Mining Dynamic Relationships From Spatio-temporal Datasets: An Application to Brain fMRI Data

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Gowtham Atluri

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor Of Philosophy

Advisor: Vipin Kumar
Co-advisor: Angus MacDonald III

May, 2014

© Gowtham Atluri 2014
ALL RIGHTS RESERVED

Acknowledgements

My advisor, Prof. Vipin Kumar, has been a beacon of light illuminating unlit corners of my mind. I am grateful for his contribution in creating a space and time where I could experiment with myself and with (the data from) the world. My co-advisor, Prof. Angus MacDonald III, has been very encouraging and supportive of all the questions and ideas I posed. His guidance in planning my research has helped me channel my curiosity towards academic milestones.

I am grateful for the opportunity to interact with my committee members and many faculty members who were generous to share their experience.

Graduate school has been a wonderful mix of space and time that has put me through inspiring, intimidating, challenging, and encouraging moments. I am thankful to my peers who shared this ride with me.

I am thankful to all my friends for their tremendous support through the years.

Words will not suffice to express my gratitude to my family for their contribution in shaping me. I bow down to them.

Dedication

To my parents

Abstract

Spatio-temporal datasets are being widely collected in several domains such as climate science, neuorscience, sociology, and transportation. These data sets offer tremendous opportunities to address the imminent problems facing our society such as climate change, dementia, traffic congestion, crime etc. One example of a spatio-temporal dataset that is the focus of this dissertation is Functional Magnetic Resonance Imaging (fMRI) data. fMRI captures the activity at all locations in the brain and at regular time intervals. Using this data one can investigate the processes in the brain that relate to human psychological functions such as cognition, decision making etc. or physiological functions such as sensory perception or motor skills. Above all, one can advance the diagnosis and treatment procedures for mental disorders.

The focus of this thesis is to study dynamic relationships between brain regions using fMRI data. Existing work in neuroscience has predominantly treated the relationships among brain regions as stationary. There is growing evidence in this community that the relationships between brain regions are transient. In the time series data mining community transient relationships have been studied and are shown to be useful for various tasks such as clustering and classification of time series data. In this work we focused on discovering combinations of brain regions that exhibit high similarity in the activity time series in small intervals. We proposed an efficient approach that can discover all such combinations exhaustively. We demonstrated its effectiveness on synthetic and real world data sets.

We applied our approach on fMRI data collected in different settings on different groups of people and studied the reliability and replicability of the combinations we discover. Reliability is the degree to which a combination that is discovered using fMRI scans from a population can be found again using a different set of scans on the same population. Replicability is the degree to which a combination discovered using scans from one set of subjects can be discovered again using scans from a different set of subjects. These two factors reflect the generality of the combinations we discover. Our results suggest that the combinations we discover are indeed reliable and replicable. This indicates the validity of the combinations and they suggest that the underlying neuronal principles drive these combinations. We also investigated the utility of the combinations in studying differences between healthy and schizophrenia subjects.

Existing work in estimating transient relationships among time series typically uses sliding time windows of a fixed length that are shifted from one end to the other using a fixed step size. This approach does not directly identify the intervals in which a pair of time series exhibit similarity. We proposed another computational approach to discover the time intervals where a given pair of time series are highly similar. We showed that our approach is efficient using synthetic datasets. We demonstrated the effectiveness of our approach on a synthetic dataset. Using this approach we provided a characterization of the transient nature of a relationship between time series and showed its utility in identifying task related transient connectivity in fMRI data that is collected while a subject is resting and while involved in a task.

In summary, the computational approaches proposed in this thesis advance the state-of-the-art in time series data mining. Whereas the extensive evaluations that are performed on multiple fMRI datasets demonstrate the validity of the findings and provide novel hypothesis that can be systematically studied to advance the state-of-the-art in neuroscience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Spatio-temporal data

Approaches for discovering useful information from data that is measured from real world objects are being studied in the data mining community. The nature of datasets in question can have very different properties and so approaches have to be designed to handle such properties. Nature of the data can be as simple as relational data, i.e., a set of attributes measured from an object. One example of relational data is electronic Health Record (EHR) data where subject's name, demographics, and clinical variables are stored. A relatively more complex form of the data is a time series data where measurements from an object at regular time intervals are collected. An example of time series data is blood sugar level and body temperature of a subject that is collected every day. A relatively more complex form of data is one where measurements have both spatial and temporal information associated with them. For example, in remote sensing data periodic measurements of vegetation at each location on the surface of the earth are measured. Note that in addition to the measurement of an attribute (vegetation), location and time of measurement are also available.

Several decades of work has been done on analyzing relational data, the simplest form of data, where canonical problems like classifying objects into classes, grouping them into clusters, identifying anomalies and discovering patterns are studied [5]. In the last decade, time series data analysis techniques have been developed [6]. They include approaches to discover

similarities between time series, to group time series that are similar, and to predict the next measurement based on previous measurements.

In the last five years there is increased interest in analyzing spatio-temporal datasets that are collected in several domains such as climate science [7], neuorscience [8], sociology, and transportation [9, 10]. In climate science, data is collected using remote sensing satellites. In neuroscience, magnetic resonance imaging tools are used to scan the brain. In transportation, sensors are placed on several streets to capture the amount of traffic and related features. Data collected in each of these domains have different properties. For example, in climate and neuroscience, data is collected from all the locations at a certain resolution whereas in transportation the sensors are placed at a selected intersections or locations on the freeway. Information mined from such datasets can potentially lead to effective solutions to many of the society's problems such as climate change, brain disorders, and traffic. Effective data mining approaches are needed to be able to analyze the increasingly available spatio-temporal data sets.

1.1.2 Opportunities in mining fMRI data

Functional Magnetic Resonance Imaging (fMRI) data indirectly measures the activity at each gray matter location in the brain at regular intervals [11]. There are typically hundreds of thousands of locations from which activity is measured, at every two second interval during a typical 6 minute scan. Locations that exhibit high activity are expected to share similar or at least related functionality.

The increasingly available fMRI data can be mined to answer a number of questions: i) How does the brain's functional network adapt while a skill is being learned ? ii) How does the brain's functional network change while the subject is working on a task from when he is not working on a task ? iii) How does the brain's functional network differ in healthy and disease subjects ? iv) Can we characterize the brain's functional state and it's transition to another state based on its functional network ?

In the last 20 years several efforts have been made to analyze fMRI data with a goal of advancing the state-of-the-art in our understanding of the brain functionality. Until early 2000s, several hypothesis that related one brain region's activity to a specific task were studied. Recently, more sophisticated hypothesis that interactions between multiple brain regions are related to a specific task are being studied. Very recently, the dynamics in the interactions of brain

regions is being studied. Despite decades of work in this area, there is a prevailing dissatisfaction in the neuroimaging community with the progress that is made in elucidating how the brain accomplishes its functions or how the brain's dysfunction can be directly identified from the fMRI Data [12].

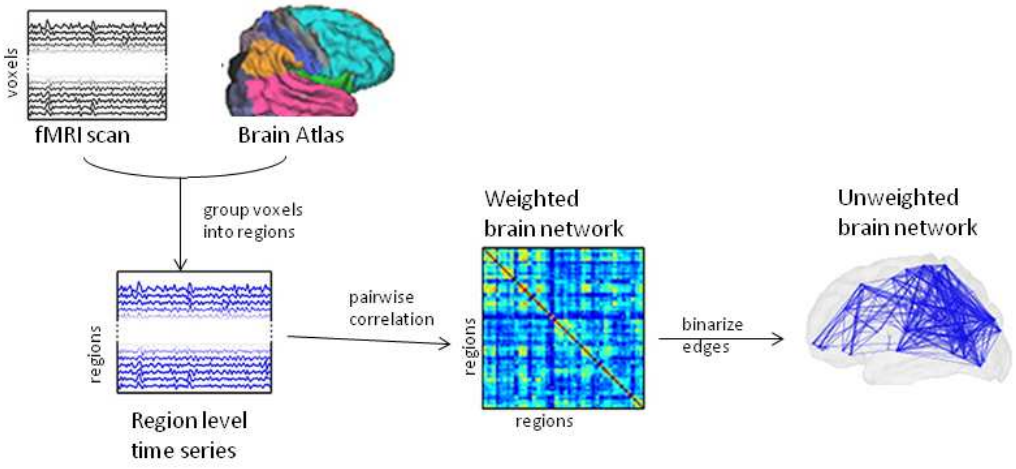

Figure 1.1: Constructing brain networks from fMRI data

One of the popular approaches to analyze fMRI data is to construct brain networks as shown in Figure 1.1 [13]. First, a brain atlas that groups contiguous sets of locations (also known as voxels) into 'brain regions' is chosen and a mean time series of voxels for each brain region is computed. Adjacent voxels are known to exhibit highly similar time series and this step of determining region level time series as a mean of its constituent voxel time series reduces the redundancy. A pairwise correlations between all brain regions is then computed. This can be represented as a matrix where rows and columns are brain regions and the elements are the pairwise correlation values. This matrix can also be treated as a weighted brain network where brain regions are nodes and the correlation values indicate the strength of the connectivity. One can also choose a correlation threshold to binarize the pairwise correlation to get a binary matrix that can be treated as a binary brain network. The weighted and the binary networks have been traditionally analyzed by computing graph theoretic properties to learn the principles of the brain network. The fundamental idea is that by learning the principles of the brain network one can arrive at the operating principles of the brain [14].

Recently it has been noticed that the functional connectivity computed from two different scans obtained from the same subject can be very different [15]. This raised questions about the reliability of the published findings. In many of the published studies the objective was to discover biomarkers that can explain differences between groups of healthy and a group of disease subjects [16]. In these studies the hope is that the findings in one study can be replicated on a different set of subjects in a different study. However, if the functional network computed for the same subject is changing depending on the scan, how can one expect the biomarkers to reliable [17]. Hence, the reported biomarkers are not of clinical value.

While there is increased interest in neuroscience community in quantifying the reliability of the findings, other studies have pointed out that the functional connectivity is transient even with the duration of one scan [18]. This raised more questions about the validity of earlier studies that assumed that the connections are stationary. A large number of studies have evaluated the properties of brain networks and their role in various mental disorders with the assimption that the connectivity is stationary. The reliability of the reported findings in these studies is unclear. Moreover, approaches to study and characterize the transient nature of connectivity are highly desired.

1.2 Thesis Contributions

The focus of this thesis is to study dynamic relationships between brain regions using fMRI data. Existing work in neuroscience has predominantly treated the relationships among brain regions as stationary. There is growing evidence in this community that the relationships between brain regions are transient. In the time series data mining community transient relationships have been studied and are shown to be useful for various tasks such as clustering and classification of time series data. In this work we focused on discovering combinations of brain regions that exhibit high similarity in the activity time series in small intervals. We proposed an efficient approach that can discover all such combinations exhaustively. We demonstrated its effectiveness on synthetic and real world data sets.

We applied our approach on fMRI data collected in different settings on different groups of people and studied the reliability and replicability of the combinations we discover. Reliability is the degree to which a combination that is discovered using fMRI scans from a population can be found again using a different set of scans on the same population. Replicability is the degree

to which a combination discovered using scans from one set of subjects can be discovered again using scans from a different set of subjects. These two factors reflect the generality of the combinations we discover. Our results suggest that the combinations we discover are indeed reliable and replicable. This indicates the validity of the combinations and they suggest that the underlying neuronal principles drive these combinations. We also investigated the utility of the combinations in studying differences between healthy and schizophrenia subjects.

Existing work in estimating transient relationships among time series typically uses sliding time windows of a fixed length that are shifted from one end to the other using a fixed step size. This approach does not directly identify the intervals in which a pair of time series exhibit similarity. We proposed another computational approach to discover the time intervals where a given pair of time series are highly similar. We showed that our approach is efficient using synthetic datasets. We demonstrated the effectiveness of our approach on a synthetic dataset. Using this approach we provided a characterization of the transient nature of a relationship between time series and showed its utility in identifying task related transient connectivity in fMRI data that is collected while a subject is resting and while involved in a task.

In summary, the computational approaches proposed in this thesis advance the state-of-the-art in time series data mining. Whereas the extensive evaluations that are performed on multiple fMRI datasets demonstrate the validity of the findings and provide novel hypothesis that can be systematically studied to advance the state-of-the-art in neuroscience.

1.3 Thesis Overview

The organization of this thesis is as follows. A broad overview of spatio-temporal data characteristics and data mining problems of interest are presented in Chapter 2. Overview of fMRI data and the specific data mining questions pertaining to the brain that can be investigated are discussed in Chapter 3. In Chapter 4 we present our approach for discovering transient relationships between brain regions. In Chapter 5 we present the application of our approach in Chapter 4 and show that the discovered relationships are reliable and replicable. We present our approach for discovering the time intervals directly from the data in Chapter 6. We conclude in Chapter 7 with a discussion on future work.

Chapter 2

Overview of Spatio-temporal Data Mining Problems

2.1 Introduction

Traditionally data mining community has analyzed relational datasets where different features from objects of interest are collected. For example, the popular Iris dataset from UCI datasets [19] has measurements *sepal length*, *sepal width*, *petal length* and *petal width* from flowers of three different types of plants. Using such datasets, several canonical data mining problems such as classification, clustering, anomaly detection and pattern mining have been thoroughly studied [5]. Classification [20] deals with the problem of discovering a group to which a new instance (set of measurements) belongs to. Clustering [20] deals with the problem grouping instances based on their similarity in measurements. The goal of anomaly detection is to find instances that are unusually dissimilar to other instances in the dataset [21]. Frequently occurring trends in the data are captured by pattern mining [22].

In the last decade, there is increased interest in analyzing data where repeated measurements are collected from the object of interest at regular time intervals. Such data is referred to as time series data. For example, blood sugar level of a patient collected on a weekly basis can be treated as time series data. Time series data has become increasingly ubiquitous in several domains including climate, medical records, bioinformatics, and social media [23, 6]. Data mining community has studied several problems pertaining to analyzing time series data [24, 6]. They include clustering [25, 26], classification [27], anomaly detection [28], forecasting [29],

and segmentation [30]. The problem of clustering is to group a given set of time series into clusters such that any pair of time series' within a cluster are highly similar than those that are between clusters. Given a set of time series and their labels, the problem of classification is to learn the characteristics of the time series' assigned to each label, so a potentially correct prediction can be made when an unseen time series' is presented. The problem of anomaly detection in a given time series is to identify the time point at which an unexpected behavior is exhibited. Given a time series of t time points, the problem of forecasting is to predict the behavior at time $t+1$. Segmentation deals with the goal of finding piece-wise segments of time points in a given time series, within which the behavior is homogeneous.

On the other hand, the data mining community has also studied data that has spatial attributes. Several events such as accidents, tornadoes or burglaries have spatial locations associated with them. To explore the patterns underlying these events several spatial clustering approaches have been studied in the data mining community [31, 32, 33].

In the recent years, several spatio-temporal datasets have become increasingly available. For example, remote sensing satellites collect data such as temperature, pressure, sea surface height from the planet every day. These measurements have space and time associated with them. Similarly, brain imaging technologies measure activity at a given location and time. Similarly events such as crimes have also space and time associated with them. Air and transportation networks also have spatial locations and the time associated with events such as arrival or departure of a flight or a vehicle. Data mining techniques to analyze such datasets can help address many of the society's problems from addressing climate change to discovering effective treatment strategies to mental disorders. In this chapter we outline the several problems that can be defined in spatio-temporal datasets and the challenges that need to be addressed to solve these problems.

2.2 Data

Data collected from any system where every measurement has a location associated with it is referred to as a spatial dataset. Note that this measurement is made only once. Amyloid PET scans measure the amount of Amyloid protein deposition at various locations in the human brain and the presence of this protein is indicative of Alzheimer's disease [34]. This dataset has only spatial context, i.e., locations in the brain.

On the other hand, data collected repeatedly from a system is referred to as a temporal dataset. Note that this measurement has no spatial attributes. Measurements recorded on a valve on a space shuttle, weekly power usage at a research plant are examples of temporal datasets.

There are many scenarios where measurements are associated with both space and time. Such datasets are referred to as spatio-temporal datasets. For example, climate variables such as temperature, humidity, pressure etc measured at various locations on the Earth's surface and at different time points every day. Other examples, include EEG and fMRI data that measure brain's activity at several locations in the brain and at different time points.

Spatio-temporal datasets can be further classified based on the uniformity in the measurement in spatial and temporal attributes:

- Uniform spatial and Uniform temporal
- Uniform spatial and Non-uniform temporal
- Non-uniform spatial and Uniform temporal
- Non-uniform spatial and Non-Uniform temporal

We refer to a dataset as Uniform spatial if the measurements are made at spatial locations that are equally spaced covering the entire space. Non-uniform spatial datasets are those in which the measurements are made at spatial locations that are do not cover entire space and the covered locations are not necessarily equally spaced. Similarly, Uniform temporal datasets are those where measurements are collected at regular intervals and Non-uniform temporal datasets are those where measurements are collected at irregular intervals.

Examples of uniform spatial and uniform temporal datasets include, fMRI data where brain's activity is measured at every cubic location and at a particular frequency, environmental variables such as vegetation index that is collected at every spatial location on Earth at a particular resolution.

Examples of uniform spatial and non-uniform temporal include Positron Emission Tomography (PET) scan obtained from Alzheimer's subjects on every visit to the hospital. PET scans are used to assess the loss of gray matter tissue in a subject. The scan collects information at all locations in the brain (uniform), however the scan is obtained only when the patient visits the hospital (non-uniform).

Examples of Non-uniform spatial and uniform temporal include EEG datasets where the measurements at regular intervals are obtained only from specific locations that are not necessarily equally spaced or that covers the entire brain space.

Examples of Non-uniform spatial and non-uniform temporal include crime data in a city, where reported crimes are attributed to spatial locations and time of event. Note that the crimes do not occur at every location and at every regular interval.

Data related challenges: The underlying nature of spatio-temporal datasets poses numerous challenges irrespective of the problem that is studied.

- **Scale** The spatial resolutions can be high enough to result in millions of spatial locations and a temporal resolution can be high enough to result in thousands of time points. This scale can potentially introduce computational challenges in most problems.

- **Noise** As with any real world datasets noise and missing values are inherent in spatio-temporal datasets and approaches have to be robust to be able to handle this challenge.

- **Autocorrelation** Adjacent spatial locations typically share similar measurements at a given time point due to spatial proximity and adjacent time points will share similar measurements at a given location. This will often introduce redundant information in to the analysis and approaches have to be cognizant of this issue.

- **Notion of similarity** Many similarity measures are considered for studying similarity between two time series. Some of these similarities work consider the relationship between two time series in lock-step fashion, where a value of time series at one time point is compared to that of a different time series at the same time point. Measures that allow for lagged similarity or warped similarity are also popular in time series data mining literature. One has to be aware of the type of similarity that is relevant in a given dataset.

- **Uniformity vs. Non-uniformity** As discussed above some datasets have uniformity in space and time and others have uniformity in either of them or none. Approaches designed for one combination will not necessarily be suitable for the other combinations and so this issue has to be taken into consideration while designing algorithms to address data analysis problems.

2.3 Problems

In this section we describe the several research questions that can be posed in the context of spatio-temporal datasets, discuss the related work, and point out the challenges that need to be addressed.

2.3.1 Generalization of canonical data mining problems for spatio-temporal data

Classification, clustering, and frequent pattern mining are treated as canonical data mining problems due to the depth of understanding the community has developed in each of these problems. These problems are typically defined on relational datasets where each instance has a multivariate features space. In the case of spatio-temporal datasets, each instance can be treated as a collection of attributes that have space and time associated with it.

In the context of classification, the goal can be defined as classifying an instance i.e., a spatio-temporal measurement into one class or the other. In neuroscience, one is interested in determining if a subject is schizophrenic or not. This can be achieved by classifying an fMRI scans collected from a subject as healthy or schizophrenic.

In the context of clustering, the goal can be defined on spatio-temporal instances where the objective is to group instances that are highly similar to each other such that instances in different groups are not similar. Given a set of climate (spatio-temporal) instances generated from different models, one may be interested in grouping these models based on the similarities among spatio-temporal instances.

In the context of frequent pattern mining, instances can be treated as items and spatio-temporal attributes can be treated as transaction, and the objective can be defined as one of finding groups of instances such that they share similarities in sufficiently many spatio-temporal attributes.

2.3.2 Anomaly detection in spatio-temporal data

A time series can be labeled as anomalous if it is characteristically different from an expected behavior determined using a set of relevant time series. Anomaly detection can be useful in identifying rare events like intrusion in computer networks, malfunction in equipment etc. Current approaches to anomaly detection in time series data typically estimate the deviation of a time series by expected behavior using any of the various time series similarity measures [35].

The deviation can be computed as the i) distance based on the entire time series or ii) as the average distance of the windows of a chosen length or iii) as the average distance at each time point. The deviation in all these cases can be computed using either a k^{th} nearest neighbor type approach or a modeling based approach that predicts expected behavior [35].

Anomaly detection is also relevant in spatio-temporal datasets where anomalous behavior can be defined in the context of it's spatial neighborhood. For example, time series measured from a cubic location in the brain can be significantly different from a majority of the time series in the neighborhood. This could happen due to motion related artifacts during scanning or due to caveats in the scanning protocols or defective equipment. Anomaly detection in such scenarios could shed light on sources of noise and will be potentially useful in ensuring data quality and informing preprocessing procedures to correct for artifacts.

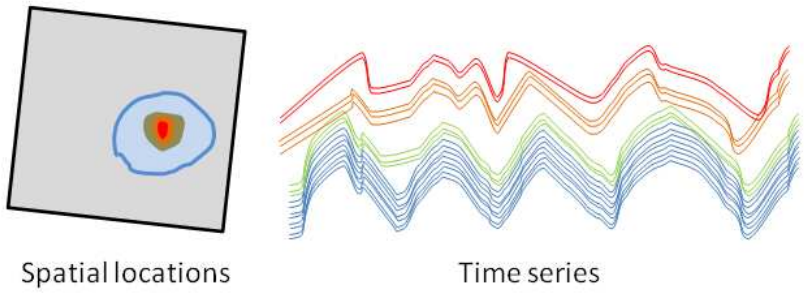

Figure 2.1: Anomaly detection problem

Consider the example shown in Figure 2.1 where anomalous behavior is localized in the red region. Notice that due to spatial autocorrelation the green region around the red region shares some anomalous properties and some properties of the blue region. Discovering the red region that has distinct time series compared to the blue neighborhood as the region of anomalous behavior is the objective. Traditional time series based anomaly detection techniques will not be able to use the spatial structure in the data as they have been designed to work with independent set of time series. Several challenges needs to be addressed to extend existing approaches for spatio-temporal datasets. First, the anomaly will be present in a local neighborhood and is expected to be different from its immediate neighborhood. The extent of the spatial of the anomaly and the relevant neighborhood from which the expected neighborhood needs to be inferred from the data. Second, unlike the existing time series approaches that determine if a given time series is anomalous, spatio-temporal anomalous detection techniques need to identify

the set of locations that have anomalous time series from the entire dataset. This requires exhaustive search through the entire space.

Another type of anomaly detection problem relevant in spatio-temporal datasets is one where multiple spatio-temporal datasets are available and one needs to identify a dataset that is anomalous. This is often relevant in fMRI time series data analysis where fMRI scans are obtained from several subjects and due to errors in imaging protocols some scans are markedly different from others.

2.3.3 Change detection in spatio-temporal data

Given a time series, the goal of identifying the time at which there is change in the nature of the time series is in general referred to as the time series change detection problem. In time series data analysis literature any of the following goals are treated as 'change detection' problems: i) In a given time series, determine at what time point its characteristics change. ii) Segment a given time series such that the time series is homogeneous within each segment. iii) Discover the top-k most unusual subsequences in a given time series

The notion of 'change' is generally defined based on the nature of the time series in the initial part of the time series. Several approaches have been proposed in the last two decades to address the above variants of the change detection problem. A common theme shared by these approaches is to estimate an expected behavior at a time point based on the time series available until the current time point and then compute the deviation. These approaches include:

- Statistical parameter based approaches that estimate a distribution for the time series and a hypothesis test is performed to determine if a change point exists.
- Segmentation approaches split a given time series into segments until a user provided segment number is reached. Often a model (linear, polynomial etc.) is fit to estimate the homogeneity of a segment.
- Predictive approaches infer a model based on the time series observed and predict an expected value for the next time point. Deviations from the expected value are treated as change points.

Consider the example shown in Figure 2.2 where the time series from the red region changes its behavior in the middle. Notice that due to spatial autocorrelation the green region around the

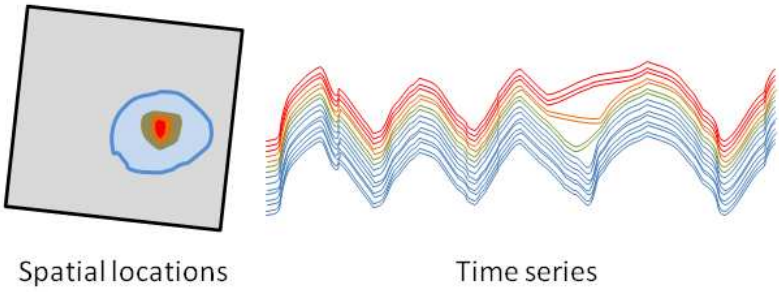

Figure 2.2: Change detection problem

red region shares this change. Discovering the red region and the time point in its time series is the objective of the change detection problem. Change detection in spatio-temporal datasets share similar properties with that of a time series based change detection problem, except that this change can also seen in the immediate spatial neighborhood. The visibility of a change point in time series obtained from adjacent locations allows one to discover change points with high statistical power. Change detection approaches can also leverage this information in order to improve the performance of traditional change detection approaches. The challenges that are relevant in spatio-temporal scenarios are: i) Existing approaches treat each time series independently. In spatio-temporal datasets they need to be cognizant of their spatial neighborhood. ii) As in the case of anomaly detection, here the goal is to discover sets of locations where change points exist. Therefore, a global search of change points is needed. iii) Different changes can potentially be visible to different extents in the neighborhood. Approaches have to be robust to these variations.

Another notion of change detection that is also partly related to anomaly detection discussed above is to identify change in the trend for time series at a given location with respect to its neighborhood. A recent work by Chen *et al.* [36] extends traditional change detection to identifying change detection with respect to a context that is also provided in addition to the time series in which change detection needs to be done. This can be further generalized to spatial context where a spatial context can be considered. Note that the difference between this problem and anomaly detection is that here the time series is largely similar to the context or the spatial neighborhood.

2.3.4 Spatial clustering in spatio-temporal datasets

The problem of spatial clustering [37, 38, 39] has been earlier studied in data mining community where objects are grouped based on their spatial proximity.

In time series data mining literature clustering problem for independent set of time series was studied. Major contributions towards this problem has been in designing computational approaches to assess similarity. These similarity measures were used in conjunction with the state-of-the-art clustering techniques to discover the time series clusters.

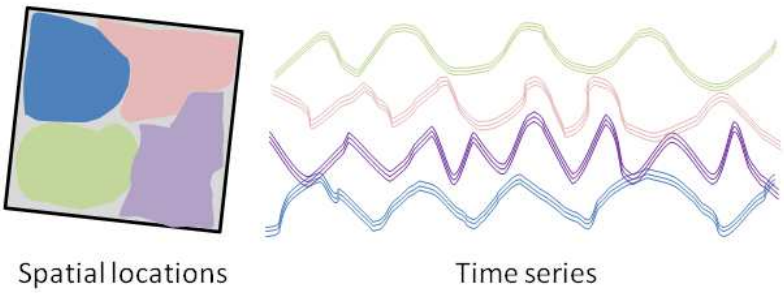

Figure 2.3: Spatial clustering problem

The goal of spatial clustering in spatio-temporal datasets is to group contiguous sets of voxels that share similar temporal behavior into clusters. Consider the example shown in Figure 2.3 where there are four different regions: green, pink, purple, and blue. The time series from there regions appear to be different from each other, but the time series within each of these regions are highly similar.

In the context of time series clustering, when additional information is available in the form of spatial structure it can be used to not only improve the performance of these clustering approaches, but also to improve the interpretability of the clusters discovered. This problem is particularly relevant in the context of fMRI datasets where the goal was to determine a data-driven brain atlas i.e., identify contiguous sets of locations as brain regions where time series from each location within a region are highly similar. This can also be relevant in other climate data sets to identify contiguous groups of locations that share similar climatic conditions.

This problem requires taking into account the spatial proximity and the temporal similarity. A number of questions need to be studied to address this problem. First, how to incorporate spatial structure into the clustering process ? Second, how much importance needs to be placed

on the spatial proximity and temporal similarity ? Third, how to handle spatial autocorrelation due to which nearby locations in adjacent clusters tend to exhibit high similarity ?

2.3.5 Time point clustering in spatio-temporal datasets

The problem of time-point clustering is one where the goal is to discover groups of time points in which the spatial data has a similar structure. This problem is particularly relevant in fMRI datasets where the activity at every location in the brain is measured at every time point and the time points can be grouped into clusters such that within a cluster the brain's activity is highly similar. This problem can also be referred to as 'discovering brain states'.

In the case time point clustering one could potentially use the information from all spatial locations or use only a relevant part of them to group time points.

Using full spatial data

Consider the example shown in Figure 2.4 where the activity measured in space is shown in four different colors at 5 different time points $\{t_1, t_2, t_3, t_4, t_5\}$. Notice that the activity profiles of t_1, t_2, and t_5 are very similar and that of t_2 and t_4 are also very similar. Discovering these two groups of time points is the objective of the time point clustering (using full space information).

When the entire spatial data can be used from each time point for clustering, this is similar to traditional clustering schemes except that there is a spatial structure among 'attributes' and adjacent attributes take highly similar values. Traditional similarity measures that are designed for independent set of attributes may not be effective in this case. Spatial maps from two time points could match well with some minor adjustments that are similar in flavor to dynamic time warping where temporal similarity is assessed allowing for some gaps. Another issue that these approaches have to be cognizant about is the temporal autorcorrelation in the data and hence similarity between time points that are non-contiguous is more interesting than similarity between contiguous time points.

Using a subset of spatial data

In many cases the measurements from a subset of spatial locations may be relevant in grouping time points. For example, only a small subset of brain locations are known to be active or relevant during most cognitive tasks and so considering all the locations may introduce more

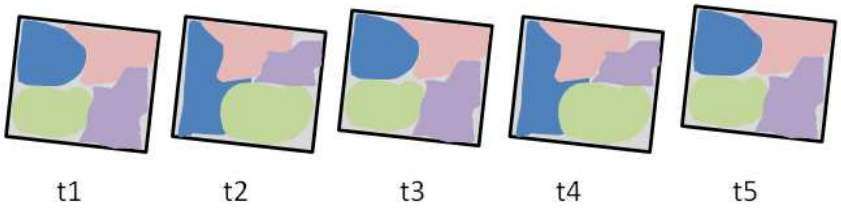

Figure 2.4: Time point clustering problem with full ‘space’ data

noise than signal. Consider the example shown in Figure 2.5 where activity measured in space is shown in two different colors at 5 different time points $\{t_1, t_2, t_3, t_4, t_5\}$, where the white and black pattern in the background is noise. Notice that when the colored groups of locations are considered, the activity profiles of t_1, t_2, and t_5 are very similar and that of t_2 and t_4 are also very similar. Discovering groups such as these where the similarity is shared only in a sub-‘space’ is the objective here. Challenges discussed in the case of using full spatial data are also relevant here. Approaches to use only a subset of spatial data for time point clustering have to deal with additional challenges such as: i) identifying the subset of the spatial locations that may be relevant ii) choosing appropriate similarity measures to assess similarity between spatial locations.

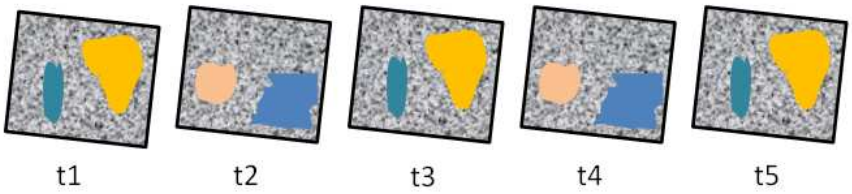

Figure 2.5: Time point clustering problem with sub ‘space’ data

2.3.6 Dynamics in contours of active/interesting locations

A spatial location can be treated as ‘active’ when a measurement at a given time point is greater than a threshold. Due to spatial auto-correlation adjacent locations for a given location tend to be active simultaneously. The spread of this activity in the neighborhood can be termed as an ‘active region’. An active region can arise a time point and it can grow, shrink or shift spatially

and finally vanish as time progresses. In several disciplines identifying dynamics in the contours of active or 'interesting regions' is of interest. Such problems include identifying the dynamics in contours of lakes over weeks that will allow one to track the status of a lake at any given time, tracking eddies in oceans to discover the progress of an eddy over its lifetime, tracking dynamic activity in brain regions to identify the span of activity in space and time.

Consider the example shown in Figure 2.6 where the active regions are shown in colors and the background noise is shown as a gray noisy pattern. Here examples of the nature of change in an active region over time is shown. These change include: growing, shrinking, merging, splitting, moving, appearing and disappearing. Tracking these changes over time is the objective of this problem. This problem involves two parts: i) identifying contiguous locations that are 'interesting' from a domain perspective. ii) tracking change over time.

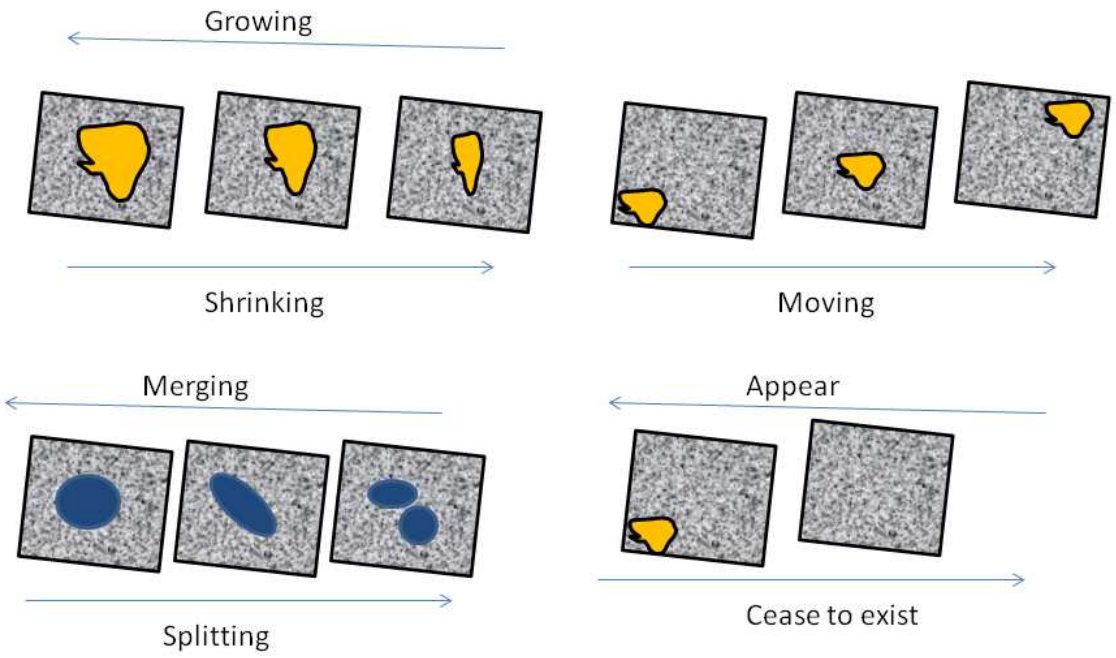

Figure 2.6: Contour dynamics in spatio-temporal data

Challenges that are relevant to this problem are: i) Determining the nature change occurring to an active region in successive time points. An active region can grow, shrink, split into multiple active regions, dislocate or vanish altogether with time. ii) Multiple active regions that are potentially present in each time point will introduce additional complexity iii) Additional

challenges arise in determining an active region from a given spatio-temporal dataset with real valued measurements. Although one could potentially use a threshold to determine locations whose measurements are significantly high at every time point, real valued data can provide additional information that can be useful in assessing the core of an active region or compare the measurements across time points to increase the performance in tracking.

This problem is similar to multi-target tracking in computer vision and surveillance [40, 41, 42]. Recently, Faghmous *et al.* [43] proposed an approach that can discover and track eddies in sea surface height datasets in the presence of noise and missing measurements.

2.3.7 Relationship between distant spatial locations

Distant locations can have highly similar time series. Identifying distant sets of contiguous locations that exhibit similarity in their time series is of interest in fMRI and climate data analysis. Such long distance relationships can capture novel relationships among distinct parts of the system and can potentially shed light on the governing relationships of the system. Long distance negative correlations in climate variables are referred to as dipoles and they are known to be useful in climate predictions. Long distance positive correlations in fMRI dataset are known to reflect synergies between distinct brain regions that are known to perform specific functions.

These long distance relationships can be estimated using either the full time series or may only be seen in smaller intervals.

Using full temporal data

Consider the example, shown in Figure 2.7 that shows time series from two distant regions colored in blue and orange. The time series are not only similar within each region, but are also similar between the two distant regions. Discovering such pairs of regions (containing contiguous voxels) is the objective of this problem.

This problem can be broken down into two subproblems: i) Discovering sets of contiguous locations that are highly similar within themselves ii) Estimating similarities between time series that belong to these sets that are distantly located. It is important to note that breaking down the problem in this fashion will not necessarily enable one to discover all interesting long distance relationships in the data, because the first step can have a significant impact on the relationships that can be discovered in the second step. Hence both these steps needs to be

performed simultaneously.

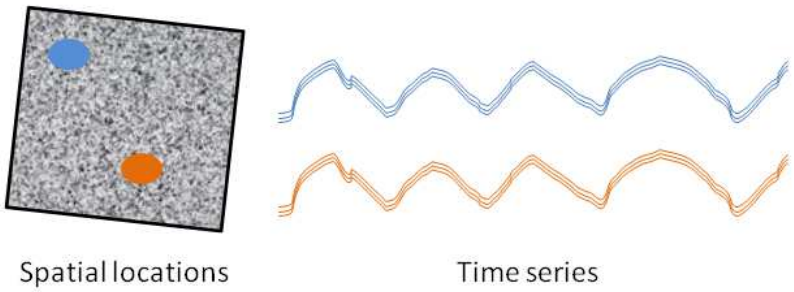

Figure 2.7: Long distance relationships using full time series

This problem is very similar to the biclustering type problems where relationships between two different sets of objects needs to be discovered. The goal of biclustering problem is to discover a subset of genes and related conditions in which the genes are highly expressed. In our case both genes and conditions are locations and 'expression' is analogous to a 'interesting relationship' between their time series. One can potentially compute pairwise relationship (e.g., similarity based on correlation or euclidean distance) between all locations and then use standard biclustering approaches to discover long distance relationships. Such an approach will have several limitations: i) It does not take into account the spatial relationship that exists between different time series ii) Biclustering approaches will not take into account spatial autocorrelation that exists among adjacent locations. iii) Biclustering is an NP-Hard problem and so the scale of the resultant similarity matrix cannot be handled by existing approaches.

Using subset of time

In many domains such as neuroscience and climate long distance relationships are known to exist in only small intervals in time. Consider the example shown in Figure 2.8 where the time series from two distant regions colored in blue and orange are shown. Notice that the time series within each region are highly similar, but across the two regions they are similar only in the indicated interval. Discovering such pairs of regions where they share similarity in only small intervals is the objective of this problem.

This is even more challenging than the above problem where the full time series is considered as the interval of interest needs to be discovered. One has to address the above challenges

as well as additional challenges such as: i) Discovering relevant intervals where long distance relationships are present iii) A region can share a similarity with two other regions in two different time intervals. Approaches have to be cognizant of issues such as this. iii) Compared to a scenario where a relationship is expected to be found in an interval in lock-step fashion, the scenario where lagged relationships may exist introduces additional computational overheads.

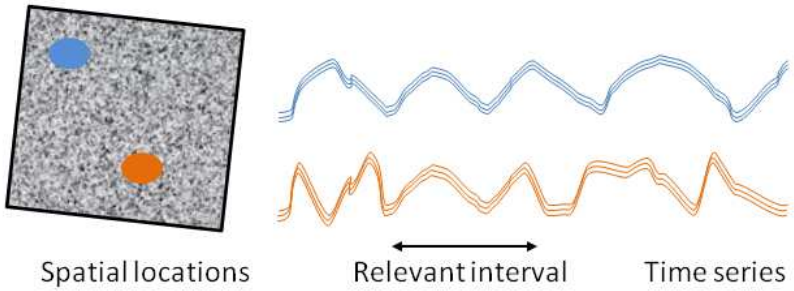

Figure 2.8: Long distance relationships found in small intervals of time series

2.3.8 Discriminatory analysis

The systems from which spatio-temporal datasets are collected are generally dynamic in nature and so the system is bound to be in one state or another at any given time. Understanding these states and identifying the differences between these states is crucial to our understanding of these systems. The spatio-temporal problems discussed above can be useful in understanding the state of a system in two different states. For example, in fMRI data, the temporal or spatial clusters can be different when the subject is resting or while the subject is working on a task. Similarly, they can be different in fMRI data obtained from healthy subjects and disease subjects. In climate data, one can use discriminatory analysis to shed light on the different physics based climatic models.

One could potentially study the difference between two states using any of the analysis described above such as : i) Spatial clustering, ii) Time point clustering, iii) Dynamics of contours, and iv) Long distance relationships. In Figure 2.9 we show an example where the blue and orange region time series share a relationship in a short interval when the hypothetical system is in a healthy state. This relationship does not exist when this system is in an unhealthy state. Discovering such scenarios is the objective of this problem.

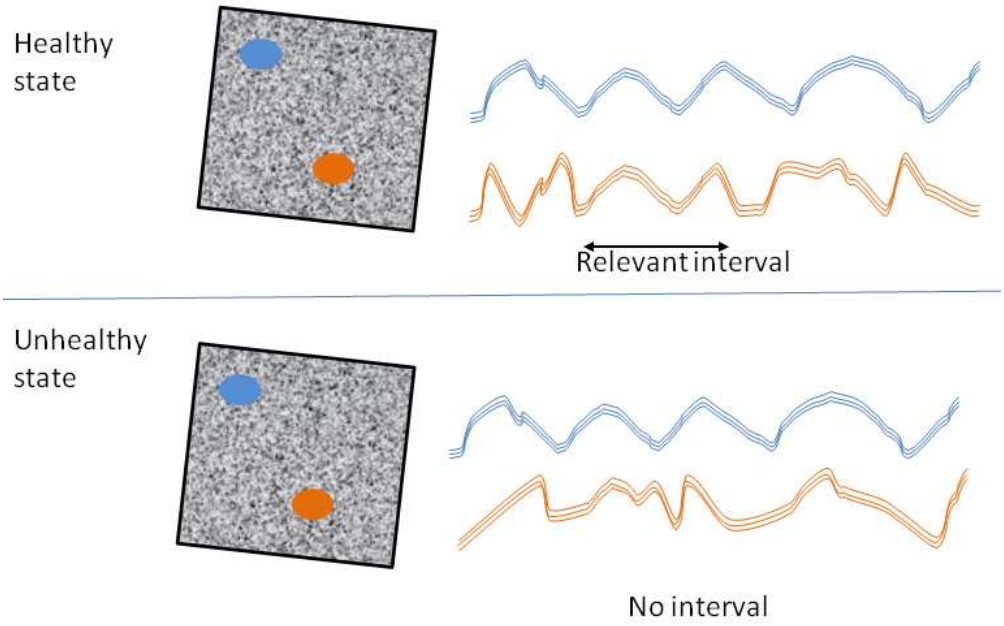

Figure 2.9: Discriminatory analysis of long distance relationships

One could use approaches for the above problems on data collected from a system when it is in healthy and unhealthy states and then compare the outcomes. However, this approach will result in multiple hypothesis testing all of which are not necessarily interesting and hence reduces the statistical power of the findings. A systematic approach to directly search for scenarios that are discriminating will not only be computationally efficient, but will also be statistically robust.

2.4 Conclusion

The problems defined above are very general and are relevant for multiple applications such as studying brain functionality, climate change and crime patterns. As the focus of this thesis is on analyzing brain fMRI data, we discuss the nature of fMRI data and the relevant neuroscience questions in the following chapter.

Chapter 3

Mining fMRI Data

3.1 Introduction

Over the last decade, neuroscientists have treated the brain as a complex system with interconnected parts that interact to achieve brain functions (learning, memory, etc.) [14, 44, 45]. Discovering the principles underlying these interactions is critical to elucidating the neuronal underpinnings of complex brain functions such as cognition and decision making as well as advancing the diagnosis and treatment procedures for mental disorders that are increasingly taking a bigger toll on human life and health care budgets [16, 46].

The increasing availability of various neuroimaging data such as Functional Magnetic Resonance Imaging (fMRI) offers great opportunity to study brain connectivity and elucidate its operating principles [47, 48, 11, 49, 50]. fMRI is a noninvasive imaging technology that measures the simultaneous amount of oxygen absorbed from hemoglobin in a human brain. The amount of oxygen absorbed at a given location reflects its underlying neuronal activity, and thus, fMRI allows to indirectly measure brain activity over time. Such brain activity data can be represented by a time-series representing the activation values for a given location over time. A typical fMRI recording consists of measurements from nearly 160,000 locations at 2 second intervals for scan duration of 5 minutes. Each recording location is refereed to as a volumetric pixel (voxel), a three-dimensional data point representing the brain activation on a regular grid in three-dimensional space. Given that each voxel's activation may be viewed as a time-series developing time-series analysis tools for mining such data would allow us to discover brain regions responsible for achieving specific tasks, discover the synergistic interactions between

brain regions to achieve a task, explore the interactions that affect faculties like memory, cognition, behavior, *etc.*, and explore the role disrupted interactions between regions may play in mental disorders.

A common practice in analyzing fMRI data is to build brain networks where nodes are either static regions from a predefined atlas or those that are derived using data driven approaches on the data collected from multiple subjects [11, 14, 46, 51, 52, 53, 54, 55, 56, 57, 44, 58, 13, 45]. Then the strength of the connection between any two nodes is defined using some similarity measure (e.g., correlation between regions time series) to capture the degree of co-activation of the two regions in question. Brain networks thus constructed (where nodes are brain regions and edges are similarities) are studied for different subjects (at rest vs. while performing a task, normal vs. subjects with mental disorder) to elucidate the underlying mechanisms driving a task or a mental disorder.

Despite 20 years of research in fMRI and related datasets there is a prevailing dissatisfaction in the community with the progress that is made in elucidating how the brain accomplishes its functions [17]. The following are the limitations with the state-of-the-art techniques that are typically used to analyze fMRI data: (i) Current approaches to construct networks rely on a static definition of nodes (e.g., an anatomically derived brain atlas [1] or an ICA based atlas that is designed from a population). In reality, these nodes could be subject specific and may possibly evolve with time. (ii) Existing approaches treat brain regions as independent entities while in reality adjacent (or nearby) locations share similar properties. (iii) Relationship between two nodes may only exist in certain intervals, while current approaches compute relationship over the entire duration of the scan for relationships between all pairs of brain regions. Recent approaches have started considering scenarios where the connections can be dynamic. Yet systematic approaches to explore such connections are lacking. (iv) Brain may undergo state transitions and the connectivity and dynamics may be state-specific. (v) Brain networks evolve with time and environmental factors. For example, when a person is learning a skill the brain networks adapt to encode for the learned skill. Existing approaches do not take into account this evolutionary behavior.

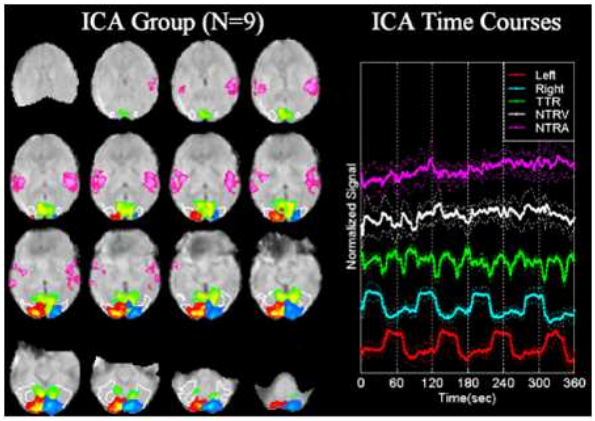

Figure 3.1: Example of ICA components (borrowed from [2]).

3.2 Problems

Significant advances can be made to the state-of-the-art in studying brain's functional architecture by discovering data driven brain regions, mining connections that can be complex and dynamic, discovering brain states, and by modeling the evolutionary behavior of the brain. In the following we present data science questions and challenges for each of these categories:

3.2.1 Defining data driven brain regions

In order to construct brain networks one needs to first define the nodes. These nodes in a spatio-temporal context are contiguous group of voxels that exhibit highly similar time series. Traditional approaches typically rely on anatomically defined brain regions or those that are derived from Independent Component Analysis (ICA) [59, 2] that looks for sets of contiguous voxels that exhibit independent time series. An example of ICA approach is shown in Figure 3.1, where the components, i.e., voxels corresponding to the time series on the right, are shown on the left size. These approaches have been shown to have high reliability when used on a group of subjects but have low to modest reliability when subject level components are considered.

This problem entails discovering brain regions from the data where the goal is to find clusters of contiguous voxels that share highly similar time series. Figure 3.2 motivates the need for the discovery of data driven brain regions, by showing that the inferred connectivity can be spurious otherwise. Using a traditional atlas, the strength of correlation between Frontal Middle and Parietal Superior regions shown in yellow in Figure 3.2(c) is 0.54. The time series are

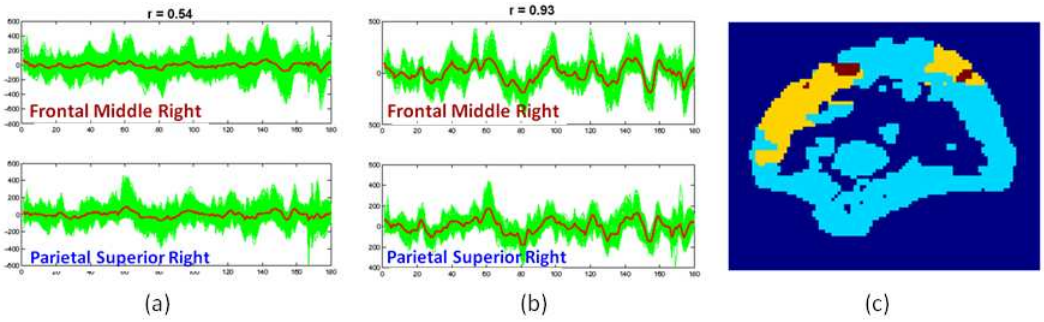

Figure 3.2: Impact of discovering data driven brain regions.

shown in Figure 3.2(a). A data-driven approach can find regions in 'red' that have a significantly stronger correlation of 0.93. The corresponding time series is shown in Figure 3.2(b).

This problem is challenging because of the spatial autocorrelation that is present in the data and so determining the boundary of any region is non-trivial. Different brain regions can have different sizes and handling this variability is also challenging since smaller regions can potentially be missed when they are in the neighborhood of a larger region due to spatial autocorrelation. Additionally the number of 'true' brain regions is unknown. Once brain regions are determined for every subject independently, determining a population level brain atlas is even more challenging, as it involves determining which regions in a subject's brain relate to which region in a different subject's brain.

3.2.2 Studying connectivity in the brain

The goal here is to design a data-driven approach that can directly discover the existing functional connections between distant brain regions in fMRI data. A number of challenges need to be addressed. First, the search space can be intractable since connectivity can exist between any two sets of brain regions. Second, spatial contiguity needs to be taken into account. Third, the nature of connectivity is not known. Brain connectivity in temporal domain can potentially be a linear, non-linear, lagged or a transient relationship. An illustration of transient connectivity between multiple brain regions is shown in Figure 3.3. Here the time series from two anatomical brain regions are found to be highly correlated ($r > 0.8$) in multiple intervals and not highly correlated in other intervals. This suggests that the two brain regions are functionally connected only transiently. Discovering such relationships and the groups of brain regions that

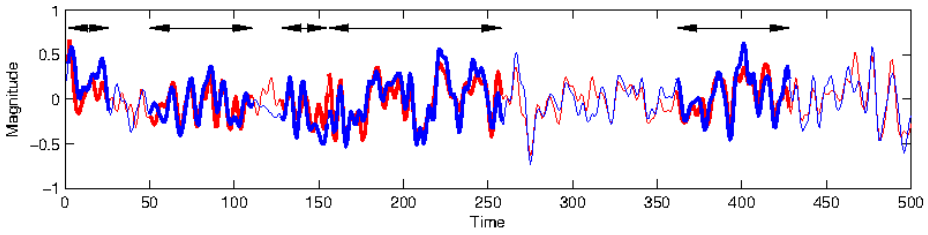

Figure 3.3: Transient connections.

exhibit such transient relationships is the focus of this thesis.

3.2.3 Discovering brain states

Approaches that can detect the time points at which a brain switches its state using spatio-temporal fMRI data can shed light on the operating principles of the brain. Novel approaches where a change point can be identified given a time series have been developed. Here there are two new aspects that need to be dealt with: i) We need to determine the change point by taking into account the time series collected from all the locations in the brain ii) We need to incorporate spatial information into the data mining techniques. Additionally, change may not necessarily manifest in the raw fMRI time series, but rather in high level properties such as connectivity or community structure. Figure 3.4 illustrates this with an example by showing that there are several sliding windows where the brain networks are highly similar. Here pair-wise euclidean distance between functional networks are constructed. Blue color in the figure represents high similarity. The black squares show the time intervals where the brain networks are highly similar and hence can be treated as brain states. The notion of brain states can also be defined based on ‘subspace’ similarities, such as similarity in the activity of a subset of voxels or brain regions; similarities in the subnetwork computed in sliding windows.

3.2.4 Modeling the evolutionary behavior

Another problem that is of interest in studying fMRI data is one of tracking how brain networks adopt to external environment with time. For example, when a skill is being learned one can be interested in how the brain encodes this skill in its network. This will aid in increasing our understanding of the brain’s cognitive functionality. This requires longitudinal analysis of the relationships among brain regions, where changes over time (inter-scan periods) need to

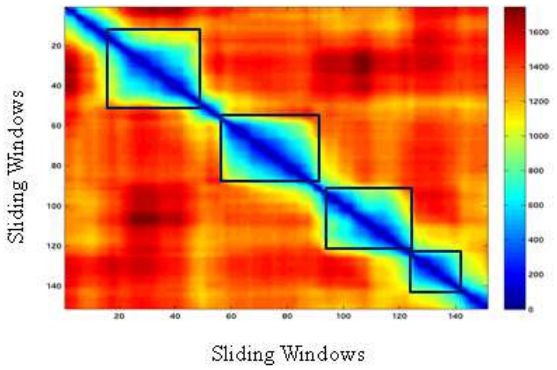

Figure 3.4: Illustrative example of brain states.

be investigated. Longitudinal analysis in this case is challenging because the functional brain networks are non-stationary in the intra-scan periods. Learning inter-scan differences while accounting for intra-scan variability is a challenging problem.

3.2.5 Discriminative analysis

A common problem that is studied in the neuroimaging community is that of discovering the properties of the brain network that is different two different samples, e.g., disease vs. healthy, resting vs. involved in a task. These difference could exist at many levels such as brain regions, connections, transient connections, network properties, sub-networks etc. In Figure 3.5 we show the correlated intervals found between brain regions 51 and 55 in rest and cartoons data with the help of double sided arrows where the left and right arrows indicate the start and end points, respectively. Between the resting state and watching cartoons there is a large difference in the amount of time the two time series are correlated. This suggests that these two regions exhibit synergy more when the subject is watching cartoons than when the subject is resting. Region 51 is a *middle occipital region* (left), referred to as the visual V1 system, is well known for its role in processing spatial information, where as region 55 is *fusiform* (left) that is known for its role in object and color information processing [60]. These two regions that handle different aspects of visual information processing can be hypothesized to work synergistically in processing visual information while the subject is watching cartoons.

Approaches to discover such differences between two sets of brain scans while subjects are resting or working on a task can help understand the role of brain regions and their interactions.

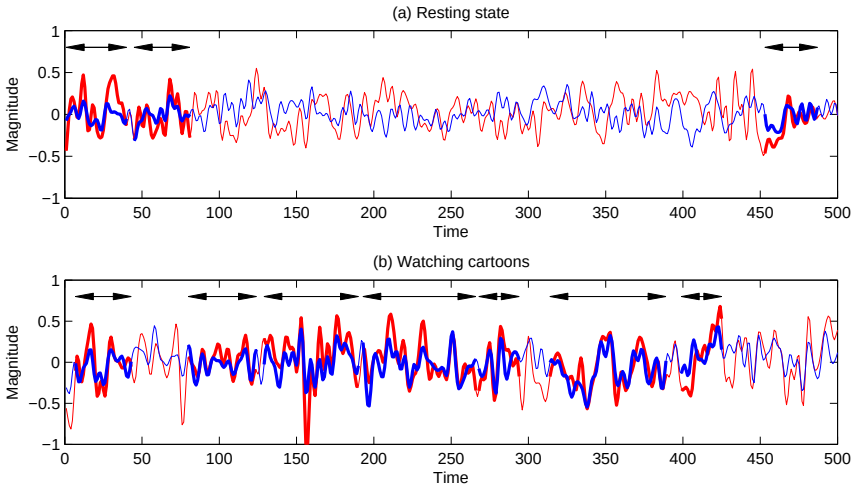

Figure 3.5: Comparison between resting state and task scenarios

In addition, they can help increase our understanding of complex mental disorders and aid in designing effective treatment strategies.

3.3 Conclusion

In this dissertation we address a subset of the problems discussed above. Specifically, we focus on designing data mining approaches to discover transient relationships between brain regions. We show that the approaches presented in this thesis are effective and efficient. We show the utility of these approaches in discovering reliable and replicable patterns by testing our approaches on multiple datasets that are collected on different sample sets and at different locations using diverse acquisition protocols. We also study the utility of these approaches in capturing differences between resting state and task scenarios. We also evaluate the use of our approaches in studying differences between healthy and schizophrenia subjects in a limited setting.

Chapter 4

Discovering groups of time series that share similarity in multiple small intervals

4.1 Introduction

Time series data has become increasingly ubiquitous during the last two decades in several domains including climate, bioinformatics, social media and neuroimaging [23, 6]. The data mining community has studied several problems pertaining to analyzing time series data [24, 6]. They include clustering [25, 26], classification [27], anomaly detection [28], forecasting [29], and segmentation [30]. The focus of this chapter is to address the problem of discovering groups of time series that share similar behavior in multiple small intervals of time. We refer to such groups as 'intermittently coherent time series' in the rest of this chapter.

In a complex dynamic system different groups of entities in the system may behave coherently for short intervals of time to achieve a specific objective. For example, in a human brain, a brain region can be treated as an entity and the amount of activity measured over time at a brain region could be treated as its behavior. Multiple brain regions are said to behave coherently for a short period of time when the time series of their activity levels become highly similar within this time period. Consider the hypothetical example shown in Figure 4.1, that depicts four time series each with 200 time points. These time series do not appear to be similar when all the 200

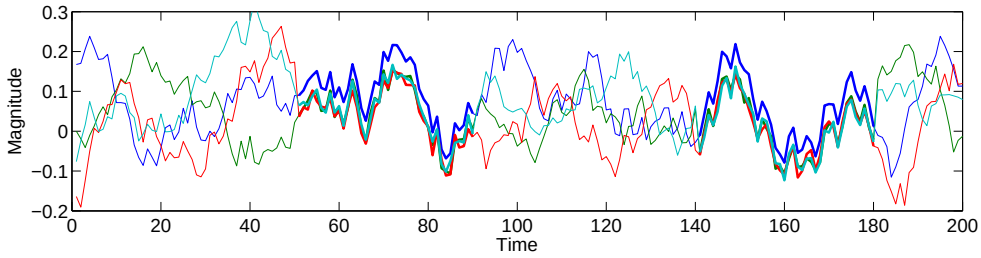

Figure 4.1: Four time series exhibiting intermittently coherent behavior. (All figures in this manuscript are best seen in color.)

time points are considered. However, in the time intervals from 51 to 90 and from 141 to 180 they exhibit high similarity. If such time series represent activity levels of brain regions over time (measured using fMRI technology) the corresponding brain regions could be hypothesized to work together to accomplish a specific task [61].

The problem of discovering groups of intermittently coherent time series from a given time series data set has two characteristics: i) There are exponentially many combinations of time series that needs to be explored to find these groups, ii) The groups of time series of interest need to have similar behavior only in some subsets of the time dimension.

Pattern mining approaches that have been studied in the context of market basket data [62, 22] address these two characteristics directly. The goal of these approaches is to find groups of items that occur together in many transactions (i.e., they are frequent itemsets). These techniques explore the combinatorial nature of the search space in a systematic fashion relying on the Apriori principle [62] that guarantees that if an item set is frequent then all of its subsets are frequent too. However, these pattern mining approaches have been designed to work with binary features, that indicate whether an item is contained in a transaction or not. Recently, they have also been explored for continuous valued datasets [63], but there is no existing framework that works with time series data.

In this chapter we generalize the well studied frequent pattern mining techniques to work with time series data in order to discover all groups of objects whose time series are intermittently coherent. Specifically we use a sliding window based approach and we propose the notion of support for time series data with a goal of capturing intermittent coherence for a candidate group of time series. Using this, we provide an Apriori based framework that can discover all groups of intermittently coherent time series such that the total length of coherent intervals for a group is longer than a given window-based threshold. We evaluate our approach on a synthetic

dataset and show its effectiveness in discovering all the desired intermittently coherent groups in comparison to that of alternative approaches. We then show the utility of our approach on a real world neuroimaging dataset, where we demonstrate that our approach can be used to discover significantly reproducible groups from independent sets of time series data collected from the same set of subjects. On the same dataset, we show its effectiveness in comparison with an alternative approach. We also demonstrate the utility of our approach on an S&P500 weekly stock prices data set.

The following are the key contributions of this chapter:

• A novel approach to quantify the duration of intermittent coherence for a given set of time series.

• A systematic framework for discovering all groups of time series that exhibit intermittently coherent behavior.

• Comparative evaluation of the proposed approach with alternative approaches to demonstrate its effectiveness on synthetic and real world datasets.

This chapter is organized as follows: In section 2, we formally define the problem. We present alternative approaches and the proposed approach in sections 3 and 4, respectively. In section 5, we present the evaluation of our approach on two real world datasets. We conclude with section 6.

4.2 Problem Formulation

Consider a set of observations made on n objects $\{I_1, I_2, \ldots I_n\}$ at m different time points $\{t_1, t_2, \ldots t_m\}$. Let the observations made on i^{th} object I_i be represented as a time series $d^i = (d^i_1, d^i_2, \ldots d^i_m)$. Let D be a matrix whose columns are the vectors $d^i, \forall i \in (1, \ldots, n)$. Consider a time window of length *window-length* ω that is moved across the time series in steps of size s. Our goal is to find those subset of objects $\{I_{j_1}, I_{j_2}, \ldots I_{j_p}\}$ such that the time series observed on these objects behave 'similarly' in at least a user provided number of windows. A number of different ways of characterizing "similarity" for time series have been studied in the literature [64, 6]. We will use Pearson's correlation as a measure of similarity between two objects for a given time interval in this chapter. A given set of objects is deemed to behave similarly if the minimum of the pairwise correlation of all the time series obtained from these

objects is above a user provided correlation threshold.

4.3 Alternative Approaches

To the best of our knowledge, there is no existing approach that can directly discover all groups of time series such that for every group there are sufficiently many time windows in which all constituent time series exhibit sufficiently high correlations among themselves. In this section we outline possible approaches that can help one find such groups.

Clustering of time series data is one way to determine groups of time series that are highly correlated. Traditional clustering approaches like k-means, hierarchical and density based clustering are often used with time series data sets by choosing an appropriate measure of similarity. Several similarity measures such as dynamic time warping, euclidean distance, and correlation have been studied in the literature [6, 64]. Note that these similarity measures have also been used to capture lagged relationships in the data which is not the focus of the problem that is being studied. Nevertheless, these techniques cannot capture the similarity (high correlation) in small time intervals, as they take into consideration the full time series available.

Frequent pattern mining techniques can be applied to time series data after binarizing the data using a suitable threshold. Consider a matrix D whose columns are the time series vectors d^i for every object i, and whose rows are time points. Using a binarization threshold this matrix can be converted to $D_{0/1}$ where an element takes a value 1, if its value in D is greater than the binarization threshold, and 0 otherwise. Frequent pattern mining on this data can explore all combinations of objects, but it is limited to capturing groups of objects whose value is beyond a threshold for a number of time points that is greater than a user provided threshold. This approach does not directly look for intermittently strong correlations, i.e., time intervals where the time series are highly correlated among them. Moreover, the binarization threshold based similarity cannot capture correlations in the full time series, let alone the intermittently strong correlations. Therefore the traditional binary pattern mining technique applied on a binarized version of time series data is not suitable to address the problem at hand.

Alternatively, one can use frequent pattern mining techniques on time series clusters obtained from sliding time windows. To achieve this, one can use a sliding time window of a chosen length and compute time series clusters within each window, by moving the window in steps of a predetermined size along the length of time series. These clusters can be used to

construct a binary matrix CT, where each row is a cluster and each column is a time series. A value of 1 in the matrix indicates the presence of time series in the corresponding cluster. Frequent pattern mining can then be used on this CT matrix to find groups of time series that participate together in the same cluster for sufficiently many time windows. This approach has the potential to recover groups of time series that share high correlations in many windows. A challenge with this approach is that it is not trivial to determine the choice of number of clusters within each sliding window. One can construct a scenario where there are different number of clusters in different sliding windows and this approach will not perform well in such a case. Moreover, in windows where there are no high correlations among the time series, this approach will find spurious clusters and so the resultant groups discovered could be potentially spurious.

4.4 Pattern Mining Framework

Discovering groups of time series that behave similarly for at least a given number of time points is a challenging problem. It requires searching through all combinations of objects as well as determining intervals in time at which the objects in question behave similarity. These challenges have been addressed in market-basket data sets by frequent pattern mining techniques. Market basket data captures the items purchased in a transaction in a binary data matrix X, whose columns are items in a market, and whose rows are transactions, and whose elements X_{ij} have a value 1 indicating the presence of an item j in a transaction i, and a value 0 otherwise. The goal of frequent pattern mining techniques is to discover all subsets of items (also referred to as itemsets) that are purchased "frequently". The ratio of the number of times a set of items are purchased together to the total number of transactions is treated as the *support* of an itemset. A user provided $support$ threshold is used to determine whether a given item-set is frequent. A transaction in which all the items in an itemset in question are present is said to "support" the itemset.

A standard pattern mining approach that is widely used with binary data sets is the Apriori algorithm [62]. At the heart of this approach is the Apriori principle that guarantees that if a set of items are not frequently purchased together, then any bigger set that includes this set is not frequent. This is due to the anti-monotonic nature of the $support$ measure, i.e., $support$ of a given set of items is less than equal to the $support$ of any of its subsets. Relying on this principle, the Apriori algorithm builds item sets bottom up, where it starts with all single items

and filters out items that are not frequent. It then groups the frequent single items to enumerate candidate item-pairs and then evaluates them to select those pairs that are truly frequent. Then candidate item-triples are enumerated from the frequent pairs by joining the pairs that share one item and the frequent triples are determined by filtering out the infrequent ones from the candidate triples. In this fashion it constructs higher-order sets until no more bigger sets can be enumerated. Note that the higher order candidate itemsets are only enumerated from the frequent itemsets at a given level. This reduces the number of candidate itemsets effectively. By systematically pruning the search space of all possible combinations of items, this approach can efficiently discover all possible itemsets that are frequently purchased together.

4.4.1 Designing a notion of support for time series data

The key difference between market basket data and time series data is that in market basket data we have a binary vector (a column in X) for every item indicating its presence in each of the transactions, while in time series data we have a time series d^i with continuous values for an object I_i. In the case of market basket data, supporting transactions for a given set of items can be determined by computing the intersection of the transactions in which each of the individual items are present. This is not trivial with time series data. Moreover, the goal is to identify the intervals during which a high correlation is exhibited.

Here we use a sliding window based approach to compute coherence between time series for each window. Specifically, we choose a *window-length* ω to determine the duration of a window and to move the window across the time series in steps of size s. For example, the first window captures the time points $(t_1, \ldots, t_\omega)$ and the second window captures the time points $(t_{s+1}, \ldots, t_{s+\omega})$. We refer to each window using the index of the ending time point. For example, the first two windows are referred to as w_ω and $w_{s+\omega}$. For a given time series d^i of length m, using a choice of window length ω and a step size s, the set of windows is referred to as $w^i = (w^i_\omega, w^i_{\omega+s}, \ldots w^i_{\frac{m-\omega}{s}+1})$.

We treat each window as a transaction in traditional frequent pattern mining. To determine if a window supports a group of time series we need to estimate if the group of time series exhibit high coherence within this window. We perform this by computing the pairwise correlations between the time series for a given window. A window is said to support a group of time series if the minimum of the pairwise correlations is greater than a user-provided correlation threshold γ. The number of time windows that support a group of time series is referred to as

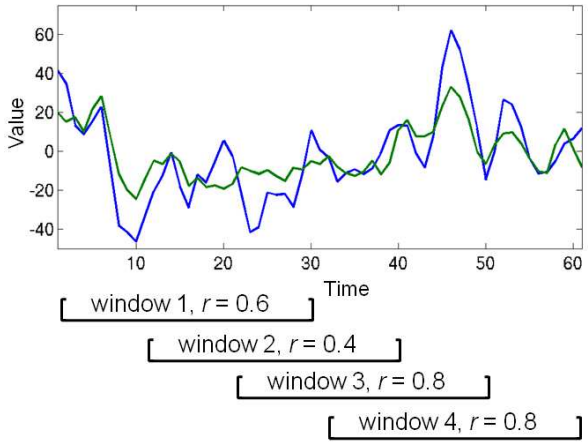

Figure 4.2: Example to illustrate the notion of $ts - support$ with $\omega = 30$, $s = 10$ and $\gamma = 0.8$.

$ts - support$. Formally, $ts - support$ for a set of objects $S \in \{I_1, \ldots I_n\}$ is defined as follows:

$$ts - support(S, \omega, s, \gamma) = \sum_{i=\omega}^{\frac{m-\omega}{s}+1} \mathbf{1}_{minpwc(S,w_i)\geq\gamma} \tag{4.1}$$

where $minpwc(S, w_i)$ is the minimum of the pairwise correlations between objects in the set S for the window w_i. $\mathbf{1}_{minpwc(S,w_i)\geq\gamma}$ is 1 when $minpwc(S, w_i)$ is greater than a user provided threshold γ, 0 otherwise. Note that the windows that support a given set of time series are the windows in which the given set exhibits sufficiently high correlations. Greater the $ts - support$ of a set of objects, longer is the duration of sufficiently high correlations among them.

We illustrate the notion of $ts - support$ with the help of an example shown in Figure 4.2. Here two time series are shown for which $ts - support$ needs to be estimated. The choice of window length $\omega = 30$, step size $s = 10$, and correlation threshold $\gamma = 0.8$ are used. In the first window w_{30} spanning $(t_1, \ldots t_{30})$ the time series has a correlation 0.6. The second window spans w_{40} spanning $(t_{11}, \ldots t_{40})$ and the two time series have a correlation 0.4 in this window. Similarly, for the third and fourth windows, w_{50} and w_{60}, the correlations are 0.82 and 0.83, respectively. Only the third and fourth windows, w_{50} and w_{60}, contribute to support as their correlation surpasses the γ threshold. Therefore, $ts - support$ for the time series in this example is 2.

Antimonotonicity of the $ts - support$ measure allows us to use the Apriori framework to enumerate all frequent groups of time series.

4.4.2 Antimonotonicity of $ts - support$

We now prove that the $ts - support$ measure we defined above is anti-monotonic so it can be used in an Apriori like framework [62] to discover all subsets of time series that satisfy a given $ts - support$ threshold, γ.

Theorem 4.4.1 *$ts - support(S, \omega, s, \gamma)$ measure decreases monotonically as new items are introduced for a given set of time series S, window length ω, step size s, and a pairwise correlation threshold γ.*

Proof Consider a new set S', such that $S' = S \cup x$.

A window w_i that does not contribute to $ts - support(S, \omega, s, \gamma)$, i.e., $minpwc(S, w_i) < \gamma$, will not contribute to $ts - support(S', \omega, s, \gamma)$ because the minimum pairwise correlation $minpwc(S, w_i)$ will not increase as a new time series x is introduced to the set S.

A window w_i that contributes to $ts-support(S, \omega, \gamma)$, i.e., $minpwc(S, w_i) \geq \gamma$, will either contribute or not contribute to $ts - support(S', \omega, s, \gamma)$ depending on how the new time series x affects the minimum pairwise correlation. If $minpwc(S, w_i) \geq \gamma$ and $minpwc(S', w_i) \geq \gamma$, then $ts-support(S, \omega, s, \gamma) = ts-support(S', \omega, s, \gamma)$, otherwise $ts-support(S, \omega, s, \gamma) > ts - support(S', \omega, s, \gamma)$.

Therefore, $ts - support(S, \omega, s, \gamma) \geq ts - support(S', \omega, s, \gamma)$

4.4.3 Apriori-based approach for time series data

Using the above notion of computing support from time series data we now describe a generalized Apriori algorithm that can work with time series data. First, we start with all pairs of objects and then evaluate their $ts-support$ to determine the pairs that are interesting. Note that the original Apriori starts with single items and determine frequent itemsets. Here we cannot filter at the first level because we need at least two time series to determine similarity and so we start by enumerating all pairs. Once the frequent pairs (i.e., pairs with $ts - support \geq \gamma$) are determined, we then enumerate the candidate triples as is done in a traditional Apriori algorithm [62] by joining interesting pairs that share one object. This approach continues until no more bigger frequent sets are found.

The algorithm is outlined here:

Step 2 enumerates all possible pairs, while steps 3-6 compute the support of a pattern and determine the frequent pairs that satisfy the support criteria, $ts-support(cs_k, \omega, \gamma) \geq \sigma$. Steps

Algorithm 1 Time Series Pattern Mining

Input:

i. D, a real valued time series data matrix of size $|m \times n|$, where columns are items $I = \{I_1, I_2, \ldots I_n\}$ and rows are time points $T = \{t_1, t_2, \ldots t_m\}$

ii. σ, a support threshold

iii. ω, window length

iv. γ, minimum correlation threshold

Output:

All subsets of objects with $ts - support \geq \sigma$

1. $k = 2$
2. $CS_k = \{(I_i, I_j) | i \neq j, I_i \in I, I_i \in I\}$
3. **for** each candidate $cs_k \in CS_k$ **do**
4. compute $ts - support(cs_k, \omega, s, \gamma)$ using Eq. 4.1
5. **end**
6. $S_k = \{cs_k | cs_k \in CS_k \wedge ts - support(cs_k, \omega, s, \gamma) \geq \sigma\}$
7. **while** $S_k \neq \emptyset$ **do**
8. $k = k + 1$
9. $CS_k = Apriori - gen(S_{k-1})$
10. **for** each candidate $cs_k \in CS_k$ **do**
11. compute $ts - support(cs_k, \omega, s, \gamma)$ using Eq. 4.1
12. **end**
13. $S_k = \{cs_k | cs_k \in CS_k \wedge ts - support(cs_k, \omega, s, \gamma) \geq \sigma\}$
14. **end**
15. Result $= \bigcup S_k$

7 through 14 enumerates candidates and determines frequent bigger patterns in an iterative way, until no bigger frequent patterns can be found.

4.4.4 Handling issues due to highly similar time series

Note that in a given dataset there could be groups of time series that are correlated when all the time points considered. For example, in stocks data many stocks that belong to a given sector (e.g., health sector) could exhibit high correlations for the entire duration of time considered. These groups will have high value for our newly defined notion of support and will make it computationally hard to discover the low support patterns that are sufficiently correlated for a relatively shorter amount of time. To avoid finding these groups (that can be more easily found using alternate techniques), we add an additional constraint to our approach that discards any

candidate set that has two objects I_i and I_j whose full time series d^i and d^j have a correlation that is greater than a user provided $full - corr - thresh$, before computing their support. This is achieved by filtering out such candidates immediately after the candidates are enumerated in steps 2 and 9 of Algorithm 2.

4.4.5 Handling artifacts due to globally similar behavior

In many cases high correlations among all the time series in an interval can be induced due to a global event in the system. For example the 2007-2008 recession induces a similar behavior in most of the stocks, and any windows that contribute to $ts - support$ in this period will inflate the support even though the event is not specific to the candidate pattern. Similarly, motion related artifacts create global patterns in neuroimaging data [65]. There is a need to control for windows that have such globally similar behavior from contributing towards the $ts - support$. One approach to address this challenge would be to discard all windows that capture a globally similar behavior and work with the remaining windows. Another approach is to weight the windows depending on how similar the behavior of a candidate set for a window is to the global behavior (e.g., correlation between mean time series for a candidate set with that of the entire set). In the context of market basket data this will be akin to developing a weighted version in which transactions that have too many items provide no support (former approach) or smaller support (later approach). We use the former approach and we show its utility in finding groups of time series that exhibit intermittent correlations not due to a global scenario in Section 5.3. This is achieved by ignoring those windows whose median of pairwise correlations between all the time series is greater than a $global - corr - thresh$ threshold. We incorporate this into our definition of $ts - support$ as follows:

$$ts - support(S, \omega, s, \gamma, global - corr - thresh) = \sum_{i=\omega}^{\frac{m-\omega}{s}+1} \mathbf{1}_{(minpwc(S,w_i)\geq\gamma)\&(mediangpwc(w_i)\leq globalcorrthresh)} \quad (4.2)$$

where $mediangpwc(w_i)$ is the median of the pairwise correlations between all objects in the set I for the window w_i.

4.5 Evaluation

Designing a thorough evaluation pipeline is a challenge for the problem at hand as is the case with many unsupervised algorithms. We used a synthetic dataset to highlight the key strength of the proposed approach and the weakness of competing approaches. The lack of ground truth in real world datasets limits us from directly comparing the groups of time series discovered using the proposed and the competing approaches. However, we performed a comparative evaluation the quality of the discovered groups. Using a neuroimaging time series data collected from same set of subjects at two different time points we studied the replicability of the findings which is necessary to test the validity of the results. In addition to this, we demonstrate the utility of our approach using a case-study on S&P stocks data.

4.5.1 Evaluation on a Synthetic Dataset

Data: We first created a random 400x10 matrix R, where rows are time points and columns are time series, by sampling each element from a uniform distribution with a range [0 1]. Each time series is further smoothed by computing the value at a time point t as the average of neighboring points from $t-5$ to $t+5$ to incorporate temporal auto-correlation that naturally exists in real world time series datasets, i.e., consecutive time points in a time series have similar values. We then impute four sets $\{(1,2,3,4),(5,6),(7,8),(9,10)\}$ of strong correlations for 120 time points (separate intervals of length 60 and 60). This is done for every set by copying the first time series for a chosen set of 60 contiguous time points in the other members of the set with a small amount of additive noise sampled from a Gaussian distribution with a mean of 0, and a standard deviation of 0.01. The four groups of time series are shown in Figure 4.3. The regions of time series shown in bold curves in each of these groups are the imputed highly correlated intervals that we expect the following approaches to capture.

Approaches: We used three other competing approaches, in addition to the proposed approach:

1. *K-means clustering (K-means)*

We clustered the set of 10 time series into four clusters using correlation as a distance metric. We clustered them into four groups as the number of groups that were imputed was also four.

2. *Apriori on binarized time series (Apriori-$R_{0/1}$)*

We first constructed a binary matrix $R_{0/1}$ using a threshold on matrix R and then found maximal frequent patterns of time series using a support threshold. We considered the following choices

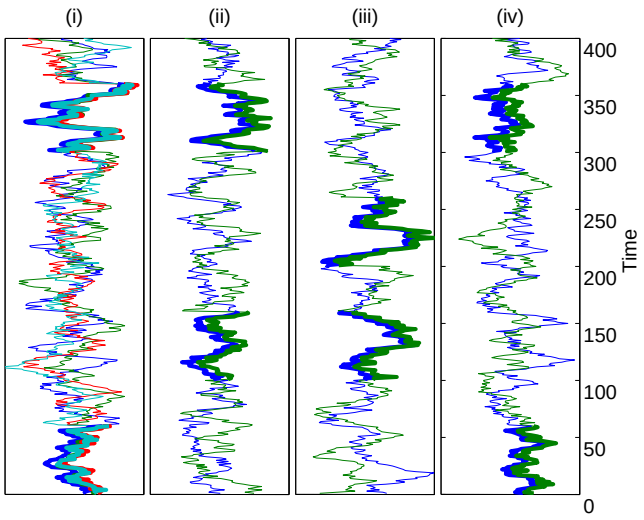

Figure 4.3: Four groups of synthetically generated intermittently correlated time series: (i) $\{1, 2, 3, 4\}$ (ii) $\{5, 6\}$ (iii) $\{7, 8\}$ (v) $\{9, 10\}$. Regions of the bold time series are the correlated intervals.

of quantile based thresholds from the matrix R: $\{0.5,\ 0.55,\ 0.6,\ 0.65,\ 0.7\}$. A value in the matrix $R_{0/1}$ was 1, only if the corresponding value in R was above the chosen quantile based threshold. We used a support threshold of 60 for consistency in comparison with the other Apriori based schemes that are described below. For the sake of interpretability, we treat number of rows supporting a pattern (not fraction) as it's $support$ for Apriori based methods.

3. *Apriori on K-means clusters (K-means+Apriori)*

We used a sliding window of length 30 that is moved along the time series in steps of size 1. This resulted in 371 sliding windows. Within each window we considered the 10 time series and clustered them into k clusters. Several choices of k were explored: $k = \{2, 3, \ldots 8\}$. Each cluster that has more than one member is then used to construct a binary CT matrix whose rows are clusters and whose columns indicate time series. A value of 1 in this matrix indicates that a time series was part of a cluster from the window in which it was discovered. We then found maximal frequent sets of time series that were part of more than 60 clusters. Note that every candidate set of time series can be supported by at most one cluster from a sliding window, because k-means clustering is partitional in nature.

4. *Time series pattern mining (TS-Apriori)*

We used a sliding window length $\omega = 30$, step size $s = 1$, minimum pairwise length threshold $\gamma = 0.8$, support threshold $\sigma = 60$.

The rationale for the choice of support $\sigma = 60$ in all the Apriori based approaches that work with sliding windows (Apriori+K-means, and TS-Apriori) was that each input group has two independent 60 time point long highly correlated intervals. With the chosen window length of 30, an interval of 60 time points will be visible in at least 30 sliding windows and together the two intervals (for a given group) will be visible for at least 60 windows. Therefore a support of 60 should suffice to discover all the imputed groups. Apriori-$R_{0/1}$ on the other hand does not use sliding windows and treats each time point independently. Therefore, a support of 60 is smaller than the sum of the duration of highly correlated intervals (120).

Comparison metrics: For each approach presented above, we evaluated two key factors: *recoverability* and *spuriousness*. Recoverability is the fraction of imputed groups that were discovered. Only when an imputed group is a subset of a discovered group, an imputed group is treated as a recovered group. Spuriousness is the fraction of discovered groups that were not imputed, i.e., those discovered groups that are not subsets of any imputed group. For an ideal approach, the recoverability is expected to be high (1) and the spuriousness is expected to be low (0).

Observations: The recoverability and the spuriousness of the groups/patterns discovered using the four approaches are shown in Table 4.1. For the full time series based approaches K-means and Apriori-$R_{0/1}$, the recoverability is poor and spuriousness is high. High spuriousness is mainly because they take the full time series into account for finding groups and low recoverability is due to fact that the locally high correlations are not apparent when correlation is assessed for the entire time series.

K-means+Apriori performs differently for different choices of k. When k is very small, the recoverability is very poor and the spuriousness is very high. This is because the clusters in each window are forced to be much bigger than the imputed groups and they support spurious patterns in the Apriori framework. When k is moderate ($k = 4, 5$), the recoverability increases, and spuriousness increases too. When k is high ($k = 6, 7$), the recoverability is relatively high, and spuriousness is relatively low. This is because the clusters become smaller as k increases. At the same time a high choice of k will not leave all the clusters intact, as it splits some real groups into smaller clusters. This is the reason recoverability is only as high as 0.75, for $k = \{4, 5, 6, 7\}$, and decreases to 0.5, for $k = 8$. In general, it can be noticed that more spurious

Approach	Parameters	Recoverability	Spuriousness
K-means	k=4	0.25	0.5
Apriori	q = 0.5	0.25	0.98
-$R_{0/1}$	q = 0.55	0.25	0.86
($\sigma = 60$)	q = 0.6	0.5	0.81
	q = 0.65	0.25	0.57
	q = 0.7	0.25	0.20
K-means	k = 2	0.25	0.96
+ Apriori	k = 3	0.25	0.96
($\sigma = 60$)	k = 4	0.5	0.89
	k = 5	0.75	0.57
	k = 6	0.75	0.36
	k = 7	0.75	0
	k = 8	0.5	0
TS -	$\gamma = 0.8$	1	0
Apriori			
($\sigma = 60$)			

Table 4.1: Comparison with competing approaches

groups are found when the choice of k is low, and some real groups are missed when k is high. Moreover, there are different number of imputed groups in different intervals. For example, from Figure 4.3 it can be seen that for the interval 301 to 360, there are three groups that are imputed, while there is only one group imputed in the interval from 201 to 260. Spuriousness could also be a result of windows where there are no imputed groups, where K-means is forced to find k groups in all windows. Therefore, using the same choice of k for all windows will not yield a recoverability of 1 and spuriousness of 0 in this synthetic dataset. Even in cases where same number of clusters are imputed in each window, choosing the right k is still nontrivial, as a high k will result in low recoverability and a low k will result in high spuriousness.

For the proposed approach, TS-Apriori, the recoverability is 1 and spuriousness is 0, which is the ideal scenario. This is mainly because it does not rely on clustering and it evaluates the relationship between candidate groups for each window independently and so it is able to recover all of the imputed groups without discovering any spurious groups.

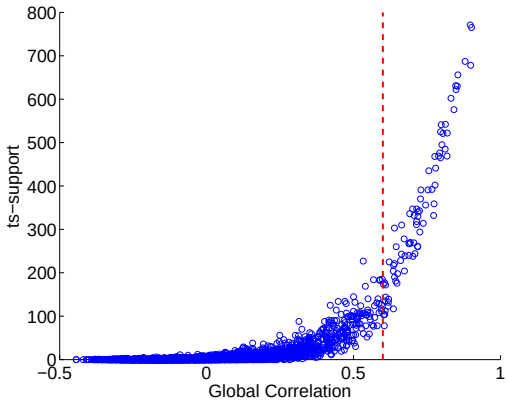

Figure 4.4: Comparison between pairwise global correlation and $ts - support$

4.5.2 Case study on Neuroimaging Data

Functional Magnetic Resonance Image (fMRI) data measures the amount of oxygen consumed at every 2x2x2 mm cubic location in the brain (referred to as a voxel) and it is known to indicate the amount of activity occurring at any location. Data from an fMRI scan can be represented in the form of a time$\times$voxel matrix B, where every element Bv_{ij} in the matrix indicates the amount of neuronal activity occurring at a time point i and at a location represented by voxel j. We used the dataset from [47] that contains 6 minute resting state fMRI scans from 27 healthy subjects obtained at two different time points that are 9 months apart. We refer to the first set of scans from 27 subjects as Scan 1 data, and the second set as Scan 2 data. The spatial resolution of each fMRI scan was 2$\times$2$\times$2 mm and the temporal resolution was 2 seconds. Several prepossessing steps have been performed on the data obtained from the scanner and they have been elaborately discussed in [47]. In addition, following the approach in [66], global mean time series is regressed from the data, as is done in most fMRI studies. The resultant time$\times$voxel matrix for each scan was of dimensions $180 \times 160,990$. We further group voxels into 90 brain regions based on an automated anatomical labelling (AAL) atlas provided by [1] (see Table A.1 for a list of the brain regions). The resultant matrix, Br, for each scan was of size 180×90. We then appended the time series from each of the 27 scans from Scan 1 data to get a 4860x90 matrix. Similarly we appended the time series from Scan 2 data to get another 4860x90 matrix.

Out of the 90 brain regions, a few brain regions that are related to visual system of the brain

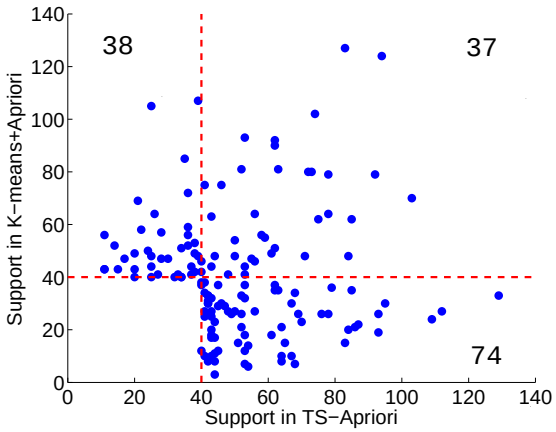

Figure 4.5: Patterns discovered using TS-Apriori and K-Means+Apriori

are found to be consistently correlated in earlier studies [67]. These set of brain regions with highly correlated time series will introduce many high support patterns in our analysis and these patterns are uninteresting in our case as they can also be discovered using time series clustering techniques. In Figure 4.4 we show the global correlation and the corresponding $ts - support$ for all pairs of brain regions. The pairs of regions that are highly correlated ($r \geq 0.6$) have a $ts - support$ ranging from 300 to 800. The strength of our approach lies in finding groups of brain regions that exhibit similar behavior in multiple small intervals in time. Therefore, we use a $full - corr - thresh = 0.6$ to prune all those candidates that have a high $ts - support$ to directly find those interesting groups that are otherwise unknown.

We used the proposed TS-Apriori with window-length $\omega = 30$, $s = 5$, $\gamma = 0.7$, $\sigma = 40$ on Scan 1 appended time series data matrix and found 111 size-3 patterns. We also used K-means+Apriori, that is the best of the competing approaches from our evaluation using synthetic data, to discover intermittently correlated groups of time series from Scan 1 data, with $k = 30$ clusters in each window using parameters $\omega = 30$, $s = 5$, and $\sigma = 40$ that are same as those used with TS-Apriori. We discovered 75 size 3 patterns. The union of the 111 and 75 patterns discovered using TS-Apriori and K-means+Apriori approaches, respectively, results in 149 patterns and their support computed using the two approaches is compared in Figure 4.5. Note that the support in K-means+Apriori and the $ts - support$ in TS-Apriori can be compared, because both of them represent the number of windows that support a group of brain regions ($\gamma > 0.7$). Out of the 75 size 3 patterns discovered from K-means+Apriori, only 37 patterns

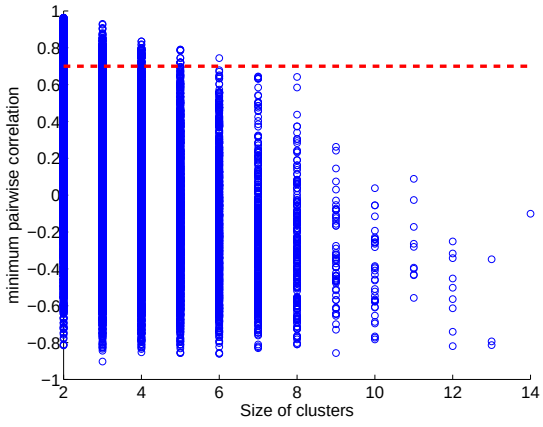

Figure 4.6: Relationship between cluster size and its quality

have a $ts - support \geq 40$ (49.3%, approximately). This suggests that the remaining 50.7% patterns are spurious according to our objective of finding group of time series that exhibit similar behavior in at least a given number of time steps. These patterns are shown above the horizontal red dashed line indicating $support \geq 40$ and to the left of the vertical dashed red line indicating $ts - support \leq 40$. This spuriousness is mainly due to the poor quality of the clusters discovered, i.e., the minimum pairwise correlation of clusters is less than the γ threshold used in TS-Apriori. Figure 4.6 shows the relationship between the clusters and their quality ($minpwc$ measure) from the windows they were discovered from. The clusters whose $minpwc$ is greater than $\gamma = 0.7$ threshold are those that lie above the dashed red line, while those that have relatively poor $minpwc$ lie below the red line. The 50.7% spurious patterns are supported by these clusters that lie beneath the dashed red line in the figure.

One could argue that a smaller k can be used to ensure that all clusters have a $minpwc \geq \gamma$. However, a smaller k could potentially result in splitting naturally existing clusters in other windows into smaller clusters. Even at the choice of $k = 30$, K-means+Apriori only recovered 37 of the 111 TS-Apriori patterns, indicating that the recoverability is only 29.7% (along with spuriousness 50.7%). This is potentially due to the different number of natural groups that exist in different windows and so these groups cannot be recovered using a uniform k for all windows. On the other hand, our approach estimates the strength of correlation between the brain regions in a set using $minpwc$ measure and determines whether a window supports a pattern or not.

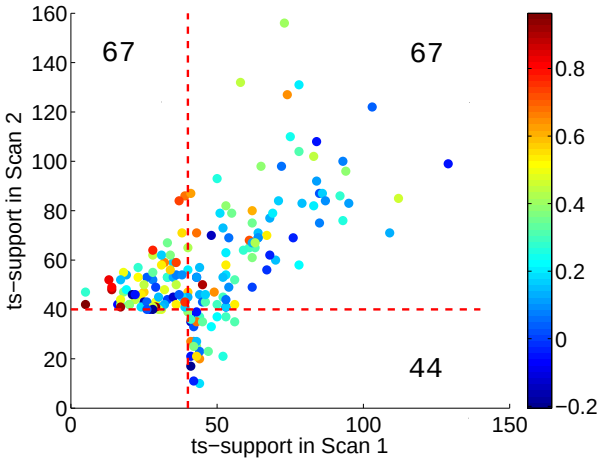

Figure 4.7: $ts - support$ of patterns found in Scan 1 and Scan 2 datasets

On Scan 2 dataset, using the same parameters as in Scan 1 dataset, we found similar observations where K-means+Apriori missed 54.5% (73 out of 134) of the patterns found by TS-Apriori and 21.8% (17 out of 78) of the patterns found by K-means+Apriori were spurious. As the recoverability and spuriousness of K-means+Apriori relies heavily on the choice of k, we tried several additional choices of k, including $k = 10, 20, 40, 50$. We found that spuriousness increases dramatically for lower choices of k, while very few of the TS-Apriori patterns were discovered for higher choices of k. These observations are similar to those demonstrated above using the synthetic dataset. These results highlight the limitations of the K-means+Apriori approach and the strengths of the proposed TS-Apriori approach on a real world dataset.

We further studied the similarity in the 111 and 134 patterns that were discovered from Scan 1 and Scan 2 datasets, respectively. In Figure4.7 we compare the $ts - support$ of the 178 patterns (union of 111 and 134 patterns) in Scan 1 and Scan 2 data. The color of each circle in this figure is the correlation between the number of windows contributed from 27 subjects in Scan 1 and Scan 2 datasets. There are 67 patterns that are common in the 178 patterns. This overlap is very significant given the large number of possible size-3 patterns ($\binom{90}{3} = 117,480$). Using a hypergeometric distribution we computed that the probability of expecting an overlap of 67 or more when 111 and 134 objects are drawn independently from a set of $117,480$ is less than 10^{-12}.

The correlations of contributions from subjects towards $ts - support$ (in Figure 4.7) are

weak. The average of the correlation of contributions for the 67 patterns that are common is approximately 0.24. This is indicating that the contribution of subjects towards patterns is different in different scans, and that both the scans do not have same information about these patterns. This is inline with observations made by many studies that the reliability of the correlations between time series computed from two scans of the same subject are poor [67, 47]. Despite this weak similarity between scans, the fact that these patterns have high support in both the datasets suggests that an underlying neurological phenomenon could be driving these patterns.

4.5.3 Case Study on Stock Market Data

We obtained the weekly closing stock prices of S&P500 companies over a 10-year period from January 2000 to December 2009 (521 weeks) from Yahoo! Finance website. We then removed those companies from this list for which only part of the data (less than 521 weeks) was available. We were left with 443 companies for which stock prices were available for all the 521 weeks. As the stock prices are at different scales, we normalized each time series d^i such that

$$d\_new_t^i = \frac{d_t^i - min(d^i)}{max(d^i) - min(d^i)} \tag{4.3}$$

where, d_t^i is the original stock price of stock i at time t, and $min(d^i)$ and $max(d^i)$ are the minimum and maximum stock prices of stock i, respectively.

Discovering groups of companies that exhibit strong correlations in small intervals from a span of 10 years could reveal novel direct or indirect relationships among companies. We found that this stocks data has two key characteristics that can lead to the discovery of uninteresting patterns: i) Two stocks that belonged to the same industry generally showed very strong correlation during the 10 year period. For example, stocks APA and APC that belong to oil and gas industry have a correlation of 0.95, approximately. Such groups can be directly discovered using traditional clustering based schemes and are uninteresting for our purpose. ii) Certain incidents affect all the stocks, e.g., the mortgage crisis, and so contribution of such windows towards $ts - support$ may lead to spurious and uninteresting patterns. Our approach addresses the first problem by building candidates involving those companies whose minimum of 10 year pairwise correlations is less than 0.6 ($full - corr - thresh$). The second problem is addressed by discarding the windows where the median of pairwise correlations for all companies is $mediangpwc$ is beyond 0.6. Under these conditions, using our time series pattern mining

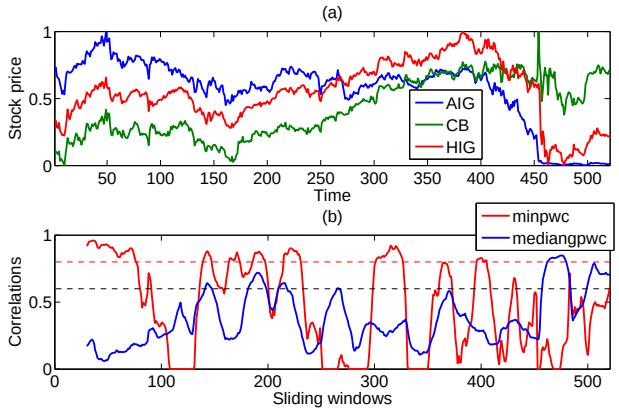

Figure 4.8: A selected Apriori-TS pattern generated from Stocks data set.

approach we found all groups of companies that share high correlations in at least $\sigma = 80$ time windows, using $\omega = 30$, $s = 2$, and $\gamma = 0.8$. There were 2965 size-2 patterns and 41 size-3 patterns.

Figure 4.8(a) shows one group of three financial sector companies American International Group (AIG), The Chubb Corporation (CB), and Hartford Financial Services Group (HIG) that was discovered in our analysis. In Figure 4.8(b) we show the minimum of pairwise correlation ($minpwc$) among these companies for each window using a red-colored curve. The horizontal dashed line in red indicates the γ threshold used to determine the windows that contribute to the $ts - support$. The $minpwc$ curve is above the γ line for windows that end in the time points from 30 to 75, 140 to 150, 160 to 170, 185 to 195, 210 to 230 and 395 to 405, suggesting that these stocks are highly correlated in these windows. It is interesting that these companies, despite belonging to the same sector, exhibit relatively weak correlations for more than half the time. The blue curve in Figure 4.8(b) indicates the median of the pairwise correlations among all companies in each sliding window ($mediangpwc(w_i)$). Note that for windows ending in time points 145 to 150, 175 to 200, 210 to 220, and 460 to 470 this curve crosses the $global-corr-thresh = 0.6$ threshold, suggesting that almost all of the companies exhibit similar behavior in these windows. Overall, 82 of the 492 sliding windows are discarded.

The three stocks AIG, CB, and HIG that belong to the finance sector are expected to behave similarly for the entire duration. However, from Figure 4.8(b) it can be seen that during the first 80 weeks starting from the January 2000 they share a strong relationship. As time progresses,

this relationship deteriorates and resurfaces due to several events that punctuate the time series. In period 2004-2005 (250 to 300 time points) AIG faced civil actions from regulatory authorities and later reached a settlement. AIG and HIG were hit by the financial crisis that occurred in late 2008 (400 to 450 time points). These events have impacted the stock prices and so they deviated from the other stocks with which they exhibited similar behavior at the beginning of the decade. The proposed approach allows one to discover such groups of intermittently correlated time series.

4.6 Conclusion and Future Work

In this chapter we presented a pattern mining based approach for discovering groups of time series that exhibit strong intermittent correlations. We have shown, using a synthetic dataset, that the proposed approach is more suited to this problem than the competing approaches. Our approach is guaranteed to discover all groups given a support threshold. We also demonstrated the reproducibility of the groups found in fMRI data using two independent sets of scans obtained from the same cohort of subjects. Using the same dataset, we also demonstrated that the proposed approach directly searches for the desired groups and so it is effective in discovering them in comparison to alternative approaches. We also show the utility of the proposed approach on S&P 500 stocks dataset.

A number of aspects of the proposed framework need further investigation. The sliding window based support is a surrogate to measure the extent of time for which a candidate set of time series exhibit high correlations and it does not always accurately reflect the duration. Consider two time series that exhibit high correlation in two non-overlapping windows. Consider another example where the two time series exhibit high correlation in successive and overlapping windows. Although the $ts - support = 2$ for both these examples, the total duration of the strong correlation in the first case can be approximately twice that of the second, when the step-size is small. To address this issue, approaches that can directly capture the time intervals in which a given set of time series are highly correlated needs to be explored. The frequent pattern mining framework introduces challenges in the context of noisy data, high dimensional nature of the data, and continuous-valued nature of time series correlations. Existing pattern mining approaches that address these challenges needs to be investigated for their use in time series data.

Chapter 5

Evaluating Reliability and Replicability of Transient Groups of Brain Regions

5.1 Introduction

Resting state functional connectivity has been widely studied in the neuroimaging community in the last decade [47, 48, 11, 49, 50]. Graph theoretic approaches that are widely used in analyzing resting state fMRI data treat brain regions as node and pairwise correlation between the time series from these nodes as edges in a functional network and study the properties of the functional network [11, 14, 46, 51, 52, 53, 54, 55, 56, 57, 44, 58, 13, 45]. These approaches have also been used to study differences between healthy and schizophrenia subjects. The underlying assumption these approaches make is that the functional connections are stationary, i.e., the strength of a functional connection does not change over the duration of the scan.

Chang and Glover [18] were among the first to report that the functional connections exhibit non-stationarity. In the last couple of years, several efforts have been made to capture the principles underlying non-stationarity of the functional relationship between brain regions. These studies generally fall into three categories: i) Dynamics between two nodes ii) Dynamics between more than two nodes iii) Dynamics at the level of a network.

Hutchison *et al.* [68] studied the non-stationarity in a selected set of functional connections in resting state fMRI data in humans and macaques. Handwerker *et al.* [69] found that there is periodicity in the strength of functional connection that changes with time. They also found that different connections exhibit different periodicities. Note that Hutchison *et al.* [68], Handwerker *et al.* [69] and Chang and Glover [18] have studied dynamics at the level of one connection (between two nodes) at a time.

Recently, Jones *et al.* [70] studied the hypothesis that at every instance brain exhibits modular architecture. They discovered modules in each time window and then they grouped these brain regions into four major categories based on their module membership. This approach captures the modularity at the level of individual sliding windows, but collapses this fine granular information into four groups of regions. Note that this approach studies dynamics at the level of modules generally involving more than two nodes.

Allen *et al.* [71] studied the hypothesis that the functional network transitions through various states over time. They clustered functional networks estimated in small time windows to shed light on the different brain states. Bassett *et al.* [3] showed that when a motor task is being learnt, the success of learning is associated with the core-periphery structure where the core is stationary and its relationship with the periphery is transient. This study explored the changes in the functional network with respect to a learning task. Note that Allen *et al.* [71] and Bassett *et al.* [3] explored dynamics at the level of a network.

The focus of this chapter is to study the dynamics that exists between two or more brain regions. One way to study dynamics that involve multiple brain regions is to discover all combinations that share similarities in at least a given number of time windows. This is illustrated with the help of a toy example in Figure 5.1. The time series data in this figure has 5 time series, corresponding to 5 hypothetical brain regions. Four combinations of these time series are shown in the figure. Consider that the time window of interest is equal to the shaded region in combination 1 and that there has to be at least one such time window for a combination to be interesting. Combination 1 has three time series (one blue and two green) that share similarity in the shaded interval. Combination 2 has three time series (one blue and two red) that share similarity in the shaded interval. Combinations 3 and 4 do not share any intervals of a length equal to the shaded intervals in combinations 1 and 2. Therefore, only combinations 1 and 2 are interesting given a fixed length interval and a constraint on the number of intervals. Note that the blue time series participates in combinations 1 and 2, i.e., it is similar to green time series in

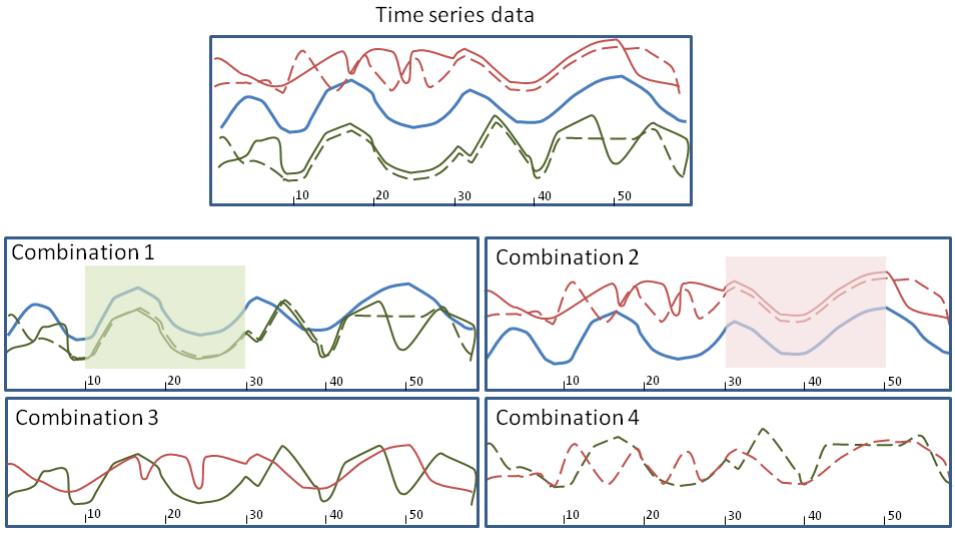

Figure 5.1: Example combinations of time series

one interval and it is similar to the red time series in a different interval. Jones *et al.*'s approach [70] can potentially find the combinations 1 and 2 as modules in the time windows where they share high similarity. However, when the group the brain regions into groups based on shared module membership the blue time series can only belong to one group. Due to this Jones *et al.*'s approach [70] cannot exhaustively find combinations of brain regions that share similarity in small intervals.

We have recently proposed a data mining approach to discover combinations of time series that share high similarity in multiple small intervals. In this chapter we study the utility of this approach in exhaustively discovering brain regions that share similarity in resting state data. We use multiple datasets: single subject multiple scans and group of subjects multiple scans for this analysis. We study reliability and replicability of these combinations in different settings. We also evaluate the utility of the combinations in explaining the differences in healthy and schizophrenia subjects. Our results suggest that the reliability and replicability of the combinations we find is significantly high. Our results also indicate that the reliability of the combinations we find in their discriminative power to separate schizophrenia cases from controls is also significantly high.

5.2 Methods

5.2.1 Data sets

Control study

i) Single subject resting state data (SSRS-T1 + SSRS-T2) 10 scans of 5 min duration were obtained from a healthy subject in five different sessions as part of the Anderson *et al.*'s [72] study. Data acquisition details were discussed in [72]. The preprocessing steps include motion correction and regression, coregistration to MPRAGE; segmentation of gray matter, white matter, and CSF; and normalization to MNI template brain, allowing removal by regression of motion; physiologic, CSF, white matter, and soft-tissue signals; bandpass filtering between 0.001 and 0.1 Hz; and linear detrend at each voxel in the brain. We use an Automated Anatomical Labeling (AAL) atlas that provides mapping from voxels to brain regions (see Table A.1 for a list of the brain regions). Using this we compute mean time series for each of the 90 brain regions in the atlas.

We append all the scans (10 scans x 5 sessions) to get 250 minutes of scan data. We split this data into two halves and refer to the first half as SSRS-T1 and the second half as SSRS-T2.

Group of subjects resting state data at T1 (GSRS-T1) + T2 (GSRS-T2) 6 minute resting scans were collected from 29 healthy subjects. 9 months later another set of 6 minute scans were collected from the same subjects. The preprocessing steps performed on these scans include motion correction, unwarping, 6mm spatial smoothing, voxelwise motion regression, registration to MNI space and High-pass temporal filtering (0.008 Hz).

We append the first set of scans to get a population level data of 174 minutes. As above, we use AAL atlas to get mean time series for 90 brain regions. We refer to this data as GSRS-T1. We repeat this with the data collected in the second set, i.e., 9 months after the first set was collected. We refer to this data as GSRS-T2.

Group of subjects S1 resting state data (GSRS-S1) + Group of subjects S2 resting state data (GSRS-S2) 6 minute scans were also collected from 62 healthy subjects. The same preprocessing steps that were used in GSRS-T1 and GSRS-T2 were used here. We append the set of 31 subjects to get a population level data of 186 minutes. AAL atlas was used to compute mean time series for 90 brain regions. We refer to this data as GSRS-S1.

Similar steps were repeated for the remaining set of 31 scans. We refer to the resultant brain region time series as GSRS-S2. Note that the preprocessing for all GSRS datasets was same and it is different from SSRS data.

Case-Control study

The preprocessing pipeline included deletion of first 3 volumes to account for magnetization stabilization; motion correction; B0 field map unwarping; slice-timing correction; non-brain removal; motion regression; registration to standard MNI space. (check if we used wavelets).

i) Healthy vs. Schizophrenia HRS-T1 vs. SRS-T1 6 minute resting state fMRI scans were obtained from 27 schizophrenia and 31 healthy subjects. The preprocessing pipeline included motion correction; B0 field map unwarping; slice time correction; non-brain removal; spatial smoothing (6mm); high pass filtering (50 s); registration to standard MNI space. We used AAL atlas to compute mean time series for each of the 90 brain regions for each scan. We append the 90 time series thus computed from each healthy subjects scan and refer to it as HRS-T1. Similar steps were used for scans from schizophrenia subjects. We refer to this data as SRS-T1.

ii) Healthy vs. Schizophrenia HRS-T2 vs. SRS-T2 6 minute resting state fMRI scans were obtained from the same set of 27 schizophrenia and 31 healthy subjects after 9 months. We process the data as above to get HRS-T2 and SRS-T2.

Quantifying dynamics

We first define the notion of a sliding window that we use in quantifying dynamics. Sliding window is characterized by its length and the step-size by which it is moved along the time axis. If the time series is of length 60 time points, using a sliding window of length 30 time points and a step-size of 10 points will result in four windows: 1 to 30, 11 to 40, 21 to 50, and 31 to 60. We first define the notion of ts-support that is useful in dynamics over a pair of time series and then generalize it to more than two time series. Given a pair of time series from two different brain regions, sliding window length and step-size, we first compute the correlations between the two time series for all possible sliding windows. We then compute the fraction of the sliding windows that have a correlation greater than a threshold (gamma). We refer to this quantity as ts-support. This is illustrated in Figure 2. In this chapter we used a sliding window

length of 15 time points (30 seconds) and a step-size of 5 time points (10 seconds). Ts-support measure can be generalized to more than two time series, by considering the minimum of the correlation among pairwise correlations between all time series within the window. The fraction of windows that have a minimum correlation greater than a threshold is treated as ts-support.

Discovering all combinations

We rely on a popular frequent pattern mining framework, Apriori, to explore the set of all possible combinations of brain regions effectively. This approach explores the combinations in a bottom-up fashion. Ts-support for pairs are estimated first and then interesting pairs (based on a ts-support threshold) are determined. Using the set of interesting pairs triples are enumerated and the interesting triples are determined. This approaches progresses towards bigger combinations in this fashion until no more bigger combinations satisfy the ts-support threshold. The algorithm is formally presented in Atluri *et al.* [73]. This is shown to outperform other competing techniques in capturing combinations effectively.

It is important to note that when a pair of brain regions (a,b) are correlated in all the windows (i.e., the connections is static and not dynamic) and if either of them is correlated with a different brain region (c) in a few windows, all three regions (a,b,c) will be found to be correlated in the windows where (a,c) are correlated. This is driven by static connection (a,b). To avoid such scenarios we ignored a combination if it consists of a pair of brain regions that are correlated in at least half of the windows.

Reliability and Replicability analysis

Reliability and replicability is non-trivial to evaluate on a set of combinations discovered by our approach due to the lack of control on the size of patterns. For example, a pattern (a,b,c) found in one dataset while a pattern (a,b,c,d,e) could be present in another dataset, where a,b,c,d, and e are hypothetical brain regions. In such cases, quantifying the similarity in discovered patterns is non-trivial. Therefore, in this analysis we study patterns of size 3 and we refer to such patterns as *triples*, henceforth. To study the reliability of triples at a group level we first appended time series from scans from all subjects at T1 and then we computed ts-support for all possible triples. We repeated the same process for T2 data. Using the ts-support in T1 and T2 data comparisons can be made in two different ways to assess reliability. The first approach is to select all triples

whose ts-support is $\geq 2\%$ in T1 and T2 data and then compute the overlap between them. The second approach is to select k triples with highest ts-support in T1 and T2, and then compute the overlap between the two sets. In both of these cases, we compared the computed overlap with that of an expected overlap that is obtained by randomly sampling the same number of triples as in the original case and computing overlap between them. Replicability analysis was performed similarly where we have sets S1 and S2. We appended the time series from all subjects in S1 and computed ts-support of all triples. We repeated this for S2. Using the two different approaches presented above we studied the replicability between S1 and S2. In this analysis we only consider those triples whose ts-support is higher than the maximum of the expected ts-support empirically computed after permuting the sliding window correlations among edges of the regions in a triple.

Reliability and Replicability of Discriminatory analysis

We studied the utility of the triples in discriminating between schizophrenia and healthy subjects. We appended the scans in HRS-T1 dataset and computed ts-support for triples in healthy population. Similarly, we computed the ts-support for SRS-T1 dataset in healthy population. We then computed the absolute difference between of ts-support for each triple in the two groups. We then ranked the triples in the descending oder of the absolute difference in ts-support. We repeated this for HRS-T2 and SRS-T2 datasets to get another order in the second dataset. We evaluate the overlap between the top-k triples in T1 and T2 datasets. We did similar analysis for ts-support in pairs and also for the correlation in pairs as is typically computed for brain networks when they are assumed to be stationary. We compared these three scenarios with each other and with respect to a corresponding null model.

5.3 Results

5.3.1 Pattern analysis

Using datasets SSRS-T1, GSRS-T1 and GSRS-S1, we discovered the patterns whose ts-support of 2% and several correlation thresholds 0.7, 0.75, 0.8 and 0.85. The number of patterns of various sizes discovered at each of the above parameters in the above datasets is shown in Figure 5.2 (a), (b), and (c). Note that as the correlation thresholds increase from 0.7 to 0.85

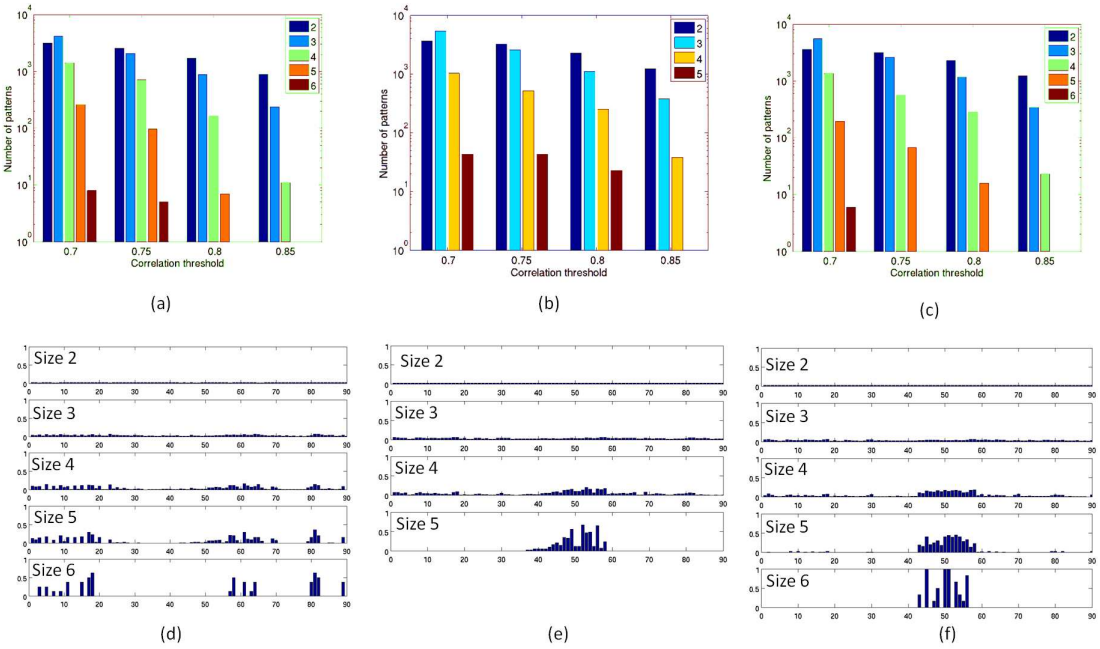

Figure 5.2: Distribution of patterns at different correlation threshold choices in three datasets.

the number of bigger patterns are not present, because a higher correlation threshold potentially reduces the number of windows that contribute to ts-support. The fraction of patterns (grouped by size) in which each brain region participated in is shown in Figures 5.2 (d), (e), and (f). The correlation threshold chosen here is 0.7, where there are bigger patterns. For patterns of size 2 and 3, all brain regions are generally equally represented. This can be seen with all the three datasets: SSRS-T1, GSRS-T1 and GSRS-S1. On the other hand, for bigger patterns, these fractions are significantly different for SSRS-T1 and GSRS-T1. For SSRS-T1 frontal, parietal and temporal regions are found to participate in patterns of size 5 and 6, while in GSRS-T1 and GSRS-S1 predominantly visual regions are found to participate in patterns of size 5 and 6. This difference in brain regions can be attributed to the fact that the comparison is in between single subject data and the group level data. Notice that the group level patterns found in GSRS-T1 and GSRS-S1 show similar brain regions, suggesting that these are population level patterns.

We also performed experiments by fixing the correlation threshold (0.8) and changing the ts-support threshold (2%, 4%, 6% and 8%). The distribution of patterns according to size at various thresholds and the brain regions participating in patterns of different sizes are shown

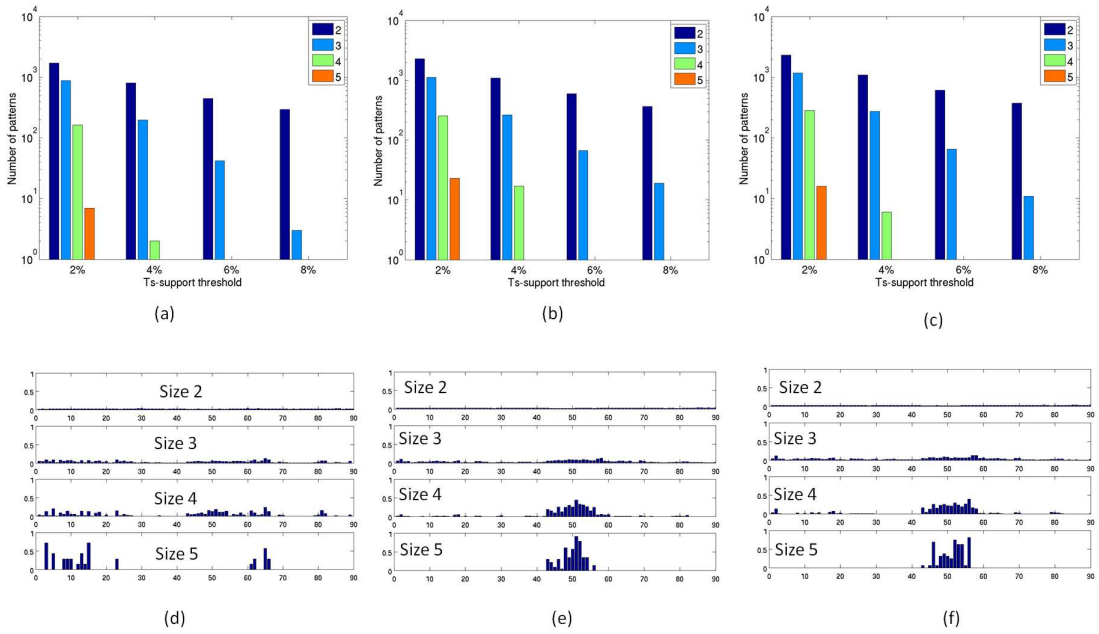

Figure 5.3: Distribution of patterns at different $ts - support$ choices in three datasets.

in Figure 5.3. As the ts-support threshold is increased the number of patterns tend to decrease, especially the bigger patterns tend to be non-existent. The difference in brain regions between single subject data and group level data are also seen here.

5.3.2 Sample patterns

Figure 5.4 shows two patterns in which a region Right Frontal Medial Orbital participates in. Figure 5.4(a) shows regions Right Frontal Medial Orbital (green), Left Frontal Superior Orbital (red), Left Rectus (blue) and Right Rectus (magenta). Figure 5.4(b) shows regions Right Frontal Medial Orbital (green), Left Frontal Superior Medial (red), Left Anteriori Cingulum (blue) and Left Posterior Cingulum (magenta). These two patterns are found in both SSRS-T1 and GSRS-T2 using a ts-support of 2% and a correlation of 0.7. The participating regions in each of these groups exhibit similarity in approximately 3% of the windows in SSRS-T1. However the two groups do not share any windows. This suggests that the Right Frontal Medial Orbital region is involved in a transient relationship with two other sets of regions independently.

Figure 5.5 shows two patterns in which two regions Left Rolandic Operculum (red), Left

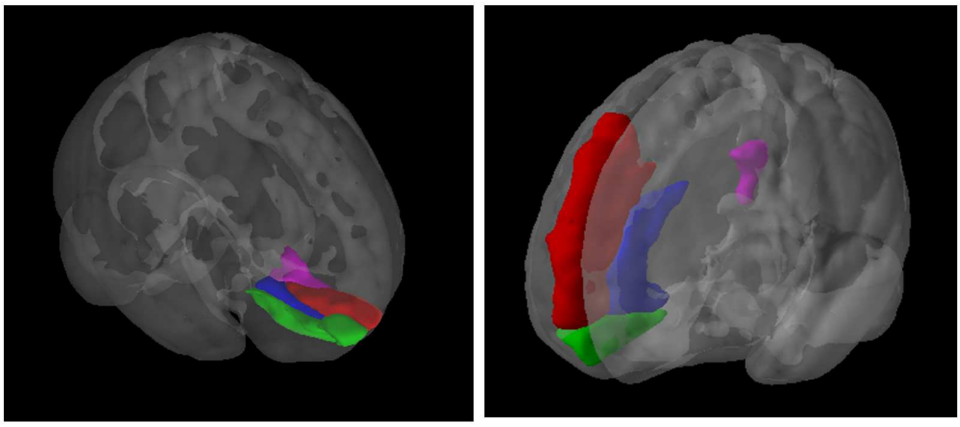

Figure 5.4: Two sample patterns that share the same region ‘Right Frontal Medial Orbital’ (green).

Post Central (green) participate in. Figure 5.5(a) shows regions Left Rolandic Operculum (red), Left Post Central (green), Left Supramarginal (blue) and Right Temporal Superior (magenta). Figure 5.5(b) shows regions Left Rolandic Operculum (red), Left Post Central (green), Left Paracentral lobule (blue) and Right Heschl gyrus (magenta). These two patterns are found in SSRS-T1 and GSRS-T2 using a ts-support of 2% and a correlation of 0.7. The participating regions in each of these groups exhibit similarity in approximately 2.7% of the windows. However the two groups share only 0.1% of the windows. This suggests that the rolandic operculum and the post central regions are involved in two different groups generally independent of each other.

These observations indicate that the transient relationships exist in groups of size greater than two and that the participating regions can be involved in more than one group.

5.3.3 Reliability results

In order to study the reliability we cannot directly compare the patterns because a pattern found in one dataset can be a subset of the pattern in another dataset. To quantify reliability and replicability we first choose a pattern size and then compared the triples that pass a ts-support threshold and a statistical significance threshold. In the following we show our results on patterns of size 3. We repeated this analysis on size 2 patterns and we found very similar observations.

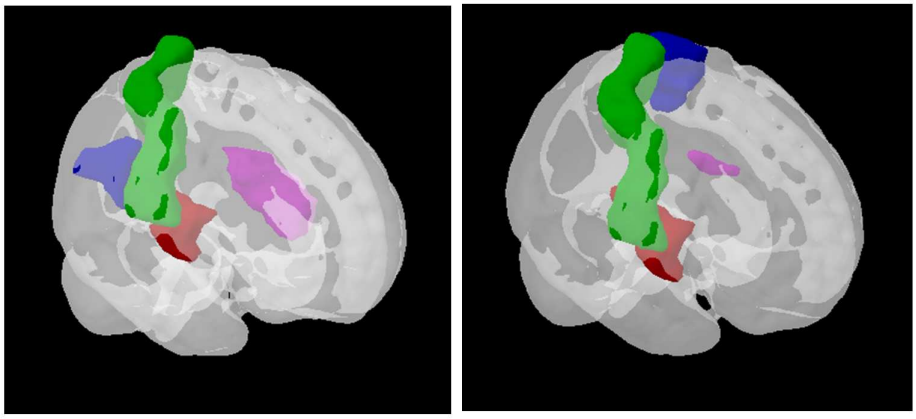

Figure 5.5: Two sample patterns that share two regions: Left Rolandic Operculum (red), Left Post Central (green).

We computed the ts-support measure for all triples among the 90 brain regions in the AAL atlas. Of all the possible triples we found those that have a ts-support of 2% or more. We also filtered out all the triples whose $p < 10^{-3}$.

There are 1325 triples in SSRS-T1 and 1270 in SSRS-T2. There are 1271 triples in GSRS-T1 and 1472 in GSRS-T2. We found that 936 triples are common between 1325 SSRS-T1 and 1270 SSRS-T2 triples. The likelihood of seeing such a huge overlap between two sets of triples due to random chance is very negligible ($p \leq 10^{-10}$). On the other hand, 969 triples are common between 1271 GSRS-T1 triples and 1472 GSRS-T2 triples. Even here the overlap is highly statistically significant ($p \leq 10^{-10}$). This suggests that there is a large agreement between the triples discovered in retest scans both at an individual level and at a population level. The number of triples and their overlap within and across datasets are shown in Figure 5.6.

5.3.4 Replicability results

We found 1398 triples in GSRS-S1 and 1466 triples in GSRS-S2 out of which 1050 triples are common. These 1050 triples can be treated as those triples that are generic in a population. Note that the GSRS datasets are obtained using the same imaging protocol and the data was processed using the same preprocessing pipeline. To study how different the triples vary across single subject and a group of subjects we computed the number of triples that are common between

	# Triples	# Common Triples
SSRS-T1	1325	936
SSRS-T2	1270	

(a)

	# Triples	# Common Triples
GSRS-T1	1271	969
GSRS-T2	1472	

(b)

	# Triples	# Common Triples
GSRS-S1	1398	1050
GSRS-S2	1466	

(c)

Set 1	Set 2	# Common Triples
SSRS-T1	GSRS-T1	564
SSRS-T2	GSRS-T2	649
SSRS-T1	GSRS-S1	595
SSRS-T2	GSRS-S2	629

(d)

Set 1	Set 2	# Common Triples
GSRS-T1	GSRS-S1	996
GSRS-T2	GSRS-S2	1038
GSRS-T1	GSRS-S2	975
GSRS-T2	GSRS-S1	1029

(e)

Figure 5.6: Reliability and replicability of the triples.

SSRS-T1 and GSRS-T1. We found that there are 564 triples that are common between and that this overlap is also statistically significant ($p \leq 10^{-10}$). Note that this overlap is smaller than that of 1050 seen between GSRS-S1 and GSRS-S2. The 564 triples that are common between SSRS-T1 and GSRS-T1 can be treated as the population level triples that are also seen in a single subject. The reduction in the number of common triples from GSRS-S1 and GSRS-S2 to SSRS-T1 and GSRS-T1 can also be attributed to the fact that these datasets are collected from different scanners and different preprocessing pipelines. We also found that the overlap between SSRS-T2 and GSRS-T2 is 649 ($p \leq 10^{-10}$). Additional comparisons are shown in Figure 5.6.

As noted earlier we found that the overlap between triples from two groups of subjects in GSRS-S1 and GSRS-S2 is 1050. We computed the overlap between GSRS-T1 and GSRS-S1 to be 996 ($p \leq 10^{-10}$) and the overlap between GSRS-T2 and GSRS-S2 to be 1038 ($p \leq 10^{-10}$). Note that these overlaps are similar in degree, suggesting the above observation that the difference in scanners, preprocessing pipelines and the population vs. individual differences can be lead to a significant reduction in the overlap in triples.

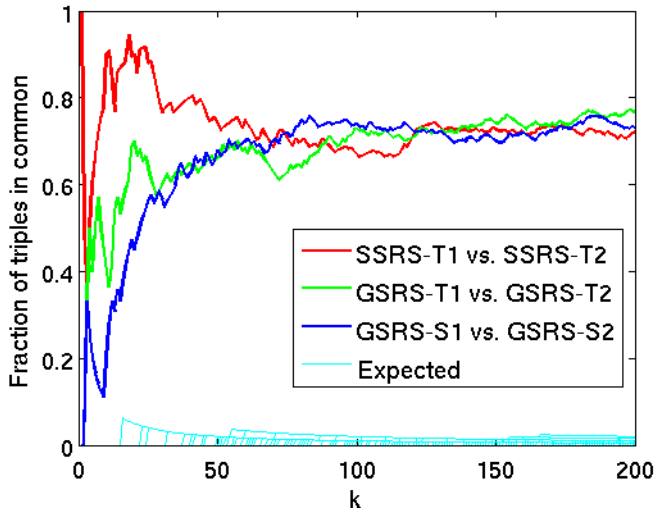

Figure 5.7: Common triples using three different datasets.

We also found that the overlap between all four sets of triples from GSRS-T1, GSRS-T2, GSRS-S1, and GSRS-S2 is 794 ($p \leq 10^{-10}$). This overlap is very highly statistically significant because even when four sets of triples are used to study the overlap the reduction from overlap of any two sets is small. Note that this overlap of 794 is more than that of 564 and 649 seen between SSRS and GSRS sets. Another approach to assess reliability and replicability is to find the overlap among top k triples in two sets of triples. Figure 5.7 shows these scores for various choices of k for different sets of triples. A null distribution is computed based on computing overlap between two arbitrarily ordered sets of triples. The real overlap curves are always higher than that of a random scenario suggesting that the overlaps are statistically significant.

5.3.5 Reliability and replicability using top-k triples

Another approach to assess reliability (T1 vs. T2) and replicability (S1 vs. S2) is to select the top-k triples with highest ts-support (and $p \leq 10^{-3}$) and to compute the overlap between the two sets. We used this approach to assess similarity in the triples between the pairs of data sets SSRS-T1 and SSRS-T2, GSRS-T1 and GSRS-T2, and GSRS-S1 and GSRS-S2 datasets. The fraction of the top-k triples that are common between the two sets are shown in Figure 5.7. We also computed an expected overlap between top-k triples in 1000 runs. Note that the expected overlap is much smaller than 0.1, whereas all the overlaps are much higher.

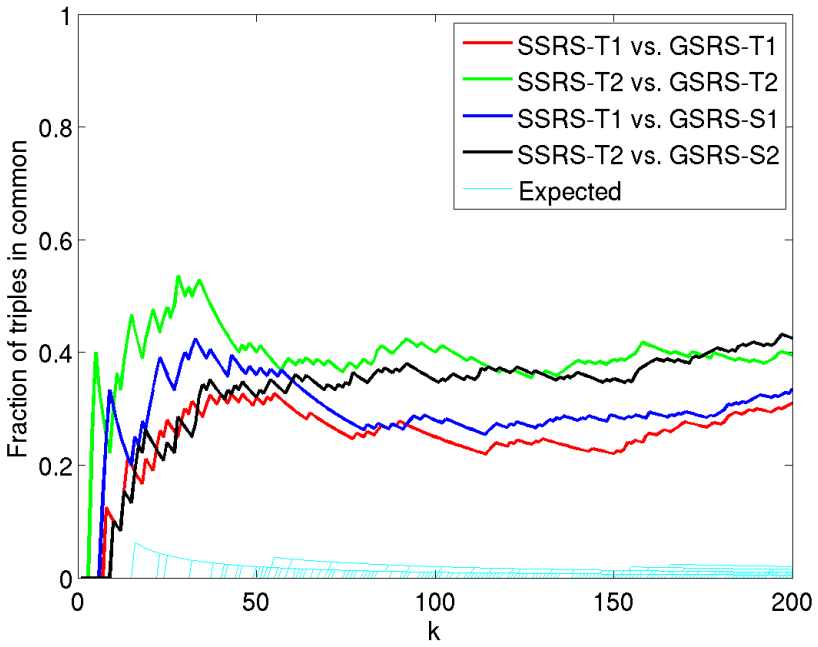

Figure 5.8: Common triples between single subject and multiple subject datasets.

We computed the overlap between single subject datasets (SSRS) and a group of subjects datasets (GSRS). The overlap curves and the expected overlap are shown in Figure 5.8. The overlap between these two sets is much smaller than that is seen in Figure 5.7. This observation is consistent with our earlier analysis of replicability of triples with ts-support $\geq 2\%$ between single subject and group of subjects data.

We computed the overlap among GSRS datasets. These curves are shown in Figure 5.9. As above, the real curves are much higher than the expected curves. The overlaps seen here are similar to that of Figure 5.7, suggesting that the population level triples are highly replicable.

5.3.6 Brain regions that participate in triples

The nodes that participate in the top-k brain regions are shown in Figure 5.10. In the single subject datasets, frontal, occipital and parietal regions appear to participate in triples. In data sets with multiple datasets occipital, temporal and some visual regions appear to participate. The difference in the single subject to multiple subject datasets could be attributed to the differences

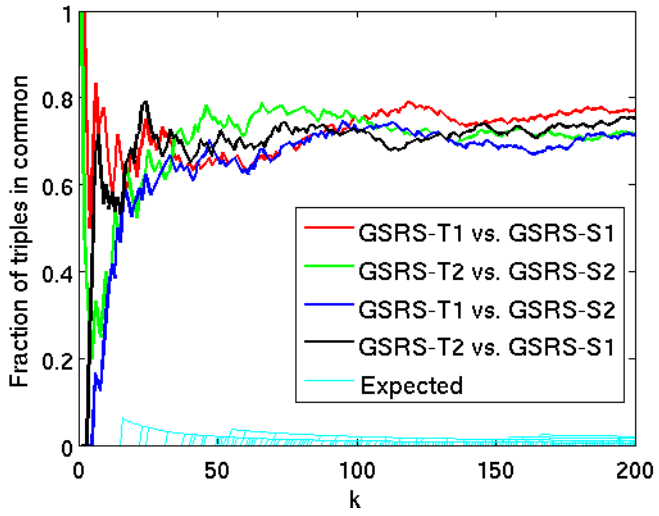

Figure 5.9: Common triples between different datasets with multiple subjects.

in scanning protocols, preprocessing pipelines and the differences in single and multiple subject scans.

5.3.7 Dynamics to explain difference between healthy and schizophrenia subjects

Similar to the GSRS-S1 and GSRS-S2 datasets we append the fMRI time series from 31 healthy subjects in HRS-T1 and then discovered the interesting triples. We compared the ts-support of the triples in time series obtained after appending fMRI data from 27 schizophrenia subjects (SRS-T1). We then ranked the triples based on the difference in ts-support in healthy and schizophrenia groups. We repeated this analysis on T2 data (HRS-T2 and SRS-T2) to get another ranking of triples. We studied the overlap in this ordering based on discriminatory nature of triples. This overlap curve is shown in Figure 5.11. In addition we provide overlap curves from discriminatory analysis of full time series correlations (static pairs) and ts-support of pairs (dynamic pairs). We compare these curves with their corresponding expected overlap curves.

All the real overlap curves in Figure 5.11 are higher than the expected curves. The dynamic pairs and dynamic triples curves are above the static pairs curve. This suggests that the replicability of discriminatory information in static connections is lower than that of dynamic connections. Note that the dynamic triples and dynamic pairs have a very similar overlap curves.

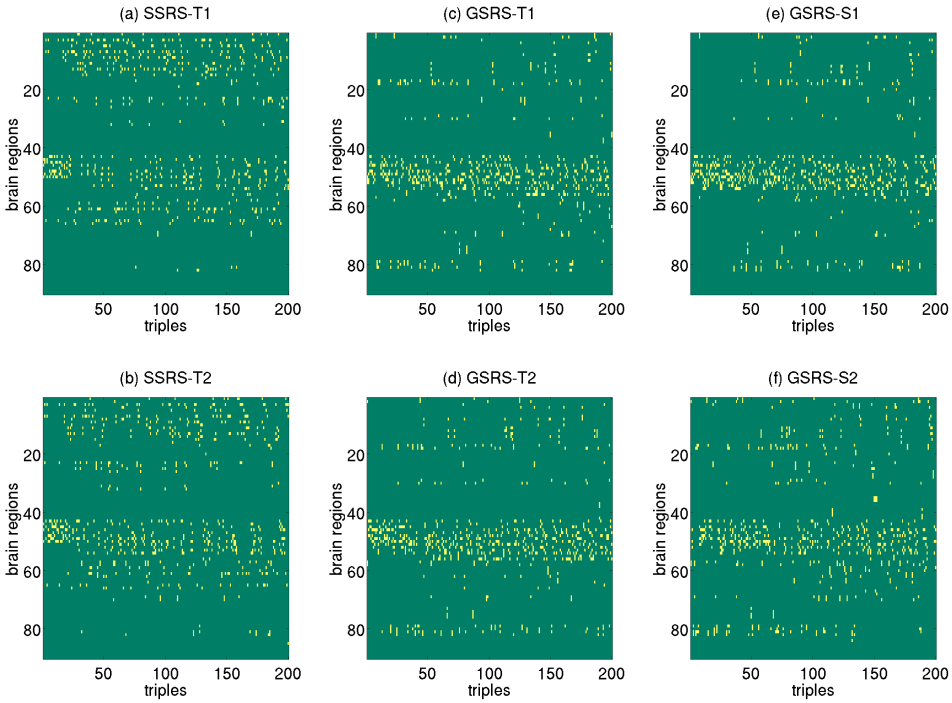

Figure 5.10: Nodes that participate in top-k triples.

However, note that the number of triples is 117,480 while the number of dynamic pairs is only 4005. Therefore, the statistical significance of overlap is triples is much more than that of the dynamic pairs.

5.4 Discussion

Patterns of brain regions that exhibit similar activity in multiple small intervals of time are studied in this chapter. We studied the nature of the patterns, their size and composition. In addition, we evaluated the reliability of the patterns at a single subject level and at a population level. We also studied the replicability of the patterns between two independent sets of samples. We found that the patterns are significantly reliable and replicable. We also found that there a significant amount of similarity in the patterns that are found in a single subject and a group of subjects. We studied the reliability of the patterns that are different in schizophrenia and healthy

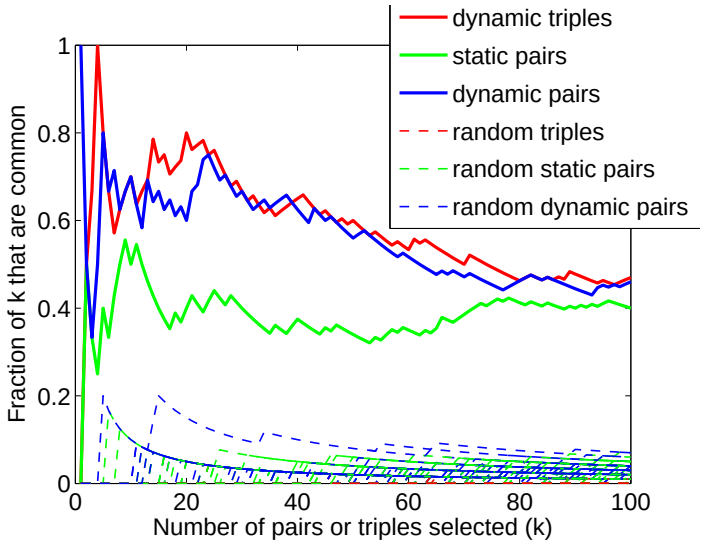

Figure 5.11: Common triples that are discriminative between healthy and schizophrenia samples.

population. Our findings suggest a statistically significant level of reliability in the patterns that are present to a different degree in schizophrenia and healthy populations.

Reliability of the patterns at a single subject and at a population level was significantly high. Recently there is an increased interest in the neuroimaging community in studying reliability in the several findings such as ICA components and connectivity differences between healthy and disease subjects. Although non-stationary connections have been studied earlier, no existing study has evaluated their reliability. Due to the non-stationary nature of relationships that are in question, the patterns are not generally to the reliable. However, our finding of significantly high reliability suggests that these patterns re-occur, indicating that some neurological principles could be driving these patterns.

Most of the patterns at a single subject level are a combination of brain regions from frontal, occipital and parietal areas. The population level patterns are predominantly a combination of occipital areas. There are some frontal and temporal regions that participate in some combinations but it is relatively rare. As the subjects are resting it is non-trivial to explain this phenomenon. However, we plan to study these patterns in the context of various tasks to be able to interpret the patterns.

Reproducibility of the patterns across different sets of samples is significantly high. This

has been studied in two perspectives: i) by considering all patterns of size-3 whose support is greater than 2%, ii) by considering patterns with the highest 'k' ts-support scores. The fact that a highly significant fraction of the patterns are reproducible suggests that they capture the general phenomenon happening in the brain. As above, we plan to study them in the context of various tasks to understand the role these patterns play in terms of brain's operations.

We found that there is also a significant overlap in triples between single subject and multiple subject data, while the overlap is smaller than that is seen between two independent samples. This suggests that some of the general patterns are not present in the single subject and that there are non-general patterns present in this subject. The implications of this can be further studied using behavioral data of the population and the single subject.

We studied overlap in the most discriminative triples across healthy and schizophrenia subjects. This overlap is not only statistically significant, but also higher than the overlap that is seen when functional correlations between brain regions are used to estimate the differences. This suggests that the patterns that capture transient relationships can potentially be useful in explaining the differences between healthy and disease subjects better than the relationships (correlations) computed assuming static or stationary relationships.

The limitations of our analysis are as follows. There are differences in acquisition protocols and preprocessing pipelines used in the single subject data and the population level data. While the overlap between the two datasets is statistically significant it is smaller than the overlap between independent sets of samples. This difference could potentially be due to the difference in acquisition protocols and preprocessing pipelines. Another limitation is that in the our discriminative analysis healthy and schizophrenia data, we could only compute the ts-support at the population level and not the single subject level due to the relatively small amount of data available in each scan. Longer scans will allow us to analyze each subject individually and to build a classification model that can be used to predict cases from controls. We used the notion of brain regions from the automated anatomical labeling (AAL) atlas. One could potentially use a different atlas or a data driven brain regions for this analysis. The impact of this choice needs to be studied. In our definition of ts-support, we do not tolerate noise, that is common in fMRI data, when we assess the similarity of time series using correlation. Another limitation is that the interpretation of the discovered patterns is non-trivial. A suitable visualization tool can be useful in reducing the complexity in the pattern space for the purpose of interpretation.

5.5 Conclusion

Statistically significant reliability and replicability of the patterns is indicative of the realistic nature of brain's regions to be functionally related intermittently. It is important to note that not all brain regions interact to the same degree. The Ts-Apriori approach discussed in the previous chapter has the potential to discover patterns that share a similarity beyond a given degree. The approach also allows one to discover the patterns that a given node participates it. This allows for a more targeted study of the combinations. There are also a few limitations with this approach that are discussed above. Addressing those limitations can advance the state-of-the-art in understanding the brain dynamics and its relevance to brain functionality.

Chapter 6

Discovering the longest set of non-overlapping maximal intervals from time series

6.1 Introduction

Complex dynamic systems are widely believed to achieve a desired function due to synergy between different components. For example, the human brain interprets an observed visual stimulus with the help of synergy between the frontal and parietal regions in the brain [74]. Here the parietal region processes the visual stimulus and it is interpreted by the frontal region. This synergy is typically reflected in high similarity (e.g., high correlation) in the activity time series measured from these two regions during the time interval when the brain is interpreting a visual stimulus. Discovering these highly similar intervals will be useful for many reasons. One can characterize the relationship between brain regions based on the frequency and the duration of such intervals. A comparison of characteristics of these intervals during rest and performance of a task has the potential to shed light on transient synergistic relationships in the human brain that are necessary for achieving a specific task.

To discover such transient relationships, similarities in time series are to be found in certain intervals and not in the entire duration of the time series. There can be many ways in which similarities in multiple time intervals can be defined. For example, given two time series one

may be interested in finding the longest interval where the two time series are similar [75]. Another example, is one where one may be interested in finding the number of intervals of a given length (time windows) in which two time series are similar [73]. As yet another example, one may be interested in finding those intervals of similarity that can be used to cluster [26] or classify time series [76, 77] effectively.

One of the objectives of this chapter is to discover all maximal correlated intervals, i.e., intervals where a given pair of time series are highly correlated (greater than a threshold) for every possible subinterval and no subsuming intervals are highly correlated. Figure 6.1 illustrates this with the help of an example that consists of two synthetically generated time series. The bold regions in the time series shown in Figure 6.1(a) indicate the time intervals where similar intervals are imputed in the two time series. These intervals are [31 80], [111 140], and [151 175]. Note that these intervals are very different in their duration: 50, 30, and 25, respectively. In Figure 6.1(b) all maximal correlated intervals (correlation ≥ 0.9) of length 4 or more are shown with double sided arrows where the left and right arrows indicate start and end points.

Sliding window based approaches have been widely used in scenarios when transient relationships between time series are of interest [78, 76, 79, 73]; however, they are not suited for addressing our objective. For example, one could use a sliding window of fixed length (say w) and find k consecutive windows that meet the correlation threshold (say β). But there is no guarantee that every subinterval (of size smaller or larger than w) in the interval of size $w+k-1$ satisfies the correlation threshold β. In this chapter we present an efficient approach to discover all maximal correlated intervals.

Once the set of all maximal correlated intervals are discovered from a given pair of time series, our next objective is to find the longest set of non-overlapping maximal correlated intervals. In the example shown in Figure 6.1(b), the two intervals [151 154] and [175 178] of length 4 shown in cyan overlap with a longer interval [154 175]. Overlapping maximal intervals such as these typically exist in real-world data and discovering the longest set of non-overlapping intervals will provide a realistic estimate of the total duration in which two time series exhibit high correlations. The set of intervals $\{$[31 82], [111 141], and [154 175]$\}$ shown in red in Figure 6.1(b) form the longest set of intervals. Note that these intervals very closely capture the imputed intervals.

In this chapter we define the notion of the Longest set of non-overlApping Maximal correlated INtervAls (LAMINA) that has the ability to capture distinct maximal correlated intervals

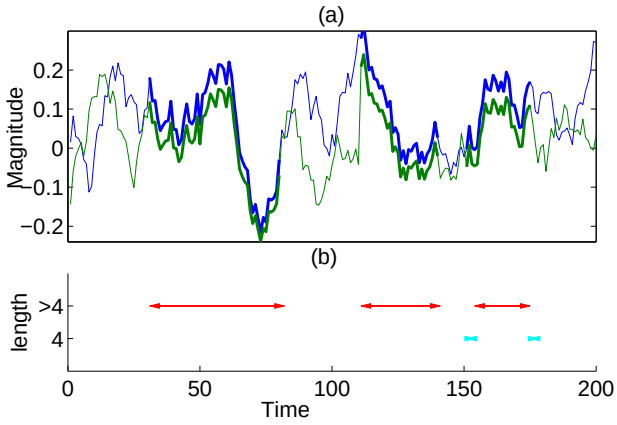

Figure 6.1: An illustrative example to demonstrate correlated intervals in time series: (a) Time Series (b) Maximal correlated intervals.

in a pair of time series such as those shown in red in Figure 6.1(b). We show, empirically, that this notion has the ability to capture the desired distinct intervals using a synthetic dataset. We provide an efficient approach to solve the LAMINA problem. The proposed notion of LAMINA can be broken down into two different subproblems: i) Finding maximal correlated intervals and ii) Discovering non-overlapping intervals such that their total length is the longest. Our proposed approach addresses the first problem using a bottom-up approach to discover all maximal correlated intervals. This allows us to prune the space of all intervals that is typically explored in a brute force approach. Using a dynamic programming approach we then discover the longest set of non-overlapping maximal correlated intervals.

We evaluate our approach to study its efficiency and effectiveness in capturing the desired correlated intervals on synthetic datasets at different parameter settings. Our results suggest that the proposed approach is very efficient compared to a brute force approach. Our effectiveness based evaluation on a synthetic dataset suggests that LAMINA based formulation has the ability to capture all imputed similar intervals in given a pair of time series.

We use the proposed approach on a real-world neuroimaging data that has activity time series collected from several brain regions in a subject in two different scenarios: i) while the subject is resting and ii) while the subject is watching cartoons. We found pairs of brain regions that have many correlated intervals in cartoons data and only a few correlated intervals in resting data. These pairs of brain regions indicate the synergies that are necessary to visualize and interpret cartoons. We also found pairs of brain regions that have significantly longer correlated

intervals than others while the subject is resting and while watching cartoons. Overall, our results highlight the utility of the proposed approach in capturing distinct correlated intervals with which one can characterize the task-related and transient synergistic relationships between different brain regions. Even though we present our evaluation on neuroimaging data, this approach has applicability in multiple domains (e.g., climate, stock-market data analysis).

The key contributions of this chapter are as follows:

- A problem formulation to capture distinct correlated intervals in a pair of time series.
- An efficient algorithm to discover the longest set of non-overlapping maximal correlated intervals.

The rest of this chapter is organized as follows. Section 2 presents our problem formulation. Section 3 discusses related work. The proposed method is presented in Section 4. Section 5 presents our evaluation and results. We conclude with Section 6.

6.2 Problem Formulation

Let X and Y be two time series of equal length n. We indicate the value of X at time i as X_i.

Definition 1: *Interval.* The set of time points $(i \ldots j)$, where $i < j,\ i \geq 1$ and $j \leq n$, is referred to as an interval $I_{(i,j)}$.

The length of an interval $I_{(i,j)}$, denoted using $l_{(i,j)}$, is the number of time points covered by the interval and is computed as $l_{(i,j)} = j - i + 1$. The set of values in a time series X in the interval $I_{(i,j)}$ is represented using $X_{(i,j)}$.

Definition 2: *Non-overlapping Intervals.* Two intervals $I_{(a,b)}$ and $I_{(c,d)}$ are said to be non-overlapping when ($a < c$ and $b < c$) or ($c < a$ and $d < a$), i.e., when the two intervals do not share any time points.

Definition 3: *Correlated Interval.* Given two time series X and Y an interval $I_{(a,b)}$ is referred to as a correlated interval if (i) length $l_{(a,b)} \geq \alpha$, (ii) Pearson's correlation between $X_{(a...b)}$ and $Y_{(a...b)}$, $r(X_{(a...b)}, Y_{(a...b)})$, exceeds a user provided threshold β, and (iii) for all subintervals $I_{(a',b')}$ of length $l_{(a',b')} \geq \alpha$, $a' \geq a$, $b' \leq b$, $r(X_{(a'...b')}, Y_{(a'...b')}) \geq \beta$.

The first constraint that ensures that every interval is longer than α is useful to avoid spurious small intervals that may exhibit high correlations due to random chance. The third constraint is useful to ensure that an interval that satisfies the correlation threshold β does not contain

an interval with a correlation lower than the threshold. For example, an interval [111 175] in Figure 6.1(a) has a correlation of 0.98, while the correlation in the interval [141 155] is only 0.56. Therefore, any interval containing [141 155] cannot be a correlated interval when $\beta > 0.56$ is used.

Definition 4: *Maximal Correlated Interval.* A correlated interval that is not subsumed by an immediately larger correlated interval is treated as a maximal correlated interval. Formally, $l_{(a,b)} \geq \alpha, r(X_{(a,b)}, Y_{(a,b)}) \geq \beta$ is a correlated interval and is maximal when $r(X_{(a-1,b)}, Y_{(a-1,b)}) < \beta$ and $r(X_{(a,b+1)}, Y_{(a,b+1)}) < \beta$.

Note that two adjacent intervals $I_{(a,b)}$ and $I_{(a+1,b+1)}$ can be maximal correlated when there is no immediately subsuming correlated interval $I_{(a,b+1)}$. These adjacent intervals overlap largely. To be able to capture distinct correlated intervals in such scenarios, we define the notion of non-overlapping maximal correlated intervals.

Definition 5: *Non-overlapping Maximal Correlated Intervals.* Given two time series X and Y and the set of all maximal correlated intervals $I = \{I_{(a,b)}, I_{(c,d)}, \ldots I_{(k,l)}\}$, non-overlapping intervals are those subset of intervals in which each interval is a maximal correlated interval and every pair of intervals are non-overlapping.

Definition 6: *Longest Set of Non-overlapping Maximal Correlated Intervals.* The set of non-overlapping maximal correlated intervals whose sum of intervals is the largest is referred to as the 'longest set of non-overlapping maximal correlated intervals', also referred to as LAMINA.

In Section 4 we provide an efficient approach to discover a LAMINA from a pair of time series.

6.3 Related work

Although subsequence based similarity in time series data was studied earlier [75, 80, 81, 82, 76, 77, 26], to the best of our knowledge there is no existing approach to discover the largest set of non-overlapping maximal correlated intervals from a pair of time series.

One recent paper by Li et al.[75] is relevant but has a different goal. They address the problem of discovering time series that share a longest correlated subsequence with a given query time series. Note that they look for one longest correlated interval between a query and a target sequence while our goal is to find a set of correlated intervals such that collectively

their duration is the longest. The interval [111 175] of length 65 with a correlation of 0.98 in Figure 6.1(a) will be discovered as the longest correlated subsequence. While we consider this as a spurious correlated subsequence, our goal is to discover the longest set of distinct maximal correlated intervals shown in red Figure 6.1(b). Hence, the problem at hand is different from the one addressed in Li et al.[75].

Another related work is by Das et al [83] where they propose a dynamic programming based approach to find the longest set of potentially non-contiguous indices in time series X and Y where the two time series are linearly related (i.e, $Y_{tr} = aX_{ts} + b$). Note that the selected indices in X and Y should have the same relative ordering and need not be the same. The lack of identity between indices in X and Y, the contiguity (interval length α) constraint, and the notion of spuriousness in correlated intervals in our problem are the main differences with the problem addressed in Das et al [83]. Moreover, their focus is on selecting time points that contribute to the linear relationship and not the intervals that are the focus of this chapter.

6.4 Proposed methods

The problem of finding the longest set of non-overlapping maximal correlated intervals can be dealt with in two parts. First, the set of maximal correlated intervals can be enumerated. Second, the longest set of non-overlapping maximal correlated intervals can be discovered from the set of maximal correlated intervals. In this section we first show a brute force method to enumerate all the maximal correlated intervals and we propose an efficient method for this problem. We present a dynamic programming based solution for discovering the longest set of non-overlapping maximal correlated intervals for the latter part of the problem.

6.4.1 Discovering Maximal Correlated Intervals

Brute-force Approach

The problem of discovering maximal correlated intervals can be divided into three subproblems: i) enumerating all intervals that satisfy the correlation threshold, ii) pruning correlated intervals whose subintervals are not correlated, and iii) discovering maximal correlated intervals among them.

Algorithm 2 BruteForceIntervalEnumeration

Input:
i. X and Y, two real valued time series of length n
ii. α, interval length threshold
iii. β, correlation threshold
Output:
$Intvl\_Mat$ where a 1 in element (i, j) indicates that interval $I_{(i,j)}$ is of length atleast α and has correlation atleast β

1. $Intvl\_Mat =$ zero matrix of size $n \times n$
2. $Corr\_Intvls = \phi$
3. **for** $win\_len = \alpha \; to \; (n-1)$
4. **for** $i = 1 \; to \; (n - win\_len + 1)$
5. $r = corr(X_{(i,i+win\_len-1)}, Y_{(i,i+win\_len-1)})$
6. **if** $(r \geq \beta)$
7. $Intvl\_Mat[i, i + win\_len - 1] = 1$
8. **end if**
9. **end for**
10. **Exit** if no intervals with correlations $\geq \beta$ are found for this win_len
11. **end for**
12. Return $Intvl\_Mat$

Enumerating All Intervals First, let us consider the problem of finding all intervals that satisfy the correlation threshold for two time series X and Y. Note that the definition of correlated intervals entails three constraints: i) minimum length of interval (α) ii) minimum correlation (β) iii) condition to avoid intervals whose subintervals are not correlated. In this part we will only address the first two constraints and the third constraint will be addressed in the next part.

A brute force approach enumerates each interval of valid length ($\geq \alpha$) and tests if the correlation threshold β is satisfied. From the first constraint, the candidate intervals are all intervals of length α that potentially start at every time point. For a chosen interval length l, the number of valid intervals to consider in a time series of length n are $n - l + 1$. Therefore the total number of valid intervals of all valid lengths are $\sum_{l=\alpha}^{n} n - \alpha + 1$, i.e., $O(n^2)$ and computing correlation for an interval of length l takes $O(l)$ time. Note that l can potentially approach n when there are longer correlated intervals. Overall, the computational complexity of brute force enumeration is $O(n^3)$.

The brute force approach for enumerating all correlated intervals is shown in Algorithm 2. $Intvl\_Mat$ is a two-dimensional array where a value 1 will be placed in element (i, j) when the

interval $I_{(i,j)}$ of length at least α and has correlation at least β and 0 otherwise. The algorithm iterates over all valid interval lengths in steps 3-11 and over all possible intervals of a chosen length win_len in steps 4-9. The correlation is tested in step 6 and $Intvl\_Mat$ is updated in step 7. Note that the algorithm exits (in step 10) when no intervals are found to be correlated for a given interval length.

There can be two types of irrelevant correlated intervals that are discovered by Algorithm 2: i) correlated intervals whose subintervals of length at least α are not correlated intervals. ii) correlated intervals that are not maximal. In the following we address each of these issues separately.

Pruning Spurious Intervals Once all the intervals that satisfy the correlation threshold are enumerated, in order to address the third constraint in the definition of *correlated intervals* we prune those intervals whose subintervals are not correlated. The pruning step is performed for each interval that is considered in step 5 with correlation $< \beta$, by listing all the enclosing intervals and filtering out all the enclosing intervals with correlation $\geq \beta$. This approach is shown in Algorithm 3. Steps 1-10 iterate over all valid interval lengths and steps 2-9 iterate over all possible intervals of a given length (win_len). For a chosen interval $(i, i + win\_len - 1)$ all enclosing intervals are determined in step 4 and their corresponding values in $Intvl\_Mat$ are set to 0 to suggest that they are not valid correlated intervals. The computational complexity of the $PurningSpuriousIntervals$ is $O(n^2)$.

Enumerating Maximal Intervals The goal of this part is to enumerate the maximal correlated intervals. This can be achieved by testing for every correlated interval if an immediately subsuming interval is also correlated. If no subsuming interval is correlated, a correlated interval can be enumerated as a maximal interval. This approach is shown in Algorithm 4. Steps 1-9 iterate over all valid interval lengths and steps 2-8 iterate over all possible intervals of a given length win_len. For a chosen interval $(i, i + win\_len - 1)$ if the correlation threshold is satisfied $(Intvl\_Mat[i, i + win\_len - 1] = 1)$ and if its immediately enclosing intervals did not satisfy the correlation threshold (step 4) then interval $I_{(i,i+win\_len-1)}$ is added to the list of maximal correlated intervals. The computational complexity of the $ListMaximalIntervals$ is $O(n^2)$.

Algorithm 3 PruneSpuriousIntervals

Input:
$Intvl\_Mat$ that encodes all correlated intervals
Output:
$Intvl\_Mat$ where a 1 in element (i, j) indicates that interval $I_{(i,j)}$ is of length at least α and has correlation at least β and all smaller intervals of length α or more are also correlated intervals

1. **for** $win\_len \ = \ \alpha \ to \ (n-1)$
2. **for** $i \ = \ 1 \ to \ (n - win\_len + 1)$
3. **if** $(Intvl\_Mat[i, i + win\_len - 1] = 0)$
4. $enclosing\_intvls =$ list all enclosing intervals
5. **for** all enclosing intervals (a, b)
6. $Intvl\_Mat[a, b] = 0$
7. **end for**
8. **end if**
9. **end for**
10. **end for**
11. Return $Intvl\_Mat$

A Bottom-up Approach

The brute force approach first enumerates all intervals of length α and longer that satisfy the correlation threshold β. It then prunes out spurious as well as non-maximal correlated intervals. Here we propose a relatively efficient approach (as we show in our evaluation section) that does not compute correlations for all intervals to determine correlated intervals.

Enumerating Correlated Intervals The third constraint in the definition of a correlated interval is that all subintervals of length α or more are also required to have a correlation $\geq \beta$. One way to address this constraint would be to start with all intervals of length α and then build bigger intervals only when immediately smaller intervals satisfy the correlation threshold. Using this observation, we propose a bottom-up enumeration scheme where we compute correlation of a longer interval only when both the immediate sub-intervals are correlated. Formally, if intervals $I_{(a,b-1)}$ and $I_{(a+1,b)}$ satisfy the correlation threshold, only then the correlation of the interval $I_{(a,b)}$ is evaluated. Using this procedure has two advantages: i) we do not have to evaluate correlations for all the candidate intervals, and ii) we can avoid the pruning spurious intervals step in the case of the brute force approach (shown in Algorithm: PruneSpuriousIntervals). The computational complexity of this approach is also $O(n^3)$. However, the number

Algorithm 4 ListMaximalIntervals

Input:
$Intvl\_Mat$ that encodes all correlated intervals
Output:
$Corr\_Intvls$ List of all maximal correlated intervals

1. **for** $win\_len = \alpha \; to \; (n-1)$
2. **for** $i = 1 \; to \; (n - win\_len + 1)$
3. **if** $(Intvl\_Mat[i, i + win\_len - 1] = 1)$
4. **if** $(Intvl\_Mat[a, b] = 0)$ for all
4. enclosing intervals (a, b)
5. $Corr\_Intvls = Corr\_Intvls \cup I_{(j,i+j+win\_len-1)}$
6. **end if**
7. **end if**
8. **end for**
9. **end for**
10. Return $Corr\_Intvls$

of correlated intervals found using the brute-force approach are lower bounded by the number of correlated intervals discovered using this bottom-up approach. When most of the candidate intervals are correlated, the number of correlations evaluated by the bottom-up approach will approach the number of correlations evaluated by the brute-force approach.

The bottom-up interval enumeration is shown in Algorithm 5. Similar to brute force enumeration scheme it iterates over all valid interval lengths and over all possible intervals for a chosen interval length in steps 2-19 and 3-18, respectively. When the interval length is the least possible (α), the correlations are computed for all the intervals. For other interval lengths l, correlation is estimated only if the two smaller subintervals of length $l-1$ were found to have a correlation of at least β (steps 11-15). Note that the algorithm exits when no intervals are found to be correlated for a given interval length.

Enumerating Maximal Intervals Once the correlated intervals are determined in a bottom up fashion the maximal intervals can then be enumerated by listing only those intervals whose immediate enclosing intervals are not correlated. Formally, if $I_{(a,b)}$ is correlated and neither of $I_{(a-1,b)}$ and $I_{(a,b+1)}$ are correlated, $I_{(a,b)}$ is listed as a maximal correlated interval. Algorithm 4 (ListMaximalIntervals) does this.

Algorithm 5 $BottomUpIntervalEnumeration$

Input:

i. X and Y, two real valued time series of length n

ii. α, interval length threshold

iii. β, correlation threshold

Output:

A matrix $Intvl\_Mat$ indicating all correlated intervals with 1's

1. $Intvl\_Mat[1 : num\_win, 1 : num\_win] = 0$
2. **for** $win\_len \ = \ \alpha \ to \ n - 1$
2. **for** $i \ = \ 1 \ to \ (n - win\_len + 1)$
5. **if** $(win\_len \ = \ \alpha)$
6. $r \ = \ corr(X_{(i,i+win\_len-1)}, Y_{(i,i+win\_len-1)})$
7. **if** $(r \ \geq \ \beta)$
8. $Intvl\_Mat[i, i + win\_len - 1] = 1$
9. **end if**
10. **end if**
11. **if** $(win\_len \ > \ \alpha$
11. $and \ Intvl\_Mat[i, i + win\_len - 2] = 1$
11. $and \ Intvl\_Mat[i + 1, i + win\_len] = 1)$
12. $r \ = \ corr(X_{(i,i+win\_len-1)}, Y_{(i,i+win\_len-1)})$
13. **if** $(r \ \geq \ \beta)$
14. $Intvl\_Mat[i, i + win\_len - 1] = 1$
15. **end if**
16. **end if**
17. **end for**
18. **Exit** if no correlated intervals are found for this win_len
19. **end for**
20. Return $Intvl\_Mat$

6.4.2 Discovering the Longest Set of Non- overlapping Maximal Correlated Intervals

Given a set of potentially overlapping intervals, the goal is to find the longest set of non-overlapping intervals. This problem can be treated as the classical dynamic programming problem of weighted interval scheduling [84] where the objective is to determine a schedule such that no two scheduled jobs overlap in time and the entire schedule maximizes the sum of weights of scheduled jobs. Therefore, by treating the each maximal correlated interval as a job and its length as the weight of the job, we can use the standard dynamic programming algorithm

to find the desired longest set.

The dynamic programming algorithm for finding the LAMINA is shown in Algorithm 6. A brief account of this approach is provided here and an interested reader is referred to [84]. This approach first sorts the intervals in ascending order of their start times. It then starts at the end of the list of intervals and traces back to the first interval by determining at each step the optimum weight from the current interval to the end of the list of intervals by considering two choices: i) include the current interval ii) ignore the current interval. Once the optimal weight for the entire list of intervals is determined, one can trace from the first interval to the last to determine if an interval was included in the optimal list. The computational complexity is $O(n \log n)$.

In summary, the brute force approach uses algorithms BruteForceIntervalEnumeration, PruneSpuriousIntervals , ListMaximalIntervals, and FindLAMINA to determine LAMINA, given a pair of time series and parameters (α and β). The bottom-up approach uses algorithms BottomUpIntervalEnumeration, ListMaximalIntervals, and FindLAMINA to determine LAMINA. Note that both these approaches share the last two algorithms. The key difference is in how the intervals are enumerated. Due to the bottom-up style of enumeration the latter approach reduces the search space and it does not have to filter out spurious correlated intervals. Although both these approaches have a complexity of $O(n^3)$, as we will show in our evaluation, the bottom-up scheme is practically much faster.

6.4.3 Proof of correctness

We now prove that the proposed bottom-up approach discovers the longest set of non-overlapping maximal correlated intervals. The proposed approach relies on three components: i) Bottom-up enumeration approach ii) List Maximal correlated intervals iii) Discovering the longest set using a dynamic programming approach. The second part is a filtering step where all non-maximal intervals are filtered out and the dynamic programming approach in the third part is proven to be correct [84]. So, we are left with proving that the bottom-up enumeration approach lists all intervals and their sub-intervals of length at least α are correlated.

Theorem 6.4.1 *Bottom-up interval enumeration algorithm lists all intervals and their sub-intervals of length at least α when they pass the correlation threshold β.*

Proof This can be proved in two parts. i) When an interval is enumerated all subintervals of length at least α are correlated. ii) All such intervals are enumerated.

We now prove the first part.

This part has two scenarios: When an interval is enumerated it has: a) either no subintervals of length α or b) all its subintervals of length at least α are correlated. The algorithm estimates the correlation of intervals of length α and enumerates it when the correlation threshold is satisfied. Hence the former scenario is addressed. Correlation for an interval $I_{(a,b)}$ is computed only when its subintervals $I_{(a,b-1)}$ and $I_{(a+1,b)}$ (where $b - a - 1 \geq \alpha$) are found to be correlated. Therefore, all subintervals starting from length α to $b - a - 1$ are bound to be correlated due to this bottom-up style of enumeration. This addresses the second scenario.

We now prove the second part.

This can be proved by contradiction. Let us assume that there is a correlated interval $(I_{a,b})$ and is not enumerated. The bottom-up approach estimates the correlation of all intervals of length α and enumerates those that satisfy the correlation threshold. Therefore, $(I_{a,b})$ cannot be of length α. For any bigger intervals $I_{(a,b)}$ ($(l_{a,b}) > \alpha$) that is not enumerated, it is is possible that a subinterval $I_{(a',b')}, l_{(a',b')} \geq \alpha$ does not satisfy the correlation threshold. Therefore, $(I_{a,b})$ is not a correlated interval as it defies the third constraint and its contradicts the original assumption. Therefore, all correlated intervals are enumerated by the algorithm.

6.5 Evaluation and results

In this section we present an evaluation of the proposed approach on synthetic and real datasets. We compared the efficiency of the proposed approach with a brute force approach. We also compared the effectiveness of the proposed approach in identifying a set of arbitrarily long imputed correlated intervals from a synthetic dataset. We also show the utility of our approach on a real-world neuroimaging dataset.

6.5.1 Efficiency comparison

In order to study the efficiency of the proposed bottom-up approach in comparison to the brute force approach we generated synthetic datasets with varying lengths of correlated intervals.

Synthetic data We first created two vectors of length 1500 whose values are sampled from a uniform distribution with range [0 1]. Each vector is now smoothed by computing each value as the average of preceding 5 values and succeeding 5 values. Following this smoothing the two

vectors are now treated as a time series with temporal auto-correlation, i.e., consecutive values in a time series often have highly similar values. We marked the duration of the time series with consecutive intervals of length that is sampled uniformly between 41 and 80 time points. Starting with the first interval every alternative interval is treated as a 'synchronous' interval and the remaining intervals are treated as 'asynchronous' intervals. For the 'synchronous' intervals the values in the first time series are copied to the second time series by adding a small amount of Gaussian noise ($\mu = 0, \sigma = 0.01$). We then created 100 such pairs of synthetic time series where synchronous intervals are of length between 41 and 80. We refer to this set as TS_{41-80}. Using the same approach we created two sets with 100 pairs of time series with synchronous intervals of length from 101 to 140 and 161 to 200, separately. We refer to these sets as $TS_{101-120}$ and $TS_{161-200}$, respectively. These datasets where each set has different lengths of synchronous intervals will be useful in studying the impact of correlated interval lengths on the performance of the proposed approach.

Parameter choices We use three different parameter choices for interval length: $\alpha = [20,\ 30,\ 40]$. We did not use an α that is longer than 40 as the smallest imputed interval is of length 41. We used $\beta = 0.8$. We also varied the length of the time series used as input by starting from the first point and ending at several points including $\{100, 300, 500, \ldots, 1500\}$ to study the impact of length of time series on the time taken by the algorithms.

Both the brute force and the bottom-up approaches were implemented in Matlab® and were executed on a node with 15 Xeon 2.40GHz processors and 100GB of main memory. Nevertheless, our implementation does not use more than one processor at a time.

Observations The comparison of the time taken to discover the longest set of non-overlapping maximal correlated intervals is shown in Figure 6.2. X-axis in this figure shows the length of the time series used and Y-axis indicates time in seconds. Note that Y-axis in the figure is in logarithmic scale. As one would expect, the amount of time taken increases dramatically with increase in the length of the time series. From Figure 6.2 it can be seen that, in general, the brute force approach takes at least 10 fold more time to discover the LAMINA. This is due to the key difference in the two approaches that is bottom-up enumeration versus exploring all possible intervals. Additionally brute force approach has to prune intervals whose sub-intervals are not correlated.

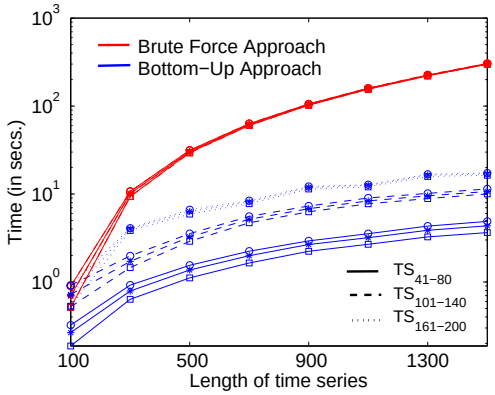

Figure 6.2: Efficiency comparison between Brute force and Bottom-up approaches. The curves with circles, stars, and squares represent $\alpha = 20, 30,$ and 40, respectively.

The time taken for the brute force approach on different datasets (TS_{41-80}, $TS_{101-140}$ and $TS_{161-200}$) and for different choices of α followed a very similar trend. On the other hand, the amount of time taken for the bottom-up approach on $TS_{161-200}$ is more than that of $TS_{101-140}$ and the time taken on $TS_{101-140}$ is more than that of TS_{41-80}. Therefore, the bottom-up approach is faster when the size of correlated intervals is shorter. This is because the bottom-up approach need not take into account the longer intervals.

Figure 6.2 also shows the impact of varying α. On TS_{41-80} dataset, as the length of the smallest interval is increased from 20 to 40, the overall time taken has reduced marginally. This change was relatively small for $TS_{101-140}$ and was much smaller for $TS_{161-200}$. This is due to the fact that the number possible intervals that need to be considered decreases with the increase of the valid interval length.

6.5.2 Effectiveness comparison

The proof of correctness provided in the methods section argues for the correctness (i.e., effectiveness) of the proposed approach in capturing the longest set of non-overlapping maximal correlated intervals. However, when the goal is to discover the longest set of distinct intervals during which two given time series are correlated we need to empirically evaluate it. We evaluated our approach on a synthetic dataset by imputing correlated intervals of different lengths.

Synthetic Data We created a set of 100 pairs of synthetic time series each of length 1000 as described in the previous section where the length of synchronous intervals can vary from 51 to 150. We refer to this set as TS_{51-150}. This large difference in the interval lengths was used to show the ability of our algorithm to capture all imputed intervals irrespective of the difference in their length.

Approaches A traditional way to assess similarity in small intervals between two time series is to use a *sliding window based approach* [78, 76, 79, 73] where a fixed length time window is moved along the time axis and a distance metric (correlation in our case) is computed in each window. One could potentially use this approach to identify the time windows of a given length that satisfy the correlation threshold (β). To study how well this traditional sliding-window approach captures the imputed intervals in TS_{51-150}, we first computed correlation for each possible interval of a chosen window length and then collected the list of all intervals that satisfy the correlation threshold. These intervals are further assessed for their effectiveness in capturing the imputed intervals. We used window lengths $\{40, 50, 60, \ldots 160\}$. Note that these chosen lengths capture the range of lengths of imputed correlated intervals and therefore every imputed interval should be captured in at least one of these intervals. We used $\beta = \{0.8, 0.85, 0.9, 0.95\}$.

We used the proposed approach with $\alpha = \{20, 30, 40\}$ and $\beta = \{0.8, 0.85, 0.9, 0.95\}$ to compare its effectiveness with the above approach.

Evaluation metrics We studied two main aspects of the problem: a) How well a discovered interval captures an imputed interval ? b) How well is an imputed interval recovered ? Note that these two aspects are similar to that of the traditional predictive model evaluation metrics of *precision* and *recall*. For evaluating the effectiveness of discovering intervals we use very similar measures.

For each discovered interval, we evaluated its precision by finding the best matching imputed interval and then computing the fraction of the discovered interval that is a part of the imputed interval. The best match is determined as the interval that has the largest overlap with a given interval. Formally,

$$Precision(D_{I_{(a,b)}}, S_{II}) = \max \frac{|overlap(I_{(a,b)}, I_{(c,d)})|}{l_{(a,b)}}, \forall I_{(c,d)} \in S_{II} \tag{6.1}$$

Where $D_{I_{(a,b)}}$ is a discovered interval $I_{(a,b)}$ and S_{II} is the set of all imputed intervals. The

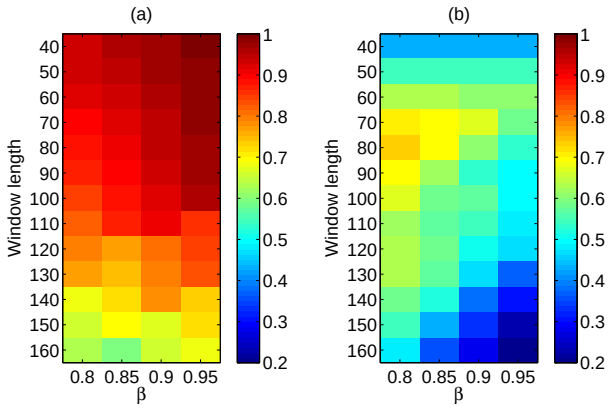

Figure 6.3: Effectiveness of the sliding-window approach: (a) Precision (b) Recall

overall precision for a set of all discovered intervals is computed as the average of the precision of each of the discovered intervals.

We evaluated recall of an imputed interval, by first identifying the best matching discovered interval and then computing the fraction of the imputed interval that matches with the discovered interval. Formally,

$$Recall(I_{I_{(a,b)}}, S_{DI}) = \max \frac{|overlap(I_{(a,b)}, I_{(c,d)})|}{l_{(a,b)}}, \forall I_{(c,d)} \in S_{DI} \tag{6.2}$$

Where $I_{I_{(a,b)}}$ is an imputed interval $I_{(a,b)}$ and S_{DI} is the set of all discovered intervals. Here as well, the recall over the set of all imputed intervals is computed as the average recall over all the imputed intervals.

Observations Figure 6.3 shows the precision and recall values obtained using the traditional sliding-window approach. The precision values were the highest for high values of β, i.e, $\beta = 0.95$ and for smaller window length (i.e., 40). However, the recall at this combination was quite low (0.43, approximately). This is because with windows as short as 40 time points the longer intervals (length ¿ 40) cannot be directly captured. However, the intervals that were captured using windows of length 40 match well with the imputed intervals. When longer intervals were considered (length ¿100) the recall was still low because the smaller intervals can no longer be captured effectively. The best recall of 0.73 was observed when $\beta = 0.8$ and when window length was 80. However, precision at this combination was 0.87.

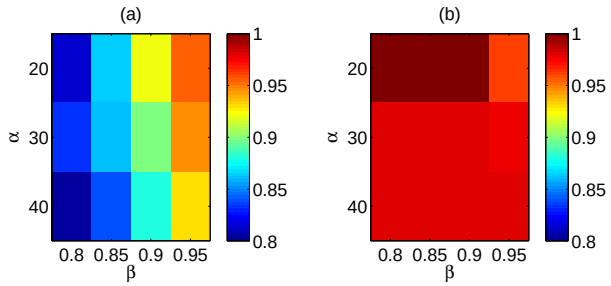

Figure 6.4: Effectiveness of the Bottom-up approach: (a) Precision (b) Recall. (The scale of color-bars in this figure is different from that of Figure 6.3.)

Figure 6.4 shows the precision and recall values obtained using the proposed approach. Note that the scale of the color-bar in this figure is different from that of Figure 6.3. As seen in the case of the traditional approach, precision was high when β was high ($\beta = 0.95$) and the minimum window length was small ($\alpha = 20$). Here, recall was quite high compared to the previous approach as almost all of the imputed intervals were recovered perfectly due to the ability of the proposed algorithm in recovering all maximal arbitrary length intervals when there was no overlap.

The best precision and recall for the proposed approach were (0.95, 0.96) and (0.92, 1) for ($\alpha = 20$, $\beta = 0.95$) and ($\alpha = 20$, $\beta = 0.9$), respectively. These are superior to those of the traditional approach (0.87, 0.73). Moreover, the proposed method does not require one to try out all the interval lengths as it is required for the above traditionally used approach. Hence, we argue that the proposed method is not only efficient but is also effective in recovering arbitrarily long correlated subsequences in a pair of time series.

6.5.3 Case study: Neuroimaging data

We now show the utility of our proposed formulation on a real world Neuroimaging dataset.

Real-world Data Functional Magnetic Resonance Image (fMRI) data measures the amount of oxygen absorbed by gray matter tissue at every tiny cubic location in the brain (referred to as a voxel), at every time instant during the scan. The amount of oxygen absorbed at a given voxel and at a given time point is known to indicate the amount of activity occurring at the voxel. Data from an fMRI scan can be represented as a set of time series, one for every voxel.

This can be represented in the form of a voxel×time matrix, where every ij^{th} element in the matrix indicates the amount of neuronal activity occurring at a location represented by voxel i and at a time point j. We used the dataset from [72] that contains 10 five minute resting state fMRI scans from one healthy subject obtained during one visit on a day. We append all these scans to get one 50 minute resting state scan. We refer to this dataset as *rest* dataset. The spatial resolution of each fMRI scan was 3×3×3 mm and the temporal resolution was 2 seconds. Several prepossessing steps have been performed on the data obtained from the scanner and they have been elaborately discussed in [72]. In addition, following the approach in [66], global mean time series is regressed from the data, as is done in most fMRI studies. The resultant voxel×time matrix for each scan was of dimensions $47,640 \times 1550$. We further group voxels into 90 brain regions based on an automated anatomical labelling (AAL) atlas provided by [1] (see Table A.1 for a list of the brain regions). We refer an interested reader to Table 2 in [85] for a list of these regions. The resultant matrix, D_{rest}, was of size 90×1550.

We also used the 10 five minute fMRI scans obtained from the same subject as above while the subject was watching cartoons [72]. This data allows us to study the differences that occur in the brain between rest and task. We refer to this data as *cartoons* data. The data was processed similarly to that of resting state data. The resultant matrix, $D_{cartoons}$ was also of size 90×1550.

Approach We used our bottom-up approach to discover the longest set of non-overlapping maximal correlated intervals between every pair of brain regions in D_{rest}. There were a total of 4005 ($\binom{90}{2}$) such pairs. We used parameters $\alpha = 25$ (minimum interval length) and $\beta = 0.7$ (correlation threshold). The reason for a relatively relaxed β compared to the one used with synthetic datasets is that fMRI data is often noisy and a relaxed correlation threshold allows us to capture potentially relevant intervals.

We repeated this analysis on $D_{cartoons}$ using the same parameters. We then characterized the intervals captured on D_{rest} and relate them with our findings on $D_{cartoons}$.

Observations • **Comparing LAMINA in rest and cartoons data** The correlated intervals in LAMINA that are discovered using our approach enables us to study the difference in overall duration of correlated intervals for a region pair in rest and cartoons data. For example, a pair of brain regions could be correlated for a small number of intervals in rest and they could be correlated for many more intervals in cartoons data. One such example where a dramatic

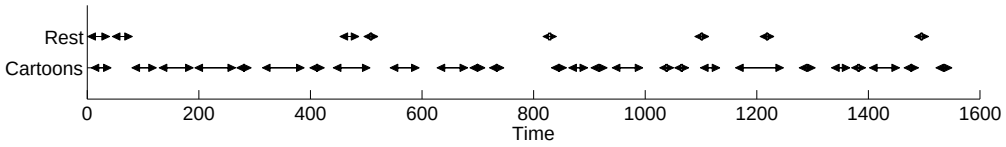

Figure 6.5: Correlated intervals between regions 51 and 55 while the subject is resting and while watching cartoons.

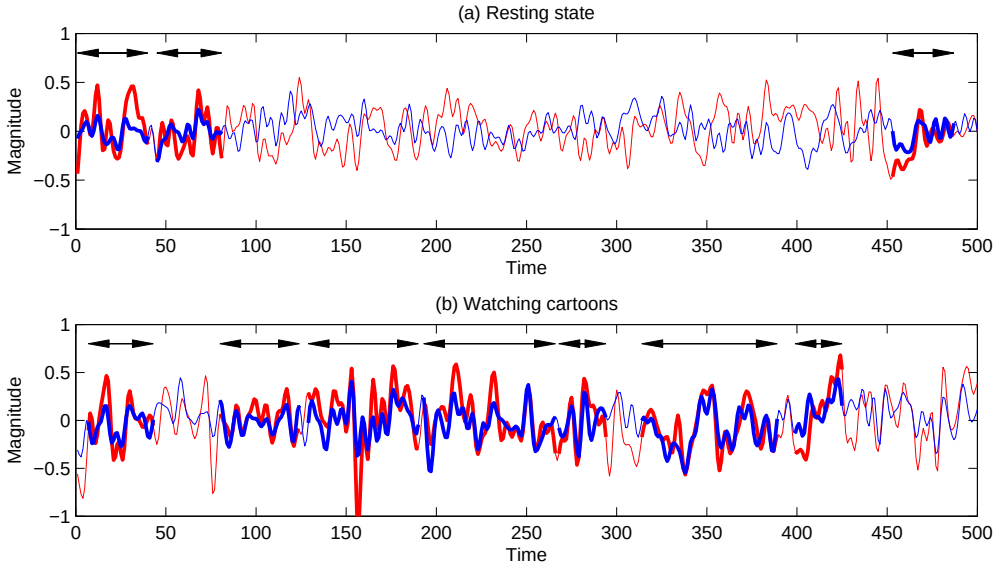

Figure 6.6: Time series for regions 51 and 55 while the subject is resting and while watching cartoons.

difference is seen between rest and cartoons data is shown in Figure 6.5. This figure shows the correlated intervals found between brain regions 51 and 55 in rest and cartoons data with the help of double sided arrows where the left and right arrows indicate the start and end points, respectively. The total duration of these intervals in rest is 241 and in cartoons is 1105. This large difference in the amount of time the two time series are correlated suggests that these two regions exhibit synergy more when the subject is watching cartoons than when the subject is resting. The time series from these two regions are shown in Figure 6.6 (only the first 500 time points are shown due to space limitation). The intervals discovered using our approach are indicated with double sided arrows as well as bold lines in time series. It is easy to see from

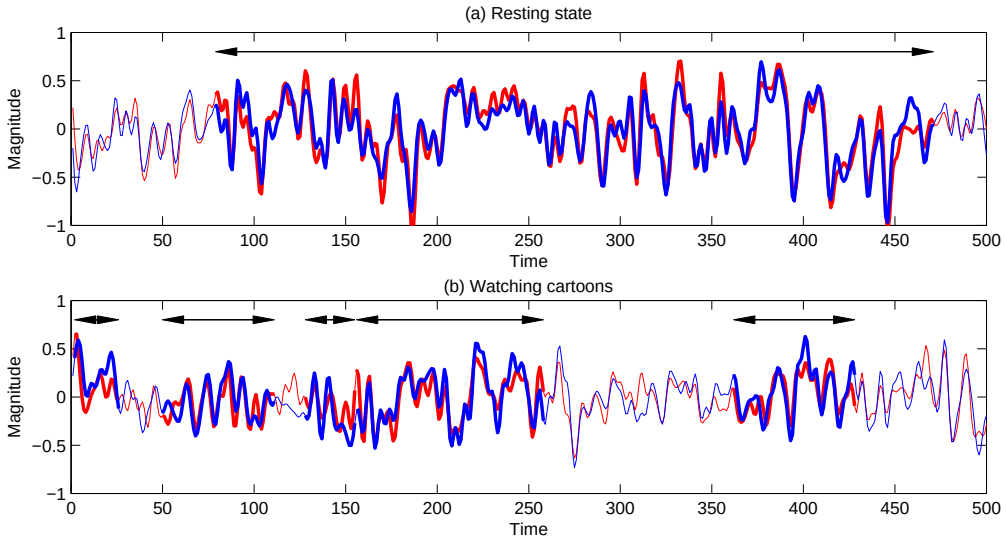

Figure 6.7: Time series for regions 3 and 7 while the subject is resting and while watching cartoons. Only first 500 time points are shown due to space limitation.

this figure that when a region is not captured in a correlated interval, the time series from the two regions exhibit very different behavior.

Region 51 is a *middle occipital region* (left), referred to as the visual V1 system, is well known for its role in processing spatial information, where as region 55 is *fusiform* (left) that is known for its role in object and color information processing [60]. These two regions that handle different aspects of visual information processing can be hypothesized to work synergistically in processing visual information while the subject is watching cartoons. Our approach enables one to discover such novel differences in the way brain regions work together to achieve a task.

• **Different type of intervals in rest and cartoons data** In addition to the differences in the total duration of intervals between brain regions there could exist more subtle differences in how two brain regions work synergistically. For instance, two brain regions could behave similarly for longer intervals on average or could only behave similar for shorter intervals. These differences can be investigated by comparing the average length of intervals when their total duration of intervals is similar.

An example of this type of scenario is shown in Figure 6.8. Here the black double sided arrows indicate the correlated intervals discovered using our approach. For the brain regions (3 7), there are three long correlated intervals [79 471], [1032 1388], and [1389 1550] of duration

393, 357, and 162, respectively. The total duration of these intervals is 912. We chose six other region pairs (58 64), (69 70), (46 47), (7 11), (25 32), and (8 62) whose total length of correlated intervals are similar to that of (3 7) for comparison. Their total duration of intervals are 906, 908, 910, 913, 913, and 915, respectively. The longest correlated interval in these six different brain regions is 113 and the mean of all intervals in them is approximately 40.8, whereas the shortest correlated interval in (3 7) is 162. Therefore the intervals in these six region pairs are much shorter than those of (3 7).

The intervals indicated in black in Figure 6.8 are bound to have correlation ≥ 0.7, however it is possible that between two successive intervals the correlation may be slightly smaller than 0.7 and they are split because of the choice of the threshold. Before concluding that there is a significant difference in average length of correlated intervals in LAMINA for the above pairs, it is necessary to ensure that these splits are due to a significant reduction in correlation and not due to the choice of β threshold. One way to achieve this is to check if there are intervals of significantly lower correlations between two correlated intervals in a LAMINA. Our approach which can find maximal intervals with correlation above a threshold β can also be used to find maximal intervals with correlation $< \beta'$ using the condition $r < \beta'$ in step 13 of Algorithm 5. We used this variant of our approach with $\beta' = 0.5$ to find maximal intervals with correlations < 0.5. The resultant intervals are shown in green in Figure 6.8. If every consecutive pair of correlated intervals shown in black are separated by a green interval it guarantees that a split was due to a significant change in correlation and was not due to the choice of β. Almost all of the consecutive pairs of black intervals in Figure 6.8 have green intervals between them. This provides support for our argument that indeed the intervals are correlated for longer duration in (3 7) than the other pairs listed above.

The dramatic difference in the duration of intervals in (3 7) compared to that of the other region pairs suggest a difference in operating principles for these two regions. These regions tend to work in a synergistic fashion for longer periods.

We also compared the duration of intervals for the pairs of regions (3 7) between rest and cartoons datasets. This comparison is shown in Figure 6.9. The length of intervals in cartoons data varies from 25 to 207 with a mean of 77.7, approximately. While the length of intervals in rest data varies from 162 to 393 with a mean of 304. Interestingly, the longest interval (207) found in cartoons for the pair (3 7) is smaller than the mean of the length of correlated intervals (304) in rest data. It is interesting to note that the total duration of intervals is 912 in rest and

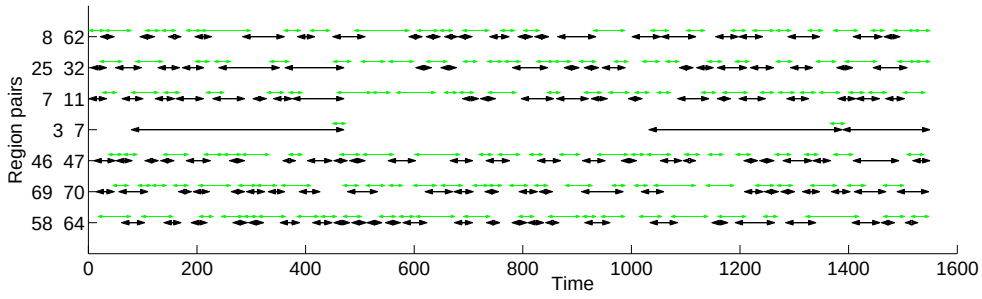

Figure 6.8: Correlated intervals for regions 3 and 7, as well as other region pairs that have a similar total length of correlated intervals. Intervals indicated in black have a correlation ≥ 0.7 and those in green have a correlation ≤ 0.5.

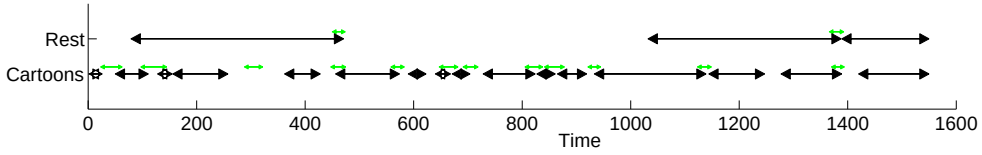

Figure 6.9: Correlated intervals for regions 3 and 7 while the subject is resting and while watching cartoons. Intervals indicated in black have a correlation ≥ 0.7 and those in green have a correlation ≤ 0.5.

1243 in cartoons data. Despite having correlated intervals for approximately 33% longer in cartoons data than in rest data, the average length of intervals in cartoons is smaller than the smallest interval length in rest data.

This indicates that these two brain regions (3 and 7) operate differently when the subject is resting and when the subject is watching cartoons. While the subject is resting the two regions work synergistically for longer periods than when cartoons are being watched. The time series from these two regions are shown in Figure 6.7. The intervals discovered using our approach are indicated with double sided arrows as well as bold lines in time series. Based on these observations we claim that our approach enables one to characterize the dynamic behavior between brain regions based on the duration of individual maximal correlated intervals.

6.6 Conclusion and Future work

In this chapter we presented a novel formulation to capture distinct maximal correlated intervals in a pair of time series. We proposed an efficient bottom-up approach to discover the longest set of such intervals. We evaluated our approach on synthetic datasets to demonstrate its efficiency and its effectiveness. We show that traditional approaches to discover these intervals are limited because of their inability to capture arbitrarily long intervals. We also present a case study on a real world dataset.

A number of aspects of this problem need to be studied further. Real world time series datasets are often noisy and due to this an originally long correlated interval could be discovered as multiple disjoint correlated intervals. Error tolerant techniques for discovering correlated intervals need to be explored. Often general trends in the data may result in intervals that do not reflect synergy between the two entities from which the time series are obtained. For example, the mortgage crisis in 2008 affected a majority of the stocks in a similar fashion and so correlated intervals that cover this period are not interesting in the context of the given pair of stocks. Approaches to address this issue need to be studied. The intervals that are discovered by our approach could be potentially used to cluster [26], classify [76] and discover patterns [73] in time series data. The suitability of LAMINA for these problems needs to be studied.

Algorithm 6 FindLAMINA

Input:
Set of k correlated intervals CI
Output:
A set of non-overlapping correlated intervals whose intervals are collectively the longest

1. $sorted\_CI = sort\_by\_starting\_time(CI)$
2. $next\_interval[1:k] = 0$
3. **for** $i = 1$ *to* k
4. $\quad foll\_intvls =$ an immediate interval starting after i ends
5. **end for**
6. $cum\_sum\_sel\_intvls[1:k] = 0$
7. $sel\_intvls[1:k] = 0$
8. **for** $i = n$ *to* 1
9. $\quad cum\_sum\_sel\_intvls[i] =$
9. $\qquad max(l_i + cum\_sum\_sel\_intvls[foll\_intvls[i]],$
9. $\qquad\quad cum\_sum\_sel\_intvls[i+1])$
10. $\quad sel\_intvls[i] = 1$, if interval i was part of solution
11. **end for**
12. $Result\_Corr\_Intvls = \phi$
13. $current =$ smallest i with $sel\_intvls[i] = 1$
14. **while** $true$
15. $\quad$ **if** $(sel\_intvls[current] == 1)$
16. $\qquad Result\_Corr\_Intvls = Result\_Corr\_Intvls \cup i$
17. $\qquad current = foll\_intvls[current]$
18. $\quad$ **end if**
19. $\quad$ **if** $(sel\_intvls[current] == 0)$
20. $\qquad current = current + 1$
21. $\quad$ **end if**
21. $\quad$ **Exit** if no more intervals can be selected
22. **end while**
23. return $Result\_Corr\_Intvls$

Chapter 7

Conclusion and future work

In this chapter we first list the future work for our contributions to computer science and to our contributions to neuroscience separately. We then give an account of other ways of advancing the state-of-the-art for analyzing fMRI data.

7.1 Contributions to Computer Science

Ts-Apriori approach presented in this paper can be potentially improved. First, fMRI datasets are known to be noisy and so tolerance for noise can be introduced. Second, similarity is assessed using a correlation threshold. There could be some windows where the correlation is marginally below the threshold. To address this issue a real valued notion of ts-support can be introduced where the contribution of every window can be assessed based on some quantile based score. Third, the approach will not be able to work with more than a few thousands of time series. Approaches such as colossal patterns need to be explored for large datasets. Fourth, sliding window based ts-support score can be replaced with the interval based scoring scheme to get a better estimate of the interval duration. Fifth, a number of patterns are typically generated with Apriori like techniques that often share a good amount of overlap. Interpreting such large sets of overlapping patterns is challenging. Tools for visualizing such patterns or summarizing them into small number of groups can be useful.

The objective of the LAMINA approach that finds longest set of non-overlapping maximal correlated intervals is one of the many ways in which correlated intervals can be discovered. Alternative objectives have to be explored and evaluated to determine the best possible objective

to capture the intervals. In addition, general trends in the data may result in intervals that do not reflect synergy between the two entities from which the time series are obtained. For example, the mortgage crisis in 2008 affected a majority of the stocks in a similar fashion and so correlated intervals that cover this period are not interesting in the context of the given pair of stocks. Approaches to address this issue need to be studied. Furthermore, the intervals that are discovered by our approach could be potentially used to cluster [26], classify [76] and discover patterns [73] in time series data. The suitability of LAMINA for these problems needs to be studied.

7.2 Contributions to Neuroscience

The frequency of measurement in the fMRI data that is used in evaluation is 2 seconds. As we are studying transient relationships, a high frequency measurement can be useful in accurately capturing such relationships. A recent dataset [86] uses data where the frequency of measurement is 0.8 seconds. The approaches proposed in this thesis can be used on such a dataset to study the nature of patterns and to compare them with those reported here.

Another potential direction is to use block design task data for studying transient relationships. In block design data a subject is involved in a task for a block of time and is rested for another block. These two phases are repeated for the entire scan. Combinations of time series that exclusively exhibit similarity within the task blocks or within the rest blocks indicate a direct relationship to the task or the rest states.

To study the role of transient relationships in mental disorders a longer fMRI scan will be potentially effective. In this thesis we presented algorithms where transient relationships are computed overall healthy subjects and disease subjects separately. Ideally, we would like to find markers that are subject specific for them to be of clinical applicability. As the relationships are transient, having a longer fMRI scan will provide a greater opportunity to accurately capture them if they are present in a subject. With a 6 minute scan, it is unclear if such a relationship is not present in the subject or if it is not just present in the 6 minute scan but is present in the subject.

7.3 Other directions

In this section we discuss a broad set of problems pertaining to analyzing fMRI data to advance our understanding of brain functionality.

7.3.1 Data driven brain regions

As in most fMRI studies that construct a brain network from fMRI data [14], in this thesis we relied on an anatomical atlas (AAL) [1] to determine the brain regions. We have found in our earlier analysis that anatomical regions defined in AAL do not generalize to every subject. In effect, building networks based on such an atlas can result in missed underlying relationships that exist in the data. One way to address this limitation is to directly find the regions in a data driven fashion.

7.3.2 Evolutionary analysis

As a brain acquires a skill the brain network is expected to adapt to encode the acquired skill. There is very little information available on how brain encodes the information it learns. To understand this, one approach is to study the brain networks from subjects at regular intervals when a skill is being learned. Some early work has been done in this direction [3]. Figure 7.1 illustrates the study setup where participants were scanned at different stages of the training process. These studies look at high level network properties such as core-periphery architecture to study the nature of changes in the brain networks. The approaches to study the dynamic nature of connectivity, that are presented in this thesis, can be used to estimate the increased or inhibited connectivity over the course of training.

7.3.3 Brain states

The functional brain network is expected to be different in different scenarios. For example, when a person is planning his day the brain network can be in a particular configuration, whereas the brain network when the person is watching a visual stimulus, it could exhibit a different configuration. The first step towards understanding these states would be to characterize brain states in resting state fMRI data and then relate it to the task based network configuration. Here inter-subject variability needs to be taken into account, as everyone's brain network can

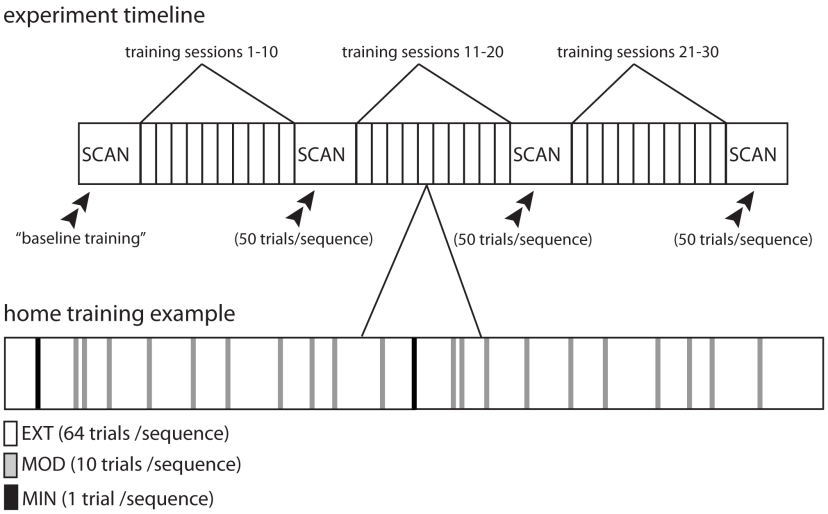

Figure 7.1: Experimental setup for studying brain networks while learning a skill (Figure borrowed from [3]). Here fMRI scans are collected after every 10 training sessions for each subject.

potentially appear to be different despite similarities in the task.

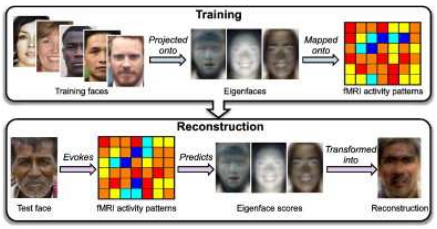

Figure 7.2: Face reconstruction from fMRI data (Figure borrowed from [4]).

7.3.4 Discovering stimulus from brain activity

Existing studies compare the brain networks of a subject from resting state and task fMRI datasets. These studies point out the differences in the brain network that potentially arise due to the task at hand. While these studies provide valuable insights into the connectivity that is desired to achieve a function or task, our ability to detect the stimulus by only looking at the brain activity is minimal. Ability to detect the stimulus by analyzing brain activity will demonstrate our understanding of the brain's working principles in different scenarios. As an

example, Cowen et al [4] recently built a model (an illustration is shown in Figure 7.2) based on brain's cortical activity to discover the faces that were presented to the subject. This is the first study to look at activity patterns of higher level cortical regions for face reconstruction, while most studies relied on activity in the visual cortex. More efforts on reconstruction of visual, auditory and sensory stimulus from cortical regions will help us understand how and where different aspects of a stimuli are processed.

In summary, brain fMRI data offers a rich landscape of spatio-temporal data mining problems. Investigations in this area can not only advance our understanding of brain functionality but will also advance the state-of-the-art in spatio-temporal data mining that has wide applicability in diverse areas such as climate science and transportation.

References

[1] N Tzourio-Mazoyer et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 2002.

[2] VD Calhoun, T Adali, LK Hansen, J Larsen, and JJ Pekar. Ica of functional mri data: An overview. In *in Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*. Citeseer, 2003.

[3] Danielle S Bassett, Nicholas F Wymbs, M Puck Rombach, Mason A Porter, Peter J Mucha, and Scott T Grafton. Task-based core-periphery organization of human brain dynamics. *PLoS computational biology*, 9(9):e1003171, 2013.

[4] Alan S Cowen, Marvin M Chun, and Brice A Kuhl. Neural portraits of perception: Reconstructing face images from evoked brain activity. *NeuroImage*, 94:12–22, 2014.

[5] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to data mining, 2006.

[6] Chotirat Ann Ratanamahatana et al. Mining time series data. In *Data Mining and Knowledge Discovery Handbook*. 2010.

[7] Auroop R Ganguly and Karsten Steinhaeuser. Data mining for climate change and impacts. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, pages 385–394. IEEE, 2008.

[8] Huda Akil, Maryann E Martone, and David C Van Essen. Challenges and opportunities in mining neuroscience data. *Science (New York, NY)*, 331(6018):708, 2011.

[9] Shashi Shekhar and Sanjay Chawla. *Spatial databases: a tour*, volume 2003. prentice hall Upper Saddle River, NJ, 2003.

[10] Harvey J Miller and Jiawei Han. *Geographic data mining and knowledge discovery*. CRC Press, 2009.

[11] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, 2010.

[12] Bruce R Rosen and Robert L Savoy. fmri at 20: Has it changed the world? *Neuroimage*, 62(2):1316–1324, 2012.

[13] Mary-Ellen Lynall, Danielle S Bassett, Robert Kerwin, Peter J McKenna, Manfred Kitzbichler, Ulrich Muller, and Ed Bullmore. Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, 30(28):9477–9487, 2010.

[14] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[15] Zarrar Shehzad, AM Clare Kelly, Philip T Reiss, Dylan G Gee, Kristin Gotimer, Lucina Q Uddin, Sang Han Lee, Daniel S Margulies, Amy Krain Roy, Bharat B Biswal, et al. The resting brain: unconstrained yet reliable. *Cerebral cortex*, 19(10):2209–2229, 2009.

[16] Gowtham Atluri, Kanchana Padmanabhan, Gang Fang, Michael Steinbach, Jeffrey R Petrella, Kelvin Lim, Angus MacDonald III, Nagiza F Samatova, P Murali Doraiswamy, and Vipin Kumar. Complex biomarker discovery in neuroimaging data: Finding a needle in a haystack. *NeuroImage: Clinical*, 3:123–131, 2013.

[17] William Pettersson-Yeo, Paul Allen, Stefania Benetti, Philip McGuire, and Andrea Mechelli. Dysconnectivity in schizophrenia: where are we now? *Neuroscience & Biobehavioral Reviews*, 35(5):1110–1124, 2011.

[18] Catie Chang and Gary H Glover. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

[19] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[20] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[21] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[22] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[23] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 2011.

[24] Alessandro Camerra et al. iSAX 2.0: Indexing and mining one billion time series. In *ICDM*, pages 58–67, 2010.

[25] Xianping Ge et al. Deformable markov model templates for time-series pattern matching. In *SIGKDD*, 2000.

[26] Jesin Zakaria et al. Clustering time series using unsupervised-shapelets. In *IEEE ICDM*, 2012.

[27] Li Wei and Eamonn Keogh. Semi-supervised time series classification. In *SIGKDD*, pages 748–753, 2006.

[28] Eamonn Keogh et al. Finding surprising patterns in a time series database in linear time and space. In *SIGKDD*, 2002.

[29] Chiung-Hon Leon Lee et al. Pattern discovery of fuzzy time series for financial prediction. *TKDE*, 2006.

[30] Jessica Lin et al. Experiencing SAX: a novel symbolic representation of time series. *DMKD*, 2007.

[31] Raymond T. Ng and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5):1003–1016, 2002.

[32] Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In *Advances in spatial databases*, pages 47–66. Springer, 1995.

[33] Krzysztof Koperski, Jiawei Han, and Nebojsa Stefanovic. An efficient two-step method for classification of spatial data. In *1998 international symposium on spatial data handling SDH*, volume 98, pages 45–54, 1998.

[34] William J Jagust, Dan Bandy, Kewei Chen, Norman L Foster, Susan M Landau, Chester A Mathis, Julie C Price, Eric M Reiman, Daniel Skovronsky, and Robert A Koeppe. The alzheimer's disease neuroimaging initiative positron emission tomography core. *Alzheimer's & Dementia*, 6(3):221–229, 2010.

[35] V Chandola, D Cheboli, and V Kumar. Detecting anomalies in a time series database. *Computer Science Department, University of Minnesota, Tech. Rep*, 2009.

[36] Xi C Chen, Karsten Steinhaeuser, Shyam Boriah, Snigdhansu Chatterjee, and Vipin Kumar. Contextual time series change detection. Technical report, SIAM, 2012.

[37] A Tung, Jean Hou, and Jiawei Han. Spatial clustering in the presence of obstacles. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 359–367. IEEE, 2001.

[38] Osmar R Zaïane and Chi-Hoon Lee. Clustering spatial data when facing physical constraints. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 737–740. IEEE, 2002.

[39] Xin Wang, Camilo Rostoker, and Howard J Hamilton. *Density-based spatial clustering in the presence of obstacles and facilitators*. Springer, 2004.

[40] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3457–3464. IEEE, 2011.

[41] Songhwai Oh, Stuart Russell, and Shankar Sastry. Markov chain monte carlo data association for general multiple-target tracking problems. In *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, volume 1, pages 735–742. IEEE, 2004.

[42] Ingemar J Cox. A review of statistical data association techniques for motion correspondence. *International Journal of Computer Vision*, 10(1):53–66, 1993.

[43] James H Faghmous, Muhammed Uluyol, Luke Styles, Matthew Le, Varun Mithal, Shyam Boriah, and Vipin Kumar. Multiple hypothesis object tracking for unsupervised self-learning: An ocean eddy tracking application. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[44] Alex Fornito, Andrew Zalesky, and Michael Breakspear. Graph analysis of the human connectome: promise, progress, and pitfalls. *Neuroimage*, 80:426–444, 2013.

[45] Olaf Sporns. The human connectome: a complex network. *Annals of the New York Academy of Sciences*, 1224(1):109–125, 2011.

[46] Danielle S Bassett and Edward T Bullmore. Human brain networks in health and disease. *Current opinion in neurology*, 22(4):340, 2009.

[47] Krista M Wisner, Gowtham Atluri, Kelvin O Lim, and Angus W MacDonald III. Neurometrics of intrinsic connectivity networks at rest using fmri: Retest reliability and cross-validation using a meta-level method. *NeuroImage*, 76:236–251, 2013.

[48] Danielle S Bassett, Brent G Nelson, Bryon A Mueller, Jazmin Camchong, and Kelvin O Lim. Altered resting state complexity in schizophrenia. *Neuroimage*, 59(3):2196–2207, 2012.

[49] Jinhui Wang, Liang Wang, Yufeng Zang, Hong Yang, Hehan Tang, Qiyong Gong, Zhang Chen, Chaozhe Zhu, and Yong He. Parcellation-dependent small-world brain functional networks: A resting-state fmri study. *Human brain mapping*, 30(5):1511–1523, 2009.

[50] Satoru Hayasaka and Paul J Laurienti. Comparison of characteristics between region- and voxel-based network analyses in resting-state fmri data. *Neuroimage*, 50(2):499–508, 2010.

[51] David Meunier, Renaud Lambiotte, and Edward T Bullmore. Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 4, 2010.

[52] Alex Fornito, Andrew Zalesky, and Edward T Bullmore. Network scaling effects in graph analytic studies of human resting-state fmri data. *Frontiers in systems neuroscience*, 4, 2010.

[53] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[54] Ed Bullmore and Olaf Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349, 2012.

[55] Petra E Vértes, Aaron F Alexander-Bloch, Nitin Gogtay, Jay N Giedd, Judith L Rapoport, and Edward T Bullmore. Simple models of human brain functional networks. *Proceedings of the National Academy of Sciences*, 109(15):5868–5873, 2012.

[56] Andrew Zalesky, Alex Fornito, and Edward T Bullmore. Network-based statistic: identifying differences in brain networks. *Neuroimage*, 53(4):1197–1207, 2010.

[57] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

[58] Megan H Lee, Carl D Hacker, Abraham Z Snyder, Maurizio Corbetta, Dongyang Zhang, Eric C Leuthardt, and Joshua S Shimony. Clustering of resting state networks. *PloS one*, 7(7):e40370, 2012.

[59] Madiha J Jafri, Godfrey D Pearlson, Michael Stevens, and Vince D Calhoun. A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *Neuroimage*, 39(4):1666–1681, 2008.

[60] Ido Davidesco et al. Spatial and object-based attention modulates broadband high-frequency responses across the human visual cortical hierarchy. *J. N.Sci.*, 2013.

[61] Hugo Spiers et al. Decoding human brain activity during real-world experiences. *Trends in cognitive sciences*, 2007.

[62] Rakesh Agrawal et al. Fast algorithms for mining association rules. In *VLDB*, volume 1215, pages 487–499, 1994.

[63] Gaurav Pandey et al. An association analysis approach to biclustering. In *ACM SIGKDD*, pages 677–686, 2009.

[64] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *DMKD*, pages 349–371, 2003.

[65] Jonathan Power et al. Methods to detect, characterize, and remove motion artifact in resting state fmri. *NeuroImage*, 2013.

[66] Kevin Murphy et al. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage*, 2009.

[67] Jin-Hui Wang et al. Graph theoretical analysis of functional brain networks: test-retest evaluation on short-and long-term resting-state functional mri data. *PLoS One*, 2011.

[68] R Matthew Hutchison, Thilo Womelsdorf, Joseph S Gati, Stefan Everling, and Ravi S Menon. Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. *Human brain mapping*, 34(9):2154–2177, 2013.

[69] Daniel A Handwerker, Vinai Roopchansingh, Javier Gonzalez-Castillo, and Peter A Bandettini. Periodic changes in fmri connectivity. *Neuroimage*, 63(3):1712–1719, 2012.

[70] David T Jones, Prashanthi Vemuri, Matthew C Murphy, Jeffrey L Gunter, Matthew L Senjem, Mary M Machulda, Scott A Przybelski, Brian E Gregg, Kejal Kantarci, David S Knopman, et al. Non-stationarity in the resting brains modular architecture. *PloS one*, 7(6):e39731, 2012.

[71] Elena A Allen et al. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, 2012.

[72] Jeffrey S Anderson, Michael A Ferguson, Melissa Lopez-Larson, and Deborah Yurgelun-Todd. Reproducibility of single-subject functional connectivity measurements. *American Journal of Neuroradiology*, 32(3):548–555, 2011.

[73] Gowtham Atluri et al. Discovering groups of time series with similar behavior in multiple small intervals of time. In *SDM*, 2014.

[74] David A Crowe et al. Prefrontal neurons transmit signals to parietal neurons that reflect executive control of cognition. *Nature neuroscience*, 2013.

[75] Yuhong Li et al. Discovering longest-lasting correlation in sequence databases. *Proceedings of the VLDB Endowment*, 2013.

[76] Lexiang Ye et al. Time series shapelets: a new primitive for data mining. In *SIGKDD*, 2009.

[77] Abdullah Mueen et al. Logical-shapelets: an expressive primitive for time series classification. In *SIGKDD*, 2011.

[78] Aoying Zhou et al. Tracking clusters in evolving data streams over sliding windows. *KIS*, 2008.

[79] Eamonn Keogh et al. Finding the unusual medical time series: Algorithms and applications. *IEEE Transactions on Information Technology in Biomedicine*, 2005.

[80] Christos Faloutsos et al. Fast subsequence matching in time series databases. In *SIGMOD*, 1994.

[81] Tassos Argyros et al. Efficient subsequence matching in time series databases under time and amplitude transformations. In *ICDM*, 2003.

[82] Huanmei Wu et al. Subsequence matching on structured time series data. In *SIGMOD*, 2005.

[83] Gautam Das et al. Finding similar time series. In *Principles of Data Mining and Knowledge Discovery*. 1997.

[84] Jon Kleinberg and Eva Tardos. *Algorithm Design*. Addison-Wesley, 2005.

[85] Zude Zhu et al. Large scale brain functional networks support sentence comprehension: Evidence from both explicit and implicit language tasks. *PloS one*, 2013.

[86] Stephen M Smith et al. Temporally-independent functional modes of spontaneous brain activity. *PNAS*, 109(8):3131–3136, 2012.

Appendix A

Appendix

A.1 List of AAL regions

Number	Region	Number	Region
1	Left precentral gyrus	47	Left lingual gyrus
2	Right precentral gyrus	48	Right lingual gyrus
3	Left superior frontal gyrus, dorsolateral	49	Left superior occipital gyrus
4	Right superior frontal gyrus, dorsolateral	50	Right superior occipital gyrus
5	Left superior frontal gyrus, orbital part	51	Left middle occipital gyrus
6	Right superior frontal gyrus, orbital part	52	Right middle occipital gyrus
7	Left middle frontal gyrus	53	Left inferior occipital gyrus
8	Right middle frontal gyrus	54	Right inferior occipital gyrus
9	Left middle frontal gyrus, orbital part	55	Left fusiform gyrus
10	Right middle frontal gyrus, orbital part	56	Right fusiform gyrus
11	Left inferior frontal gyrus, opercular part	57	Left postcentral gyrus
12	Right inferior frontal gyrus, opercular part	58	Right postcentral gyrus
13	Left inferior frontal gyrus, triangular part	59	Left superior parietal gyrus
14	Right inferior frontal gyrus, triangular part	60	Right superior parietal gyrus
15	Left inferior frontal gyrus, orbital part	61	Left inferior parietal, but supramarginal and angular gyri
16	Right inferior frontal gyrus, orbital part	62	Right inferior parietal, but supramarginal and angular gyri
17	Left rolandic operculum	63	Left supramarginal gyrus
18	Right rolandic operculum	64	Right supramarginal gyrus
19	Left supplementary motor area	65	Left angular gyrus
20	Right supplementary motor area	66	Right angular gyrus
21	Left olfactory cortex	67	Left precuneus
22	Right olfactory cortex	68	Right precuneus
23	Left superior frontal gyrus, medial	69	Left paracentral lobule
24	Right superior frontal gyrus, medial	70	Right paracentral lobule
25	Left superior frontal gyrus, medial orbital	71	Left caudate nucleus
26	Right superior frontal gyrus, medial orbital	72	Right caudate nucleus
27	Left gyrus rectus	73	Left lenticular nucleus, putamen
28	Right gyrus rectus	74	Right lenticular nucleus, putamen
29	Left insula	75	Left lenticular nucleus, pallidum
30	Right insula	76	Right lenticular nucleus, pallidum
31	Left anterior cingulate and paracingulate gyri	77	Left thalamus
32	Right anterior cingulate and paracingulate gyri	78	Right thalamus
33	Left median cingulate and paracingulate gyri	79	Left heschl gyrus
34	Right median cingulate and paracingulate gyri	80	Right heschl gyrus
35	Left posterior cingulate gyrus	81	Left superior temporal gyrus
36	Right posterior cingulate gyrus	82	Right superior temporal gyrus
37	Left hippocampus	83	Left temporal pole: superior temporal gyrus
38	Right hippocampus	84	Right temporal pole: superior temporal gyrus
39	Left parahippocampal gyrus	85	Left middle temporal gyrus
40	Right parahippocampal gyrus	86	Right middle temporal gyrus
41	Left amygdala	87	Left temporal pole: middle temporal gyrus
42	Right amygdala	88	Right temporal pole: middle temporal gyrus
43	Left calcarine fissure and surrounding cortex	89	Left inferior temporal gyrus
44	Right calcarine fissure and surrounding cortex	90	Right inferior temporal gyrus
45	Left cuneus		
46	Right cuneus		

Table A.1: List of AAL regions [1]