependence is estimated with model-based or data-driven methods. Model-based methods can be classified as cross-correlation analysis [13], coherence analysis [64] and statistical parametric mapping (SPM) [23], based on the connectivity metric used. On the other hand, data-driven methods are either decomposition based methods like principle component analysis and singular value decomposition (PCA/SVD) [21] or independent component analysis (ICA) [31] or clustering based methods like fuzzy [22] or hierarchical [15] clustering analysis [37].

2.2.3 Effective Connectivity

Effective connectivity characterizes the influence that a neural system may exert over another [21]. It measures the directional effect of whether an activation in a region may trigger an activation in another region. It can be measured by causality metrics like Granger causality or transfer entropy. Granger causality is a statistical concept which is based on prediction [58]. According to Granger causality, if a signal S_1 Granger-causes signal S_2, then S_1 values provide statistically significant information about future values of S_2. Another concept used to measure effective connectivity between a pair of signals is transfer entropy. It is an information theoretic measure of time directed information transfer between jointly dependent processes [11]. When the past values of a process S_1 given past values of another process S_2 is known, transfer entropy from S_1 to S_2 is the amount of uncertainty reduced in future values of S_2.

Notice that, if a process S_1 Granger-causes S_2, it does not mean that S_2 also Granger-causes S_1. Similarly, transfer entropy from S_1 to S_2 is not equal to transfer entropy from S_2 to S_1. Therefore, effective connectivity measures are not symmetric.

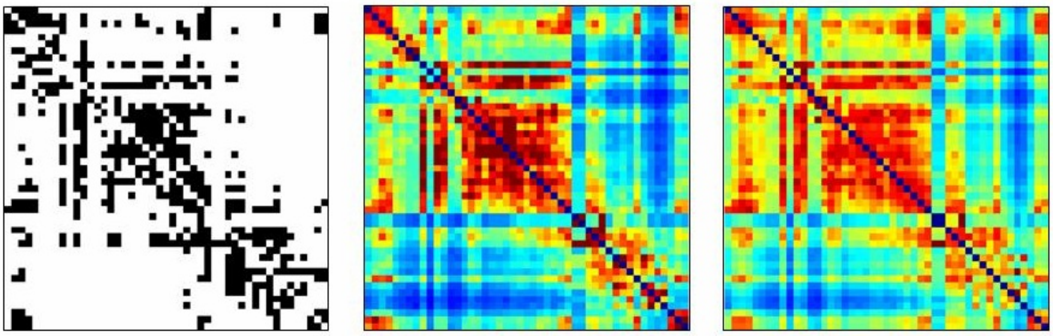


Figure 2.5: Anatomical connectivity showing binary structural connections(left), Functional connectivity representing pairwise correlations among voxels(middle) and Effective connectivity showing pairwise transfer entropy (right) [62].

.

Figure 2.5 represents anatomical, functional and effective connectivity matrices. Supposing there exist M voxels, the size of each connectivity matrix is MxM. Hence, rows and columns of each matrix correspond to voxels at coordinates $\bar{s_j}$. Anatomical connectivity matrix is a binary matrix in which each row represents whether there exists an anatomical link between the seed voxel (at row index) and all other voxels. If a link exists between two voxels, the corresponding cell is colored in black and otherwise, it is colored in white. On the other hand, functional connectivity matrix is a colored matrix where each row represents the correlations between the seed voxel (at row index) and all other voxels. In this matrix, red colors represent positive correlation whereas blue colors represent negative correlation. Moreover, the darker the color, the more the value of correlation. For example, a dark red cell means that voxels are highly positively correlated whereas a light red color indicates a low correlation for this pair of voxels. Notice that, functional connectivity matrix is a full symmetric matrix since correlation has a symmetric property. Similar to functional one, effective connectivity matrix is also colored and each row represents transfer entropy between the seed voxel (at row index) and all other voxels. Here, red colors represent positive transfer entropy whereas blue colors represent negative transfer entropy. Moreover, the darker the color, the more the voxel effects the corresponding voxel. Unlike functional one, effective connectivity matrix is not symmetric since an activation in a voxel may activate another one but the other way may not be true.

2.3 Multi-voxel Pattern Analysis (MVPA)

The functional magnetic resonance imaging (fMRI) technology has enhanced the ability of researchers to observe human brain activities in a non-invasive way. Once fMRI is proven to be a powerful brain imaging tool, at first many studies focused on identifying the activated brain regions under a cognitive task. [39]. In these studies, participants are exposed to the stimuli belonging to the same cognitive task. During the experiments, fMRI measurements are recorded for each trial. Then, the regions of brain that become active when the participant is exposed to a cognitive task can be identified by averaging the fMRI responses. This approach is called univariate (voxel-wise) approach, since the focus is on the individual voxels. Conventional fMRI methods having univariate approach aim to identify the voxels with significant response on average to the stimuli of the same cognitive task.

Unlike location-based, univariate approach, recent studies take into account the full spatial pattern of brain activity and use pattern classification algorithms to decode the subtle information represented in that pattern [41]. This approach is called multi-voxel pattern analysis (MVPA). Rather than focusing on "where" the information is encoded in the brain for the cognitive task, MVPA methods focus on "how" the information is encoded. Therefore, MVPA methods allow for the detection of non-local relationships between the cognitive task and brain activity[51]. The basic MVPA methods construct fMRI analysis as a pattern classification problem where the patterns are vectors of voxel intensity values.

Four main steps of MVPA methods are [41]:

- **Feature selection:** In this phase the voxels to be used in the classification are determined. Voxels with noise may reduce the performance of the classifier. Via feature selection methods, the voxels having noise are eliminated and the ones carrying information are kept.

- **Pattern assembly:** Data is discretisized into brain patterns and each pattern is associated with a label based on the experimental condition generated the pattern.

- **Classifier training:** A subset of samples belonging to different cognitive tasks, called training set, is used to train a classifier with the brain patterns. With these patterns and corresponding labels, a classification algorithm learns a function mapping between the brain pattern and the experimental condition.

- **Generalization:** Finally, the model is tested with samples, called test samples, whose labels are to be predicted via the model [69]. The accuracy of the classifier indicates the generalization performance of the classifier.

MVPA methods have some advantages over univariate methods. First, voxels having non-significant response to a cognitive task are discarded in univariate methods. However, these voxels might carry some information about the presence or absence of the cognitive task [41]. Therefore, an information loss is inevitable in the univariate approaches. Conversely, in MVPA methods the weak information in different locations can be collected in an efficient way. Secondly, most univariate approaches employ spatial smoothing to increase sensitivity. Hence, this smoothing causes spatial patterns that might carry discriminative information to be spoiled. Thirdly, although two different brain regions do not carry information about the cognitive task individually, their combination might be informative about the task. While, univariate approaches disregard this information, MVPA methods can detect it. Fourth disadvantage of univariate approaches is that, they average voxel intensity values over samples to detect regions taking part in the cognitive task [27]. Hence, sample size should be large for statistical significance, which may not be the case for fMRI experiments.

Several studies that use MVPA methods to decode cognitive state in different domains have been conducted in the last decades. Study of Haxby et. al. [25] was the pioneering work to show how multi-voxel patterns of brain activity in ventral temporal cortex can be used to discriminate between different cognitive tasks. In their study, subjects viewed objects from different categories which are faces, cats, non-sense objects and five categories of man-made objects. This study proved that, representations of objects and faces are overlapping and distributed in ventral temporal cortex. However, by using MVPA methods they were able to find a distinct pattern for each of these categories which are not resulting from the distinct responses in different regions. In addition to decoding object categories, Kamitani et. al. [32] proved the ability of MVPA to decode which of eight orientations the subject was viewing. Moreover, in their following study, Haynes et. al. [26] showed that, multi-voxel patterns of brain activity in visual cortex can be used to discriminate between unconscious representation of orientations. Davatzikos et. al. [18] used MVPA approach to discriminate patterns of brain activities measured by fMRI during truth-telling or lying experiment. Mitchell et. al. [40] proved that MVPA methods can be used to decode whether subject is presented a sentence or a picture, whether the subject is presented an ambiguous or non-ambiguous sentence and which of the twelve categories (fruits, tools, etc.) does the presented picture belong. In their work, Polyn et. al. [50] showed that, MVPA methods can be used in memory retrieval tasks. Subjects were presented pictures from three different categories namely faces, houses and objects in the encoding phase. Then, they were expected to recall them in the retrieval phase. In this work, MVPA is used to detect how similar the patterns of brain activity belonging to a category are in the encoding and retrieval phases.

In several studies, MVPA methods were used as a powerful diagnosis tool. Craddock et. al. [17] trained Support Vector Machine (SVM) classifiers with resting state functional connectivity patterns from healthy participants and patients having major depression. They proved that, activations of patients and healthy participants can be discriminated using MVPA methods. Moreover, Shen et. al. [59] proved that, using patterns of resting state functional connectivity in machine learning tools can discriminate schizophrenic patients from healthy participants. Furthermore, brain activations measured by fMRI of patients having Autism Spectrum Disorder (ASD) and that of healthy people were discriminated using MVPA methods [16].

In all of the aforementioned studies, brain activation patterns measured by fMRI are used to decode a cognitive state. Unlike univariate methods, the aim in these studies is not to detect the brain regions responsible for the cognitive task. Instead, in most of the studies MVPA methods are used to test hypotheses about whether brain activity patterns of cognitive task A are discernable from that of cognitive task B.

Human brain consists of massively coupled dynamic interactions at all scales [19]. However, fMRI has the capability to measure and image the individual voxel intensity values. Remember that, conventional MVPA methods employ voxel intensity values as features to a classification algorithm. Yet, it is not sufficient to fully understand these interactions by employing only the individual voxel intensity values measured via fMRI. Therefore, unlike in conventional MVPA methods, the relationship among voxels should be modeled as in [19, 20, 46, 47]

2.4 Local Mesh Model for Classifying Cognitive States

Generally, in order to classify a predefined set of cognitive states, the voxel intensity values measured by fMRI for each sample are concatenated to form a feature vector. Then, a well-know classifier is trained with feature vectors of all training samples formed that way. Finally, a feature vector belonging to a test sample is asked to classifier to learn which class the sample belongs to. In this procedure, each voxel is assumed to be a feature fed to the classifier and for a good classification accuracy, voxel intensity values are expected to carry a high discriminative power among different classes. In their study Özay et. al. [47] investigated whether the raw voxel intensity values present discriminative behaviour among different classes. Their observations indicated that, voxel intensity values are nearly constant for each time instant. Hence, they do not exhibit a significant change for samples belonging to distinct classes and can not be used to discriminate the classes. Rather they observed that, there are slight variations in the voxel intensity values of neighboring voxels for a time instant. Moreover, they found out that the distribution of voxel intensity values in space show slight variations in time. Therefore, the relationships between the voxel and its neighbors are modeled in the study [47] and the discriminative power of such relations are also investigated. The results indicated that, relationships among voxels carry more discriminative information than raw voxel intensity values measured by fMRI. The model, that presents the relationships among voxels is named as "local mesh model".

In local mesh model, voxels $v(t_i, \bar{s_j})$ at time instant t_i, where $i = 1, 2, ...N$ and location $\bar{s_j}$, where $j = 1, 2, ...M$ are used to model cognitive states. Here, N represents the number of samples and M represents the number of voxels. Since each voxel takes places in a three dimensional space, the voxel location $\bar{s_j}$ is a three dimensional vector, where $\bar{s_j} = (x_j, y_j, z_j)$. Around each voxel $v(t_i, \bar{s_j})$, called seed voxel, a local mesh is formed with the *p-nearest neighbors* $\{v(t_i, \bar{s_k})\}_{k=1}^{p}$ of seed voxel. In the study of Özay et. al. [47], *p-neighborhood* is defined spatially as the set of p number of voxels having smallest Euclidean distance to the seed voxel. In another approach, Fırat et. al. [19] defined the *p-neighborhood* functionally, in which *p-nearest neighbors* are selected based on the functional connectivity between the seed voxel and the surrounding voxels. Note that, in this approach *p-nearest neighbors* $\{v(t_i, \bar{s_k})\}_{k=1}^{p}$ of seed voxel are selected as the voxels whose zero-order correlations with the seed voxel are maximum among others.

While voxel intensity values $v(t_i, \bar{s_j})$ represent the vertices in the mesh, the relationship between these voxels in the local mesh are represented with the arc weights $a_{i,j,k}$ between vertices. (Fig. 2.6). Therefore, a local mesh consists of seed voxel and *p-nearest neighbors* of the seed voxel as vertices and arc weights $a_{i,j,k}$ as edges. These arc weights $a_{i,j,k}$ are estimated using the linear regression equation (Equation 2.2),

$$v(t_i, \bar{s_j}) = \sum_{\bar{s_k} \in \eta_p} a_{i,j,k} v(t_i, \bar{s_k}) + \varepsilon_{i,j} \ , \tag{2.2}$$

where $\varepsilon_{i,j}$ is the error obtained during the estimation of the arc weights $a_{i,j,k}$ of the local mesh at time instant t_i, where the seed voxel is $v(t_i, \bar{s_j})$ and the *p-nearest neighbors* are $\{v(t_i, \bar{s_k})\}_{k=1}^{p}$. In Equation 2.2, the arc weights are estimated by minimizing the squared error $\varepsilon_{i,j}^2$ using Levinson - Durbin recursion [65]. The arc weights $a_{i,j,k}$ of the local mesh, representing the relationships among the seed voxel $v(t_i, \bar{s_j})$ and its *p-nearest neighbors* $\{v(t_i, \bar{s_k})\}_{k=1}^{p}$ is used to form a mesh arc vector $\bar{a}_{i,j} = [a_{i,j,1}, a_{i,j,2}, ...a_{i,j,p}]$ of size $1xp$. Note that, each voxel is now represented as its relationships with its neighbors $a_{i,j,k}$ instead of its own fMRI intensity value $v(t_i, \bar{s_j})$. Then, this mesh arc vectors are combined for a sample at time t_i to form a mesh arc matrix for a sample $A_i = [\bar{a}_{i,1}, \bar{a}_{i,2}, ...\bar{a}_{i,M}]$ having size $1xp.M$. Finally, all the mesh arc vectors are concatenated to form a feature matrix $F = [A_1^T, A_2^T, ...A_M^T]^T$ of size $Nxp.M$ for a participant.

2.5 Information Theoretic Approaches for Model Order Selection

In many problems of signal processing, it is possible to model the vector of observations as a superposition of finite number of signals with an additive noise [67]. Therefore, in such problems, besides estimating the vector of parameters, it is often necessary to estimate the dimension of the parameter vector (the number of parameters). Examples in which the estimation of model order is necessary, includes the order of regression equation, the number of sinusoidal components in a sinusoid in noise signal or the number of source signals impinging on a sensor array. In all these examples, the number of unknown parameters should also be estimated. Suppose that:

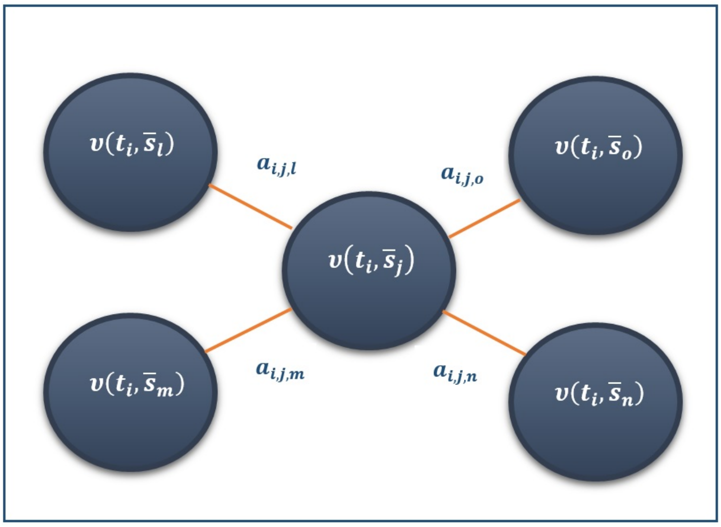


Figure 2.6: A local mesh representing the relationships among the seed voxel $v(t_i, \bar{s}_j)$ and its *p-nearest neighbors* $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$ with the arc weights $a_{i,j,k}$

.

$G = [g(1), ..g(M)]$ is a vector of time series data having size M ($G \in R^M$),
θ is the real valued parameter vector ($\theta \in R^p$),
p is the dimension of parameter vector θ.

In the above notation, a method estimating p from the time series data G besides estimating θ is called a *model order selection* method [63]. Some information theoretic criteria can be used to to estimate the optimal order of a regression equation defined as

$$g(\tau) = \sum_{k=1}^{p} \theta_k g(\tau - k) + \epsilon(\tau) , \tag{2.3}$$

where p is the model order, $\{\theta_1, \theta_2, ...\theta_p\}$ are the parameters of the model, $\epsilon(\tau)$ is the error calculated at time τ. In this model, the output $g(\tau)$ depends linearly on its own previous values $\{g(\tau - 1), g(\tau - 2), ...g(\tau - p)\}$ and in order to find the optimal value of p, or to find an answer to the question *"How many of the previous values of output should be used to model the output itself?"*, some information theoretic criteria may be used. Employing information theoretic criteria for model order selection is a powerful method, since one does not require any subjective judgment in the decision process. Instead, from the available data, the optimal model order is selected as the one that minimizes the criterion.

In the following sections, four well known informaton criteria namely Akaike's Final Prediction Error (FPE) [5, 6], Akaike Information Criterion (AIC) [7], Bayesian Information Criterion (BIC) [57] and Rissanen's Minimum Description Length (MDL) [52] will be explained in detail.

2.5.1 Final Prediction Error (FPE)

In the classical time series analysis, when an autoregressive model is fitted to the present series of $g(\tau)$ and this model is applied to another independent realization of $g(\tau)$, or applied to another process which is independent of $g(\tau)$ but has the same covariance characteristics as that of $g(\tau)$, a prediction error is obtained. In his pioneering work [5], Akaike defined *Final Prediction Error (FPE)* as the expected variance of this prediction error. This approach overcomes the difficulty of determining constants as in the solution of Anderson et. al. [9] to model order selection. Akaike states FPE as a figure of merit of a predictor, which is calculated for each model being fitted and the model with the best figure is chosen to be the predictor. Final prediction error estimated for an autoregressive function is defined as,

$$FPE(p) = \hat{\sigma}_p^2 \frac{(N + p + 1)}{(N - p - 1)}, \tag{2.4}$$

where N is the number of samples, p is the model order of autoregressive function and σ_p^2 is the mean squared error and also the maximum likelihood estimate of error variance which is approximated by,

$$\hat{\sigma}_p^2 \cong \frac{1}{N} \sum_{\tau=p}^{N-1} \epsilon(\tau)^2 \ , \tag{2.5}$$

where $\epsilon(\tau)$ is the error at time τ from Equation 2.3).

2.5.2 Akaike Information Criterion (AIC)

Suppose that data is generated by some unknown stochastic process $h(\tau)$ and the aim is to find a model that best fits the data, in other words find a model that best approximates to the unknown distribution $h(\tau)$. Let $\mathcal{F}(p) = \{f(\tau|\theta_p) \mid \theta_p \in \Theta_p\}$, where θ_p is a p-dimensional parameter vector, Θ_p is a class of p-dimensional parameter vectors (θ_p)'s and $f(\tau|\theta_p)$ is the likelihood function. By maximizing the likelihood $f(\tau|\theta_p)$ with respect to θ_p, the estimated parameter vector $(\hat{\theta}_p)$ that best fits the data can be found. Hence $f(\tau|\hat{\theta}_p)$ denotes the corresponding fitted model. However, in the model order selection problems, p is not fixed and the problem returns to select the model among the set $\hat{\mathcal{F}} = \{f(\tau|\hat{\theta}_1), f(\tau|\hat{\theta}_2), ...f(\tau|\hat{\theta}_N)\}$. Therefore, the model is selected that best approximates the unknown distribution $h(\tau)$ from the set $\hat{\mathcal{F}}$.

If $h(\tau)$ were known, the information loss acquired by representing $h(\tau)$ with each member of $\hat{\mathcal{F}}$ would be measured using the Kullback - Leibler divergence. Furthermore, the model whose Kullback - Leibler divergence from $h(\tau)$ is the smallest would be selected as the best model in the set and the corresponding p would be selected as the model order. Since $h(\tau)$ is unknown, it is not possible to measure Kullback - Leibler divergence of a model from $h(\tau)$. Akaike proposed an information criterion *Akaike Information Criterion (AIC)* [7] to determine the model order and showed that the information loss caused by selecting a model among $\hat{\mathcal{F}}$ to approximate $h(\tau)$ can be estimated [8].

Akaike Information Criteria (AIC) is defined to select the model order of an autoregressive function (Equation 2.3) as,

$$AIC(p) = N \ln(\hat{\sigma}_p^2) + 2p, \tag{2.6}$$

where N is the number of samples, $\hat{\sigma}_p^2$ is the mean squared error defined in Equation 2.5 and p is the order of the autoregressive function [60]. Low AIC indicates that the model is a better approximate of the unknown process. As the model order increase, the first term of Equation 2.6, the variance of error ($N \ln(\hat{\sigma}_p^2)$) decreases. Therefore, the selected model would fit the data better as the model order increases. However, this time the complexity of the model increases and the second term of the equation ($2p$) is the penalty term that penalizes the increase in complexity. Hence, AIC acts as a penalized log-likelihood criterion, trying to balance between good fit and complexity.

2.5.3 Bayesian Information Criterion (BIC)

In time series analysis, it is discovered that, as sample size increases AIC tends to select more complex models as the best model. This is mainly caused by the fact that the first term of Equation 2.6 ($N \ln(\hat{\sigma}_p^2)$) increases linearly with the sample size. However, the second term responsible for penalizing the complexity ($2p$), is only proportional to the model order p [55]. Hence, as the sample size increases, AIC tends to over-fit the data. In order to overcome this over-fitting problem, Schwarz proposed *Bayesian Information Criterion (BIC)* [57] where unlike AIC, the penalty term is also proportional to the sample size,

$$BIC(p) = N \ln(\hat{\sigma}_p^2) + p \ln(N), \tag{2.7}$$

where N is the number of samples, $\hat{\sigma}_p^2$ is the he maximum likelihood estimate of error variance defined in Equation 2.5 and p is the order of the autoregressive function. Although the equations of AIC and BIC are similar, the idea behind these two information theoretic approaches are totally different. As it is stated in section 2.5.2, AIC estimates a model that approximates to the unknown data generating process. Hence, it does not assume the selected model to be the "true" model. On the other hand, BIC aims to find the "true" model among many alternatives. BIC is a function estimate of the posterior probability of a model being true under a certain Bayesian setup. It answers how likely the data is generated by the model by estimating the posterior probability. Therefore, a lower BIC implies that a model is considered to be more likely to be the true model.

2.5.4 Rissanen's Minimum Description Length (MDL)

Inspired by the idea of AIC, Rissanen developed an information theoretic estimation principle called *Minimum Description Length (MDL)* [52, 53]. It aims to find the model that best represents the information in a compact form. MDL is a formalization of Occam's razor such that, it assumes the best model that represents the information as the one that leads to the best compression of data, yet retains the salient features in the present data. Grünwald [24] states that the idea behind MDL is learning as data compression. Rissanen defined MDL representing the information in [52] as,

$$RIS(p) = -\ln(\hat{\sigma}_p^2) + \frac{1}{2} p \ln(N), \tag{2.8}$$

where p represents the order of the autoregressive function, N is the number of samples and $\hat{\sigma}_p^2$ is the he maximum likelihood estimate of error variance defined in Equation 2.5. In his further study [54], Rissanen showed how to select optimum model for autoregressive processes by proving it on Gaussian autoregressive moving average (ARMA) processes. He claimed that in the class of Gaussian ARMA processes, the mean prediction error of any measurable predictor of the past data is bounded by,

$$MDL_p = \sigma_p^2 (1 + (\frac{p+1}{N}) \ln(N)). \tag{2.9}$$

Moreover, in [54] he showed that owing to the similarity of information and prediction bounds, the model where MDL_p is minimum is the optimum model. Note that, unless stated otherwise, in this thesis MDL represents the Minimum Description Length criterion used to estimate the optimal model order for autoregressive process. Hence, further derivations in this thesis will be based on MDL_p, not RIS_p.

In all of the abovementioned information theoretic criteria, the first term (error term) tends to decrease as the model order p increases providing a better fit. On the other hand, the second term is an increasing function of model order p indicating an increase in the complexity. Hence, all of these criteria make a trade-off between the degree of fit and the degree of complexity.

2.6 k - Nearest Neighbor for Classification

As defined in [38], *k-Nearest Neighbor (kNN)* algorithm is the most basic instance-based method. In this algorithm, all instances are assumed to correspond to a point in an *m-dimensional space*. The nearest neighbors of an instance is the ones having the smallest Eucledean distance to it among other instances. Let an arbitrary instance *ins* be described by the $m - dimensional$ feature vector:

$$< f_1(ins), f_2(ins), .., f_m(ins) > \quad , \tag{2.10}$$

where $f_r(ins)$ correspond to r^{th} attribute of instance ins. The distance between two instances ins_1 and ins_2 is defined as:

$$dist(ins_1, ins_2) = \sqrt{\sum_{r=1}^{m} (f_r(ins_1) - f_r(ins_2))^2} \ , \tag{2.11}$$

In the nearest neighbor algorithm, the target function may be discrete or real-valued. In discrete case, target function $f : R^m \rightarrow V$ maps an instance to a class label where $V = \{v_1, v_2, .., v_c\}$ and $c = 1, 2, .., C$. In the training phase, training examples are added to a list *training examples*. When a new instance ins_q is queried, the classification phase is computed as in Algorithm 1 and the corresponding class label is obtained.

```
Algorithm 1 k-Nearest Neighbor ($k - NN$) algorithm
Require: Given a query instance $ins_q$ to be classified
  Let $ins_1, ins_2, .., ins_k$ to be $k$ instances from training examples that are nearest to $ins_q$
  $\hat{f}(ins_q)) \leftarrow \underset{v \in V}{\arg\max}(\sum_{i=1}^{k} \delta(v, f(ins_i)))$
  where $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ otherwise.
Ensure: $\hat{f}(ins_q))$
```

Many studies [12, 61, 66] used k-NN algorithm to classify cognitive states using fMRI.

2.7 Summary

In this chapter, firstly the history of functional Magnetic Resonance Imaging (fMRI) and how it measures the activations in the brain are presented. Then, available voxelwise connectivity metrics of brain namely anatomical, functional and effective connectivity are explained. In the third subsection, MVPA methods in the literature, which are used in the classification of various types of cognitive states, are surveyed. After that, local mesh model, which states relationships among voxels are more discriminative than the voxel intensity values, is introduced. Then, the ideas behind four main information criteria namely Final Prediction Error, Akaike Information Criterion, Bayesian Information Criterion and Rissanen's Minimim Description Length are presented. Moreover, how these criteria are used to estimate the optimal order of linear regression equation is overviewed. Finally, k-Nearest Neighbor algorithm for the classification is explained.

CHAPTER 3

AN INFORMATION THEORETIC APPROACH FOR ESTIMATING BRAIN CONNECTIVITY USING FMRI MEASUREMENTS

3.1 Local Relational Brain Connectivity

In Section 2.2, the main connectivity metrics namely anatomical, functional and effective connectivity are introduced. When the units of these connectivities are voxels and the brain connectivity is established via the interconnections between voxels, anatomical, functional and effective brain networks can be constructed at voxel level. In all of these networks, the nodes of network are the voxel intensity values. However, the edges of the networks and their weights differ based on the connectivity type.

In order to represent connectivity, the number of connections among voxels should be determined. In other words, considering all the pairwise connections between voxels, it should be estimated that which pairwise connections are taken into account in the brain connectivity and which of them should be discarded so that corresponding pair of voxels are counted as disconnected. In the anatomical brain networks, the network is constructed using the anatomical links between voxels. Hence, if two voxels are connected with anatomical links, an edge is constructed in the anatomical brain network. Yet, for functional and effective connectivity, the edges among voxels are either statistical dependencies (correlation, coherence etc.) or causal interactions. Although the pairwise correlations or causal interactions are computed among all voxels, only the voxels whose correlations or causal interactions are above some threshold are considered to be connected. Hence, in the network only those corresponding edges are established. As it can be seen, the number of edges or the number of connections for a voxel depends on a threshold and may vary with this threshold.

A new type of voxel-wise connectivity, called "local relational connectivity" is proposed in this study as a hypothesis. Similar to aforementioned connectivity metrics, the nodes of local relational connectivity are the individual voxels whereas the edges among voxels are the arc weights of local mesh model proposed by Ozay et. al.[47]. Hence, the voxels are connected to each other with the coefficients of linear regression model 2.2 in local relational connectivity model where these coefficients represent the relationships among voxels in a local neighborhood. Unlike in functional or effective connectivity, pairwise connections to be taken into account are not determined using a threshold in local relational connectivity. Instead, size of each local mesh formed around each voxel is estimated using information theoretic criteria as model order selection methods.

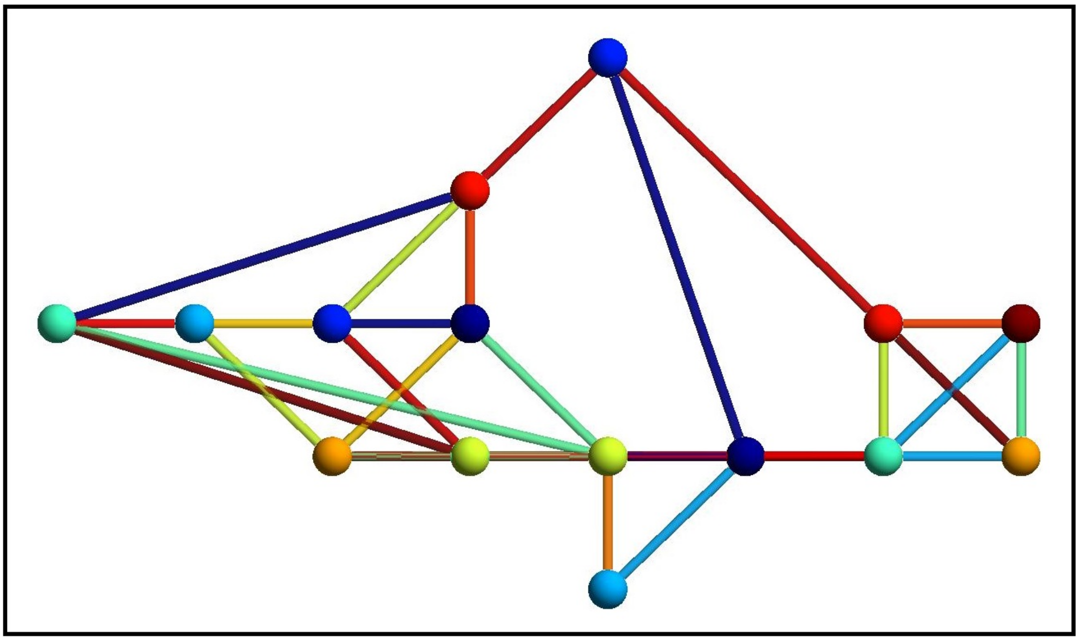


Figure 3.1: A sample local relational network where nodes correspond to voxel intensity values and edges correspond to arc weights of local mesh model.

Figure 3.1 represents a sample local connectivity network, where nodes of the network correspond to voxel intensity values and edges of the network correspond to arc weights of the local mesh model. In this connectivity network, the color of node states the intensity of voxel. In the colormap used, a dark red implies a high intensity value of voxel whereas a light blue implies that the intensity value of voxel is low. Hence, in this connectivity network, the activations in the voxels can be observed.

As it can be seen from Figure 3.1, the degree of node, in other words the number of edges connected to the node varies. The degree of each node is determined using the optimal mesh size so that some nodes have massive connections among its neighbors while others make only a few connections. Notice that, the color of edges also vary with the magnitude of arc weight of local mesh model. In other words, if the magnitude of an arc weight is high, it is represented with a dark red edge whereas a light blue edge indicates that the magnitude of the arc weight is small. However, the magnitude of arc weight does not indicate a higher connectivity or correlation as in the functional connectivity case. Rather, the arc weights are the coefficients of the linear regression equation.

In this hypothetical connectivity metric, based on the local relationships among voxels, the optimal mesh size of local mesh model (optimal number of neighbors in the local mesh) is assumed to determine the degree of connectivity, where small optimal mesh size indicates that the voxels are connected to a few number of voxels. On the other hand, a large optimal mesh size is an indicator of massively interconnected voxels. Therefore, in the proposed local relational connectivity metric, the optimal mesh size defines the connectivity.

In this study, optimal mesh size defining the local relational connectivity is estimated for:

- Each participant of the underlying experiment
- Each class representing a cognitive state
- Each sample measured during a cognitive state
- Each voxel of the fMRI measurement

The optimal mesh size is estimated by maximizing some information theoretic criteria in Section 3.3. By estimating the optimal mesh size for each participant, how local relational connectivity varies for each participant can be analyzed. Moreover, optimal mesh size estimated for each class reflects whether connectivity changes for different categories of samples or not even for the same participant. On the other hand, without considering the class it belongs, optimal mesh size can be estimated for each sample. Notice that, a sample represents the measurements recorded during a cognitive state from all voxels for a single time instant. Finally, in a more detailed manner it can be analyzed how the optimal mesh size varies for each voxel. In other words for each voxel of the sample, different neighborhoods are formed and each local mesh is formed using variable number of neighbors. This way, the voxels which tend to connect more to the others and the ones forming less connections can be examined.

3.2 The Role of Error Term

As it can be seen from Section 2.5, in all information theoretic criteria, maximum likelihood estimate of error variance, in other words, expected value of squared error is used to estimate the model order. In our method, the aim is to find the optimal mesh size. Therefore, in order to estimate the number of neighbors in a mesh, the expected value of squared error parameter in the local mesh model is used (Equation 2.2).

Let ρ represent the participant, $\rho = \{1, 2, .., P\}$, where P is the number of participants. Since further analysis and comparisons based on the results of different participants will be made in the following sections, besides time instant t_i and voxel coordinates $\bar{s_j}$, the data is represented in terms of the participant ρ it belongs to, for clarification. Hence, the Equation 2.2 can be reformulated as,

$$v_\rho(t_i, \bar{s_j}) = \sum_{\bar{s_k} \in \eta_p} a_{i,j,k} v_\rho(t_i, \bar{s_k}) + \varepsilon_{i,j,\rho} \ , \tag{3.1}$$

where $v_\rho(t_i, \bar{s}_j)$ represent the voxel included in the data belonging to participant ρ at time t_i, coordinates $\bar{s}_j$. Moreover, $\varepsilon_{i,j,\rho}$ is the error resulting from the linear regression equation, estimated for the data belonging to participant ρ.

In this study, fMRI data is represented as an NxM matrix, where N represents the number of samples and M represents the number of voxels. Around each voxel $v(t_i, \bar{s}_j)$ in this NxM matrix, a local mesh is created and the expected value of the error term $\varepsilon_{i,j,\rho}$ is minimized with respect to the arc weights $a_{i,j,k}$. By concatenating the minimum errors $\varepsilon_{i,j,\rho}$ for all voxel coordinates, error vector for a time sample at t_i, $err_{i,\rho} = [\varepsilon_{i,1,\rho}, \varepsilon_{i,2,\rho}, .., \varepsilon_{i,M,\rho}]$ is obtained. Finally, all error vectors for all samples are combined to form the error matrix for participant ρ, $Err_\rho = [err^T_{1,\rho}, err^T_{2,\rho}, ..err^T_{N,\rho}]^T$. Notice that, Err_ρ is an NxM matrix.

Although the error terms $\varepsilon_{i,j,\rho}$'s are the same for all types of local relational connectivity estimations (for each participant, class, sample or voxel), the way to approximate the expected value of this squared error varies for each type. For all types the squared error can be easily derived from Equation 3.1 as,

$$\varepsilon^2_{i,j,\rho} = (v_\rho(t_i, \bar{s}_j) - \sum_{\bar{s}_k \in \eta_p} a_{i,j,k} v_\rho(t_i, \bar{s}_k))^2 \ . \tag{3.2}$$

In the following subsections, how the expected value of squared error terms are estimated for each participant, each class, each sample and each voxel will be explained in detail.

3.2.1 Error for Each Participant

In this section a technique to estimate an optimal mesh size for a participant is suggested. Error matrix for each participant err_ρ is an NxM matrix, where N represents the number of samples and M represents the number of voxels (Figure 3.2). Since the purpose of this work is to estimate the error for each participant, it is assumed that an optimal mesh size exists for a participant. Therefore, around each voxel in each sample which belongs to the same participant, a local mesh of same size is formed.

The expected value of squared error ($\hat{E}_\rho$) is estimated for each participant ρ by taking the expectation over all time instants t_i and all voxels at coordinates $\bar{s}_j$ as:

$$\hat{E}_\rho = E_{i,j}(\varepsilon^2_{i,j,\rho}) \cong \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\varepsilon^2_{i,j,\rho} \ , \tag{3.3}$$

where $E_{i,j}(.)$ is the expectation on all possible time instants t_i and all voxels at coordinates $\bar{s}_j$, M is the number of voxels, N is the number of samples, $\varepsilon^2_{i,j,\rho}$ is the squared error at time instant t_i for voxel coordinates s_j and for participant ρ (see Equation 3.2). Moreover, $\hat{E}_\rho$ represents the resulting expected value of squared error for a participant. Therefore, by averaging the error over all time instants and all voxels, the expected value of squared error can be approximated for participant ρ (Equation 3.3). During the selection of optimal mesh size for a participant using information theoretic criteria, $\hat{E}_\rho$ will be used in the first terms of the equations 3.7 3.15 3.23 3.31.

In Figure 3.2, all rows corresponding to all samples are colored in purple. It does not mean that all samples belong to same class. Indeed, there may be multiple number of classes for experiments. The reason why all samples are colored in the same color (purple) is that when expected value of squared error is to be estimated for a participant, the error is averaged over all the samples without considering their class. Hence, the class label of a sample does not make any difference during the calculations of the expected error variance for a participant.

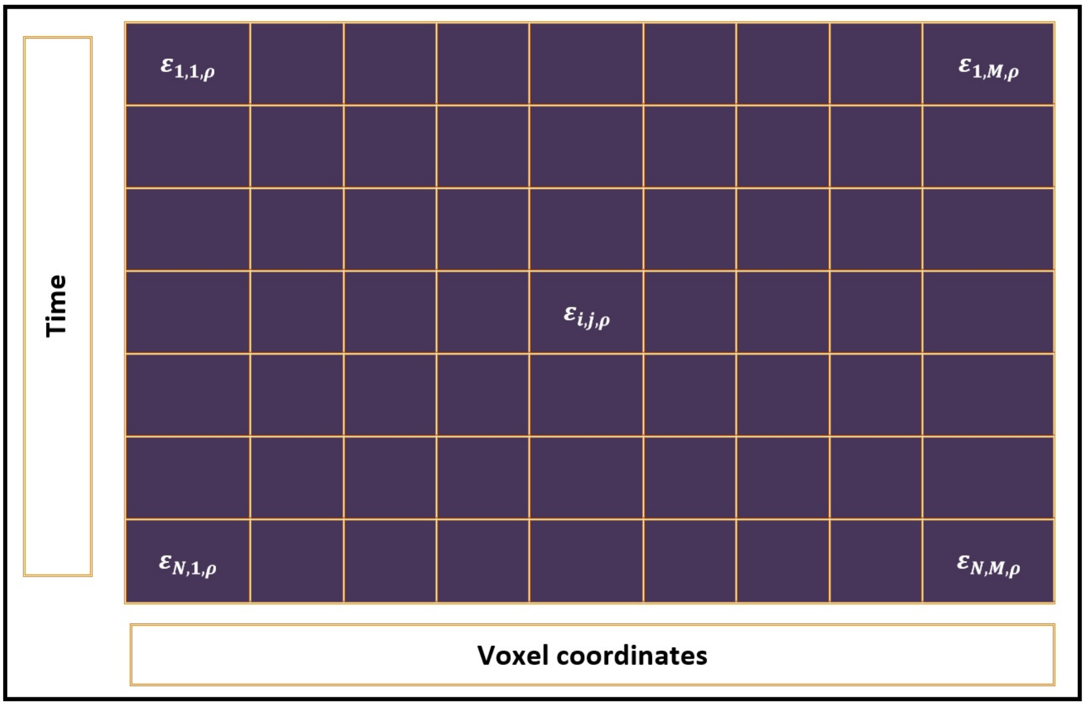


Figure 3.2: Organization of error matrix Err_ρ belonging to participant ρ, for all time instants t_i and all voxel coordinates $\bar{s}_j$.

3.2.2 Error for Each Class

In this section a technique to estimate an optimal mesh size for each class is suggested. Therefore, the fMRI data is recorded as a time series, where each time instant t_i is associated with a class label c_i where $c_i = 1, 2, ..C$. Assume that, as in the previous section, Section 3.2.1, error matrix for each participant Err_ρ is an NxM matrix, where N represents the number of samples and M represents the number of voxels. However, this time the class of each sample at t_i, (c_i) is important. For each class cl, where $cl \in \{1, 2, .., C\}$ the expected value of squared error is approximated using only the samples at t_i belonging to class cl, $\forall c_i = cl$.

The expected value of squared error is estimated for each class ($\hat{E}_{cl,\rho}$) by taking the expectation over all samples belonging the class cl and over all voxels at coordinates $\bar{s}_j$ as:

$$\hat{E}_{cl,\rho} = E^{cl}_{i,j}(\varepsilon^2_{i,j,\rho}) \cong \frac{1}{N_{cl}}\frac{1}{M}\sum_{\forall c_i=cl}\sum_{j=1}^{M}\varepsilon^2_{i,j,\rho}\ , \tag{3.4}$$

where $E^{cl}_{i,j}(.)$ is the expectation on all time instants belonging to class cl ($\forall t_i$, $c_i = cl$) and over all voxels at coordinates $\bar{s}_j$, M is the number of voxels, cl represents the class of the sample, N_{cl} is the number of samples where $c_i = cl$ and $\varepsilon^2_{i,j,\rho}$ is the squared error at time instant t_i for voxel coordinates s_j and for participant ρ (see Equation 3.2). Furthermore, $\hat{E}_{cl,\rho}$ represents the resulting expected value of squared error for a class, which will be used in further derivations to select optimal mesh size for a class. Note that, by averaging the error over all samples belonging to same class and over all voxels, the expected value of squared error can be approximated for a class. (Equation 3.4).

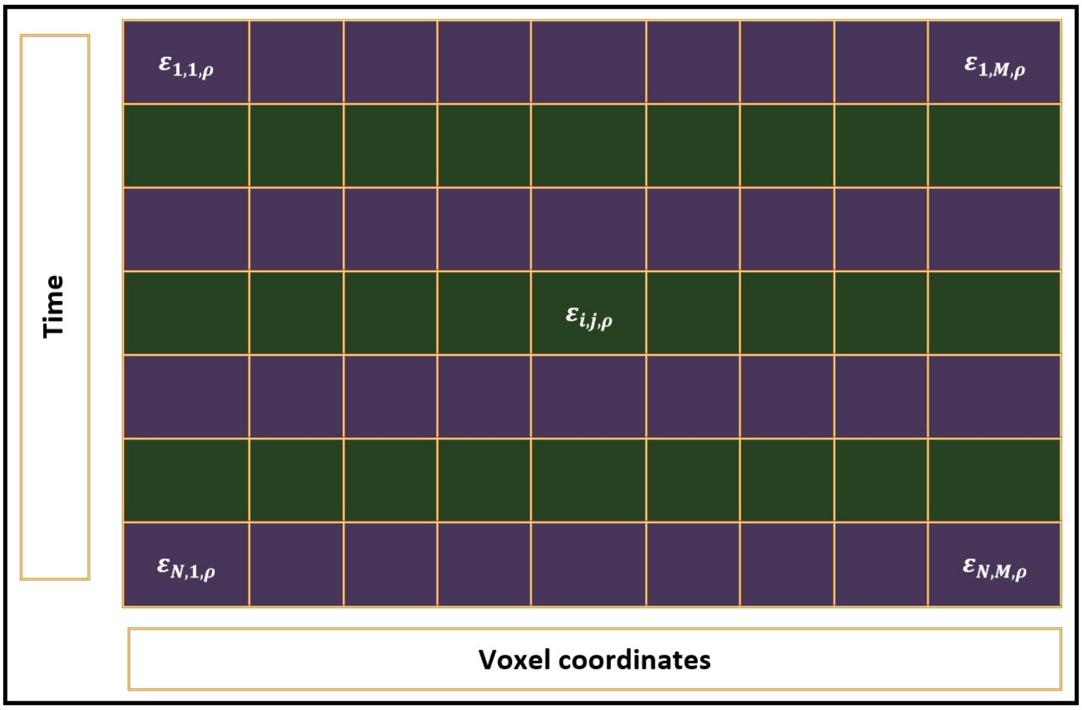


Figure 3.3: Organization of error matrix Err_ρ belonging to participant ρ, for all time instants t_i and all voxel coordinates $\bar{s}_j$ and color of each sample represents its class. Samples having green color belong to one class while samples having purple color belong to another class.

Figure 3.3 represents the errors obtained from fMRI samples belonging to two different classes. Assume that samples having red color belong to class *purple* and samples having green color belong to class *green*. In order to estimate the expected value of squared error for class *purple*, errors coming from voxels of samples coloured in purple are averaged. For class *green*, the expected value of squared error is estimated in the same way.

3.2.3 Error for Each Sample

Remember that, a sample represents the measurements recorded during a cognitive state from all voxels for a single time instant t_i. In previous two subsections, expected value of squared error is estimated for each participant and for each class. In both cases, since many samples are included for both a participant and a class, errors are averaged over many samples. In this section, expected value of squared error is estimated for each sample, where sample represents the cognitive state of a participant at a time instant. Therefore, this time the expected value of squared error is dependent on i for the sample at time instant t_i. Moreover, it is approximated by averaging the error over all voxels belonging to the sample.

The expected value of squared error for each sample at time instant t_i and for participant ρ ($\hat{E}_{i,\rho}$) is estimated by taking the expectation over all voxels at coordinates $\bar{s}_j$ at the same time instant t_i as:

$$\hat{E}_{i,\rho} = E_j(\varepsilon^2_{i,j,\rho}) \cong \frac{1}{M} \sum_{j=1}^{M} \varepsilon^2_{i,j,\rho} \ , \tag{3.5}$$

where $E_j(.)$ is the expectation over all voxels at coordinates $\bar{s}_j$, M is the number of voxels and $\varepsilon^2_{i,j,\rho}$ is the squared error at time instant t_i for voxel coordinates s_j and for participant ρ (see Equation 3.2). While selecting the optimal mesh size for a sample, $\hat{E}_{i,\rho}$ term will be used in derivations. Notice that, for a sample at time instant t_i, the expected value of squared error is approximated by averaging error term over all voxels belonging to the sample. In Section 3.3, while selecting the optimal mesh size for each sample, $E(\varepsilon^2_{i,\rho})$ will be used in the information criteria estimations as the first term of equations 3.11, 3.19, 3.27, 3.35.

From Figure 3.4, it can be observed that, error vector for a sample at time instant t_i corresponds to a row in the NxM error matrix for a participant (Figure 3.2). While estimating the expected value of squared error for a sample, the class of the sample is not taken into account.

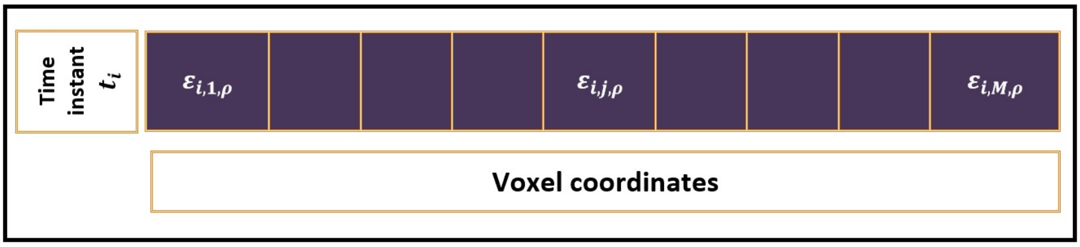


Figure 3.4: i^{th} row of error matrix Err_ρ belonging to participant ρ, for a single time instant t_i for all voxel coordinates $\bar{s}_j$. The class label of the sample is not considered during the estimation.

3.2.4 Error for Each Voxel

In this section the expected value of squared error is approximated for each voxel by averaging the error over all time instants of a single voxel. Note that, when estimated for a voxel, the expected value of error depends on both j, where $\bar{s}_j$ represents the voxel coordinates, and the participant ρ.

The expected value of squared error for each voxel at coordinates $\bar{s}_j$ and for participant ρ ($\hat{E}_{j,\rho}$) is estimated by averaging the error over all time instants ($\forall t_i$) belonging to a single voxel as:

$$\hat{E}_{j,\rho} = E_i(\varepsilon^2_{i,j,\rho}) \cong \frac{1}{N}\sum_{i=1}^{N}\varepsilon^2_{i,j,\rho} \, , \tag{3.6}$$

where $E_i(.)$ is the expectation operator taking the average over all time instants t_i for a single voxel, N is the number of samples and $\varepsilon^2_{i,j,\rho}$ is the squared error at time instant t_i for voxel coordinates s_j and for participant ρ (see Equation 3.2). When the optimal mesh size is selected for each voxel using information theoretic criteria, $\hat{E}_{j,\rho}$ will be used as the first term of the model order selection equations 3.13, 3.21, 3.29, 3.37.

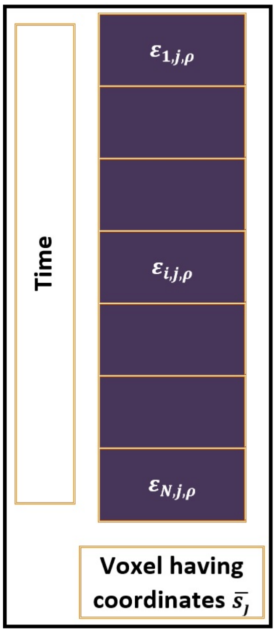


Figure 3.5: j^{th} column of error matrix Err_ρ for a voxel at coordinates $\bar{s}_j$, for all time instants t_i. The class belonging to the sample including the voxel is not considered.

Figure 3.5 represents all time instants of a voxel at coordinates $\bar{s}_j$. Therefore, time series of a voxel at coordinates $\bar{s}_j$ corresponds to a column in the NxM matrix for a participant (Figure 3.2). Although the whole column is colored in red, each time instant of voxel at coordinates $\bar{s}_j$ is a member of a sample belonging to various classes. However, during the estimation of the expected value of squared error for a voxel, the class of the sample including the voxel is not taken into account and error is averaged over all time instants for a voxel belonging to the same participant, not over a subset of time instants belonging to same class.

In this section how the expected value of squared error is approximated for a participant, a class, a sample and a voxel is explained. Although all of them uses the error term in a linear regression equation used in local mesh model, different averaging over errors will lead to different expected values of squared error. Later on, these expected values will be used in various information theoretic criteria in the following sections, in order to select the optimal mesh size for a participant, a class, a sample and a voxel, respectively.

3.3 An Information Theoretic Approach for Modeling Brain Connectivity

Among many others, information theoretic criteria, which are used to estimate the order of the linear regression models are introduced in Section 2.5. Recall that, output of the autoregressive model depends linearly on its own previous values and information criteria are used to estimate how many of the previous values of output should be considered to estimate the output value itself.

In this study, the output is the voxel intensity value $v(t_i, \bar{s_j})$. However, this intensity value $v(t_i, \bar{s_j})$ does not linearly depend on its own previous values, $(\{v(t_{i-1}, \bar{s_j}), v(t_{i-2}, \bar{s_j}), ...\})$, instead, the voxel $v(t_i, \bar{s_j})$ is represented as a linear combination of its nearest neighbors at the same time instant t_i $(\{v(t_i, \bar{s_l}), v(t_i, \bar{s_o}), ...\})$. Unless stated otherwise, the nearest neighbors of a voxel $v(t_i, \bar{s_j})$ are the ones having the smallest Euclidean distance to the voxel $v(t_i, \bar{s_j})$ among others. Here, while estimating the coefficients of linear regression model (local mesh model), the dimension of coefficients, i.e. the optimal number of neighbors, should also be estimated. This brings the idea of using information theoretic criteria in a spatial manner to estimate the model order. Note that, in this model, model order corresponds to the optimal mesh size and the purpose of using information theoretic criteria is to select the optimal mesh size, in other words, to select the optimal number of neighboring voxels that best represents the seed voxel of the mesh.

At this point, we assume that there is a trade-off between the mesh size p, which represents the model complexity, and expected value of squared error, which represents the degree of fit among the voxels in a neighborhood. As the mesh size p increases, the expected value of squared error decreases, implying that model fits better. On the other hand, with an increase in mesh size p, the complexity of the local mesh model increases, since each voxel is then represented as a linear combination of more number of its neighbors. Therefore, the degree of local relational connectivity can be represented by optimizing this trade-off between the mesh size p and error term. During this optimization, four well-known information theoretic criteria, namely Final Prediction Error (FPE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Rissanen's Minimum Description Length (MDL) are adopted in a spatial manner to select the optimal mesh size for each participant, class, sample and voxel.

In the following subsections, how these information theoretic criteria are adopted to estimate the degree of connectivity in human brain by estimating the optimal mesh size will be explained in detail.

3.3.1 Final Prediction Error (FPE)

In his pioneering work, Akaike proposed an information theoretic criterion called Final Prediction Error (FPE) to select the model order of a linear regression function, in which the output value of a time series depends on its own previous values. Therefore, by using FPE not only the coefficients of linear regression equation, but also the number of coefficients can be estimated. In local mesh model relationships among the voxels are also estimated using a linear regression function and this time the question is *"How many neighbors should be used to form the mesh around a voxel?"*. Hence, in our work, model order of linear regression equation should also be estimated. Note that, the linear regression functions used in Akaike's work and local mesh model are rather different. In local mesh model a voxel is represented as a linear combination of its spatial neighbors, not as a linear combination of its own previous values. In other words, the linear regression function of local mesh model is not time based as in its original use, rather it is spatial. As a result in this study FPE is adopted to a regression model to select the number of neighbors.

In the following subsections, optimal mesh size is estimated for each participant, class, sample and voxel using FPE.

3.3.1.1 FPE for Each Participant

In order to select the optimal mesh size for a participant, FPE is used ($FPE_\rho(p)$) in the following way:

$$FPE_\rho(p) = \hat{E}_\rho \left(\frac{M + p + 1}{M - p - 1} \right) . \tag{3.7}$$

In this equation, $\hat{E}_\rho$ represents the expected value of error derived in Equation 3.3, p is the mesh size, M is the number of voxels and $FPE_\rho(p)$ is a function of mesh size p. Note that, $FPE_\rho(p)$ has only one minimum with respect to p, since $\hat{E}_\rho$ is a monotonically decreasing function of p in the ideal case whereas $\left(\frac{M + p + 1}{M - p - 1} \right)$ is a monotonically increasing one. Therefore, using Equation 3.7, FPE is computed for the mesh sizes p where $p \in [p_{min}, p_{max}]$. Then, the mesh size p minimizing FPE is selected as the optimal mesh size ($\hat{p}_\rho^{MDL}$) for participant ρ using the equation below:

$$\hat{p}_\rho^{FPE} = \underset{p}{\arg\min}(FPE_\rho(p), p \in [p_{min}, p_{max}]) . \tag{3.8}$$

The interval $[p_{min}, p_{max}]$ is selected large enough such that the mesh size p minimizing $FPE_\rho(p)$ lies in this interval. In other words, $FPE_\rho(p)$ should decrease with an increase in p up to some value of p and then it should start to increase with p. In order to observe such behavior, the interval $[p_{min}, p_{max}]$ is taken large enough so that optimal mesh size $\hat{p}_\rho^{FPE}$ lies in this interval. In all of the following sections, the interval $[p_{min}, p_{max}]$ is selected with this manner. Moreover, in all of the experiments based on either participant, class, sample or voxel using either one of the information criteria, the intervals $[p_{min}, p_{max}]$ are the selected as the same.

3.3.1.2 FPE for Each Class

FPE is used to analyze whether optimal mesh size differs for each class, where samples of a class are acquired from the same participant, using the following equation:

$$FPE_{cl,\rho}(p) = \hat{E}_{cl,\rho}\left(\frac{M+p+1}{M-p-1}\right) , \tag{3.9}$$

where p is the mesh size, M is the number of voxels and ρ represents the participant. Note that, the optimal mesh size estimated using FPE for each class ($\hat{p}_{cl,\rho}^{FPE}$) depends on both the class cl and the participant ρ. In other words, during the estimation of optimal mesh size for a class, only the samples belonging to same participant are considered.

$FPE_{cl,\rho}(p)$ is estimated for the integer values of p in the interval $[p_{min}, p_{max}]$ and the mesh size minimizing the $FPE_{cl,\rho}(p)$ is selected as the optimal mesh size ($\hat{p}_{cl,\rho}^{FPE}$) for class cl.

$$\hat{p}_{cl,\rho}^{FPE} = \arg\min_p(FPE_{cl,\rho}(p), p \in [p_{min}, p_{max}]) . \tag{3.10}$$

3.3.1.3 FPE for Each Sample

FPE is also adopted to estimate the optimal mesh size for a sample at time instant t_i belonging to a participant ρ as:

$$FPE_{i,\rho}(p) = \hat{E}_{i,\rho}\left(\frac{M+p+1}{M-p-1}\right) , \tag{3.11}$$

where p is the mesh size, M is the number of voxels and ρ represents the participant and $\hat{E}_{i,\rho}$ is the expected value of squared term for a sample in Equation 3.5. After calculating the $FPE_{i,\rho}(p)$ for all the mesh sizes in the interval $[p_{min}, p_{max}]$, optimal mesh size for the sample at t_i is estimated as the one minimizing $FPE_{i,\rho}(p)$, by using:

$$\hat{p}_{i,\rho}^{FPE} = \underset{p}{\arg\min}(FPE_{i,\rho}(p), p \in [p_{min}, p_{max}]) \ , \tag{3.12}$$

where $\hat{p}_{i,\rho}^{FPE}$ is the optimal mesh size estimated using FPE for sample at t_i of participant ρ.

3.3.1.4 FPE for Each Voxel

Expected value of squared error term for each voxel (see Equation 3.6) is used in the adopted FPE criterion ($FPE_{j,\rho}(p)$) to select the optimal mesh size for each voxel at coordinates $\bar{s_j}$ belonging to participant ρ as:

$$FPE_{j,\rho}(p) = \hat{E}_{j,\rho}\left(\frac{N+p+1}{N-p-1}\right) \ , \tag{3.13}$$

where N is the number of samples. After $FPE_{j,\rho}(p)$ is estimated for many different mesh sizes p in the interval $[p_{min}, p_{max}]$, the optimal mesh size for voxel at coordinates $\bar{s_j}$ is selected as the one that minimizes $FPE_{j,\rho}(p)$ by using:

$$\hat{p}_{j,\rho}^{FPE} = \underset{p}{\arg\min}(FPE_{j,\rho}(p), p \in [p_{min}, p_{max}]) \ , \tag{3.14}$$

where $\hat{p}_{j,\rho}^{FPE}$ refers to the optimal mesh size for voxel at $\bar{s_j}$, where the experiment is conducted on participant ρ.

Variants of FPE for a participant, class, sample and voxel indicate that, FPE is always a function of mesh size p. In order to select the optimal mesh size for any one of them, FPE is estimated for various mesh sizes and the p minimizing the FPE is chosen to be the optimal mesh size. Moreover, the interval $[p_{min}, p_{max}]$ should be selected in a way that, the optimal mesh size minimizing the information criterion equation lies in this interval.

3.3.2 Akaike Information Criterion (AIC)

If the data generating process, in our case the generation of fMRI data during a cognitive process were known, information loss of the local mesh model of size p would be found using the Kullback–Leibler (KL) divergence between the model and the information distribution with certainty. Hence, the optimal mesh size would be selected as the one having the smallest KL divergence with the underlying cognitive process. However, the information distribution in the human brain is yet unknown and we approximate this unknown by using AIC for a local mesh formed around each voxel. Therefore, we assume that the mesh size p which leads to the minimum AIC is the one that best approximates the unknown information distribution in the brain and this p is selected as the optimal mesh size.

In the following subsections, optimal mesh size is estimated for each participant, class, sample and voxel using AIC.

3.3.2.1 AIC for Each Participant

AIC criterion is adopted to select the optimal mesh size for each participant ($AIC_\rho(p)$) using the expected value of squared error term approximated in Equation 3.3:

$$AIC_\rho(p) = \ln(\hat{E}_\rho) + \frac{2p}{M} \ . \tag{3.15}$$

where p represents the mesh size, M represents the number of voxels, ρ is the participant, $\hat{E}_\rho$ is the expected value of squared error approximated for a participant and $AIC_\rho(p)$ is a function of mesh size p.

For the values of p in the interval $[p_{min}, p_{max}]$, $AIC_\rho(p)$ is estimated for participant ρ and the one minimizing the $AIC_\rho(p)$ is selected as the optimal mesh size estimated using AIC for participant ρ as follows:

$$\hat{p}_\rho^{AIC} = \underset{p}{\arg\min}(AIC_\rho(p), p \in [p_{min}, p_{max}]) \ . \tag{3.16}$$

3.3.2.2 AIC for Each Class

In order to select the optimal mesh size for a class, where samples belong to the same participant, AIC is used ($AIC_{cl,\rho}(p)$) in the following way:

$$AIC_{cl,\rho}(p) = \ln(\hat{E}_{cl,\rho}) + \frac{2p}{M} \ , \tag{3.17}$$

where p represents the mesh size, M represents the number of voxels and ρ is the participant. In this section, the optimal mesh size estimated with AIC ($\hat{p}_{cl,\rho}^{AIC}$) depends on both the class cl and the participant ρ. Using the mesh sizes p in an interval, $AIC_{cl,\rho}(p)$'s are estimated for a class and the mesh size minimizing this term is selected as the optimal as in Equation 3.18.

$$\hat{p}_{cl,\rho}^{AIC} = \underset{p}{\arg\min}(AIC_{cl,\rho}(p), p \in [p_{min}, p_{max}]) \ . \tag{3.18}$$

3.3.2.3 AIC for Each Sample

By using the expected value of squared term for a sample in Equation 3.5, AIC is adopted to estimate the optimal mesh size for a sample at time instant t_i belonging to a participant ρ as:

$$AIC_{i,\rho}(p) = \ln(\hat{E}_{i,\rho}) + \frac{2p}{M} \,. \tag{3.19}$$

After calculating the $AIC_{i,\rho}(p)$ for many different mesh sizes and selecting the one that minimizes $AIC_{i,\rho}(p)$, optimal mesh size for the sample at t_i is estimated using:

$$\hat{p}_{i,\rho}^{AIC} = \arg\min_{p}(AIC_{i,\rho}(p), p \in [p_{min}, p_{max}]) \,, \tag{3.20}$$

where $\hat{p}_{i,\rho}^{AIC}$ is the optimal mesh size estimated using AIC for sample at t_i of participant ρ.

3.3.2.4 AIC for Each Voxel

Optimal mesh size might differ for each voxel even in the same sample of the same participant. In order to estimate this optimal mesh size for a voxel at coordinates $\bar{s}_j$, AIC is adopted as:

$$AIC_{j,\rho}(p) = \ln(\hat{E}_{j,\rho}) + \frac{2p}{N} \,, \tag{3.21}$$

where N is the number of samples. After calculating the $AIC_{j,\rho}(p)$ for many different mesh sizes, the optimal mesh size for voxel at coordinates $\bar{s}_j$ is selected as the one that minimizes $AIC_{j,\rho}(p)$,

$$\hat{p}_{j,\rho}^{AIC} = \arg\min_{p}(AIC_{j,\rho}(p), p \in [p_{min}, p_{max}]) \,, \tag{3.22}$$

where $\hat{p}_{j,\rho}^{AIC}$ refers to the optimal mesh size for voxel at $\bar{s}_j$.

As it can be seen from the above derivations of AIC for a participant, class, sample and voxel, AIC is always a function of mesh size p. So as to select the optimal mesh size, AIC is estimated for various mesh sizes and the p that makes the AIC minimum is chosen to be the optimal mesh size. Moreover, the interval $[p_{min}, p_{max}]$ should be selected in a way that, the optimal mesh size minimizing the information criterion equation lies in this interval.

3.3.3 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) attempts to estimate a true model among the candidates. In our case, BIC is used to find the local mesh model of optimal mesh size among all the candidate local mesh models of size p. BIC answers to the question *"How likely the data is generated by the local mesh model of size p?"* by estimating the posterior probability and selecting the p which gives the highest posterior probability among all as the optimal mesh size. Therefore, the BIC values, which will be adopted to select the optimal mesh size for a participant, class, sample and a voxel, are proportional to the likelihood of local mesh model of size p generating the data. Unlike AIC, BIC uses prior probability, hence the prior selection affects the accuracy. In this study, error terms are assumed to have a normal distribution and the corresponding BIC model [57] is adopted.

In the following subsections, optimal mesh size is estimated for each participant, class, sample and voxel using BIC.

3.3.3.1 BIC for Each Participant

BIC is adopted to select the optimal mesh size of the local mesh model for each participant ρ ($BIC_\rho(p)$) using the expected value of squared error term approximated in Equation 3.3 as:

$$BIC_\rho(p) = \ln(\hat{E}_\rho) + \frac{p\ln(M)}{M}\ , \tag{3.23}$$

where p is the mesh size, M is the number of voxels, ρ represents the participant, $\hat{E}_\rho$ is the expected value of squared error approximated for a participant and $BIC_\rho(p)$ is a function of mesh size p.

$BIC_\rho(p)$'s are estimated for integer values of mesh size p, where $p \in [p_{min}, p_{max}]$ and the one minimizing $BIC_\rho(p)$ is selected as the optimal mesh size by using the following equation:

$$\hat{p}_\rho^{BIC} = \underset{p}{\arg\min}(BIC_\rho(p), p \in [p_{min}, p_{max}])\ , \tag{3.24}$$

where $\hat{p}_\rho^{BIC}$ represents the optimal mesh size estimated using BIC for participant ρ.

3.3.3.2 BIC for Each Class

BIC is used to analyze whether optimal mesh size differs for each class, where samples of a class belong to the same participant using the following equation:

$$BIC_{cl,\rho}(p) = \ln(\hat{E}_{cl,\rho}) + \frac{p\ln(M)}{M}\ , \tag{3.25}$$

where p is the mesh size, M is the number of voxels and ρ represents the participant. Notice that, the optimal mesh size estimated with BIC for each class ($\hat{p}_{cl,\rho}^{BIC}$) depends on both the class cl and the participant ρ. In other words, samples having the same class but belonging to different participants are not mixed to select the optimal mesh size for a class.

For various values of p in the interval $[p_{min}, p_{max}]$, $BIC_{cl,\rho}(p)$ is estimated for a class cl and the mesh size minimizing the $BIC_{cl,\rho}(p)$ is selected as the optimal mesh size for class cl.

$$\hat{p}^{BIC}_{cl,\rho} = \underset{p}{\arg\min}(BIC_{cl,\rho}(p), p \in [p_{min}, p_{max}]) \, . \tag{3.26}$$

3.3.3.3 BIC for Each Sample

BIC is also used to analyze whether optimal mesh size differs for each sample even belonging to the same participant. This time, the expected value of squared term for a sample in Equation 3.5 is used in BIC to estimate the optimal mesh size for a sample at time instant t_i belonging to a participant ρ:

$$BIC_{i,\rho}(p) = \ln(\hat{E}_{i,\rho}) + \frac{p\ln(M)}{M} \, . \tag{3.27}$$

$BIC_{i,\rho}(p)$ is calculated for a number of mesh sizes where $p \in [p_{min}, p_{max}]$ and the one that minimizes $BIC_{i,\rho}(p)$ is selected as the optimal mesh size for the sample at t_i using:

$$\hat{p}^{BIC}_{i,\rho} = \underset{p}{\arg\min}(BIC_{i,\rho}(p), p \in [p_{min}, p_{max}]) \, , \tag{3.28}$$

where $\hat{p}^{BIC}_{i,\rho}$ is the optimal mesh size estimated using BIC for sample at t_i belonging to participant ρ.

3.3.3.4 BIC for Each Voxel

Expected value of squared error term for each voxel (see Equation 3.6) is used in the adopted BIC ($BIC_{j,\rho}(p)$) to select the optimal mesh size for each voxel at coordinates $\bar{s}_j$ belonging to participant ρ as:

$$BIC_{j,\rho}(p) = \ln(\hat{E}_{j,\rho}) + \frac{p\ln(N)}{N} \, , \tag{3.29}$$

where N is the number of samples. After $BIC_{j,\rho}(p)$ is estimated for many different mesh sizes, the optimal mesh size for voxel at coordinates $\bar{s}_j$ is selected as the one that minimizes $BIC_{j,\rho}(p)$,

$$\hat{p}^{BIC}_{j,\rho} = \underset{p}{\arg\min}(BIC_{j,\rho}(p), p \in [p_{min}, p_{max}]) \, , \tag{3.30}$$

where $\hat{p}^{BIC}_{j,\rho}$ refers to the optimal mesh size for voxel at $\bar{s}_j$, where the experiment is conducted on participant ρ.

This subsection shows that BIC is a function of mesh size p for a participant, class, sample and voxel. BIC is estimated for various mesh sizes and minimized over the mesh sizes p in the interval $[p_{min}, p_{max}]$. Since this interval is selected large enough, the optimal mesh size is guaranteed to lie in this interval.

3.3.4 Rissanen's Minimum Description Length (MDL)

In this study Minimum Description Length (MDL) criterion is used to find the local mesh model of size p that best represents the relationship among voxels. It assumes that, the best model, i.e. the local mesh model having the optimal mesh size, requires smallest description length. A local mesh model of size M (where M is the number of voxels), representing the relationship between a voxel and all other voxels would include redundant information. Moreover, it would cause high dimensionality problem since instead of representing a voxel with its own intensity value, the voxel is now represented in terms of all other voxels. Therefore, MDL is used to find how the information is represented with the minimum number of relationships among voxels without a high information loss.

In the following subsections, optimal mesh size is estimated for each participant, class, sample and voxel using MDL.

3.3.4.1 MDL for Each Participant

MDL is adopted to select the optimal mesh size for each participant ρ ($MDL_\rho(p)$) using the expected value of squared error term approximated for each participant (see Equation 3.3) in the following way:

$$MDL_\rho(p) = \hat{E}_\rho \left(1 + \left(\frac{p+1}{M}\right) \ln(M)\right) , \tag{3.31}$$

where p represents the mesh size, M represents the number of voxels, ρ is the participant, $\hat{E}_\rho$ is the expected value of squared error approximated for a participant and $MDL_\rho(p)$ is a function of mesh size p.

$MDL_\rho(p)$ is estimated for participant ρ for various values of p in the interval $[p_{min}, p_{max}]$ and the one minimizing the $MDL_\rho(p)$ is selected as the optimal mesh size estimated using MDL for participant ρ.

$$\hat{p}_\rho^{MDL} = \arg\min_p (MDL_\rho(p), p \in [p_{min}, p_{max}]) . \tag{3.32}$$

3.3.4.2 MDL for Each Class

MDL is adopted to analyze how the optimal mesh size differs for each class, where samples of a class belong to the same participant. $MDL_{cl,\rho}(p)$ is estimated using the following equation:

$$MDL_{cl,\rho}(p) = \hat{E}_{cl,\rho}\left(1 + \left(\frac{p+1}{M}\right)\ln(M)\right) , \tag{3.33}$$

where p is the mesh size, M is the number of voxels and ρ represents the participant. Note that, the optimal mesh size estimated using MDL for each class ($\hat{p}^{MDL}_{cl,\rho}$) depends on both the class cl and the participant ρ. Therefore, only the samples of same class belonging to the same participant are considered during the estimation of optimal mesh size for a class.

For various values of p in the interval $[p_{min}, p_{max}]$, $MDL_{cl,\rho}(p)$ is estimated for a class cl and the mesh size minimizing the $MDL_{cl,\rho}(p)$ is selected as the optimal mesh size for class cl.

$$\hat{p}^{MDL}_{cl,\rho} = \underset{p}{\arg\min}(MDL_{cl,\rho}(p), p \in [p_{min}, p_{max}]) . \tag{3.34}$$

3.3.4.3 MDL for Each Sample

MDL is also adopted to estimate the optimal mesh size for a sample at time instant t_i belonging to a participant ρ as:

$$MDL_{i,\rho}(p) = \hat{E}_{i,\rho}\left(1 + \left(\frac{p+1}{M}\right)\ln(M)\right) , \tag{3.35}$$

where p is the mesh size, M is the number of voxels and ρ represents the participant and $\hat{E}_{i,\rho}$ is the expected value of squared term for a sample in Equation 3.5, After calculating the $MDL_{i,\rho}(p)$ for many different mesh sizes, optimal mesh size for the sample at t_i is estimated as the one minimizing $MDL_{i,\rho}(p)$, using:

$$\hat{p}^{MDL}_{i,\rho} = \underset{p}{\arg\min}(MDL_{i,\rho}(p), p \in [p_{min}, p_{max}]) , \tag{3.36}$$

where $\hat{p}^{MDL}_{i,\rho}$ is the optimal mesh size estimated using MDL for sample at t_i of participant ρ.

3.3.4.4 MDL for Each Voxel

Expected value of squared error for each voxel (see Equation 3.6) is used in the adopted MDL ($MDL_{j,\rho}(p)$) to select the optimal mesh size for each voxel at coordinates $\bar{s}_j$ belonging to participant ρ using the following equation:

$$MDL_{j,\rho}(p) = \hat{E}_{j,\rho}\left(1 + \left(\frac{p+1}{N}\right)\ln(N)\right) , \quad (3.37)$$

where N is the number of samples. After $MDL_{j,\rho}(p)$ is estimated for mesh sizes where $p \in [p_{min}, p_{max}]$, the optimal mesh size for voxel at coordinates $\bar{s_j}$ is selected as the one that minimizes $MDL_{j,\rho}(p)$,

$$\hat{p}_{j,\rho}^{MDL} = \underset{p}{\arg\min}(MDL_{j,\rho}(p), p \in [p_{min}, p_{max}]) , \quad (3.38)$$

where $\hat{p}_{j,\rho}^{MDL}$ refers to the optimal mesh size for voxel at $\bar{s_j}$, and the data belongs to the participant ρ.

As it can be seen from the above derivations of MDL for a participant, class, sample and voxel, MDL is always a function of mesh size p. In order to select the optimal mesh size, MDL is estimated for various mesh sizes in the interval $[p_{min}, p_{max}]$ and the p minimizing MDL is chosen to be the optimal mesh size. Moreover, this interval should be selected in a way that, the optimal mesh size minimizing the information criterion equation lies in it.

Algorithm 2, Algorithm 3 and Algorithm 4 represent, how the optimal mesh sizes are estimated for a participant, sample and voxel, respectively.

3.4 Classification

Now that after Section 3.3, how to select the optimal mesh size for each participant, class, sample and voxel is known. Using this information, corresponding hypothetical local relational brain networks are constructed with all meshes formed around all voxels. In this section, how the connections in these local relational brain networks, i.e. the arc weights of the local meshes, form a feature vector and how they are used in classification will be explained in detail. Notice that, optimal mesh size for each class is estimated only to see how the model order varies from class to class and will not be used in the classification of samples. Therefore, in the following subsections, classification steps using optimal mesh size for a participant, sample and voxel will be explained.

3.4.1 Classification using optimal mesh size for a participant

Selecting optimal mesh size for a participant ρ, means that a local mesh of size $\hat{p}_{\rho}^{IC}$ (where IC is either FPE, AIC, BIC or MDL) is formed around all the voxels belonging to all time samples of participant ρ. Remember that, voxel intensity values $v(t_i, \bar{s_j})$, $i = 1, 2, ..N$ and $j = 1, 2, ..M$ for a participant is represented as an NxM matrix where N is the number of samples and M is the number of voxels coordinates. During the construction of feature matrix,

Algorithm 2 Estimate optimal mesh size for a participant ρ using an information criterion (IC)

```
Require: Error matrix for a participant $Err_\rho$
  Approximate $\hat{E}_\rho$ using Equation 3.3 with all items of $Err_\rho$
  for all $p \in [p_{min}, p_{max}]$ do
    if IC is FPE then
      Compute $FPE_\rho(p)$
    else if IC is AIC then
      Compute $AIC_\rho(p)$
    else if IC is BIC then
      Compute $BIC_\rho(p)$
    else if IC is MDL then
      Compute $MDL_\rho(p)$
    end if
  end for
  if IC is FPE then
    Estimate $\hat{p}_\rho^{FPE}$ using Equation 3.8
  else if IC is AIC then
    Estimate $\hat{p}_\rho^{AIC}$ using Equation 3.16
  else if IC is BIC then
    Estimate $\hat{p}_\rho^{BIC}$ using Equation 3.24
  else if IC is MDL then
    Estimate $\hat{p}_\rho^{MDL}$ using Equation 3.32
  end if
Ensure: $\hat{p}_\rho^{IC}$ where IC is either FPE, AIC, BIC or MDL
```

Algorithm 3 Estimate optimal mesh size for a sample at t_i belonging to participant ρ using an information criterion (IC)

```
Require: Error matrix for a participant $Err_\rho$
  Approximate $\hat{E}_{i,\rho}$ using Equation 3.5 with $i^{th}$ row of $Err_\rho$
  for all $p \in [p_{min}, p_{max}]$ do
    if IC is FPE then
      Compute $FPE_{i,\rho}(p)$
    else if IC is AIC then
      Compute $AIC_{i,\rho}(p)$
    else if IC is BIC then
      Compute $BIC_{i,\rho}(p)$
    else if IC is MDL then
      Compute $MDL_{i,\rho}(p)$
    end if
  end for
  if IC is FPE then
    Estimate $\hat{p}_{i,\rho}^{FPE}$ using Equation 3.12
  else if IC is AIC then
    Estimate $\hat{p}_{i,\rho}^{AIC}$ using Equation 3.20
  else if IC is BIC then
    Estimate $\hat{p}_{i,\rho}^{BIC}$ using Equation 3.28
  else if IC is MDL then
    Estimate $\hat{p}_{i,\rho}^{MDL}$ using Equation 3.36
  end if
Ensure: $\hat{p}_{i,\rho}^{IC}$ where IC is either FPE, AIC, BIC or MDL
```

Algorithm 4 Estimate optimal mesh size for a voxel at coordinates $\bar{s}_j$ belonging to participant ρ using an information criterion (IC)

```
Require: Error matrix for a participant $Err_\rho$
  Approximate $\hat{E}_{j,\rho}$ using Equation 3.6 with $j^{th}$ column of $Err_\rho$
  for all $p \in [p_{min}, p_{max}]$ do
    if IC is FPE then
      Compute $FPE_{j,\rho}(p)$
    else if IC is AIC then
      Compute $AIC_{j,\rho}(p)$
    else if IC is BIC then
      Compute $BIC_{j,\rho}(p)$
    else if IC is MDL then
      Compute $MDL_{j,\rho}(p)$
    end if
  end for
  if IC is FPE then
    Estimate $\hat{p}^{FPE}_{j,\rho}$ using Equation 3.14
  else if IC is AIC then
    Estimate $\hat{p}^{AIC}_{j,\rho}$ using Equation 3.22
  else if IC is BIC then
    Estimate $\hat{p}^{BIC}_{j,\rho}$ using Equation 3.30
  else if IC is MDL then
    Estimate $\hat{p}^{MDL}_{j,\rho}$ using Equation 3.38
  end if
Ensure: $\hat{p}^{IC}_{j,\rho}$ where IC is either FPE, AIC, BIC or MDL
```

instead of using voxel intensity values $v(t_i, \bar{s}_j)$, the arc weights of the local mesh model $a_{i,j,k}$ (where $k = 1, 2, ..p$ and p is the mesh size) are used.

If optimal mesh size is estimated for a participant ρ, all the voxels belonging to a participant will form a local mesh of same size. Hence, a single voxel $v(t_i, \bar{s}_j)$ is now represented in terms of its relationships with its neighbors $\{v(t_i, \bar{s}_k)\}_{k=1}^{\hat{p}_\rho^{IC}}$ using a mesh arc vector $\bar{a}_{i,j} = [a_{i,j,1}, a_{i,j,2}, ...a_{i,j,\hat{p}_\rho^{IC}}]$ of size $1x\hat{p}_\rho^{IC}$ (where $\hat{p}_\rho^{IC} \in \{\hat{p}_\rho^{FPE}, \hat{p}_\rho^{AIC}, \hat{p}_\rho^{BIC}, \hat{p}_\rho^{MDL}\}$). (see Figure 3.6)

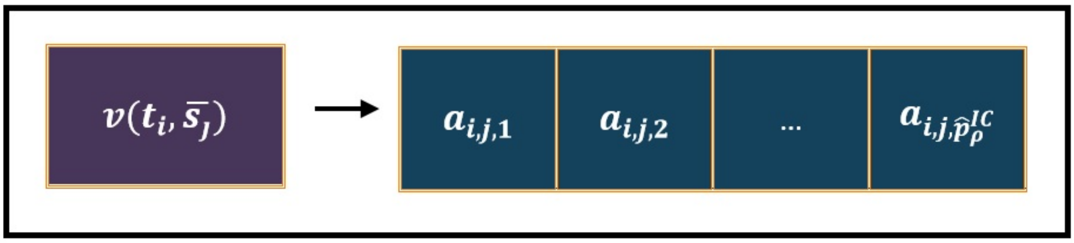


Figure 3.6: Representation of a voxel $v(t_i, \bar{s}_j)$ in terms of arc weights $a_{i,j,k}$.

By concatenating mesh arc vectors for a sample at t_i, $A_i = [\bar{a}_{i,1}, \bar{a}_{i,2}, ...\bar{a}_{i,M}]$ having size $1x\hat{p}_\rho^{IC}.M$ is formed. Finally, all the mesh arc vectors are concatenated to form a feature matrix $F = [A_1^T, A_2^T, ...A_M^T]^T$ of size $Nx\hat{p}_\rho^{IC}.M$ for a participant (see Figure 3.7)

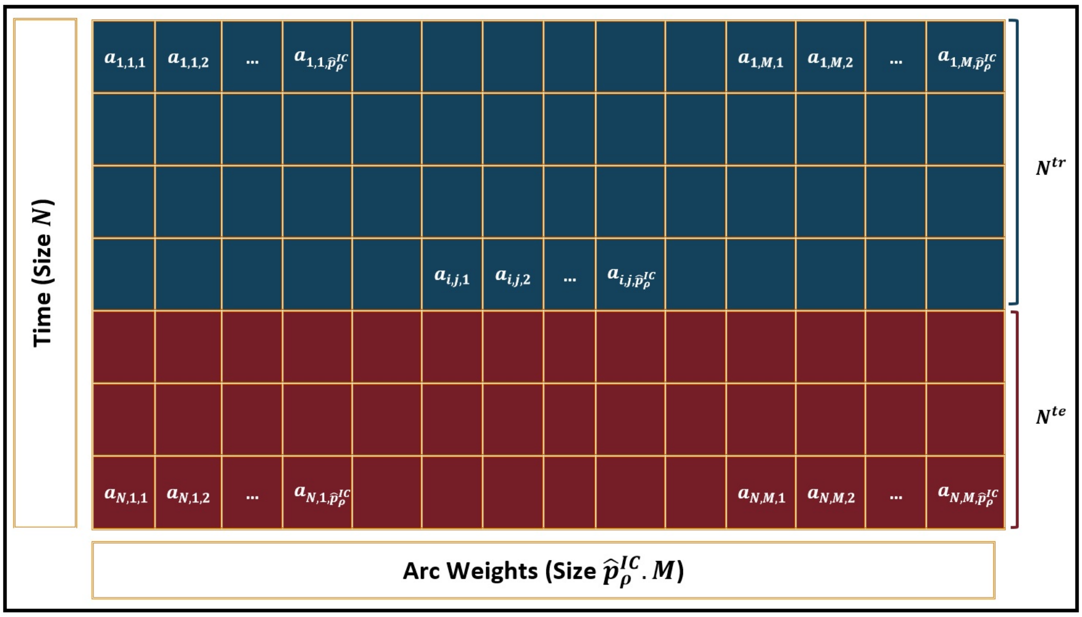
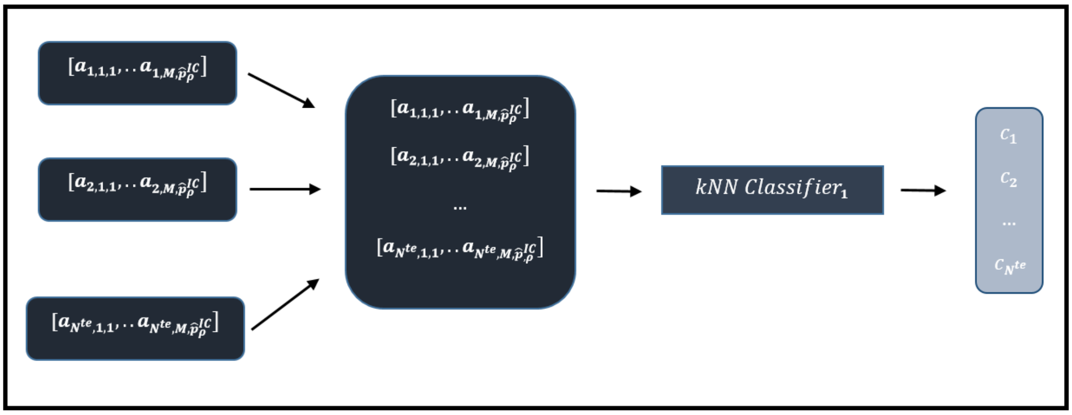

Time (Size N)
$a_{1,1,1}$
$a_{1,1,2}$
...
$a_{1,1,\hat{p}_\rho^{IC}}$
$a_{1,M,1}$
$a_{1,M,2}$
...
$a_{1,M,\hat{p}_\rho^{IC}}$
$a_{i,j,1}$
$a_{i,j,2}$
...
$a_{i,j,\hat{p}_\rho^{IC}}$
$a_{N,1,1}$
$a_{N,1,2}$
...
$a_{N,1,\hat{p}_\rho^{IC}}$
$a_{N,M,1}$
$a_{N,M,2}$
...
$a_{N,M,\hat{p}_\rho^{IC}}$
N^{tr}
N^{te}
Arc Weights (Size $\hat{p}_\rho^{IC}.M$)

Figure 3.7: Feature matrix F of a participant.

After the feature matrix F is formed, the data can be separated as training and test data by selecting N^{tr} of them for training the system and N^{te} of them to test the system, (N^{tr} is the number of training samples and N^{te} is the number of test samples). Therefore, k-NN classifier [38] is trained with training data of size $N^{tr}x\hat{p}_\rho^{IC}.M$ and tested with test data of size

$N^{te} x \hat{p}_\rho^{IC}.M$. As a result, a vector of class labels ($\hat{c}_i$) belonging to test data at t_i is obtained after this process (Figure 3.8).



$[a_{1,1,1},..a_{1,M,\hat{p}_\rho^{IC}}]$
$[a_{2,1,1},..a_{2,M,\hat{p}_\rho^{IC}}]$
$[a_{N^{te},1,1},..a_{N^{te},M,\hat{p}_\rho^{IC}}]$
$[a_{1,1,1},..a_{1,M,\hat{p}_\rho^{IC}}]$
$[a_{2,1,1},..a_{2,M,\hat{p}_\rho^{IC}}]$
...
$[a_{N^{te},1,1},..a_{N^{te},M,\hat{p}_\rho^{IC}}]$
$kNN\ Classifier_1$
c_1
c_2
...
$c_{N^{te}}$

Figure 3.8: Overview of classification process where feature vectors are constructed using the optimal mesh size for each participant. Each mesh arc vector of a test sample at t_i ($[a_{i,1,1}, .., a_{i,M,\hat{p}_{i,\rho}^{IC}}]$) is fed to the same classifier. As a result, the class label (c_i) of all test samples are estimated using the same classifier.

The steps to extract local relational features (LRF) of local mesh model are provided in Algorithm 5.

Algorithm 5 Extraction of Local Relational Features (LRF); lrf

```
Require: Dataset: $D_\rho = \{v_\rho(t_i, \bar{s}_j)\}$
  Order of LRF: $p$
  for $i = 1 \rightarrow N$ do
    for $j = 1 \rightarrow M$ do
      Compute $p - neighborhood\ \eta_p[v_\rho(t_i, \bar{s}_j)]$ of $v_\rho(t_i, \bar{s}_j)$
      Compute $\bar{a}_{i,j}$ optimizing 3.1
    end for
    Construct $A_i$ using $\bar{a}_{i,j}$
  end for
  Construct $F$ using $A_i$
Ensure: Feature Matrix $F$
```

The steps of classification using the optimal mesh size for each participant are represented in Algoritm 6

3.4.2 Classification using optimal mesh size for a sample

Selecting optimal mesh size for a sample at t_i belonging to participant ρ, means that a local mesh of size $\hat{p}_{i,\rho}^{IC}$ (where IC is either FPE, AIC, BIC or MDL) is formed around all the voxels belonging to the time sample t_i of participant ρ. In that case, $\hat{p}_{i,\rho}^{IC}$ may differ from sample to

Algorithm 6 Classification using optimal mesh size for a participant ρ using an information criterion (IC)

```
Require: Training Dataset: $D^{tr}$, Test Dataset $D^{te}$
  training class labels: $C^{tr}$
  Compute optimal mesh size for a participant $\hat{p}_{\rho}^{IC}$
  $F^{tr} \leftarrow lrf(D^{tr}, \hat{p}_{\rho}^{IC})$
  $F^{te} \leftarrow lrf(D^{te}, \hat{p}_{\rho}^{IC})$
  - Perform classification on $F^{tr}$ and $F^{te}$ using k-NN, with the algorithm parameters $\Omega$
  $\hat{C}^{te} = \{\hat{c}_i\}_{i=1}^{N^{te}} \leftarrow classify(F^{tr}, C^{tr}, F^{te}, \Omega)$
Ensure: $\hat{C}^{te} = \{\hat{c}_i\}_{i=1}^{N^{te}}$
```

sample even for the same participant ρ. Although a local mesh of same size is formed around voxels belonging to same sample at t_i, voxels belonging to different samples are represented with a different number of neighbors using arc weights $a_{i,j,k}$ of size $1x\hat{p}_{i,\rho}^{IC}$. Since the size of each mesh arc vector for a sample is $1x\hat{p}_{i,\rho}^{IC}.M$ and $\hat{p}_{i,\rho}^{IC}$ varies for each sample, it is not possible to construct a feature matrix F as suggested in Section 3.4.1. Therefore, a different approach should be followed to classify each sample.

In previous subsection, after the optimal mesh size for a participant is estimated, each sample was guaranteed to have the same size. Therefore, a subset of samples could be used to train the classifier and then the classifier could be tested with another subset of samples. However, these time samples have different sizes so, the classification process cannot be performed using a single classifier. Instead, various classifiers should be trained for each sample at t_i belonging to test data.

When a test sample comes, first the optimal mesh size $\hat{p}_{i,\rho}^{IC}$ is estimated for that sample as in Section 3.3 using either one of the FPE, AIC, BIC or MDL. Then, training data is constructed so that, each mesh arc vector corresponding to a sample in the training data has size $1x\hat{p}_{i,\rho}^{IC}.M$. Therefore, a classifier is trained with training samples having size $1x\hat{p}_{i,\rho}^{IC}$. Finally, the class label($\hat{c}_i$) of the mesh arc vector belonging to the test sample is estimated using the classifier. This procedure is performed for each test sample separately, meaning that as long as $\hat{p}_{i,\rho}^{IC}$ differs for each test sample, a different classifier is trained for it (see Figure 3.9). All the class labels $\hat{c}_i$ corresponding to mesh arc vectors belonging to t_i are combined to form resulting label vector $[\hat{c_1}, \hat{c_2}, ..\hat{c_{N^{te}}}]^T$.

The steps of classification using the optimal mesh size for each sample are provided in Algoritm 7

3.4.3 Classification using optimal mesh size for a voxel

Selecting optimal mesh size for a voxel at coordinates $\bar{s}_j$ belonging to participant ρ, means that a local mesh of size $\hat{p}_{j,\rho}^{IC}$ (where IC is either FPE, AIC, BIC or MDL) is formed around

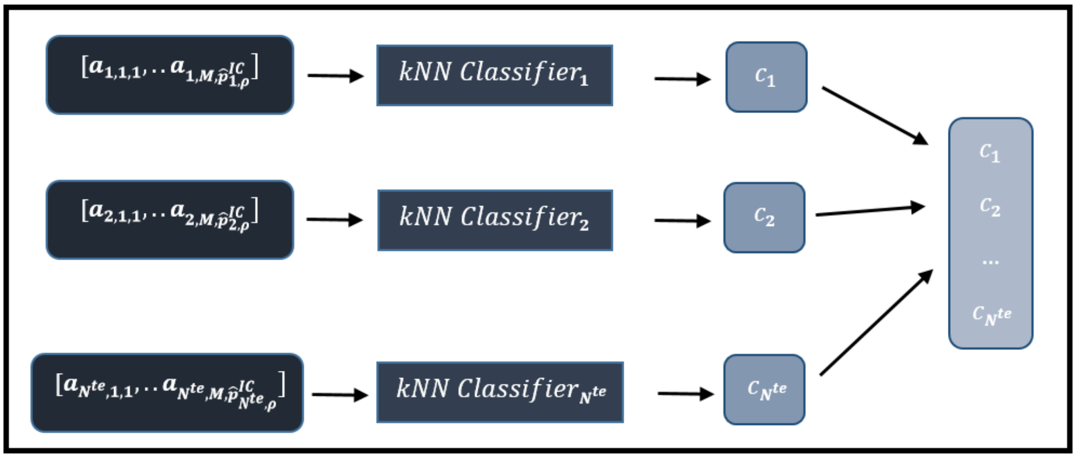

$[a_{1,1,1},..a_{1,M,\hat{p}^{IC}_{1,\rho}}]$
$kNN\ Classifier_1$
c_1
$[a_{2,1,1},..a_{2,M,\hat{p}^{IC}_{2,\rho}}]$
$kNN\ Classifier_2$
c_2
$[a_{N^{te},1,1},..a_{N^{te},M,\hat{p}^{IC}_{N^{te},\rho}}]$
$kNN\ Classifier_{N^{te}}$
$c_{N^{te}}$
c_1
c_2
...
$c_{N^{te}}$

Figure 3.9: Overview of classification process where feature vectors are constructed using the optimal mesh size for each sample. Each mesh arc vector of a test sample at t_i ($[a_{i,1,1}, .., ai, M, \hat{p}^{IC}_{i,\rho}]$) is fed to different classifier ($Classifier_i$). As a result, the class label (c_i) of test sample is estimated using each $Classifier_i$.

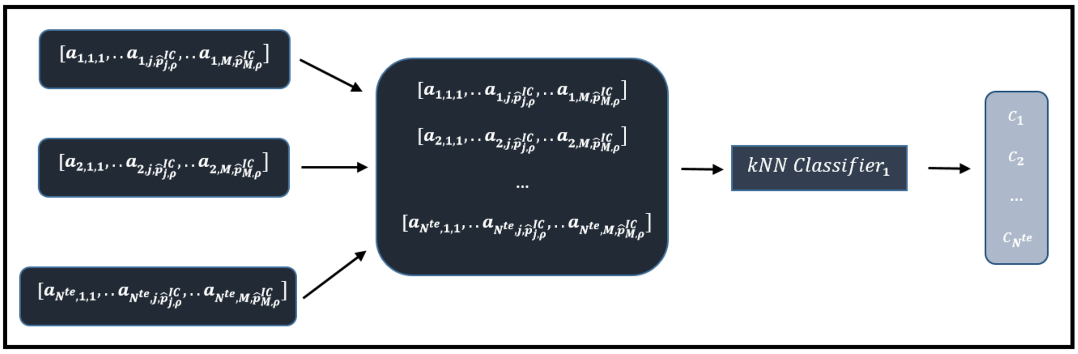
Algorithm 7 Classification using optimal mesh size for a sample using an information criterion (IC)

```
Require: Training Dataset: $D^{tr}$, Test Dataset $D^{te}$
  training class labels: $C^{tr}$
  for $i = 1 \to N$ do
    Compute optimal mesh size for a participant $\hat{p}^{IC}_{i,\rho}$
    $F^{tr} \leftarrow lrf(D^{tr}, \hat{p}^{IC}_{i,\rho})$
    $F^{te} \leftarrow lrf(A_i, \hat{p}^{IC}_{i,\rho})$
    - Perform classification on $F^{tr}$ and $F^{te}$ using k-NN, with the algorithm parameters $\Omega$
    $\hat{c}_i \leftarrow classify(F^{tr}, C^{tr}, F^{te}, \Omega)$
  end for
  Combine $\hat{c}_i$ of all test samples to form $\hat{C}^{te} = \{\hat{c}_i\}_{i=1}^{N^{te}}$
Ensure: $\hat{C}^{te} = \{\hat{c}_i\}_{i=1}^{N^{te}}$
```

each voxel. Since $\hat{p}^{IC}_{j,\rho}$ differs for each voxel even belonging to the same sample, around each voxel a local mesh of different size is formed. Therefore, voxels even belonging to same sample of the same participant ρ are represented with a different number of neighbors using arc weights $a_{i,j,k}$ of size $1x\hat{p}^{IC}_{j,\rho}$.

Since the size of each mesh arc vector for a voxel is $1x\hat{p}^{IC}_{j,\rho}$ and $1x\hat{p}^{IC}_{j,\rho}$ differs for each voxel, unlike in previous two subsections the size of constructed mesh arc vector for a sample is not $1x\hat{p}^{IC}_{j,\rho}.M$. Rather, each mesh arc vector for a sample has size $1x\sum_{\forall j}\hat{p}^{IC}_{j,\rho}$. Since each sample has the same size, they can be combined to form a feature matrix F of size $Nx\sum_{\forall j}\hat{p}^{IC}_{j,\rho}$.

For the classification, N^{tr} of samples are selected to form training data whereas remaining N^{te} of them are used to form test data (N^{tr} represents the number of training samples and N^{te} represents the number of test samples). In this section, a single k-NN classifier is trained with training data of size $N^{tr}x\sum_{\forall j}\hat{p}^{IC}_{j,\rho}$ and in order to test the classifier, test data of size $N^{te}x\sum_{\forall j}\hat{p}^{IC}_{j,\rho}$ are used. As a result, a vector of class labels ($\hat{c}_i$) belonging to test data at t_i is obtained after this process.



$[a_{1,1,1},..a_{1,j,\hat{p}^{IC}_{j,\rho}},..a_{1,M,\hat{p}^{IC}_{M,\rho}}]$
$[a_{2,1,1},..a_{2,j,\hat{p}^{IC}_{j,\rho}},..a_{2,M,\hat{p}^{IC}_{M,\rho}}]$
$[a_{N^{te},1,1},..a_{N^{te},j,\hat{p}^{IC}_{j,\rho}},..a_{N^{te},M,\hat{p}^{IC}_{M,\rho}}]$
$[a_{1,1,1},..a_{1,j,\hat{p}^{IC}_{j,\rho}},..a_{1,M,\hat{p}^{IC}_{M,\rho}}]$
$[a_{2,1,1},..a_{2,j,\hat{p}^{IC}_{j,\rho}},..a_{2,M,\hat{p}^{IC}_{M,\rho}}]$
...
$[a_{N^{te},1,1},..a_{N^{te},j,\hat{p}^{IC}_{j,\rho}},..a_{N^{te},M,\hat{p}^{IC}_{M,\rho}}]$
$kNN\ Classifier_1$
c_1
c_2
...
$c_{N^{te}}$

Figure 3.10: Overview of classification process where feature vectors are constructed using the optimal mesh size for each voxel. Each mesh arc vector of a test sample at t_i ($[a_{i,1,1},..,ai,M,\hat{p}^{IC}_{i,\rho}]$) is fed to the same classifier. As a result, the class label ($\hat{c}_i$) of all test samples are estimated using the same classifier.

3.5 Calculating Accuracy

In the previous three subsections, a resulting vector of class labels $[\hat{c_1},\hat{c_2},..\hat{c_{N^{te}}}]^T$ of size $N^{te}x1$ is obtained. By comparing this vector with the ground truth vector of class labels corresponding to the test samples at t_i, ($[c_1,c_2,..c_{N^{te}}]^T$ of size $N^{te}x1$), accuracy of the classifier (or accuracy of the classifiers for classification using optimal mesh size for a sample), acc is calculated using:

$$acc = \frac{\sum_{i=1}^{N^{te}}\delta(c_i,\hat{c}_i)}{N^{te}}, \tag{3.39}$$

where c_i is the class label of sample at t_i, $\hat{c}_i$ is the estimated class label of sample at t_i and N^{te} is the number of test samples and $\delta(c_i, \hat{c}_i) = 1$ if $c_i = \hat{c}_i$ and $\delta(c_i, \hat{c}_i) = 0$ otherwise.

In the experiments, arc weights obtained using optimal mesh size for a participant, sample or voxel are used for classification. For each of the experiments, the accuracy is calculated using 3.39.

The steps to extract local relational features (LRF) for variable size of local meshes are provided in Algorithm 8.

Algorithm 8 Extraction of Local Relational Features (LRF) for variable size of local meshes; *lrf.variable*

```
Require: Dataset: $D_\rho = \{v_\rho(t_i, \bar{s}_j)\}$
  for $i = 1 \rightarrow N$ do
    for $j = 1 \rightarrow M$ do
      Compute optimal mesh size for a voxel $\hat{p}^{IC}_{j,\rho}$
      $p \leftarrow \hat{p}^{IC}_{j,\rho}$
      Compute $p - neighborhood\ \eta_p[v_\rho(t_i, \bar{s}_j)]$ of $v_\rho(t_i, \bar{s}_j)$
      Compute $\bar{a}_{i,j}$ optimizing 3.1
    end for
    Construct $A_i$ using $\bar{a}_{i,j}$
  end for
  Construct $F$ using $A_i$
Ensure: Feature Matrix $F$
```

The steps of classification using the optimal mesh size for each voxel are represented in Algoritm 9.

Algorithm 9 Classification using optimal mesh size for a voxel ρ using an information criterion (IC)

```
Require: Training Dataset: $D^{tr}$, Test Dataset $D^{te}$
  training class labels: $C^{tr}$
  $F^{tr} \leftarrow lrf.variable(D^{tr})$
  $F^{te} \leftarrow lrf.variable(D^{te})$
  - Perform classification on $F^{tr}$ and $F^{te}$ using k-NN, with the algorithm parameters $\Omega$
  $\hat{C}^{te} = \{\hat{c}_i\}_{i=1}^{N^{te}} \leftarrow classify(F^{tr}, C^{tr}, F^{te}, \Omega)$
Ensure: $\hat{C}^{te} = \{\hat{c}_i\}_{i=1}^{N^{te}}$
```

3.6 Chapter Summary

In this chapter, a new type of hypothetical connectivity, called "local relational connectivity" is proposed and how local relational connectivity network is constructed using an information theoretic approach is explained. In this approach, first the expected value of squared error term obtained from linear regression equation should be approximated for a participant, class, sample and voxel. Therefore, Section 3.2 explains how the error term, which will be used to select optimal mesh size, is approximated for a participant, class, sample and voxel. Then, estimated error terms are used to obtain four information theoretic criteria namely FPE, AIC, BIC and MDL for all mesh sizes p in the interval $[p_{min}, pmax]$. The mesh sizes p minimizing the information theoretic criteria are selected as optimal mesh size estimated using corresponding criterion. Around each voxel, a local mesh having optimal mesh size is formed. Then, the optimal mesh size is used to form the feature vectors from training and test samples. Since optimal mesh size varies for a participant, sample and voxel, the size of feature vectors may, also vary. For a participant a local mesh having same size is formed around each voxel for all time instants. Similarly for a sample, a local mesh of same size is formed around all voxels, but the size of feature vector varies for each sample. On the other hand, if the optimal mesh size is estimated for a voxel, a local mesh of varying sizes are formed around each voxel and for all time instants feature vector of same size is formed. Finally, the extracted LRF features are used in the classification of cognitive states. Due to the variations in the dimensions of the resulting feature vector, different approaches are used in classification for participant, sample and voxel. Yet, all aim to classify the cognitive states.

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Experimental Setup to Collect the fMRI Data

In this study, neural activation during two working memory tasks, namely item recognition (IR) and judgment of recency (JOR) are recorded via fMRI [48]. For both IR and JOR tasks, each trial began with the presentation of a centered fixation point for 500 ms. Then, a study list including five consonants were presented one at a time for 500 ms each. After the presentation of the study list, a task cue was presented to indicate the upcoming memory judgment (IR or JOR) for 750 ms. Following the presentation of task cues, two probe consonants were presented for both tasks for 3000 ms. In IR trials, one consonant is from the study list where the other one was new. Participants were requested to indicate the one belonging to the study list in this task with a button press. In JOR trials on the other hand, both probes were from the study list, and participants were asked to select the probe that was more recent in the study list 4.1.

Preprocessing of neuroimaging data steps included slice acquisition timing across slices, realignment of images to the first volume for head movement correction, normalization of anatomical and functional images to a standard template EPI and smoothing of images with a 6-mm full-width half-maximum isotropic Gaussian kernel.

In this study, dataset consists of 320 time instants, (160 of them belongs to IR and 160 of them belongs to JOR). Among them, 240 samples (120 IR and 120 JOR samples) are used to train the system whereas 80 of them (40 IR and 40 JOR samples) are used to test the accuracy of the classifier. Note that, each sample includes measurements from 2030 voxels. Abovementioned dataset is for a single participant and in this study dataset belonging to 8 participants (aged 18 - 28) are used.

4.2 Optimal mesh size analysis

During the experiments, mesh size p is selected in the interval $[2, 100]$. In all experiments, as mesh size increases, the information criteria decreases up to a value in this interval and then

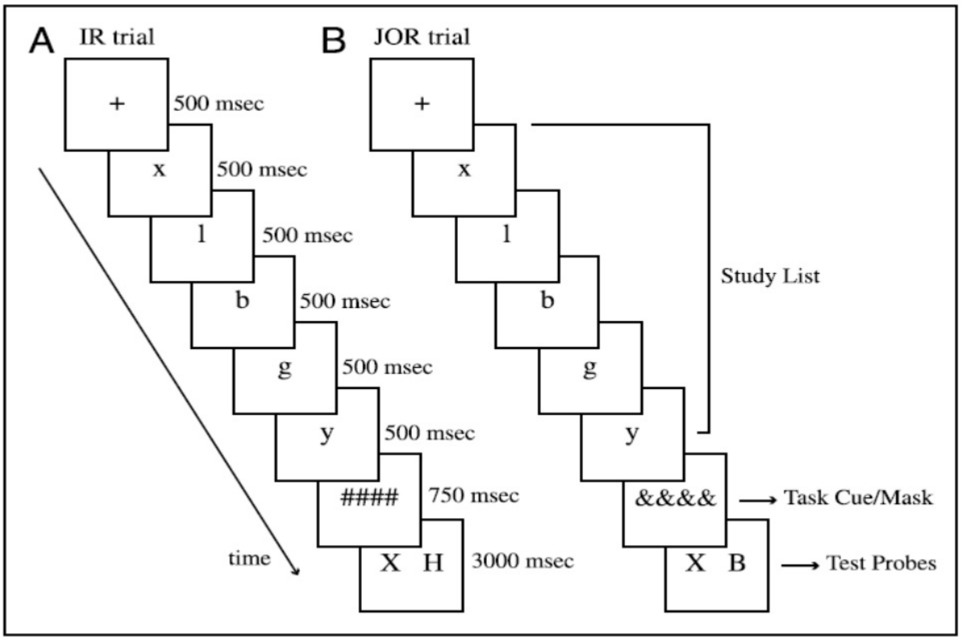


Figure 4.1: A sample sequence for an item recognition (IR) trial (shown on panel A) and a judgment of recency (JOR) trial (shown on panel B). After the presentation of fixation point, a study list consisting of five consonants were presented to the participant. Then, different visual masks that cued different tasks were presented. Finally, two test probes were shown in both tasks and either JOR or IR judgment was performed [48].

starts to increase after this value. Hence, the minimum of information theoretic criteria is always in this interval for this study. Therefore, the interval $[2, 100]$ is large enough so that the optimal mesh size is guaranteed to lie into this interval. In this thesis, optimal mesh size is estimated for each participant, class, sample and voxel using four information theoretic criteria namely, Final Prediction Error (FPE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Rissanen's Minimum Description Length (MDL). How optimal mesh size varies for each method will be presented and discussed in the following subsections.

4.2.1 Optimal mesh size for a participant

For each participant $\rho = [1, ..8]$, $FPE_\rho(p), AIC_\rho(p), BIC_\rho(p)$ and $MDL_\rho(p)$ are calculated in the interval $p = [2, 100]$ and the mesh size minimizing the corresponding criterion is selected as the optimal mesh size ($\hat{p}_\rho^{IC}$) for participant ρ (where IC is either FPE, AIC, BIC or MDL).

Table 4.1 represents estimated optimal mesh sizes for each participant using FPE ($\hat{p}_\rho^{FPE}$), AIC ($\hat{p}_\rho^{AIC}$), BIC ($\hat{p}_\rho^{BIC}$) and MDL ($\hat{p}_\rho^{MDL}$). As it can be seen, for some participants $\rho \in [1, 5, 7, 8]$ optimal mesh sizes estimated using four different criteria are the same, whereas for the remaining participants, estimated optimal mesh sizes differ based on the information criteria used. Moreover, notice that for all participants $\hat{p}_\rho^{FPE} = \hat{p}_\rho^{AIC}$. Furthermore, Table 4.1 indicates that, the optimal mesh size greatly varies from participant to participant.

Table4.1: Optimal mesh size for participants $\rho = \{1, ..8\}$ estimated using FPE, AIC, BIC and MDL.

	$(\hat{p}_{\rho}^{FPE})$	$(\hat{p}_{\rho}^{AIC})$	$(\hat{p}_{\rho}^{BIC})$	$(\hat{p}_{\rho}^{MDL})$
$\rho = 1$	17	17	17	17
$\rho = 2$	34	34	23	23
$\rho = 3$	39	39	23	29
$\rho = 4$	70	70	40	42
$\rho = 5$	23	23	23	23
$\rho = 6$	16	16	14	16
$\rho = 7$	25	25	25	25
$\rho = 8$	17	17	17	17

Suppose that, optimal mesh size estimated for a participant is small. Then, the local relational connectivity network, created using single sample at time instant t_i would be similar to the one in Figure 4.2 which is drawn using BrainNet Viewer Toolbox [68].

4.2.2 Optimal mesh size for a class

For each class of samples cl, where $cl = [IR, JOR]$, belonging to the participant $\rho = \{1, ..8\}$, $FPE_{cl,\rho}(p), AIC_{cl,\rho}(p), BIC_{cl,\rho}(p)$ and $MDL_{cl,\rho}(p)$ are calculated in the interval $p = [2, 100]$ and the mesh size minimizing the corresponding criterion is selected as the optimal mesh size $(\hat{p}_{cl,\rho}^{IC})$ for class cl (where IC is either FPE, AIC, BIC or MDL).

Note that, only the samples having class cl are used to estimate the optimal mesh size for each class belonging to a participant. As a result, for each participant ρ, single optimal mesh size is obtained for each class cl.

Table4.2: Optimal mesh size for for classes IR and JOR where samples belong to participants $\rho = \{1, ..8\}$ estimated using FPE, AIC, BIC and MDL

.

	FPE		AIC		BIC		MDL	
	$\hat{p}_{IR,\rho}^{FPE}$	$\hat{p}_{JOR,\rho}^{FPE}$	$\hat{p}_{IR,\rho}^{AIC}$	$\hat{p}_{JOR,\rho}^{AIC}$	$\hat{p}_{IR,\rho}^{BIC}$	$\hat{p}_{JOR,\rho}^{BIC}$	$\hat{p}_{IR,\rho}^{MDL}$	$\hat{p}_{JOR,\rho}^{MDL}$
$\rho = 1$	17	17	17	17	17	17	17	17
$\rho = 2$	34	34	34	34	23	23	23	23
$\rho = 3$	39	39	39	39	23	23	29	29
$\rho = 4$	70	70	70	70	40	40	42	42
$\rho = 5$	23	23	23	23	23	23	23	23
$\rho = 6$	16	17	16	17	14	16	14	16
$\rho = 7$	30	25	30	25	25	25	25	25
$\rho = 8$	17	17	17	17	17	17	17	17

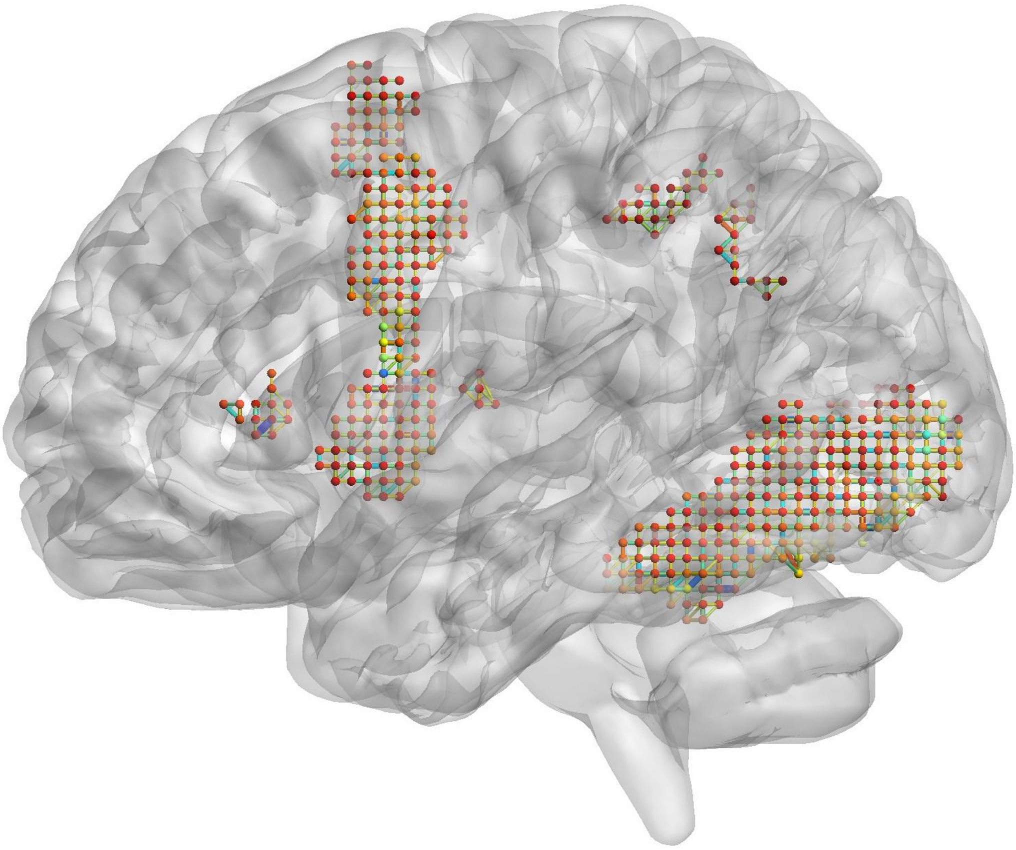


Figure 4.2: Local relational connectivity network constructed using a small optimal mesh size for a participant.

Table 4.2 represents estimated optimal mesh sizes for each class cl using FPE ($\hat{p}^{FPE}_{cl,\rho}$), AIC ($\hat{p}^{AIC}_{cl,\rho}$), BIC ($\hat{p}^{BIC}_{cl,\rho}$) and MDL ($\hat{p}^{MDL}_{cl,\rho}$), where class is either IR or JOR. The main finding of this experiment can be detected when Table 4.1 and Table 4.2 are analyzed together. As it can be seen, if the samples belonging to same class are used to estimate the optimal mesh size instead of all samples belonging to a participant, the resulting optimal mesh size does not change for 6 participants. Only in two participants ($\rho \in \{6, 7\}$), optimal mesh size estimated for two different classes differ. Therefore, it can be stated that the class of samples does not play a major role in estimating the optimal mesh size.

Recall that, this experiment is conducted to observe the effect of class on the optimal mesh size. These findings will not be used for classification of samples as classes of IR or JOR.

4.2.3 Optimal mesh size for a sample

For each sample at t_i, where $i = \{1, ..80\}$ belonging to participant $\rho = \{1, ..8\}$, $FPE_{i,\rho}(p)$, $AIC_{i,\rho}(p)$, $BIC_{i,\rho}(p)$ and $MDL_{i,\rho}(p)$ are calculated in the interval $p = [2, 100]$ and the mesh size minimizing the corresponding criterion is selected as the optimal mesh size ($\hat{p}^{IC}_{i,\rho}$) for the sample at t_i belonging to participant ρ (where IC is either FPE, AIC, BIC or MDL).

Notice that, this time for a participant ρ, various optimal mesh sizes are obtained for different samples so, this time mean (μ_ρ^{IC}) of optimal mesh sizes for a participant can be estimated as follows:

$$\mu_\rho^{IC} \cong \frac{1}{N^{te}} \sum_{i=1}^{N^{te}} \hat{p}_{i,\rho}^{IC} \,, \tag{4.1}$$

where N^{te} represents the number of test samples and $\hat{p}_{i,\rho}^{IC}$ is the optimal mesh size estimated for sample at t_i belonging to participant ρ. Similarly, the standard deviation σ_ρ^{IC} can be estimated using,

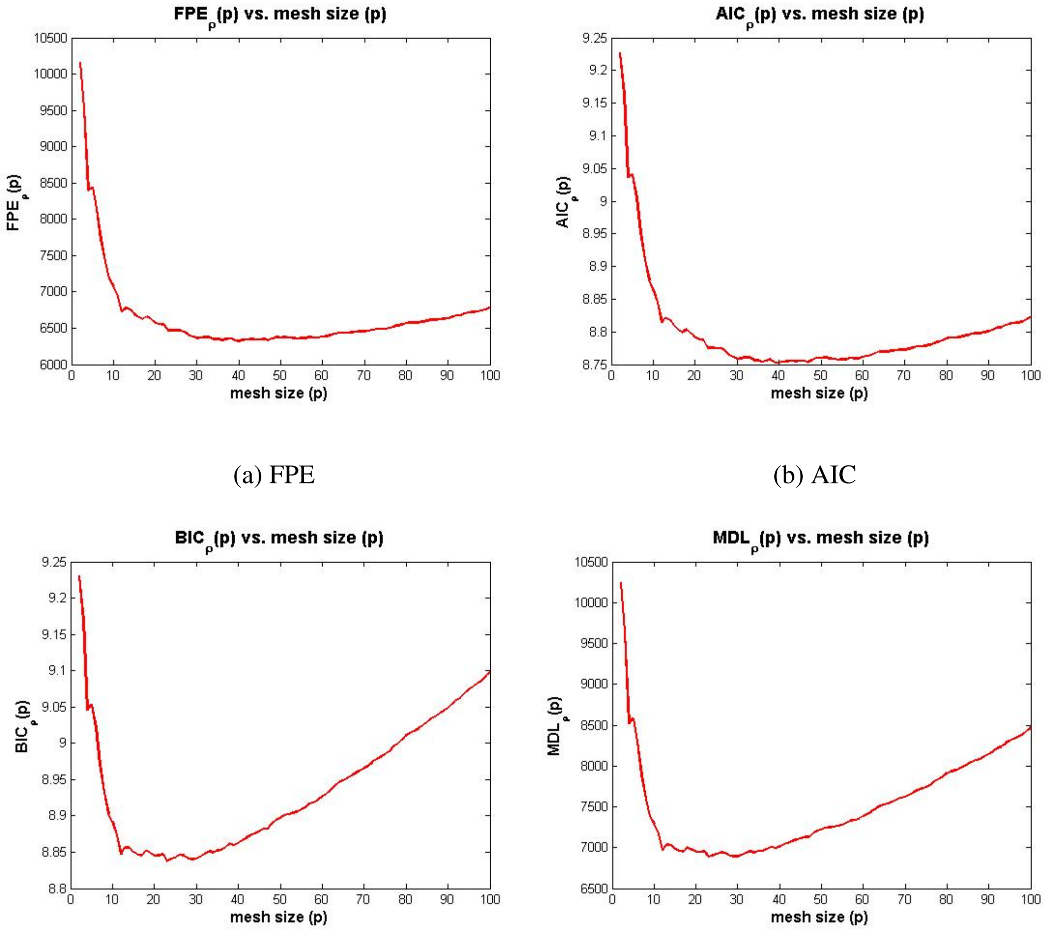
$$\sigma_\rho^{IC} \cong \sqrt{\frac{1}{N^{te}} \sum_{i=1}^{N^{te}} (\hat{p}_{i,\rho}^{IC} - \mu_\rho^{IC})^2} \,. \tag{4.2}$$

In the dataset, the number of test samples N^{te} is 80 for each participant. The resulting mean and standard deviations of optimal mesh sizes estimated using FPE, AIC, BIC and MDL are shown in Table 4.3. Moreover, this table also presents the intervals, which cover optimal mesh sizes belonging to all samples for a participant ρ.

The first finding revealed by Table 4.3 is that, optimal mesh size estimation using FPE and AIC give exactly the same intervals, mean values and standard deviations of optimal mesh sizes. Since both criteria are proposed by Akaike and AIC is proposed to overcome the inconsistency of FPE, one would expect an improvement by using AIC than using FPE. However, both criteria estimate the same optimal mesh sizes for each sample. Surprisingly, similar results, in which the same model orders are estimated using FPE and AIC are obtained in the study [33].

Moreover, Table 4.3 also reveals that for some participants, optimal mesh sizes estimated for all samples are the same so that corresponding standard deviation is 0. For example, optimal mesh size estimated for all samples belonging to participant 7 using BIC ($\rho = 7$) is 25. Notice that, in such situations optimal mesh size estimated for a participant (see Table 4.1) is the same as the mean of optimal mesh sizes estimated for each sample belonging to the same participant. For example, optimal mesh size estimated for participant 7 using BIC also equals to 25 in Table 4.1.

For most of the participants, optimal mesh sizes estimated for each sample belonging to a participant are not the same. How it varies can be understood from the corresponding mean and standard deviation values of 4.3. Notice that, these findings reveal that, unlike the situation for a class, optimal mesh size varies for each sample. Therefore, it can be concluded that the sample plays an important role in estimating the optimal mesh size.



(a) FPE

(b) AIC

(c) BIC

(d) MDL

Figure 4.3: A Sample of $FPE_{\rho}(p)$, $AIC_{\rho}(p)$, $BIC_{\rho}(p)$ and $MDL_{\rho}(p)$ distributions for a participant ρ.

Table4.3: Interval, mean and standard deviations of optimal mesh sizes over all samples for participants $\rho = [1, ..8]$ estimated using FPE (top left), AIC (top right), BIC (bottom left) and MDL (bottom right)

	FPE		
	Interval	μ_ρ^{FPE}	σ_ρ^{FPE}
$\rho = 1$	17 - 39	18.59	3.66
$\rho = 2$	23 - 35	33.41	1.53
$\rho = 3$	39 - 47	40.11	2.35
$\rho = 4$	69 - 70	69.88	0.31
$\rho = 5$	23 - 29	26.03	2.99
$\rho = 6$	16 - 17	16.43	0.49
$\rho = 7$	25 - 37	27.42	3.20
$\rho = 8$	16 - 30	20.05	5.33

	AIC		
	Interval	μ_ρ^{AIC}	σ_ρ^{AIC}
$\rho = 1$	17 - 39	18.59	3.66
$\rho = 2$	23 - 35	33.41	1.53
$\rho = 3$	39 - 47	40.11	2.35
$\rho = 4$	69 - 70	69.88	0.31
$\rho = 5$	23 - 29	26.03	2.99
$\rho = 6$	16 - 17	16.43	0.49
$\rho = 7$	25 - 37	27.42	3.20
$\rho = 8$	16 - 30	20.05	5.33

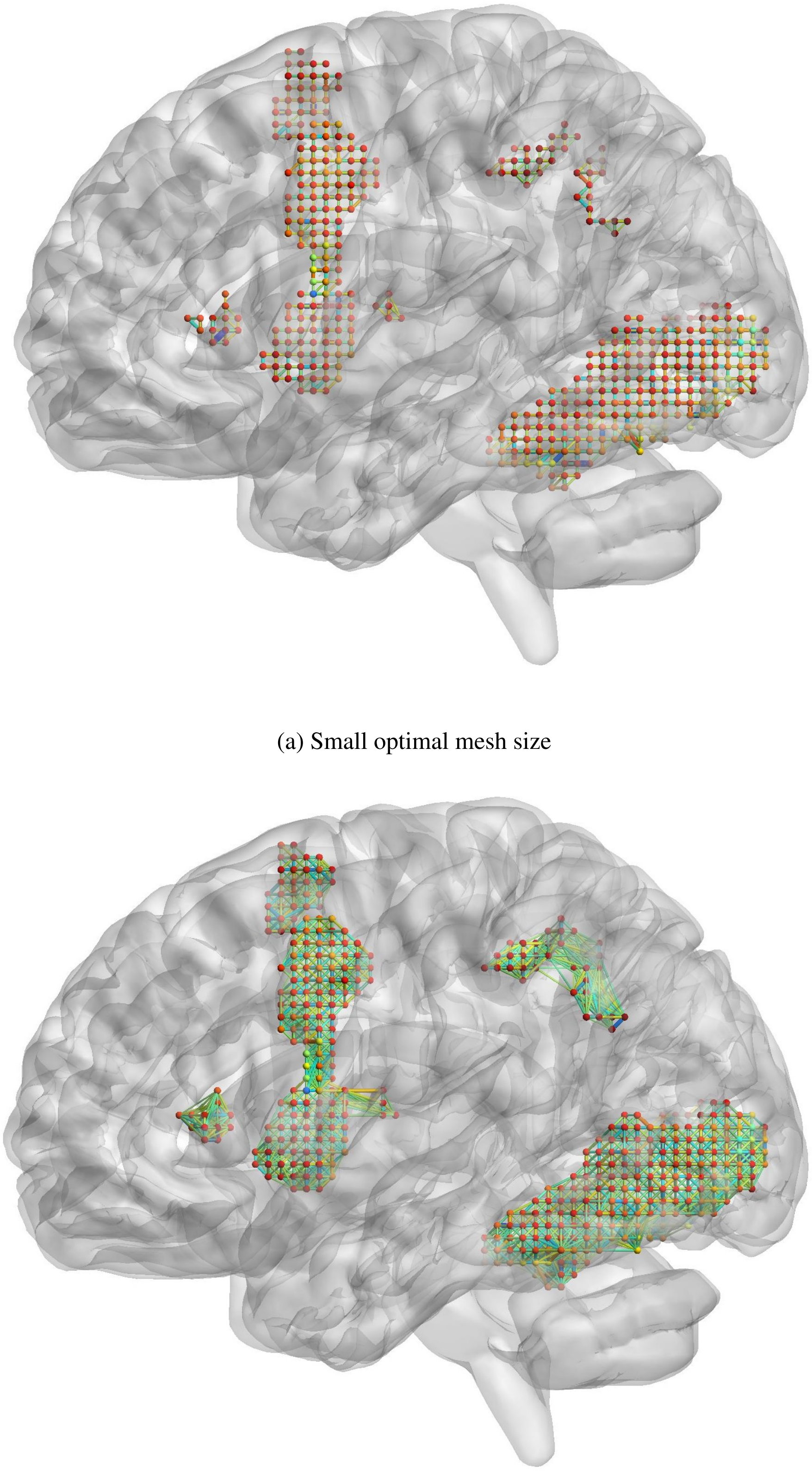
	BIC		
	Interval	μ_ρ^{BIC}	σ_ρ^{BIC}
$\rho = 1$	17 - 17	17	0
$\rho = 2$	16 - 23	22.76	1.00
$\rho = 3$	23 - 30	24.64	2.54
$\rho = 4$	40 - 42	40.83	0.98
$\rho = 5$	20 - 23	22.96	0.33
$\rho = 6$	12 - 17	14.69	1.37
$\rho = 7$	25 - 25	25	0
$\rho = 8$	12 - 17	16.93	0.55

	MDL		
	Interval	μ_ρ^{MDL}	σ_ρ^{MDL}
$\rho = 1$	17 - 17	17	0
$\rho = 2$	16 - 23	22.80	0.95
$\rho = 3$	23 - 30	28.44	2.09
$\rho = 4$	40 - 42	41.49	0.86
$\rho = 5$	23 - 23	23	0
$\rho = 6$	12 - 17	15.16	1.24
$\rho = 7$	25 - 25	25	0
$\rho = 8$	12 - 17	16.93	0.55

The $FPE_\rho(p)$, $AIC_\rho(p)$, $BIC_\rho(p)$ and $MDL_\rho(p)$ distributions over all mesh sizes p vary based on the information theoretic criterion used. As a result, for the same participant, the minimum of these distributions are different so that the optimal mesh sizes estimated for the same participant using four criteria become different.

Figure 4.3 reveals how these distributions and their minima differ based on the information theoretic criteria. Notice that, since FPE and AIC estimates the same optimal sizes for a sample, their distributions are similar and their optimal mesh sizes are equal ($\hat{p}_\rho^{FPE} = \hat{p}_\rho^{AIC} = 39$). On the other hand, the optimal mesh size estimated for the same sample using BIC equals to 23 ($\hat{p}_\rho^{BIC} = 23$) and using MDL equals to 29 ($\hat{p}_\rho^{MDL} = 29$). Notice that, all the distributions are similar, in other words the information criteria decreases up to a minimum and then starts to increase. Yet, the minima vary based on the information theoretic criteria.

Suppose that, an example optimal mesh size estimated for a sample is small and for another sample it is large. Then, the local relational connectivity network, created using single sample



(a) Small optimal mesh size

(b) Large optimal mesh size

Figure 4.4: Local Relational Connectivity Network constructed using small and large optimal mesh sizes for 2 different samples

at 2 different time instants would be similar to the one in Figure 4.2 Notice that, around every voxel a local mesh of small size or large size is constructed for 2 different samples. As it can be seen, the upper local relational connectivity network in 4.4 is rather sparse, whereas local relational connectivity network on the bottom is dense. The figure 4.4 is drawn using BrainNet Viewer Toolbox [68].

Moreover, the $FPE_{i,\rho}(p)$, $AIC_{i,\rho}(p)$, $BIC_{i,\rho}(p)$ and $MDL_{i,\rho}(p)$ distributions over all mesh sizes p vary based on the information theoretic criterion used. As a result, for the same sample, the minimum of these distributions are different so that the optimal mesh sizes estimated for the same sample using four criteria become different.

Figure 4.5 reveals how these distributions and their minima differ based on the information theoretic criteria. Notice that, since FPE and AIC estimates the same optimal sizes for a sample, their distributions and optimal mesh sizes are equal ($\hat{p}_{i,\rho}^{FPE} = \hat{p}_{i,\rho}^{AIC} = 47$). On the other hand, the optimal mesh size estimated for the same sample using BIC equals to 23 ($\hat{p}_{i,\rho}^{BIC} = 23$) and using MDL equals to 29 ($\hat{p}_{i,\rho}^{MDL} = 29$). Notice that, all the distributions are similar, in other words the information criteria decreases up to a minimum and then starts to increase. Yet, the minima vary based on the information theoretic criteria.

4.2.4 Optimal mesh size for a voxel

For each voxel at coordinates $\bar{s_j}$, where $j = 1, ..2030$ belonging to participant $\rho = \{1, ..8\}$, $FPE_{j,\rho}(p)$, $AIC_{j,\rho}(p)$, $BIC_{j,\rho}(p)$ and $MDL_{j,\rho}(p)$ are calculated in the interval $p = [2, 100]$ and the mesh size minimizing the corresponding criterion is selected as the optimal mesh size ($\hat{p}_{j,\rho}^{IC}$) for the voxel at coordinates $\bar{s_j}$ belonging to participant ρ (where IC is either FPE, AIC, BIC or MDL).

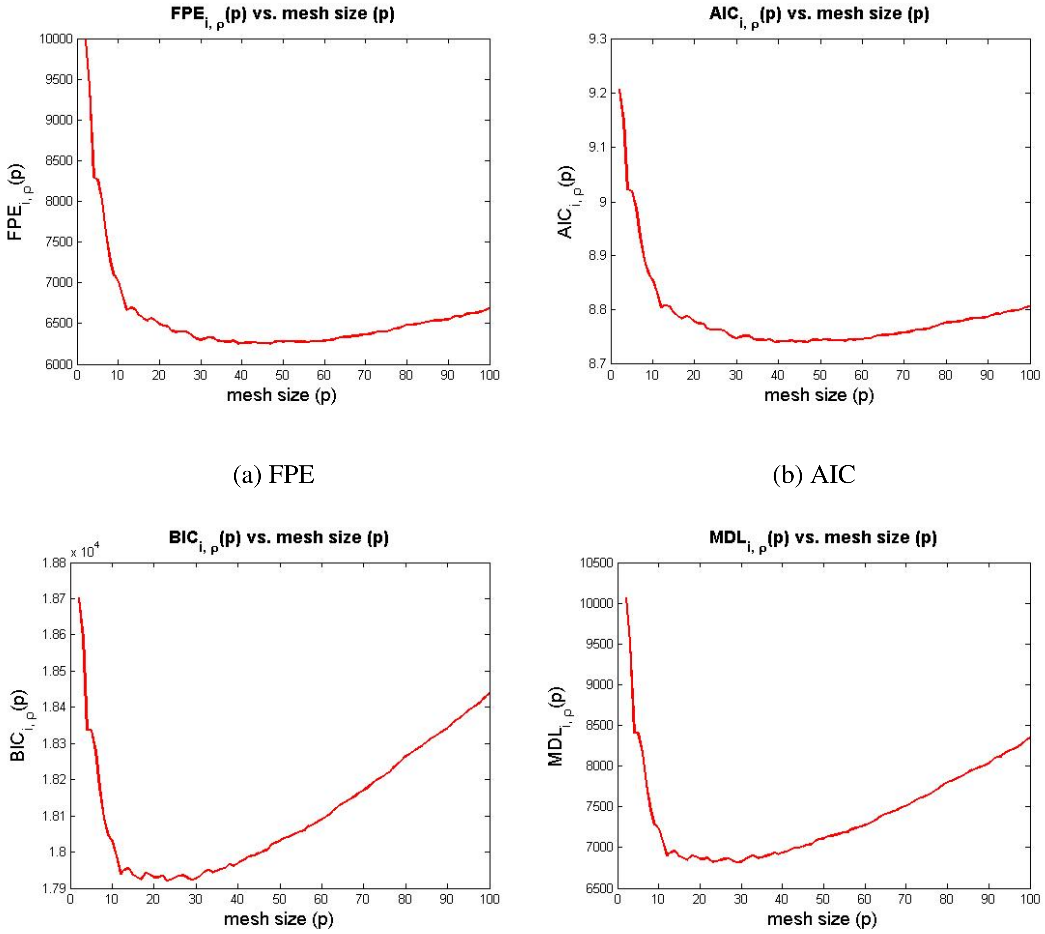
Notice that, in this case around each voxel in a sample, various sizes of meshes are created. Therefore, the mean (${\mu_{vox_\rho}}^{IC}$) of optimal mesh sizes for a participant ρ can be estimated as follows:

$$
{\mu_{vox_\rho}}^{IC} \cong \frac{1}{M} \sum_{j=1}^{M} \hat{p}_{j,\rho}^{IC} \,, \tag{4.3}
$$

where M represents the number of voxels and $\hat{p}_{j,\rho}^{IC}$ is the optimal mesh size estimated for a voxel at coordinates $\bar{s_j}$ belonging to participant ρ. After the mean value of optimal mesh sizes (${\mu_{vox_\rho}}^{IC}$) is calculated, their standard deviation can be estimated using,

$$
{\sigma_{vox_\rho}}^{IC} \cong \sqrt{\frac{1}{M} \sum_{j=1}^{M} (\hat{p}_{j,\rho}^{IC} - {\mu_{vox_\rho}}^{IC})^2} \,. \tag{4.4}
$$

Note that, ${\mu_{vox_\rho}}^{IC}$ and ${\sigma_{vox_\rho}}^{IC}$ represent the mean and standard deviation obtained over all op-



(a) FPE

(b) AIC

(c) BIC

(d) MDL

Figure 4.5: A Sample of $FPE_{i,\rho}(p)$, $AIC_{i,\rho}(p)$, $BIC_{i,\rho}(p)$ and $MDL_{i,\rho}(p)$ distributions at t_i belonging to participant ρ.

timal mesh sizes for voxels, whereas μ_{ρ}^{IC} and σ_{ρ}^{IC} in Section 4.2.3 are the mean and standard deviation estimated over all optimal mesh sizes for samples.

Table4.4: Interval, mean and standard deviations of optimal mesh sizes over all voxels for participants $\rho = [1, ..8]$ estimated using FPE (top left), AIC (top right), BIC (bottom left) and MDL (bottom right)

	FPE		
	Interval	$\mu_{vox_{\rho}}^{FPE}$	$\sigma_{vox_{\rho}}^{FPE}$
$\rho = 1$	2 - 88	11.43	12.58
$\rho = 2$	2 - 86	11.75	12.33
$\rho = 3$	2 - 90	12.85	12.81
$\rho = 4$	2 - 86	17.05	14.47
$\rho = 5$	2 - 85	11.72	12.70
$\rho = 6$	2 - 78	11.01	11.20
$\rho = 7$	2 - 84	13.58	13.11
$\rho = 8$	2 - 84	11.85	11.82

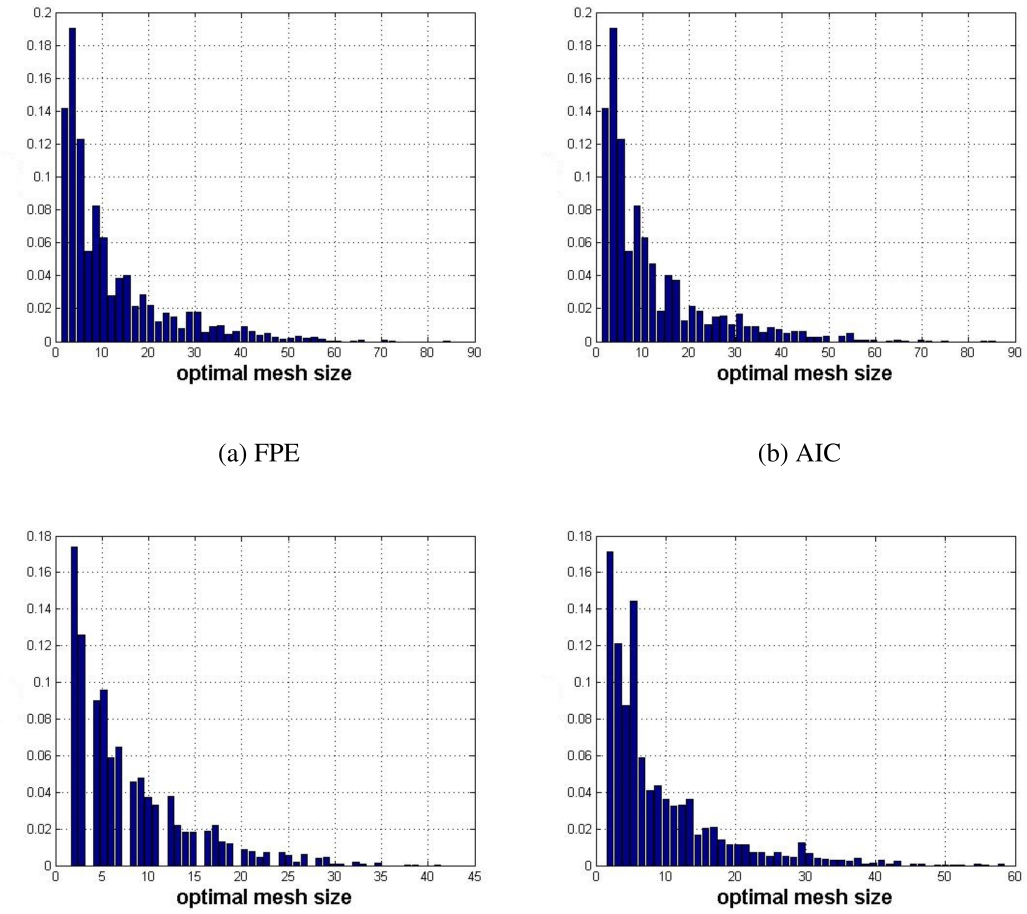
	AIC		
	Interval	$\mu_{vox_{\rho}}^{AIC}$	$\sigma_{vox_{\rho}}^{AIC}$
$\rho = 1$	2 - 91	11.58	13.03
$\rho = 2$	2 - 86	11.89	12.70
$\rho = 3$	2 - 90	12.98	13.10
$\rho = 4$	2 - 86	17.35	14.99
$\rho = 5$	2 - 90	11.80	12.98
$\rho = 6$	2 - 83	11.15	11.68
$\rho = 7$	2 - 84	13.70	13.36
$\rho = 8$	2 - 85	11.92	12.06

	BIC		
	Interval	$\mu_{vox_{\rho}}^{BIC}$	$\sigma_{vox_{\rho}}^{BIC}$
$\rho = 1$	2 - 37	7.30	5.68
$\rho = 2$	2 - 38	7.66	5.95
$\rho = 3$	2 - 37	8.26	6.50
$\rho = 4$	2 - 39	10.66	7.44
$\rho = 5$	2 - 40	7.61	6.11
$\rho = 6$	2 - 42	7.43	5.86
$\rho = 7$	2 - 45	8.62	6.62
$\rho = 8$	2 - 41	7.92	6.40

	MDL		
	Interval	$\mu_{vox_{\rho}}^{MDL}$	$\sigma_{vox_{\rho}}^{MDL}$
$\rho = 1$	2 - 88	8.76	9.34
$\rho = 2$	2 - 86	9.47	9.85
$\rho = 3$	2 - 90	10.14	9.97
$\rho = 4$	2 - 86	13.21	11.16
$\rho = 5$	2 - 83	9.27	9.97
$\rho = 6$	2 - 71	8.71	8.59
$\rho = 7$	2 - 84	10.55	10.03
$\rho = 8$	2 - 58	9.26	8.84

Table 4.4 shows that, for each participant, distribution of optimal mesh size for a voxel estimated using FPE, AIC, BIC and MDL varies. While the interval is rather large for FPE, AIC and MDL, optimal mesh sizes estimated for a voxel lie in rather smaller interval. Notice that, each mean value is small compared to the ones in Table 4.3 indicating that around majority of the voxels a local mesh of small size is formed. In order to understand how the optimal mesh sizes for all voxels are distributed, and how they differ based on the information criterion, histograms of optimal mesh sizes estimated for each voxel belonging to participant ρ can be observed as in Figure 4.6.

The $FPE_{j,\rho}(p)$, $AIC_{j,\rho}(p)$, $BIC_{j,\rho}(p)$ and $MDL_{j,\rho}(p)$ distributions over all mesh sizes p vary based on the information theoretic criterion used. As a result, for the same voxel, the minimum of these distributions are different so that the optimal mesh sizes estimated for the same voxel using four criteria become different.



(a) FPE

(b) AIC

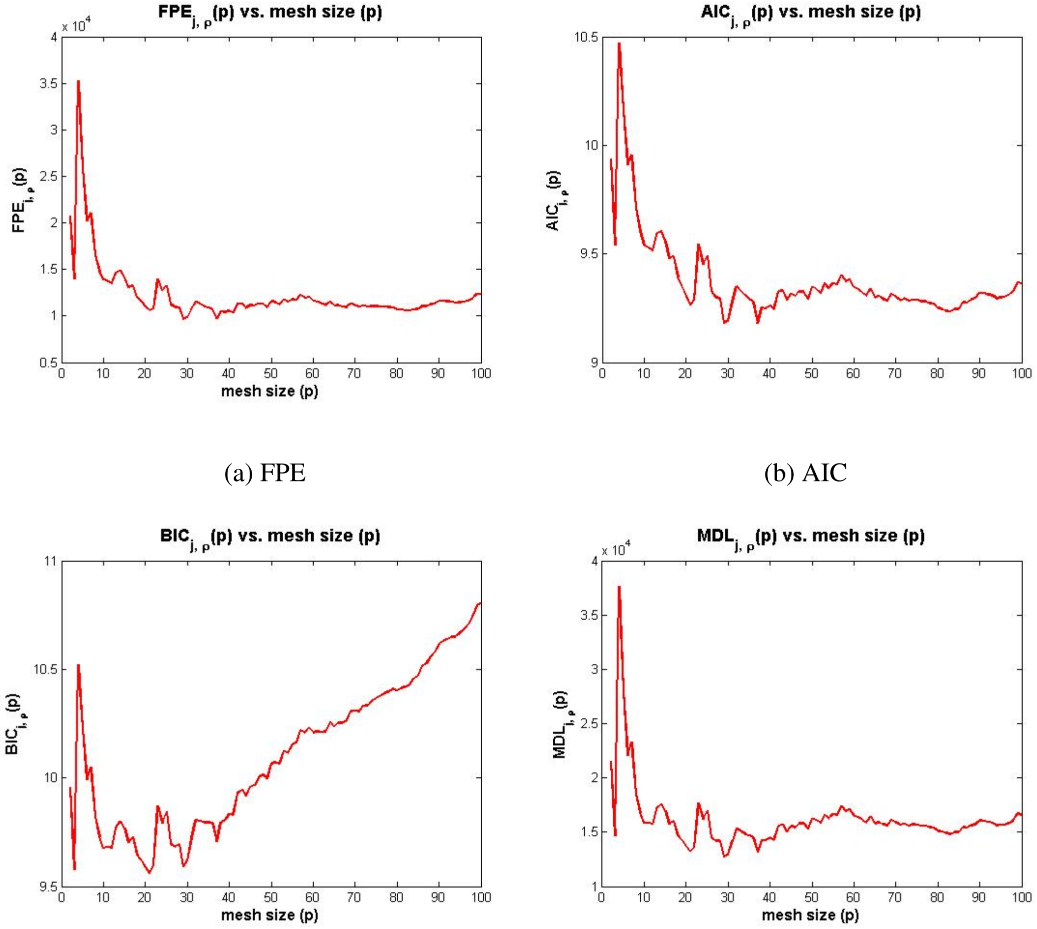
(c) BIC

(d) MDL

Figure 4.6: Histograms of optimal mesh sizes computed for each voxel of participant 8, using FPE, AIC, BIC and MDL

Figure 4.6 represents histograms of optimal mesh sizes computed for all voxels belonging to participant 8 using four different information theoretic criteria namely, FPE (top left), AIC (top right), BIC (bottom left) and MDL (bottom right). It can be seen that, for all cases, majority of the voxels form a local mesh with small number of its neighbors. On the other hand, the maximum number of optimal mesh size estimated using BIC is lower compared to other three criteria. Hence, if BIC is used, voxels tend to connect to small number of its neighbours whereas if any of other three criteria is used, voxels tend to connect to a large number of its neighbors.

Figure 4.7 reveals how these distributions and their minima differ based on the information theoretic criteria. Notice that, since FPE and AIC estimates the same optimal sizes for a sample, their optimal mesh sizes are equal ($\hat{p}_{j,\rho}^{FPE} = \hat{p}_{j,\rho}^{AIC} = 36$). However, there are slight changes between their distributions On the other hand, the optimal mesh size estimated for the same sample using BIC equals to 20 ($\hat{p}_{j,\rho}^{BIC} = 20$) and using MDL equals to 28 ($\hat{p}_{j,\rho}^{MDL} = 28$).



(a) FPE

(b) AIC

(c) BIC

(d) MDL

Figure 4.7: A Sample of $FPE_{j,\rho}(p)$, $AIC_{j,\rho}(p)$, $BIC_{j,\rho}(p)$ and $MDL_{j,\rho}(p)$ distributions for voxel at coordinates at $\bar{s}_j$ belonging to participant ρ.

Notice that, in all of the distributions, the information criteria decreases up to a minimum and then starts to increase. Yet, the minima vary based on the information theoretic criteria.

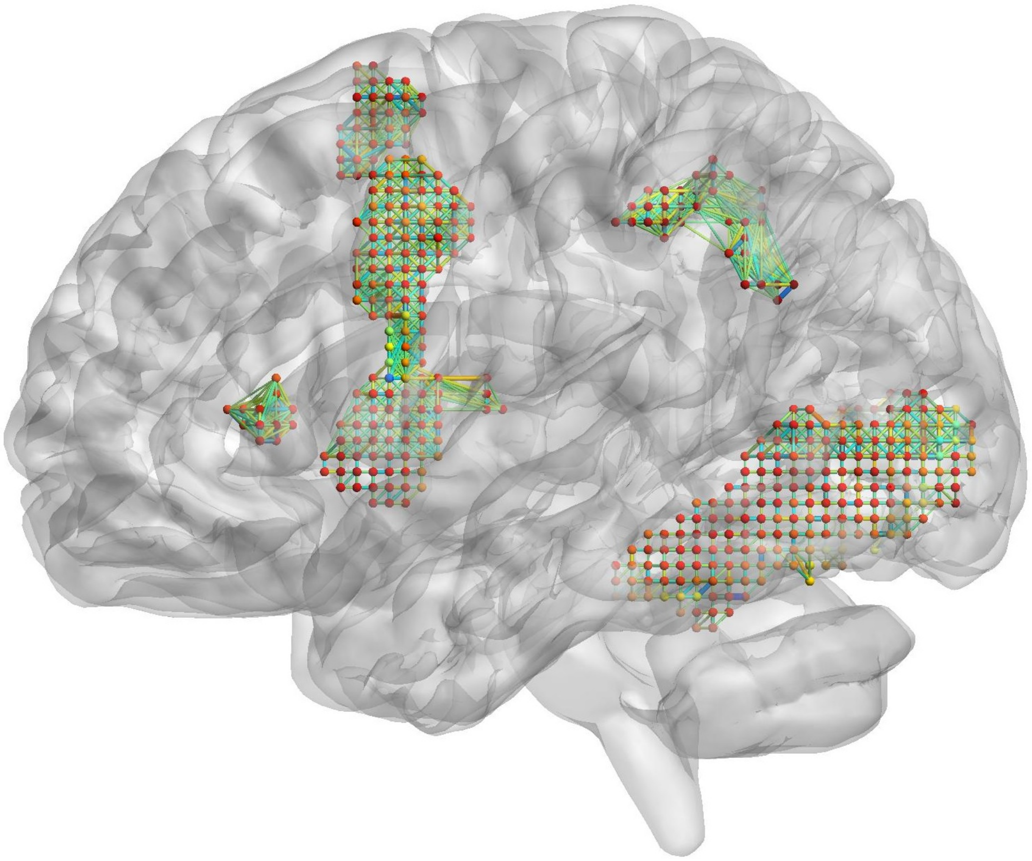


Figure 4.8: Local relational connectivity network constructed using two different optimal mesh sizes (one of them is small and the other one is large) for two example subsets of voxels.

Suppose that, in the same time instant t_i, two different optimal mesh sizes (one of them is small, the other one is large) are estimated for two different subsets of voxels. Then, the local relational connectivity network, created using single sample for a time instant would be similar to the one in Figure 4.8 which is drawn using BrainNet Viewer Toolbox [68].

Notice that, for a subset of voxels, optimal mesh size is small and for another subset, optimal mesh size is large. As it can be seen, for the same time instant, optimal mesh size of different sizes can be constructed when te optimal mesh size for each voxel is variable.

4.3 Classification Performances

Classification is computed in three ways, using optimal mesh size for a participant, sample and voxel. In all three methods k-Nearest Neighbour (k-NN) classifiers are trained with training data and the k parameter of k-NN classifiers is determined using cross-validation. After the training phase, model is tested using test data and corresponding accuracies are calculated using Equation 3.39.

4.3.1 Classification Performances using optimal mesh size for a participant

After optimal mesh size for each participant is estimated using four information criteria, FPE ($\hat{p}_{\rho}^{FPE}$), AIC ($\hat{p}_{\rho}^{AIC}$), BIC ($\hat{p}_{\rho}^{BIC}$) and MDL ($\hat{p}_{\rho}^{MDL}$), feature matrix of size $240x\hat{p}_{\rho}^{IC}.2030$ is constructed as training data and feature matrix of size $80x\hat{p}_{\rho}^{IC}.2030$ is constructed as test data (where IC is either FPE, AIC, BIC or MDL). Therefore, single k-NN classifier is trained with training data for a participant ρ. Then, the classifier is tested with test data of size $80x\hat{p}_{\rho}^{IC}.2030$ so that the corresponding accuracy is calculated using Equation 3.39.

Table 4.5 represents mesh sizes p, where $p \in [2 - 100]$, corresponding MDL values for the second participant ($MDL_2(p)$) and the corresponding classification accuracies. As it can be seen from the table, first, $MDL_2(p)$ decreases with the increase in p and then at some point it starts to increase. Notice that, $MDL_2(p)$ has its minimum value (6267) among all other values when $p = 34$. Hence, the proposed method estimates the optimal mesh size as 34 for that participant ($\rho = 2$). The third column of Table 4.5 presents classification accuracies when MDL is used. The accuracy corresponding to $p = 23$ is 67.50%, which is the highest performance among all performances calculated for $p \in [2 - 100]$. Therefore, the proposed approach selects the mesh size leading to the best accuracy as optimal for this participant.

However, the situation is not the same for all participants. In other words, the proposed method does not always estimate the optimal mesh size which leads to the best classification performance. Although this study does not claim to select the optimal mesh size giving best accuracy for each participant, it claims that using the proposed method performs better than classical MVPA methods on the average. More detailed results can be found in Table 4.6.

Table 4.6 represents classification performances where feature matrix is constructed by selecting optimal mesh size for each participant using FPE, AIC, BIC and MDL. Moreover, in the last column, performances of classical MVPA method, in which voxel intensity values are directly fed to the classifier, are presented.

As it can be seen from Table 4.6, on the average, selecting optimal mesh size for a participant using either one of the information theoretic criteria performs (3% - 4%) better than the classical MVPA method. Among them, the average performance of using BIC (60.47%) is supeior to that of other three criteria. If the results for each participant is analyzed one by one, all of these criteria perform better than classical MVPA method on 5 participants ($\rho \in \{1, 2, 4, 5, 6\}$) and has equal performance with MVPA for 1 participant ($\rho = 8$). Therefore, the results indicate that estimating optimal mesh size for a participant to form the feature matrix using either FPE, AIC, BIC or MDL is a promising method to classify cognitive states since it performs better than the classical MVPA method.

Notice that, performances of using FPE and AIC are totally equal, resulting from the fact that

Table4.5: Mesh sizes p in the interval $[2-100]$, corresponding MDL for participant 2 (ρ = 2), and corresponding accuracies (acc^{MDL})

p	$MDL_2(p)$	acc^{FPE}	p	$MDL_2(p)$	acc^{MDL}	p	$MDL_2(p)$	acc^{MDL}
2	9185	63.75	35	6427	61.25	68	7212	60.00
3	8524	63.75	36	6461	58.75	69	7238	63.75
4	7595	63.75	37	6482	66.25	70	7255	62.50
5	7519	62.50	38	6502	58.75	71	7270	62.50
6	7410	57.50	39	6520	63.75	72	7291	65.00
7	6979	65.00	40	6531	61.25	73	7321	62.50
8	6809	65.00	41	6567	63.75	74	7343	57.50
9	6608	65.00	42	6591	60.00	75	7367	62.50
10	6530	61.25	43	6610	61.25	76	7390	63.75
11	6524	62.50	44	6630	63.75	77	7413	58.75
12	6436	58.75	45	6648	63.75	78	7440	58.75
13	6390	62.50	46	6660	56.25	79	7456	58.75
14	6365	63.75	47	6664	60.00	80	7483	61.25
15	6384	61.25	48	6707	61.25	81	7516	62.50
16	6324	58.75	49	6706	61.25	82	7547	60.00
17	6365	58.75	50	6721	62.50	83	7566	60.00
18	6332	60.00	51	6764	62.50	84	7575	60.00
19	6327	57.50	52	6768	62.50	85	7595	61.25
20	6293	60.00	53	6797	67.50	86	7629	58.75
21	6311	57.50	54	6841	62.50	87	7639	61.25
22	6306	57.50	55	6872	60.00	88	7662	58.75
23	**6267**	**67.50**	56	6894	62.50	89	7691	61.25
24	6296	61.25	57	6926	58.75	90	7722	58.75
25	6315	57.50	58	6954	62.50	91	7750	63.75
26	6314	57.50	59	6989	60.00	92	7772	60.00
27	6340	57.50	60	7018	60.00	93	7794	63.75
28	6345	57.50	61	7044	61.25	94	7827	61.25
29	6350	63.75	62	7068	60.00	95	7868	60.00
30	6363	60.00	63	7084	58.75	96	7904	61.25
31	6371	58.75	64	7123	63.75	97	7932	65.00
32	6378	61.25	65	7150	62.50	98	7954	60.00
33	6398	66.25	66	7162	60.00	99	7968	62.50
34	6403	63.75	67	7194	61.25	100	7997	61.25

Table4.6: Classification performances among 8 participants where optimal mesh size is estimated for each participant

	acc^{FPE}	acc^{AIC}	acc^{BIC}	acc^{MDL}	MVPA
$\rho = 1$	65.82	65.82	65.82	65.82	58.23
$\rho = 2$	63.75	63.75	67.50	67.50	58.23
$\rho = 3$	58.23	58.23	58.23	59.49	62.03
$\rho = 4$	54.43	54.43	56.96	54.43	53.16
$\rho = 5$	59.49	59.49	59.49	59.49	54.43
$\rho = 6$	59.49	59.49	63.29	59.49	53.16
$\rho = 7$	55.00	55.00	55.00	55.00	57.50
$\rho = 8$	57.50	57.50	57.50	57.50	57.50
avg	59.21	59.21	60.47	59.84	56.78

optimal mesh sizes estimated using these two criteria are equal.

A thorough analysis reveals that, for the participants, whose estimated optimal mesh size performs worse than or equal to classical MVPA method in the classification ($\rho \in \{3, 7, 8\}$), using arc weights of local mesh model as features perform better than individual voxel intensities for some mesh size. However, information theoretic criteria was not successful to select the mesh size, which provides higher accuracy than MVPA, as optimal. As it can be seen from Table 4.7, for some mesh size p_{bet}, the accuracy of using optimal mesh size for a participant acc_{bet}, performs better than MVPA. Yet, information theoretic approach does not estimate optimal mesh size as p_{bet} so, the resulting accuracy becomes lower than MVPA.

Table4.7: Classification performances using mesh sizes providing better accuracies than MVPA and performances (with corresponding mesh sizes) using information theoretic approaches and classical MVPA method

	$\hat{p}_\rho^{FPE}$, $\hat{p}_\rho^{AIC}$	acc^{FPE}, acc^{AIC}	$\hat{p}_\rho^{BIC}$	acc^{BIC}	$\hat{p}_\rho^{MDL}$	acc^{MDL}	p_{bet}	acc_{bet}	MVPA
$\rho = 3$	39	58.23	39	58.23	23	58.23	29	59.49	62.03
$\rho = 7$	25	55.00	25	55.00	25	55.00	29	61.25	57.50
$\rho = 8$	17	57.50	17	57.50	17	57.50	20	60.00	57.50

4.3.2 Classification Performances using optimal mesh size for a sample

After optimal mesh size for each sample is estimated using four information criteria, FPE ($\hat{p}_{i,\rho}^{FPE}$), AIC ($\hat{p}_{i,\rho}^{AIC}$), BIC ($\hat{p}_{i,\rho}^{BIC}$) and MDL ($\hat{p}_{i,\rho}^{MDL}$), feature matrix of size $240x\hat{p}_{i,\rho}^{IC}.2030$ is constructed as training data for the classification of each sample. This time for each sample, different k-NN classifier is trained and only the feature vector belonging to test sample is classified using the trained classifier. Therefore, feature vector of size $1x\hat{p}_{i,\rho}^{IC}.2030$ is constructed as test data (where IC is either FPE, AIC, BIC or MDL). As a result of each step, the estimated class label of test sample is obtained and all of the estimated class labels, resulting from different classifiers, are concatenated. Finally, accuracies of using FPE, AIC, BIC and MDL in the estimation of optimal mesh size for a sample are separately are calculated using Equation 3.39.

Table4.8: Classification performances among 8 participants where optimal mesh size is estimated for each sample

	acc^{FPE}	acc^{AIC}	acc^{BIC}	acc^{MDL}	MVPA
$\rho = 1$	65.82	65.82	65.82	65.82	58.23
$\rho = 2$	62.50	62.50	62.50	66.25	58.23
$\rho = 3$	67.09	67.09	55.69	60.75	62.03
$\rho = 4$	55.69	55.69	56.96	55.69	53.16
$\rho = 5$	55.69	55.69	59.49	59.49	54.43
$\rho = 6$	62.02	62.02	63.29	62.02	53.16
$\rho = 7$	50.00	50.00	55.00	55.00	57.50
$\rho = 8$	62.50	62.50	57.50	57.50	57.50
avg	60.16	60.16	60.16	60.32	56.78

Table 4.8 represents classification performances where feature matrix is constructed by selecting optimal mesh size for each sample using FPE, AIC, BIC and MDL. Moreover, in the last column, performances of classical MVPA method, in which voxel intensity values are directly fed to the classifier, are presented.

As it can be seen from Table 4.8, on average, selecting optimal mesh size for each sample using either one of the information theoretic criteria performs (3% - 4%) better than the classical MVPA method. On average, performances of using FPE, AIC and BIC (60.16%) are equal. On the other hand, there is a slight increase in the performance on average if MDL is used (60.32%). If the results for each participant is analyzed one by one, FPE and AIC perform better than classical MVPA method on 7 participants (except for $\rho = 7$). Furthermore, using BIC or MDL gives better accuracies than classical MVPA method for 5 participants ($\rho \in \{1, 2, 4, 5, 6\}$) and has equal performance with MVPA for 1 participant ($\rho = 8$). Therefore, the results indicate that estimating optimal mesh size for a sample using either FPE, AIC, BIC or MDL and combining the results coming from separate classifiers for each sample is a promising method to classify cognitive states since it performs better than the classical MVPA method.

Notice that, performances of using FPE and AIC are also the same, resulting from the fact that even for each sample, optimal mesh sizes estimated using these two criteria are equal. The equality of intervals, mean and standard deviations of using both methods can be confirmed from Table 4.3.

4.3.3 Classification Performances using optimal mesh size for a voxel

After optimal mesh size for each voxel is estimated using four information criteria, FPE ($\hat{p}_{j,\rho}^{FPE}$), AIC ($\hat{p}_{j,\rho}^{AIC}$), BIC ($\hat{p}_{j,\rho}^{BIC}$) and MDL ($\hat{p}_{j,\rho}^{MDL}$), feature matrix of size $240x \sum_{\forall j} \hat{p}_{j,\rho}^{IC}$ is constructed as training data for the classification of each sample. Moreover, using test samples, feature matrix of size $80x \sum_{\forall j} \hat{p}_{j,\rho}^{IC}$ (where IC is either FPE, AIC, BIC or MDL) is formed so that this matrix is used to test the classifier. This procedure results in a label vector, which is used to calculate the accuracy as in Equation 3.39.

Table4.9: Classification performances among 8 participants where optimal mesh size is estimated for each voxel

	acc^{FPE}	acc^{AIC}	acc^{BIC}	acc^{MDL}	MVPA
$\rho = 1$	59.49	59.49	55.69	56.96	58.23
$\rho = 2$	58.75	57.50	62.50	63.75	58.23
$\rho = 3$	63.29	62.03	63.29	64.55	62.03
$\rho = 4$	56.96	56.96	56.96	55.69	53.16
$\rho = 5$	60.75	62.03	56.96	56.96	54.43
$\rho = 6$	60.75	62.03	51.89	53.16	53.16
$\rho = 7$	52.50	52.50	55.00	55.00	57.50
$\rho = 8$	53.75	55.00	53.75	55.00	57.50
avg	58.28	58.44	56.38	57.32	56.78

Table 4.9 represents classification performances where feature matrix is constructed by selecting optimal mesh size for each voxel using FPE, AIC, BIC and MDL. Moreover, in the last column, performances of classical MVPA method, in which voxel intensity values are directly fed to the classifier, are presented.

As it can be seen from Table 4.9, on average, selecting optimal mesh size for each sample using FPE, AIC or MDL performs (1% - 2%) better than the classical MVPA method whereas the average performance of using BIC is slightly worse than classical MVPA. Among these criteria, using AIC performs better than others (58.44%).

When the results for each participant is analyzed one by one, FPE perform better than classical MVPA method on 6 participants ($\rho = \{1, 2, 3, 4, 5, 6\}$) and has worse performance than MVPA for 2 participant ($\rho \in \{7, 8\}$). Furthermore, using AIC or MDL gives better accuracies than classical MVPA method for 4 participants and have equal performance with MVPA for 1 participant. Finally, using BIC performs better than MVPA on 4 participants, whereas performs worse than MVPA also on 4 participants. Therefore, the results indicate that estimating optimal mesh size for a voxel using either FPE, AIC or MDL is a promising method to classify cognitive states since it performs better than the classical MVPA method. However, classification using optimal mesh size for each voxel performs slightly worse than using optimal mesh size for each participant or sample.

Note that, in this case, performances of using FPE and AIC are not the same, optimal mesh sizes estimated for each sample using these two criteria are equal for all voxel. The change in intervals, mean and standard deviations of using both methods can be seen in Table 4.4.

4.4 Discussion

This chapter covers the results of various experiments performed on the proposed model. First type of experiments are participant based, in which optimal mesh size is estimated for each participant using four information theoretic criteria namely, FPE, AIC, BIC and MDL. The results indicate that, optimal mesh size differs even for the same participant based on the information criterion used. Consequently, information criterion affects the classification performance since feature matrix is constructed by employing the optimal mesh size for each participant estimated using different criteria. Estimating optimal mesh size for a participant using either one of FPE, AIC, BIC and MDL and employing it in classification performs better than the classical MVPA method. Moreover, FPE and AIC share exactly the same optimal mesh sizes and consequently give the same classification performances.

Second type of experiments are sample based, in which optimal mesh size is estimated for each sample belonging to a participant. Optimal mesh size is estimated using either one of the FPE, AIC, BIC and MDL for each sample and it is observd that, optimal mesh size differs for each sample belonging to the same participant. In other words, the sample plays an important role in the optimal mesh size and classification performance. Since the optimal mesh size differs for each sample, the size of feature vector corresponding to each sample differs. Hence, a feature matrix cannot be formed. In these experiments, different classifiers are trained (with training feature matrix of different sizes) for each sample and the optimal mesh size estimated for the sample determines the size of feature vector (and the training feature matrix). Results indicate that, using either one of the information theoretic criteria performs better than the classical MVPA method and can be successfully used to classify cognitive states. As in the participant case, optimal mesh sizes estimated for each sample are exactly the same when FPE or AIC are used.

Third type of experiments are voxel based, in which optimal mesh size is estimated for each voxel belonging to a participant. Using FPE, AIC, BIC or MDL, optimal mesh size is estimated for each voxel in a sample. Results indicate that, the optimal mesh size differs for each voxel. In other words, the local mesh formed around each voxel is different. After the training data is used to estimate optimal mesh size for each voxel, a training feature matrix is formed to train the classifier. Since each feature vector corresponding to a test sample has the same size, a test feature matrix is formed in order to test the accuracy of the classifier. Classification performances indicate that, using BIC performs worse than classical MVPA method whereas other information criteria outperform MVPA method. Furthermore, this time optimal mesh sizes estimated for each voxel are not equal for FPE and AIC. Hence, classification performances of FPE and AIC, using the optimal mesh size for each voxel, differ for each participant.

Recall that, optimal mesh size is also estimated for each class using FPE, AIC, BIC or MDL. However, optimal mesh sizes for each class are estimated only to see the effect of class on the optimal mesh size and are not used for classification purposes. It is observed that, class of a sample does not play an important role on estimating the optimal mesh size.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This chapter provides a discussion on the outcomes of proposed methods to classify cognitive states. Moreover, possible steps to be followed in the future work to reach the ultimate goal, "mind reading" are also presented.

5.1 Conclusion

In this work, an information theoretic approach to classify cognitive states, which is measured by fMRI, is presented. The proposed method employs the error term obtained from the linear regression equation in the local mesh model to select the optimal mesh size. The optimal mesh size represents the number of neighbors to which a voxel is connected. In other words estimating the optimal mesh sizes around all voxels may lead to a connectivity network in the brain. In this study, this connectivity called local relational connectivity is introduced and how it varies based on the participant, class, sample and voxel is analyzed.

This study assumes a local relational connectivity among the active voxels. In other words, whether a voxel is anatomically connected to its neighbors, where the number of neighbors equal to the optimal mesh size, is not known. Moreover, this thesis does not provide experimental evidence about the existence of this hypothetical connectivity proposed in this study. Rather the purpose of this thesis is to classify different cognitive states, specifically Item Recognition (IR) and Judgment of Recency (JOR) tasks and proposed method to classify cognitive states requires the estimation of optimal mesh size around each voxel so that the arc weights of the local mesh model are selected as features to be used for classification. The classification performances indicate that resulting features represent cognitive state better than the raw data itself.

In the proposed method, expected value of squared error term is approximated in different ways for each participant, class, sample and voxel. Then, using the error term in four information theoretic criteria namely Final Prediction Error (FPE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Rissanen's Minimum Description Length (MDL), the optimal mesh sizes are selected. Consequently, estimation of optimal mesh size also differs for participant, class, sample and voxel. If the optimal mesh size is estimated for a participant, the number of neighbors forming a mesh are equal for all voxels in the all samples belonging to the participant. The estimation of optimal mesh size for a class is similar to that of a participant but this time the class of each sample is important. Furthermore, if optimal mesh size is estimated for a sample, again all voxels in a time instant form a local mesh of same size yet in for different samples, this size varies. Finally, estimation of optimal mesh size for a voxel leads to a connectivity, in which around each voxel, a local mesh of different sizes are formed.

The results indicate that, distributions of optimal mesh sizes for each voxel varies based on the information criteria. However, most of the time two criteria proposed by Akaike, FPE and AIC, estimate the same optimal mesh size.

Resulting from the differences in optimal mesh sizes, the sizes of feature vectors corresponding to each sample also differs for participant, sample and voxel. Remember that, optimal mesh size estimated for a class is only used to analyze whether it varies based on the class and is not used in classification. When optimal mesh size is estimated for a participant or a voxel, feature vectors formed using the arc weights corresponding to each sample have the same size for each time instant. Therefore, single classifier is trained and tested with feature matrices corresponding to training and test datasets. On the other hand, when optimal mesh size is estimated for each sample, the size of feature vector varies for all time instants. Hence, each sample is tested with a different classifier.

The classification performances indicate that the proposed method is comparable and even superior to classical MVPA methods most of the time. By classical MVPA methods it is meant that voxel intensity values are directly fed to the classifier. Since the experiments are conducted on 8 participants, the results are guaranteed not te be random. When feature vectors are formed based on the optimal mesh size for each participant, using all information criteria performs $(3\% - 4\%)$ better than the classical MVPA method on average and among them BIC performs the best (60.47% on average). If optimal mesh size is estimated for each sample, all criteria perform nearly the same and (4%) better than the MVPA methods. Finally, when the optimal mesh size is estimated for each voxel and the corresponding feature vectors are employed in the classification, only using BIC performs (0.5%) worse than the MVPA method whereas other three criteria perform $(1\% - 2\%)$ better than the classical MVPA method. This time using AIC is the best among all information theoretic criteria.

Notice also that, for different participants, different type of feature vector construction perform better than MVPA. For example for participant 3 $(\rho = 3)$, when optimal mesh size is estimated for a voxel or for a sample, the performance is better than the MVPA and worse if optimal mesh size is estimated for a participant. On the other hand, a better performance than MVPA is obtained for participant 8 $(\rho = 8)$ when optimal mesh size is estimated for each sample whereas if the optimal mesh size is estimated for a voxel, the performance is worse than the MVPA.

Based upon the observations mentioned above, it can be concluded that none of the information criteria is superior to the others and none of them provides the best accuracy all the time. Moreover, it cannot be concluded that estimating optimal mesh size for either one of the participant, sample or voxel performs always better than the other two. Therefore, this thesis indicates that there is no such best criteria, or best way to estimate optimal mesh size in the classification of cognitive states. However, all the experiments reflect that using the information theoretic approach to classify cognitive states proposed in this study is always superior to classical MVPA methods.

Another finding of this study is that, the dataset is composed of samples belonging to two different categories (IR and JOR). However, although better than the performance of classical MVPA, all the performances obtained using the proposed approach in this thesis are rather low for a 2-class classification task. The main reason why the performances are low lies in the nature of experiments. In both IR and JOR tasks, the participants are shown five letters in the encoding phase and they are only cued about the task by different masks. Furthermore, in the retrieval phase in both classes two letters are presented to the participant. Therefore, the nature of the experiment is not designed for a classification task rather to understand whether specific parts of the brain are responsible for IR and JOR tasks. Hence, fMRI measurements belonging to both classes carry similar behavior, which is the main reason why all classification performances are low.

5.2 Future Work

In this thesis, a local mesh is formed around each voxel, called seed voxel, and the nearest neighbors of that voxel are selected as the ones having the smallest Euclidean distance to the seed voxel. Therefore, in the linear regression equation of local mesh model, the spatially nearest neighbors are used. Firat et. al. [19] proved that selecting functionally nearest neighbors to form a mesh and using the corresponding arc weights to classify cognitive states performs better than using local mesh model where neighbors are selected spatially. In [19], nearest neighbors of a seed voxel are selected as the ones whose correlations are higher with the seed voxel among others. Hence, the first step to follow is to use this information theoretic approach to select optimal mesh size, where functional neighborhood is defined and local mesh is formed functionally.

Moreover, in this work, the information theoretic approach is used to classify two different cognitive tasks namely item recognition (IR) and judgment of recency (JOR). In the future, the same approach may be applied to other types of experiments, which still intend to classify cognitive states, so that the proposed method would be more generalized to perform better than classical MVPA methods during the classification of all tasks.

In general use, information theoretic criteria are used to estimate the output value of a signal from its previous values. However, this approach is modified with a spatial manner to select the number of spatially nearest neighbors instead of selecting number of time points. The results indicate that the proposed approach works well on modeling the relationships among voxels and classifying cognitive states. In other areas (for example in bioinformatics), where the relationships among units (for example relationships among genes) are important and believed to carry discriminative information, the proposed method may be used for classification.

REFERENCES

[1] Activation Maps. http://www2.fmrib.ox.ac.uk/education/fmri/introduction-to-fmri/activation-maps/. Last accessed June, 2013.

[2] FMRI Functional Magnetic Resonance Imaging Lab. http://www.csulb.edu/~cwallis/482/fmri/fmri.html. Last accessed June, 2013.

[3] The Basics of MRI. http://www.cis.rit.edu/htbooks/mri/. Last accessed June, 2013.

[4] Echo-planar time course mri of cat brain oxygenation changes. *Magnetic Resonance in Medicine*, 22:159 – 166, 1992.

[5] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243 – 247, 1969.

[6] H. Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22(2):203 – 217, 1970.

[7] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267 – 281, 1973.

[8] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723, 1974.

[9] T. W. Anderson. *Determination of the Order of Dependence in Normally Distributed Time Series*. John Wiley, 2008.

[10] P. A. Bandettini, E. C. Wong, R. S. Hinks, R. S. Tikofsky, and J. S. Hyde. Time course epi of human brain function during task activation. *Magnetic Resonance in Medicine*, 1992.

[11] L. Barnett, A. B. Barnett, and A. Seth. Granger causality and transfer entropy are equivalent for gaussian variables, 2009.

[12] C. Cabral, M. Silveria, and P. Figueiredo. Decoding visual brain states from fmri using an ensemble of classifiers. *Pattern Recognition*, 45(6), 2011.

[13] J. Cao and K. Worsley. The geometry of correlation fields with an application to functional connectivity of the brain. *Annals of Applied Probability*, 9(4):1021 – 1057, 1999.

[14] M. K. Caroll. Fmri “mind readers”: Sparsity, spatial structure, and reliability, 2011.

[15] D. Cordes, V. Haughtonb, J. D. Carewc, K. Arfanakisd, and K. Maravillaa. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic Resonance Imaging*, 20(4):305 – 317, 2002.

[16] M. N. Coutanche, S. L. Thompson-Schill, and robert T. Schultz. Multi-voxel pattern analysis of fmri data predicts clinical symptom severity. *NeuroImage*, 57(1):113 – 123, 2011.

[17] R. C. Craddock, P. E. 3rd Holtzheimer, X. P. Hu, and H. S. Mayberg. Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine*, 62:1619 – 1628, 2009.

[18] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughead, R. C. Gur, and D. D. Langleben. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28(3):663 – 668, 2005.

[19] O. Firat, M. Ozay, I. Onal, İlke Oztekin, and F. T. Y. Vural. Functional mesh learning for pattern analysis of cognitive processes. In *12th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC)*, 2003.

[20] O. Firat, M. Ozay, I. Onal, I. Oztekin, and F. T. Y. Vural. Representation of cognitive processes using the minimum spanning tree of local meshes. In *Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2013.

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

[22] X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valavanis, and P. Boesiger. A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic Resonance in Medicine*, 40(2):249 – 260, 2005.

[23] 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. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 100, pages 253 – 258, 2003.

[24] P. D. Grünwald, J. I. Myung, and M. A. Pitt. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2004.

[25] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425 – 2429, 2001.

[26] J.-D. Haynes and G. Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5):686 – 691, 2005.

[27] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Natural Reviews Neuroscience*, 7:523 – 534, 2006.

[28] B. Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19(2):466 – 470, 2003.

[29] X. Hu and E. Yacoub. The story of the initial dip in fmri. *NeuroImage*, 62(2):85 – 158.17, 2012.

[30] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, Inc., 2008.

[31] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Network*, 13(4 - 5):411 – 430, 2000.

[32] Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679 – 685, 2005.

[33] V. Khim and S. Liew. On autoregressive order selection criteria, 2004.

[34] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskopf, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, R. Turner, H.-M. Cheng, T. J. Brady, and B. R. Rosen. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. In *Proc. Natl. Acad. Sci. U.S.A.*, volume 89, pages 5675 – 5679, 1992.

[35] L. J. Lanyon. *Diffusion Tensor Imaging: Structural Connectivity Insights, Limitations and Future Directions*, pages 137 – 162. 2011.

[36] P. C. Lauterbur. Image formation by induced local interactions: Examples of employing nuclear magnetic resonance. *Nature*, 242:190 – 191, 1973.

[37] K. Li, L. Guo, J. Nie, G. Li, and T. Liu. Review of methods for functional brain connectivity detection using fmri. *Computerized Medical Imaging and Graphics*, 33(2):131 – 139, 2009.

[38] T. Mitchell. *Machine Learning*. McGraw - Hill, 1997.

[39] T. M. Mitchell, R. Hutchinson, M. A. Just, R. S. Niculescu, F. Pereira, and X. Wang. Classifying instantaneous cognitive states from fmri data. In *AMIA Annual Symposium Proceedings Archive*, pages 465 – 469, 2003.

[40] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wand, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145 – 175, 2004.

[41] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *TRENDS in Cognitive Sciences*, 10(9):424 – 430, 2006.

[42] S. Ogawa and T.-M. Lee. Magnetic resonance imaging of blood vessels at high fields: In vivo and in vitro measurements and image simulation. *Magnetic Resonance in Medicine*, 1990.

[43] S. Ogawa, T.-M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. In *Proc. Natl. Acad. Sci. U.S.A.*, volume 87, pages 9868 – 9872, 1990.

[44] S. Ogawa, T.-M. Lee, A. S. Nayak, and P. Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 1990.

[45] S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. gi Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. In *Proc. Natl. Acad. Sci. U.S.A.*, volume 89, pages 5951 – 5955, 1992.

[46] I. Onal, M. Ozay, O. Firat, I. Oztekin, and F. T. Y. Vural. Analyzing the information distribution in the fmri measurements by estimating the degree of locality. In *Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2013.

[47] M. Ozay, I. Oztekin, U. Oztekin, and F. T. Y. Vural. Mesh learning for classifying cognitive processes. *Pattern Recognition*, 2011.

[48] I. Oztekin, B. McElree, B. P. Staresina, and L. Davachi. Working memory retrieval: contributions of the left prefrontal cortex, the left posterior parietal cortex, and the hippocampus. *Journal of Cognitive Neuroscience*, 21(3):581 – 593, 2009.

[49] L. Pauling and C. D. Coryell. The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. In *Proc. Natl. Acad. Sci. U.S.A.*, volume 22, pages 210 – 236, 1936.

[50] S. M. Polyn, V. S. Natu, J. D. Cohen, and K. A. Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963 – 1966, 2005.

[51] P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, and S. C. Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, 45(6):2085 – 2100, 2012.

[52] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465 – 471, 1978.

[53] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416 – 431, 1983.

[54] J. Rissanen. Universal coding, information prediction and estimation. *IEEE Transactions on Information Theory*, 30(4):629 – 636, 1984.

[55] N. Rohan and T. V. Ramanathan. A study on the focused information criterion for order selection in arma models, 2011.

[56] C. S. Roy and C. S. Sherrington. On the regulation of the blood-supply of the brain. *Journal of Physiology*, 242(1 -2):85 – 158.17, 1890.

[57] G. E. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

[58] A. Seth. Granger causality. *Scholarpedia*, 2(7):1667, 2007.

[59] H. Shen, L. Wang, Y. Liu, and D. Hu. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fmri. *NeuroImage*, 49(4):3110 – 3121, 2010.

[60] R. Shibata. Selection of the order of an autoregressive model by akaike's information criterion. *Biometrika*, 63(1):117 – 126, 1976.

[61] V. Singh, K. P. Miyapuram, and R. S. Bapi. Detection of cognitive states from fmri data using machine learning techniques. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2007.

[62] O. Sporns. Brain connectivity. *Scholarpedia*, 2(10):4695, 2007.

[63] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *Signal Processing Magazine*, 21(4):36 – 47, 2004.

[64] F. T. Sun, L. M. Miller, and M. D'Esposito. Measuring interregional functional connectivity using coherence and partial coherence analyses of fmri data. *NeuroImage*, 21(2):647 – 658, 2004.

[65] P. P. Vaidyanathan. *The Theory of Linear Prediction*. Morgan and Claypool Publishers, 2008.

[66] X. Wang, R. Hutchinson, and T. M. Mitchell. Training fmri classifiers to detect cognitive states across multiple human subjects. In *Neural Information Processing Systems (NIPS)*, 2003.

[67] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):387 – 392, 1985.

[68] M. Xia, J. Wang, and Y. He. Brainnet viewer: A network visualization tool for human brain connectomics. *PLoS ONE*, 8(7), 2013.

[69] Z. Yang, F. Fang, and X. Weng. Recent developments in multivariate pattern analysis for functional mri. *Neuroscience Bulletin*, 28(4):399 – 408, 2012.